SIAM J. Numer. Anal.
Vol. 41, No. 1, pp. 1–36

BOUNDARY CONCENTRATED FINITE ELEMENT METHODS*

B. N. KHOROMSKIJ[†] AND J. M. MELENK[†]

Abstract. A method with optimal (up to logarithmic terms) complexity for solving elliptic problems is proposed. The method relies on interior regularity, but the solution may have globally low regularity due to rough boundary data or geometries. Elliptic regularity results, high order approximation results, and an efficient preconditioner are presented.

The method is utilized to realize, with linear-logarithmic complexity, an accurate and data-sparse approximation to the associated elliptic Poincaré–Steklov operators. Further applications include the treatment of exterior boundary value problems and the solution of problems in the framework of domain decomposition methods.

Key words. *hp*-finite element methods, preconditioning, data-sparse approximation to Poincaré–Steklov operator, meshes refined toward boundary

AMS subject classifications. 65N35, 65F10, 35D10

PII. S0036142901391852

1. Introduction.

In this paper, we present the *boundary concentrated finite element method*. This method is designed to solve numerically elliptic boundary value problems with low global Sobolev regularity. The coefficients of the underlying PDE, however, are assumed to be smooth so that, owing to interior elliptic regularity, the low global Sobolev regularity is due to boundary effects such as low-regularity boundary data or geometries. The key idea of the method is to exploit this interior regularity in the framework of the *hp*-version of the finite element method (*hp*-FEM) by using low order elements on refined meshes near the boundary and high order polynomials on large elements in the interior of the domain. The combination of mesh refinement near the boundary and polynomial degree distribution proposed in this paper concentrates most degrees of freedom in a narrow neighborhood of the boundary, which explains the name boundary concentrated FEM.

Since the boundary concentrated FEM may be viewed as a generalization of the boundary element method (BEM), we illustrate its most important properties by a side-by-side comparison with the classical BEM. In the BEM (see, e.g., [18] for an introduction to the topic), an elliptic boundary value problem on a domain $\Omega \subset \mathbb{R}^d$ is reduced to a problem posed on the boundary $\partial\Omega$, thereby effecting a dimensional reduction. This dimensional reduction immediately leads to a reduction of the problem size of the discrete problems. In the present paper, we show that from the "error vs. degrees of freedom" perspective, the boundary concentrated FEM achieves the same rate of convergence as the classical, low order Galerkin h-BEMs that are formulated on quasi-uniform boundary triangulations. In this respect, therefore, the boundary concentrated FEM is comparable to the classical BEM. However, it represents a generalization of the BEM, in that it can be formulated for equations with variable (albeit piecewise analytic) coefficients, while the BEM is effectively restricted to equations with constant coefficients because explicit knowledge of a fundamental solution is required.

*Received by the editors July 5, 2001; accepted for publication (in revised form) July 30, 2002; published electronically February 12, 2003.

http://www.siam.org/journals/sinum/41-1/39185.html

[†]Max-Planck-Institute for Mathematics in the Sciences, Inselstr. 22-26, D-04103 Leipzig, Germany (bokh@mis.mpg.de, melenk@mis.mpg.de).

A second difference between the classical BEM and the boundary concentrated FEM manifests itself in the structure of the resulting linear system of equations. The boundary concentrated FEM, being an FEM, naturally leads to sparse stiffness matrices. In contrast, the stiffness matrix in BEM is in general fully populated. We mention, however, that this drawback of the classical BEM has been successfully overcome in recent years by various compression schemes, notably the panel clustering techniques [21], multipole expansions (see the survey [14] and the references therein), and wavelet compression methods [12]. A generalization of the clustering techniques are the recently introduced $\mathcal{H}$-matrices [19, 20].

A further interesting point of comparison is the cost of setting up and solving the linear system. We show in this paper that the stiffness matrix of the boundary concentrated FEM can be computed with optimal complexity $O(N)$, where N is the problem size. The classical BEM, which, as we mentioned above, achieves a comparable accuracy with the same number of degrees of freedom N, requires $O(N^2)$ operations to set up the linear system due to the fact that the stiffness matrix is fully populated. Again, only recent progress in compression schemes for the BEM has led to methods with complexity $O(N \log^q N)$ for suitable $q \in \mathbb{N}_0$.

Another important observation is that our technique leads to the accurate and data-sparse approximation of complexity $O(N \log N)$ to Poincaré–Steklov operators associated with elliptic equations with variable coefficients. This generalizes previously known methods for equations with piecewise constant coefficients in polygonal domains such as [24, 27, 28, 25].

In this paper we present a complete theory in the two-dimensional setting. Many results, however, have analogues in higher dimensions. In particular, the regularity assertion (Theorem 1.4) and *hp*-approximation results on shape-regular meshes (Theorem 2.13) can be extended in a straightforward way to three dimensions. Preconditioning techniques for *hp*-FEM/*hp*-BEM in three dimensions have recently been proposed [1, 16, 38], and we expect that these ideas can be employed for the successful development of preconditioners for the boundary concentrated FEM in three dimensions.

The paper is organized as follows. We start with a formulation of the model problem and provide analytic regularity results for the solution. In section 2, we present convergence results for the *hp*-FEM applied to the model problem and show that the method yields the same optimal convergence rate as the *h*-BEM on quasi-uniform meshes. In section 3, we address the question of efficiently solving the resulting linear system. For Dirichlet problems we show that the condition number of the linear system grows only polylogarithmically with the problem size. For Neumann problems, we exhibit a block-diagonal preconditioner such that the condition number of the preconditioned system grows again polylogarithmically. We show in section 4 how the boundary concentrated FEM can be employed to realize an application of the Poincaré–Steklov operator with linear-logarithmic complexity with respect to the boundary degrees of freedom (both in operation count and memory requirement). So far, for the sake of simplicity, our discussion has been mainly restricted to interior problems with analytic coefficients. However, the boundary concentrated FEM can also be employed for exterior problems and domain decomposition problems (piecewise smooth data with respect to a regular geometric decomposition); these applications are briefly addressed in section 5. Numerical experiments in section 6 illustrate the theoretical results of sections 2 and 3.

1.1. Notation. For a Lipschitz domain $\Omega \subset \mathbb{R}^2$, the Sobolev spaces $H^k(\Omega)$, $H^k_0(\Omega)$, $k \in \mathbb{N}_0$, are defined in the standard way. Fractional order Sobolev spaces $H^s(\Omega)$ are defined by interpolation (the real method) between integer order Sobolev spaces. Negative order spaces such as $H^{-1}(\Omega)$ are defined by duality: $H^{-s}(\Omega) = (H^s_0(\Omega))'$. Spaces on the boundary $\partial\Omega$ are defined in the usual way:

$$H^s(\partial\Omega) = \begin{cases} \{u|_{\partial\Omega} \,|\, u \in H^{s+1/2}(\Omega)\} & \text{if } s > 0, \\ L^2(\partial\Omega) & \text{if } s = 0, \\ (H^{-s}(\partial\Omega))' & \text{if } s < 0. \end{cases}$$

We mention at this point the important fact that, for *polygonal domains* Ω, the spaces $H^s(\partial\Omega)$, $|s| < 3/2$, are invariant under piecewise smooth changes of parametrization of $\partial\Omega$. In particular, the parametrization $\varphi : [0, L) \to \partial\Omega$ by arc length provides an isomorphism $u \mapsto u \circ \varphi$ from $H^s(\partial\Omega)$ to $H^s_{per}([0, L))$ for $|s| < 3/2$. Here, for $s > 0$ we set $H^s_{per}([0, L)) := \{u \in H^s(\mathbb{R}) \,|\, u \text{ is } L\text{-periodic}\}$ with the corresponding topology and $H^{-s}_{per}([0, L)) = (H^s_{per}([0, L)))'$ for $s < 0$.

Duality pairings will be denoted by $\langle \cdot, \cdot \rangle$, with subscripts indicating the spaces with respect to which the pairing is taken. Since the spaces $H^{1/2}(\partial\Omega)$, $H^{-1/2}(\partial\Omega)$ arise frequently in this paper, we abbreviate

$$Y := H^{1/2}(\partial\Omega), \qquad Y' := H^{-1/2}(\partial\Omega). \tag{1.1}$$

1.2. Problem class. For simplicity of exposition, we will restrict our attention to problems formulated on polygons, and we will not consider the case of mixed boundary conditions. That is, we consider for a *polygonal* Lipschitz domain $\Omega \subset \mathbb{R}^2$ either the Dirichlet problem

$$\mathcal{L}u = f \in L^2(\Omega) \qquad \text{in } \Omega, \tag{1.2a}$$

$$\gamma_0 u = \lambda \in H^{1/2}(\partial\Omega) \qquad \text{on } \partial\Omega, \tag{1.2b}$$

or the Neumann problem

$$\mathcal{L}u = f \in L^2(\Omega) \qquad \text{in } \Omega, \tag{1.3a}$$

$$\gamma_1 u = \psi \in H^{-1/2}(\partial\Omega) \qquad \text{on } \partial\Omega. \tag{1.3b}$$

Here, the differential operator $\mathcal{L}$ is given by

$$\mathcal{L}u := -\nabla \cdot (A\nabla u) + b \cdot \nabla u + a_0 u, \tag{1.4}$$

with uniformly (in $x \in \overline{\Omega}$) symmetric positive definite matrix $A = (a_{ij})^2_{i,j=1}$; the vector-valued function b and the scalar-valued function a_0 are assumed to be analytic on $\overline{\Omega}$. The operator γ_0 is the trace operator $\gamma_0 : H^1(\Omega) \to H^{1/2}(\partial\Omega)$, and $\gamma_1 := \sum_{i,j=1}^2 n_i a_{ij} \partial_j$ is the conormal derivative operator. We assume that the operator $\mathcal{L}$ generates an $H^1(\Omega)$-elliptic bilinear form

$$B(u, v) = \int_\Omega \sum_{i,j=1}^2 a_{ij} \partial_j u \partial_i v + \sum_{i=1}^2 b_i \partial_i u v + a_0 u v \, dx, \tag{1.5}$$

i.e.,

$$c_0 \|u\|^2_{1,\Omega} \leq B(u, u) \leq c_1 \|u\|^2_{1,\Omega} \qquad \forall u \in V, \tag{1.6}$$

where we introduced the space $V \subset H^1(\Omega)$ in the standard way as

$$V := \begin{cases} H_0^1(\Omega) & \text{if the Dirichlet problem (1.2) is considered,} \\ H^1(\Omega) & \text{if the Neumann problem (1.3) is considered.} \end{cases} \tag{1.7}$$

The boundary value problems (1.2), (1.3) are understood in the usual, variational sense. Solving (1.2) is equivalent to

$$\text{Find } u \in H^1(\Omega) \text{ with } \gamma_0 u = \lambda \text{ and } B(u,v) = \int_\Omega f\, v\, dx \qquad \forall v \in H_0^1(\Omega). \tag{1.8}$$

Solving (1.3) reads as

$$\text{Find } u \in H^1(\Omega) \text{ s.t. } B(u,v) = \int_\Omega f\, v\, dx + \langle \psi, \gamma_0 v \rangle_{Y' \times Y} \qquad \forall v \in H^1(\Omega). \tag{1.9}$$

1.3. Assumptions on the data. In this paper we make the following assumptions on the data:

$$\text{The coefficients } A,\ b,\ a_0 \text{ and the right-hand side } f \text{ are analytic on } \overline{\Omega} \text{ and the solution } u \in H^{1+\delta}(\Omega) \text{ for some } \delta \in (0,1]. \tag{1.10}$$

Such a situation arises, for example, if the boundary data λ, ψ are not smooth and/or if the domain Ω is merely a Lipschitz domain.

The problem class under consideration may be viewed as a generalization of the setting of the classical BEM in that, while the boundary input data are allowed to be rough, the coefficients of the differential equation are smooth on Ω. The particular case of constant coefficients and homogeneous right-hand side, which is the setting of the BEM, is a special case.

Remark 1.1. The boundary concentrated FEM could also be adapted to the case of piecewise analytic coefficients A, b, a_0 and right-hand side f; see also section 5.3.

Remark 1.2. Methodologically, the analysis of the present paper is closely related to the classical *hp*-FEM [40, 42]. In the classical *hp*-FEM, stronger regularity assumptions are made, namely, piecewise analyticity of the boundary $\partial\Omega$, and the boundary data λ, ψ is stipulated. These stronger regularity assumptions imply stronger regularity results for the solution u. In the classical *hp*-FEM, these stronger regularity assertions for u are exploited to design exponentially convergent methods by using meshes that are graded geometrically towards few singularities located at the boundary. Our weaker regularity assumptions (1.10) require geometric refinement towards the whole boundary and lead to algebraic rates of convergence only. Nevertheless, the algebraic rates obtained in this paper are optimal (in the sense of n-widths) for the class of problems characterized by the regularity assumptions (1.10).

Remark 1.3. Our regularity assumption (1.10) makes strong smoothness assumptions on the right-hand side f. However, the techniques presented in this paper could be employed for methods for solving

$$\mathcal{L}u = f, \qquad \gamma_1 u = \psi,$$

with $f \in H^{-1+\delta}(\Omega)$ and $u \in H^{1+\delta}(\Omega)$, $\delta \in (0,1/2)$. In that case, let $u_0 \in H_0^{1+\delta}(\Omega)$ be the particular solution of $\mathcal{L}u = f$ in Ω solving

$$B(u_0, v) = \int_\Omega f(x) v dx \qquad \forall v \in H_0^1(\Omega). \tag{1.11}$$

For the remaining $\mathcal{L}$-harmonic component of the solution $u_H = u - u_0 \in V_H$, where

$$V_H = \{v \in V : B(v,z) = 0 \quad \forall z \in H^1_0(\Omega)\}, \tag{1.12}$$

we have the equation

$$B(u_H, v) = \int_{\partial\Omega} \psi v ds + \langle \gamma_1 u_0, v\rangle_{H^{-1/2}(\partial\Omega)\times H^{1/2}(\partial\Omega)} \qquad \forall v \in H^1(\Omega) \tag{1.13}$$

and note that u_H solves an equation satisfying the regularity assumptions (1.10). For finite element discretizations, we may use different ansatz spaces to approximate the solutions to (1.11) and (1.13). The only constraint is that these spaces have to provide *the same trace space on* $\partial\Omega$. In particular, we obtain $u = u_0 + u_H$ with $u_0 \in H^{1+\delta}(\Omega)$, $u_H \in \widetilde{\mathcal{B}}^2_{1-\delta}$, where the countably normed space $\widetilde{\mathcal{B}}^2_{1-\delta}$ is defined in (1.17) below.

1.4. Regularity of the solution. The key to efficiently treating (1.2), (1.3) numerically is precise regularity assertions for their solutions. In the case analyzed in the classical *hp*-FEM (see Remark 1.2), the regularity of the solution u is best described in terms of the countably normed spaces $\mathcal{B}^2_\beta$ [4, 5]. This regularity assertion allows for a rigorous proof of exponential convergence of the *hp*-FEM on suitably chosen meshes [40]. We are interested in the case of the weakened regularity assumptions (1.10). That is, we study regularity properties of the solution u to the differential equation

$$\mathcal{L}u = f \qquad \text{on } \Omega, \tag{1.14}$$

where f is analytic on $\overline{\Omega}$; the boundary conditions—of Dirichlet, Neumann, or mixed type—however, may be rough. By standard interior regularity [35, Chapter 5], any solution u to (1.14) is analytic on Ω, but control of higher order derivatives is lost as one approaches the boundary $\partial\Omega$. Nevertheless, it is possible to measure the blow-up of higher order derivatives near the boundary in terms of weighted spaces. A very precise control, which is suitable for the *hp*-FEM error analysis below, is achieved with the countably normed space $\widetilde{\mathcal{B}}^2_\beta$ that we define as follows: For the distance function

$$r(x) := \operatorname{dist}(x, \partial\Omega) \tag{1.15}$$

and $\beta \in [0,1)$ the space $H^2_\beta(\Omega)$ is the completion of $C^\infty(\overline{\Omega})$ under the norm

$$\|u\|^2_{H^2_\beta(\Omega)} := \|u\|^2_{H^1(\Omega)} + \|r^\beta \nabla^2 u\|^2_{L^2(\Omega)}. \tag{1.16}$$

For analytic coefficients in the differential operator $\mathcal{L}$, the regularity of the solutions to (1.14) can be described in terms of countably normed spaces, akin to the spaces $\mathcal{B}^2_\beta(C,\gamma)$ introduced in [4, 5]. Specifically, for C, $\gamma > 0$, $\beta \in [0,1)$ we define $\widetilde{\mathcal{B}}^2_\beta(C,\gamma)$ by

$$\widetilde{\mathcal{B}}^2_\beta(C,\gamma) = \{u \in H^2_\beta(\Omega) \,|\, \|u\|_{H^2_\beta(\Omega)} \le C,\ \|r^{\beta+p}\nabla^{p+2}u\|_{L^2(\Omega)} \le C\gamma^p p!\ \forall p \in \mathbb{N}\}. \tag{1.17}$$

We then have the following result. (See Theorem A.1 for the proof, where in fact the assumptions on the right-hand side f are slightly weaker.)

THEOREM 1.4. *Let Ω be a Lipschitz domain. Let A, b, a_0, f be analytic on $\overline{\Omega}$, and assume that $u \in H^{1+\delta}(\Omega)$, $\delta \in (0,1]$, solves* (1.14). *Then u is analytic on Ω, and there exist C, $\gamma > 0$ depending only on Ω, A, b, a_0, δ, and $\|u\|_{H^{1+\delta}(\Omega)}$ such that*

$$u \in \widetilde{\mathcal{B}}^2_{1-\delta}(C,\gamma).$$

Remark 1.5. An analogous result can be formulated if the data A, b, c, f are piecewise analytic.

It is of interest to state conditions under which a solution u to (1.14) satisfies $u \in H^{1+\delta}(\Omega)$. For example, for a general Lipschitz domain Ω the solution u of the Dirichlet problem (1.2) satisfies the following shift theorem (see [36]):

$$\|u\|_{H^{1+\delta}(\Omega)} \leq C_\delta \left[\|f\|_{H^{-1+\delta}(\Omega)} + \|\lambda\|_{H^{1/2+\delta}(\partial\Omega)}\right], \qquad \delta \in [0, 1/2), \tag{1.18}$$

provided that the right-hand side of (1.18) is finite. Equation (1.18) represents a shift theorem with restriction $\delta \in [0, 1/2)$. Shift theorems where one can shift further (i.e., $\delta \geq 1/2$) are known for piecewise smooth boundaries (e.g., polygons). Using the techniques of [15, 6] shows that for a polygon $\Omega \subset \mathbb{R}^2$ there exists $\delta_0 \in (1/2, 1]$ (depending on Ω and A) such that the solution u of the Dirichlet problem (1.2) satisfies

$$\|u\|_{H^{1+\delta}(\Omega)} \leq C_\delta \left[\|f\|_{H^{-1+\delta}(\Omega)} + \|\lambda\|_{H^{1/2+\delta}(\partial\Omega)}\right], \qquad \delta \in [0, \delta_0). \tag{1.19}$$

We recall further that for convex polygons $\delta_0 = 1$.

2. Discretization by *hp*-FEM.

2.1. Abstract FEM.

2.1.1. Formulation. The FEM is obtained from the weak formulations (1.8), (1.9) by replacing the space V with a finite-dimensional space. For a space $V_N \subset H^1(\Omega)$ the FEM for the Neumann problem (1.9) reads as

$$\text{Find } u_N \in V_N \text{ s.t. } B(u_N, v) = \int_\Omega f\, v\, dx + \langle \psi, \gamma_0 v \rangle_{Y' \times Y} \quad \forall v \in V_N. \tag{2.1}$$

For the Dirichlet problem (1.2), we introduce the space

$$Y_N := V_N|_{\partial\Omega} = \{\gamma_0 v \,|\, v \in V_N\} \subset H^{1/2}(\partial\Omega). \tag{2.2}$$

For an approximation $\lambda_N \in Y_N$ to λ we can then define the FEM for (1.8) as

$$\text{Find } u_N \in V_N \text{ s.t. } u_N = \lambda_N \text{ and } B(u_N, v) = \int_\Omega f\, v\, dx \quad \forall v \in V_N \cap H^1_0(\Omega). \tag{2.3}$$

The coercivity assumption (1.6) ensures existence of the finite element approximation u_N. Furthermore, by Céa's lemma there is $C > 0$ independent of V_N such that

$$\|u - u_N\|_{H^1(\Omega)} \leq C \inf_{v \in V_N} \|u - v\|_{H^1(\Omega)} \tag{2.4}$$

for the solution u_N of (2.1) and

$$\|u - u_N\|_{H^1(\Omega)} \leq C \inf_{\substack{v \in V_N \\ \gamma_0 v = \lambda_N}} \left\{\|u - v\|_{H^1(\Omega)} + \|\lambda_N - \lambda\|_{H^{1/2}(\partial\Omega)}\right\} \tag{2.5}$$

for the solution u_N of the Dirichlet problem (2.3).

In practice, the approximations λ_N are obtained with the aid of a linear operator $P_N : H^{1/2}(\partial\Omega) \to Y_N$ by setting $\lambda_N = P_N\lambda$. In most of the present paper, we will choose this operator P_N to be the L^2-projection Q_N; i.e., for $\lambda \in L^2(\partial\Omega)$ the function $Q_N\lambda$ is defined by

$$\langle Q_N\lambda, v\rangle_{0,\partial\Omega} = \langle \lambda, v\rangle_{0,\partial\Omega} \qquad \forall v \in Y_N. \tag{2.6}$$

2.1.2. Discrete harmonic extension. In this paper, we will consider only families of approximation spaces V_N that are sufficiently large in the following sense: There exists $\widetilde{C} > 0$ such that for all $u \in H^1(\Omega)$ with $u|_{\partial\Omega} \in Y_N$

$$\inf\{\|u - v\|_{H^1(\Omega)} \,|\, v \in V_N \text{ and } v|_{\partial\Omega} = u|_{\partial\Omega}\} \leq \widetilde{C}\|u\|_{H^1(\Omega)}. \tag{2.7}$$

Remark 2.1. Condition (2.7) is satisfied for all approximation spaces considered in this paper. For the standard piecewise linear spaces, condition (2.7) may be verified with the aid of the Clément-type interpolation operator of [41]. For the high order spaces employed in the present paper, the corresponding *hp*-Clément-type interpolation operator is constructed in [34].

Condition (2.7) is required for the discrete harmonic extension to have the following stability properties.

LEMMA 2.2. *Let $\Omega \subset \mathbb{R}^2$ be a Lipschitz domain. Assume that a family of approximation spaces $V_N \subset H^1(\Omega)$ satisfies* (2.7)*. Then there exists $C > 0$ such that the discrete harmonic extension operator $E_N : Y_N \to V_N$ given by*

$$B(E_N u, v) = 0 \qquad \forall v \in V_N \cap H^1_0(\Omega) \tag{2.8}$$

is stable, i.e.,

$$\|E_N u\|_{H^1(\Omega)} \leq C\|u\|_{H^{1/2}(\partial\Omega)} \qquad \forall u \in Y_N.$$

Moreover, the Galerkin orthogonality implies

$$B(z, z) = B(z - E_N(\gamma_0 z), z - E_N(\gamma_0 z)) + B(E_N(\gamma_0 z), E_N(\gamma_0 z)) \qquad \forall z \in V_N.$$

2.2. Geometric meshes and *hp*-FEM spaces.

2.2.1. The geometric mesh. For simplicity of notation, we will restrict our attention to triangulations consisting of *affine triangles*. We emphasize, however, that an extension to quadrilateral elements is possible. The triangulation $\mathcal{T} = \{K\}$ of Ω consists of elements K. Each element K is the image $F_K(\hat{K})$ of the equilateral reference triangle

$$\hat{K} = \left\{(x, y) \,|\, 0 < x < 1,\ 0 < y < \sqrt{3}\left(\frac{1}{2} - \left|x - \frac{1}{2}\right|\right)\right\}$$

under the *affine* map F_K. We furthermore assume that the triangulation $\mathcal{T}$ is γ-*shape-regular*, i.e.,

$$h_K^{-1}\|F'_K\|_{L^\infty(T)} + h_K\|(F'_K)^{-1}\|_{L^\infty(T)} \leq \gamma \qquad \forall K \in \mathcal{T}. \tag{2.9}$$

Here, h_K denotes the diameter of the element K. Of particular importance to us will be the "geometric meshes," which are strongly refined meshes near the boundary $\partial\Omega$, defined as follows.

DEFINITION 2.3 (geometric mesh). *A γ-shape-regular (cf.* (2.9)*) mesh $\mathcal{T}$ is called a* geometric mesh *with* boundary mesh size h *if there exist c_1, $c_2 > 0$ such that for all $K \in \mathcal{T}$ the following hold:*

1. *If $\overline{K} \cap \partial\Omega \neq \emptyset$, then $h \leq h_K \leq c_2 h$;*
2. *if $\overline{K} \cap \partial\Omega = \emptyset$, then $c_1 \inf_{x\in K} \operatorname{dist}(x, \partial\Omega) \leq h_K \leq c_2 \sup_{x\in K} \operatorname{dist}(x, \partial\Omega)$.*

A typical example of a geometric mesh is depicted in Figure 2.1. Note that the restriction to the boundary $\partial\Omega$ of a geometric mesh is a quasi-uniform mesh, which justifies speaking of a "boundary mesh size h."

Remark 2.4. An important algorithmic issue is the automatic generation of geometric meshes. Such meshes can be generated with the algorithm of [39].

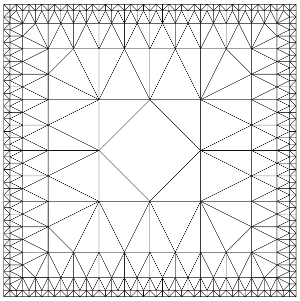


FIG. 2.1. *Example of a geometric mesh in the sense of Definition* 2.3.

2.2.2. *hp*-FEM spaces. In order to define *hp*-FEM spaces on a mesh $\mathcal{T}$, we associate a polynomial degree $p_K \in \mathbb{N}$ with each element K, collect these p_K in the polynomial degree vector $\mathbf{p} := (p_K)_{K\in\mathcal{T}}$, and set

$$S^{\mathbf{p}}(\Omega,\mathcal{T}) := \{u \in H^1(\Omega)\,|\, u\circ F_K \in \mathcal{P}_{p_K}(\hat{K}) \quad \forall K \in \mathcal{T}\}, \tag{2.10}$$

$$S^{\mathbf{p}}_0(\Omega,\mathcal{T}) := S^{\mathbf{p}}(\Omega,\mathcal{T}) \cap H^1_0(\Omega), \tag{2.11}$$

where for $p \in \mathbb{N}$ we introduced the space of all polynomials of degree p as

$$\mathcal{P}_p(\hat{K}) = \operatorname{span}\{x^i y^j \,|\, 0 \le i+j \le p\}.$$

For the approximation of solutions to (1.14) on geometric meshes (in the sense of Definition 2.3), the so-called *linear degree vector* is a particularly useful polynomial degree distribution.

DEFINITION 2.5 (linear degree vector). *Let $\mathcal{T}$ be a geometric mesh in the sense of Definition* 2.3. *A polynomial degree vector $\mathbf{p} = (p_K)_{K\in\mathcal{T}}$ is said to be a* linear degree vector with slope $\alpha > 0$ *if*

$$1 + \alpha c_1 \log\frac{h_K}{h} \le p_K \le 1 + \alpha c_2 \log\frac{h_K}{h}.$$

Here, $h := \min_{K\in\mathcal{T}}$ is a measure for the mesh size of the quasi-uniform mesh $\mathcal{T}|_{\partial\Omega}$.

Remark 2.6. Linear degree vectors $\mathbf{p}$ have the additional property that the polynomial degree varies slowly; i.e., there exists $C > 0$ such that

$$C^{-1}p_{K'} \le p_K \le Cp_{K'} \qquad \forall K, K' \in \mathcal{T} \quad \text{with } \overline{K} \cap \overline{K'} \neq \emptyset. \tag{2.12}$$

We conclude this section by showing that for geometric meshes in the sense of Definition 2.3, the number of elements of $\mathcal{T}$ is proportional to the number of elements on the boundary. Similarly, for linear degree vectors (Definition 2.5), the dimension $\dim S^{\mathbf{p}}(\Omega,\mathcal{T})$ is proportional to the number of unknowns on the boundary as is seen in the following result.

PROPOSITION 2.7. *Let $\mathcal{T}$ be a geometric mesh with boundary mesh size h. Let $\mathbf{p}$ be a linear degree vector with slope $\alpha > 0$ on $\mathcal{T}$. Then there exists $C > 0$ depending only on Ω, the shape-regularity constant γ, and the constants c_1, c_2, α of Definitions* 2.3, 2.5 *such that*

$$\sum_{K\in\mathcal{T}} 1 \le Ch^{-1},$$

$$\dim S^{\mathbf{p}}(\Omega, \mathcal{T}) \sim \sum_{K \in \mathcal{T}} p_K^2 \leq C h^{-1},$$
$$\max_{K \in \mathcal{T}} p_K \leq C |\log h|.$$

Proof. We will prove only the second estimate, as the first one is proved similarly.

$$\sum_{K \in \mathcal{T}} p_K^2 = \sum_{K \in \mathcal{T} : \overline{K} \cap \partial\Omega \neq \emptyset} p_K^2 + \sum_{K \in \mathcal{T} : \overline{K} \cap \partial\Omega = \emptyset} p_K^2. \tag{2.13}$$

For the first sum, we note that the assumptions on a geometric mesh $\mathcal{T}$ and the linear degree vector give that $p_K \leq C$ for all $K \in \mathcal{T}$ with $\overline{K} \cap \partial\Omega \neq \emptyset$. Thus,

$$\sum_{K \in \mathcal{T} : \overline{K} \cap \partial\Omega \neq \emptyset} p_K^2 \leq C \sum_{K \in \mathcal{T} : \overline{K} \cap \partial\Omega \neq \emptyset} 1 \leq C h^{-1}.$$

For the second sum in (2.13) we bound

$$\begin{aligned} \sum_{K \in \mathcal{T} : \overline{K} \cap \partial\Omega = \emptyset} p_K^2 &\leq C \sum_{K \in \mathcal{T} : \overline{K} \cap \partial\Omega = \emptyset} \int_K \frac{1 + |\ln(r(x)/h)|^2}{r^2(x)}\, dx \\ &\leq C \int_{x \in \Omega, r(x) \geq ch} \frac{1 + |\ln(r(x)/h)|^2}{r^2(x)}\, dx \leq C \int_{c'h}^{\infty} \frac{1 + |\ln(z/h)|^2}{z^2}\, dz \leq C h^{-1}, \end{aligned}$$

where in the penultimate step we have locally flattened the boundary with Lipschitz maps. The integral represents the integration normal to the boundary, whereas the integration in the tangential direction was absorbed in the constant C. □

2.3. *hp*-FEM approximation on geometric meshes.

2.3.1. Approximation on the boundary $\partial\Omega$. If $\mathcal{T}$ is a geometric mesh and $V_N = S^{\mathbf{p}}(\Omega, \mathcal{T})$ with linear degree vector $\mathbf{p}$, then the space Y_N defined in (2.2) is a space of piecewise polynomials of fixed, low degree (depending on α and the constants c_1, c_2 appearing in Definition 2.5) on a quasi-uniform mesh. It can be shown (with the aid of Proposition C.3) that for $0 \leq s < 3/2$ the L^2-projector Q_N is stable on $H^s(\partial\Omega)$; in particular, therefore, $Y_N \subset H^s(\partial\Omega)$ for $0 \leq s < 3/2$. This allows us to extend the operator Q_N by duality to an operator $H^{-s}(\partial\Omega) \to Y_N$ with $0 \leq s < 3/2$ by

$$\langle Q_N u, v \rangle_{0,\partial\Omega} = \langle u, v \rangle_{H^{-s}(\partial\Omega) \times H^s(\partial\Omega)} \qquad \forall u \in H^{-s}(\partial\Omega) \quad \forall v \in Y_N. \tag{2.14}$$

By a slight abuse of notation, the extended operator is again denoted by Q_N. It has the following properties.

LEMMA 2.8. *Let Ω be a polygon, let $\mathcal{T}$ be a geometric mesh with boundary mesh size h in the sense of Definition* 2.3, *and let $\mathbf{p}$ be a linear degree vector given by Definition* 2.5. *Set $V_N := S^{\mathbf{p}}(\Omega, \mathcal{T})$, and let Y_N be defined by* (2.2), *and the L^2-projection Q_N be given by* (2.6) *and* (2.14). *Then*

$$\|Q_N u\|_{H^s(\partial\Omega)} \leq C_s \|u\|_{H^s(\partial\Omega)} \quad \forall u \in H^s(\partial\Omega), \quad 0 \leq |s| < 3/2, \tag{2.15}$$

$$\|u - Q_N u\|_{H^s(\partial\Omega)} \leq C_{s,s'} h^{s'-s} \|u\|_{H^{s'}(\partial\Omega)} \quad \forall u \in H^{s'}(\partial\Omega), \tag{2.16}$$

where $-3/2 < s \leq s' < 3/2$. The constants C_s, $C_{s,s'}$ depend only on Ω, s, s', and the constants appearing in Definitions 2.3, 2.5.

Proof. Let $\varphi : [0, L) \to \partial\Omega$ be a parametrization by arclength. The fact that Ω is a polygon together with Lemma C.1 implies that the map $u \mapsto u \circ \phi$ is an isomorphism between the Sobolev spaces $H^s(\partial\Omega)$ and $H^s_{per}([0, L))$, $0 \leq s < 3/2$. The stability result (2.15) for $s \geq 0$ now follows from Proposition C.3 and by duality for $s \in (-3/2, 0)$. For $0 \leq s$, the approximation result (2.16) follows from (2.15) and standard approximation results in the usual way. The case $s \in (-3/2, 0)$ is again obtained by duality. □

2.3.2. Approximation of $\widetilde{\mathcal{B}}^2_\beta$-functions from $S^{\mathbf{p}}(\Omega, \mathcal{T})$. Our *hp*-FEM approximation results for functions of $\widetilde{\mathcal{B}}^2_\beta$ will be based on the following lemma.

LEMMA 2.9. *Let $\hat{K}$ be the reference triangle with edges Γ_i, $i \in \{1, 2, 3\}$. Let $\hat{u}$ be analytic on $\overline{\hat{K}}$, and assume that*

$$\|\nabla^{n+2}\hat{u}\|_{L^2(\hat{K})} \leq C_u \gamma_u^n n! \qquad \forall n \in \mathbb{N}_0$$

for some C_u, $\gamma_u > 0$. Let $c \in (0, 1]$. Then for each p, p_1, p_2, $p_3 \in \mathbb{N}$ with

$$cp \leq p_i \leq p, \qquad i \in \{1, 2, 3\},$$

there exists a polynomial $\pi_p \in \mathcal{P}_p(\hat{K})$ with

1. *$\pi_p|_{\Gamma_i} = i_{p_i,\Gamma_i}(u|_{\Gamma_i})$ for $i \in \{1, 2, 3\}$. Here, i_{p_i,Γ_i} denotes the Gauss–Lobatto interpolant of degree p_i on edge Γ_i.*
2. *$\|u - \pi_p\|_{W^{1,\infty}(\hat{K})} \leq CC_u e^{-bp}$.*

The constants C, $b > 0$ depend only on c and γ_u.

Proof. The case $p_1 = p_2 = p_3 = p$ is considered in [32, Theorem 3.2.20, Proposition 3.2.21]. The extension to the present case is straightforward. □

PROPOSITION 2.10. *Let $\mathcal{T}$ be a geometric mesh with boundary mesh size h, as defined in Definition* 2.3*. Let* $\mathbf{p}$ *be a linear degree vector on $\mathcal{T}$ with slope $\alpha > 0$. Let $u \in \widetilde{\mathcal{B}}^2_\beta(C_u, \gamma_u)$ for some $\beta \in [0, 1)$, C_u, $\gamma_u > 0$. Then there exist C, $b > 0$ depending only on Ω, the shape-regularity constant γ, and the constants c_1, c_2 of Definition* 2.3 *as well as C_u, γ_u, β such that*

$$\inf\left\{\|u - v\|_{H^1(\Omega)} \,|\, v \in S^{\mathbf{p}}(\Omega, \mathcal{T})\right\} \leq Ch^{1-\beta} + Ch^{b\alpha}. \tag{2.17}$$

In terms of degrees of freedom, we have $N = \dim S^{\mathbf{p}}(\Omega, \mathcal{T}) \sim h^{-1}$.

Proof. For $u \in \widetilde{\mathcal{B}}^2_\beta(C_u, \gamma_u)$ we define

$$C_K^2 := \sum_{n=0}^{\infty} \frac{1}{(2\gamma_u)^{2n}(n!)^2} \|r^{n+\beta}\nabla^{n+2}u\|^2_{L^2(K)}.$$

The assumption $u \in \widetilde{\mathcal{B}}^2_\beta(C_u, \gamma_u)$ guarantees

$$\sum_{K\in\mathcal{T}} C_K^2 \leq \sum_{n=0}^{\infty} \frac{1}{(2\gamma_u)^{2n}(n!)^2} \|r^{n+\beta}\nabla^{n+2}u\|^2_{L^2(\Omega)} \leq C_u^2 \sum_{n=0}^{\infty} \frac{1}{2^{2n}} = \frac{4}{3}C_u^2.$$

Hence, we conclude that $u \in \widetilde{\mathcal{B}}^2_\beta(C_u, \gamma_u)$ implies

$$\|r^{n+\beta}\nabla^{n+2}u\|_{L^2(K)} \leq C_K(2\gamma_u)^n n! \qquad \forall n \in \mathbb{N}_0, \quad \forall K \in \mathcal{T}, \tag{2.18a}$$

$$\sum_{K\in\mathcal{T}} C_K^2 \leq \frac{4}{3}C_u^2. \tag{2.18b}$$

We explicitly construct an element of $S^{\mathbf{p}}(\Omega, \mathcal{T})$ with the desired approximation properties. To that end, we first assume, as we may, that $p_K = 1$ for all elements abutting $\partial\Omega$. Next, we associate with each edge e of $\mathcal{T}$ a polynomial degree $p_e := \min\{p_K \,|\, e \text{ is an edge of element } K\}$. After these preparations, we construct the approximant element by element. We distinguish the cases $\overline{K} \cap \partial\Omega \neq \emptyset$ and $\overline{K} \cap \partial\Omega = \emptyset$.

Using Theorem B.4 and a scaling argument, we obtain for all elements K abutting on $\partial\Omega$ that the linear interpolant Iu satisfies

$$\|u - Iu\|_{H^1(K)} \leq C h_K^{1-\beta} \|r^\beta \nabla^2 u\|_{L^2(K)} \leq C h^{1-\beta} C_K.$$

For the elements not abutting on $\partial\Omega$, we employ Lemma 2.9. Using (2.18a), we see that the pull-back $\hat{u} := u \circ F_K$ satisfies

$$\begin{aligned} \|\nabla^{n+2} \hat{u}\|^2_{L^2(\hat{K})} &\leq C h_K^{2(n+1)} \|\nabla^{n+2} u\|^2_{L^2(K)} \leq C h_K^{2-2\beta} \|r^{n+\beta} \nabla^{n+2} u\|^2_{L^2(K)} \\ &\leq C C_K \, h_K^{2(1-\beta)} (2\gamma_u)^{2n} (n!)^2, \end{aligned}$$

with $C > 0$ independent of n and K. The approximant Iu of Lemma 2.9 then satisfies

$$\|u - Iu\|_{H^1(K)} \leq C C_K h_K^{1-\beta} e^{-b p_K}$$

for some C, $b > 0$ independent of the element K. We note that the interpolant constructed elementwise in this fashion is indeed an element of $S^{\mathbf{p}}(\Omega, \mathcal{T})$. (The edge polynomial degrees p_i in Lemma 2.9 are taken to be the polynomial degrees p_e of the corresponding edges e.) Using $p_K \geq c\alpha \ln(h_K/h)$, we arrive at

$$\|u - Iu\|_{H^1(K)} \leq C C_K h_K^{1-\beta-\alpha b'} h^{\alpha b'}$$

for some $b' > 0$. Exploiting that $h_K \geq ch$, a simple calculation reveals that

$$h_K^{1-\beta-\alpha b'} h^{\alpha b'} \leq C h^{\min\{1-\beta, \alpha b'\}}.$$

We thus conclude in view of (2.18b) that

$$\sum_{K \in \mathcal{T}} \|u - Iu\|^2_{H^1(K)} \leq C \sum_{K \in \mathcal{T}} C_K^2 h^{\min\{1-\beta, \alpha b'\}} \leq C h^{\min\{1-\beta, \alpha b'\}},$$

which is the desired estimate. The bound for the dimension of $S^{\mathbf{p}}(\Omega, \mathcal{T})$ follows from Proposition 2.7. □

Remark 2.11. The meshes $\mathcal{T}$ considered here consist of triangles only. Likewise, the approximation result in Proposition 2.10 is formulated for triangles only. This restriction is not essential and was done for simplicity of exposition only. The approximation results can be formulated for meshes consisting of nonaffine elements (quadrilaterals, curved elements) as well. To handle this case, it is required that the element maps F_K for elements K not abutting on the boundary be analytic (with a controlled domain of analyticity; see, e.g., [32]) and that the error on elements abutting $\partial\Omega$ be $O(h^{1-\beta})$.

Proposition 2.10 is a result for unconstrained approximation in $H^1(\Omega)$. For treating Dirichlet problems, constrained approximation as in (2.5) is required. This is accomplished in the following corollary.

COROLLARY 2.12. *Assume the hypotheses of Proposition* 2.10 *and additionally* $u \in H^{2-\beta}(\Omega) \cap \widetilde{B}^2_\beta(C_u, \gamma_u)$ *for some* C_u, $\gamma_u > 0$, $\beta \in (0, 1)$. *Let* Y_N *be the restriction* *of $S^{\mathbf{p}}(\Omega, \mathcal{T})$ to $\partial\Omega$ as given by* (2.2)*. Let Q_N be the L^2-projection into Y_N (cf.* (2.6)*). Then*

$$\inf\{\|u - v\|_{H^1(\Omega)} \mid v \in S^{\mathbf{p}}(\Omega, \mathcal{T}) \text{ with } \gamma_0 v = Q_N(\gamma_0 u)\} \le C\left[h^{1-\beta} + h^{b\alpha}\right]. \tag{2.19}$$

The constants C, $b > 0$ depend only on Ω, the shape-regularity constant γ, the constants c_1, c_2 appearing in Definition 2.3*, and C_u, γ_u, β, $\|u\|_{H^{2-\beta}(\Omega)}$.*

Proof. We first observe that the trace theorem gives $\|\gamma_0 u\|_{H^{3/2-\beta}(\partial\Omega)} \le C\|u\|_{H^{2-\beta}(\Omega)}$. Lemma 2.8 and Proposition 2.10 therefore imply the existence of $v_N \in S^{\mathbf{p}}(\Omega, \mathcal{T})$ such that

$$\|u - v_N\|_{H^1(\Omega)} \le C\left[h^{1-\beta} + h^{b\alpha}\right],$$
$$\|\gamma_0 u - Q_N(\gamma_0 u)\|_{H^{1/2}(\partial\Omega)} \le Ch^{1-\beta}\|u\|_{H^{2-\beta}(\Omega)}.$$

The desired result now follows with the aid of the discrete harmonic extension operator E_N of Lemma 2.2: Since $\gamma_0 v_N - Q_N(\gamma_0 u) \in Y_N$, we get that the function $\widetilde{v}_N := v_N - E_N(\gamma_0 v_N - Q_N(\gamma_0 u)) \in V_N$ satisfies $\gamma_0 \widetilde{v}_N = Q_N(\gamma_0 u)$ and

$$\begin{aligned}\|\widetilde{v}_N - u\|_{H^1(\Omega)} &\le \|u - v_N\|_{H^1(\Omega)} + \|E_N(\gamma_0 v_N - Q_N(\gamma_0 u))\|_{H^1(\Omega)}\\ &\le \|u - v_N\|_{H^1(\Omega)} + C\|\gamma_0 v_N - Q_N(\gamma_0 u)\|_{H^{1/2}(\partial\Omega)}\\ &\le \|u - v_N\|_{H^1(\Omega)} + C\left[\|\gamma_0(v_N - u)\|_{H^{1/2}(\partial\Omega)} + \|\gamma_0 u - Q_N(\gamma_0 u)\|_{H^{1/2}(\partial\Omega)}\right].\end{aligned}$$

The result now follows. □

Corollary 2.12 allows us to finally formulate an approximation result for the *hp*-FEM on geometric meshes applied to (1.2) and (1.3) as follows.

Theorem 2.13. *Let $\mathcal{T}$ be a geometric mesh with boundary mesh size h, as defined in Definition* 2.3*. Let $\mathbf{p}$ be a linear degree vector on $\mathcal{T}$ with slope $\alpha > 0$ (cf. Definition* 2.5*). Let Q_N be the L^2-projection onto $Y_N = S^{\mathbf{p}}(\Omega, \mathcal{T})|_{\partial\Omega}$.*

Let $u \in H^{1+\delta}(\Omega)$, $\delta \in (0, 1)$, be the solution to (1.2) *(resp., the solution to* (1.3)*) with coefficients A, b, a_0, and right-hand side f analytic on $\overline{\Omega}$. Then the FE-solution u_N given by* (2.3) *(resp.,* (2.1)*) satisfies*

$$\|u - u_N\|_{H^1(\Omega)} \le C\left[h^{\delta} + h^{b\alpha}\right]. \tag{2.20}$$

The constants C, $b > 0$ depend only on the shape-regularity constant γ, the constants c_1, c_2 appearing in Definition 2.3*, and the data A, b, c, f, Ω.*

Proof. Theorem 1.4 implies that the solution $u \in H^{1+\delta}(\Omega)$ is in $\widetilde{B}^2_{1-\delta}(C_u, \gamma_u)$ for some C_u, $\gamma_u > 0$. In view of the best approximation properties (2.5), (2.4), the assertion (2.20) now follows from Corollary 2.12. □

For α sufficiently large the boundary concentrated FEM achieves the optimal rate of convergence

$$\|u - u_N\|_{H^1(\Omega)} \le Ch^{\delta} = O(N^{-\delta}).$$

3. *hp*-FEM solution procedure.

Choosing the slope α of the polynomial degree vector sufficiently large, we obtain for the FEM approximation the optimal rate $\|u - u_N\|_{H^1(\Omega)} \le CN^{-\delta}$. In the present section we discuss how the FE solution u_N can be computed with complexity $O(N \log^2 N)$. We mention in passing that, in the two-dimensional situation, a direct solver with complexity $O(N \log^8 N)$ was constructed in [26]. We consider iterative methods for solving the Dirichlet and Neumann problems

Table 3.1
Conditioning of the hp-FEM stiffness matrices: $N = \#$ elements, $p = \max_{K\in\mathcal{T}} p_K$.

Bdy. cond.	$p = p(N)$	DOF	cond (A^c)	cond $(C^{-1}A)$
Dirichlet	$O(\log N)$	N	$O(p(1+\log p))$	$O(1+\log^2 p)$
Neumann	$O(\log N)$	N	$O(Np(1+\log p))$	$O(1+\log^2 p)$

on geometric meshes in the sense of Definition 2.3. We restrict our attention to the symmetric positive definite case; i.e., in (1.4) we take

$$b_0 = 0, \quad a_0 > 0.$$

The main results of this section are collected in Table 3.1. A is the stiffness matrix, A^c stands for the statically condensed stiffness matrix (with the shape functions discussed in Example 3.1), and C stands for the preconditioners proposed here.

We focus here on the design of preconditioners for Neumann problems. The reason for our concentrating on this case can be seen in Table 3.1: While the Dirichlet problem is fairly well conditioned (the condition number without preconditioning grows only polylogarithmically in N since $p = O(\log N)$), the Neumann problem leads to at least linearly (in N) growing conditioning numbers, thus requiring preconditioning.

Our approach for the design of a preconditioner for the Neumann problem is based on the results of [3] (see also [40, section 4.7]).

3.1. Shape functions and assembling.

3.1.1. Element shape functions. In order to set up the stiffness matrix, bases of the polynomial spaces $\mathcal{P}_p$ have to be chosen. It is customary in p- and hp-FEM to split a basis of $\mathcal{P}_{p_K}$ into *vertex*, *side*, and *internal* shape functions. These three types of shape functions are characterized as follows:

1. *Vertex shape functions $\mathcal{V}$.* These are the usual linear nodal shape functions, which are equal to one in one node and vanish on the edge opposite that node. We write $\widetilde{\mathcal{V}} = \operatorname{span} \mathcal{V}$.
2. *Side shape functions $\mathcal{S} = \mathcal{S}_1 \cup \mathcal{S}_2 \cup \mathcal{S}_3$.* The side shape functions from $\mathcal{S}_i$ are associated with the edge Γ_i of $\partial\hat{K}$ and vanish on the edges Γ_j for $j \neq i$. We write $\widetilde{\mathcal{S}}_i = \operatorname{span} \mathcal{S}_i$ and $\widetilde{\mathcal{S}} := \operatorname{span}\{\mathcal{S}_1 \cup \mathcal{S}_2 \cup \mathcal{S}_3\}$.
3. *Internal shape functions $\mathcal{I}$.* The functions from $\mathcal{I}$ vanish on $\partial\hat{K}$. We write $\widetilde{\mathcal{I}} = \operatorname{span} \mathcal{I}$.

The side and internal shape functions are not uniquely defined with the above separation. An important consideration for an actual choice is the conditioning of the resulting stiffness matrix. We will discuss this point further below. One possible choice of a basis of $\mathcal{P}_p$ is based on Lagrange interpolation polynomials with respect to the Gauss–Lobatto points on the sides, which we elaborate in the following example.

Example 3.1. Denote by v_i, $i = 1, \dots, 3$, the three vertices of $\hat{K}$, and by Γ_i, $i = 1, \dots, 3$, the three edges (we assume $\Gamma_1 = \{(x, 0) \in \mathbb{R}^2 \,|\, 0 < x < 1\}$). Let $p_i \in \mathbb{N}$ be polynomial degrees associated with the edges Γ_i, and let $p \in \mathbb{N}$ be the polynomial degree of the internal shape functions. We then define vertex shape functions $\mathcal{V}$, side shape functions $\mathcal{S} = \mathcal{S}_1 \cup \mathcal{S}_2 \cup \mathcal{S}_3$, and internal shape functions $\mathcal{I}$ as follows:

$$\mathcal{V} := \text{the usual linear nodal shape functions } n_i \text{ with } n_i(v_j) = \delta_{ij},$$

$$\mathcal{S}_1 := \left\{ l_{j,p_1}(x) \frac{y - \sqrt{3}x}{x} \frac{y + \sqrt{3}(x-1)}{1-x} \,\Big|\, j = 1, \dots, p_1 - 1 \right\},$$

$$\mathcal{I} := \left\{ y(y - \sqrt{3}x)(y + \sqrt{3}(x-1))L_i(x)L_j(y) \,|\, 0 \le i + j \le \frac{(p-3)(p-2)}{2} \right\}.$$

Here, the polynomials L_i are the Legendre polynomials scaled to the interval $[0,1]$. The polynomials l_{j,p_1} are the Lagrange interpolation polynomials with respect to the $p_1 + 1$ Gauss–Lobatto points on $[0,1]$; letting $0 = x_0 < x_1 < \cdots < x_{p_1} = 1$ be the zeros of the polynomials $x \mapsto x(1-x)L'_{p_1}(x)$, the functions $l_{j,p_1}(x)$ are defined by

$$l_{j,p_1}(x) := \prod_{\substack{i=0\\ i\neq j}}^{p_1} \frac{x - x_i}{x_j - x_i}, \qquad j = 0, \dots, p_1. \tag{3.1}$$

The side shape functions $\mathcal{S}_2$, $\mathcal{S}_3$ are obtained similarly, with p_1 replaced with p_2 (resp., p_3) and a suitable coordinate transformation.

3.1.2. Global bases and assembling. The decomposition of a basis $\mathcal{B}_K$ of $\mathcal{P}_{p_K}$ facilitates the assembly process in *hp*-FEM with variable polynomial degree. For a detailed discussion of this procedure, we refer, for example, to [37, 13]. The basis of Example 3.1, however, may serve to illustrate the main point. The topological entities "edge" and "element" carry a polynomial degree: The polynomial degree associated with an element K is p_K, whereas the polynomial degree p_e associated with an edge $e = \overline{K} \cap \overline{K'}$ is $p_e := \min\{p_K, p_{K'}\}$. These edge polynomial degrees p_e then correspond to the polynomial degrees p_i, $i = 1, \dots, 3$, in Example 3.1; since the side shape functions are Lagrange interpolation polynomials on the edges, the assembly process is straightforward.

We introduce the *assembly operator* $\mathcal{A}_{K\in\mathcal{T}}$ of [23] to combine the bases $\mathcal{B}_K$ of the spaces $\mathcal{P}_{p_K}$ into a global basis $\mathcal{B}$ of the FE-space V_N:

$$\mathcal{B} = \mathop{\mathcal{A}}_{K\in\mathcal{T}} \mathcal{B}_K.$$

One can also assemble only the functions from $\mathcal{V}$, $\mathcal{S}$, or $\mathcal{I}$:

$$V_{\mathcal{V}} = \mathop{\mathcal{A}}_{K\in\mathcal{T}} \mathcal{V}_K, \qquad V_{\mathcal{S}} = \mathop{\mathcal{A}}_{K\in\mathcal{T}} \mathcal{S}_K, \qquad V_{\mathcal{I}} = \mathop{\mathcal{A}}_{K\in\mathcal{T}} \mathcal{I}_K. \tag{3.2a}$$

Of course, the functions of $V_{\mathcal{V}}$ are just the standard piecewise linear hat functions spanning the space $S^1(\Omega, \mathcal{T})$. The shape functions of $V_{\mathcal{I}}$ are supported by a single element, and the shape functions of $V_{\mathcal{S}}$ are supported by at most two adjacent elements. We can further split $V_{\mathcal{S}}$ into

$$V_{\mathcal{S}} = \oplus_{\text{edges } e} V_e, \qquad V_e = \{v \in V_{\mathcal{S}} \,|\, v_{|_{e'}} = 0 \quad \text{for all edges } e' \neq e\}. \tag{3.2b}$$

3.2. Cost of setting up the stiffness matrix and local condensation. We first show that setting up the stiffness matrix and the optional local static condensation can be performed with optimal complexity on geometric meshes with linear degree vectors. To see that, we introduce the elementwise bilinear form B_K by restricting B to the element K:

$$B_K(u, v) := \int_K (A\nabla u \cdot \nabla v + a_0 uv)\, dx.$$

Furthermore, for functions u, $v \in V_N$, we write $u_K := u|_K$, $v_K := v|_K$. With this understanding we can write

$$B(u, v) = \sum_{K\in\mathcal{T}} B_K(u_K, v_K). \tag{3.3}$$

The actual evaluation of $B_K(u_K, v_K)$ is performed by integrating on the reference element $\hat{K}$ instead of K. That is, writing $\hat{u} = u_K \circ F_K$, $\hat{v} = v_K \circ F_K$, we set $B_K(u,v) := \hat{B}_K(\hat{u},\hat{v})$, where

$$\begin{aligned}
&\hat{B}_K(\hat{u},\hat{v}) := \int_{\hat{K}} \hat{A}\nabla\hat{u}\cdot\hat{v} + \hat{a}_0\hat{u}\hat{v}\,dx,\\
&\hat{A} := \left(F_K'\right)^{-1}(A\circ F_K)\left(F_K'\right)^{-\top}|\det F_K'|, \qquad \hat{a}_0 = a_0\circ F_K\cdot|\det F_K'|.
\end{aligned}$$

In practice, the integration over $\hat{K}$ that is required for the evaluation of B_K is performed with a quadrature rule with $O(p_K^2)$ points. In a standard *hp*-FEM code, the bilinear form B_K defines the element stiffness matrix

$$A_K = \left(B_K(u,v)\right)_{u,v\in\mathcal{V}\cup\mathcal{S}\cup\mathcal{I}},$$

which is an $O(p_K)\times O(p_K)$-matrix. Finally, in the assembly the local stiffness matrices A_K are combined into the global stiffness matrix A. As in [23], we write this assembly process as $A = \mathcal{A}_{K\in\mathcal{T}} A_K$.

In *hp*-FEM it is also customary to perform local static condensation. The partitioning of the basis of $\mathcal{P}_K$ in vertex, side, and internal shape functions implies a corresponding block structure of A_K:

$$A_K = \begin{pmatrix} A_K^{\mathcal{V}\mathcal{V}} & A_K^{\mathcal{V}\mathcal{S}} & A_K^{\mathcal{V}\mathcal{I}} \\ & A_K^{\mathcal{S}\mathcal{S}} & A_K^{\mathcal{S}\mathcal{I}} \\ sym. & & A_K^{\mathcal{I}\mathcal{I}} \end{pmatrix}.$$

Since $A_K^{\mathcal{I}\mathcal{I}}$ is invertible, one can form the following Schur complement:

$$A_K^c := A_K^{\mathcal{E}\mathcal{E}} - A_K^{\mathcal{E}\mathcal{I}}\left(A_K^{\mathcal{I}\mathcal{I}}\right)^{-1}\left(A_K^{\mathcal{E}\mathcal{I}}\right)^{\top},$$

where we introduce the notion of *external shape functions*

$$\mathcal{E} := \mathcal{V}\cup\mathcal{S}. \tag{3.4}$$

The condensed global stiffness matrix A^c is obtained by assembling the condensed local matrices A_K^c:

$$A^c := \underset{K\in\mathcal{T}}{\mathcal{A}} A_K^c. \tag{3.5}$$

An important observation is that the condensed stiffness matrix A^c is the stiffness matrix corresponding to elementwise discrete harmonic external shape functions.

PROPOSITION 3.2. *Let $\mathcal{T}$ be a geometric mesh with boundary mesh size h, and let $\mathbf{p}_K$ be a linear degree vector with slope α in the sense of Definition* 2.5. *Assume that for all elements quadrature rules with $O(p_K^2)$ points are used. Then the stiffness matrix A is generated with work*

$$W(A) = O(N),$$

where $N \sim h^{-1}$. Additionally, the local static condensation, i.e., forming the Schur complement with respect to the internal shape functions, is performed with work $W(A^c) = O(N)$.

Proof. The cost of setting up the local stiffness matrix with $O(p_K^2 \times p_K^2)$ entries using numerical quadrature with $O(p_K^2)$ points is $O(p_K^6)$. Thus, the total cost to set up the global stiffness matrix is

$$W \sim \sum_{K \in \mathcal{T}} (p_K^2)^3 = O(N),$$

where the last bound is obtained by arguments similar to those in the proof of Proposition 2.7. Also computing the condensed stiffness matrix A_K^c is done with work $O((p_K^2)^3) = O(p_K^6)$. Again, summing this work estimate over all elements K of the geometric mesh, we arrive at $W(A^c) = O(N)$. □

Remark 3.3. The presence of numerical quadrature introduces variational crimes. It can be shown (see [33]) that the bilinear form $\tilde{B}$ obtained by numerical quadrature induces an inner product on V_N that is equivalent to the inner product generated by B, and the approximation result Theorem 2.13 still holds.

3.3. The Dirichlet problem. The aim of the present section is to show that the stiffness matrix resulting from the discretization of a Dirichlet problem is fairly well conditioned. We have the following result.

PROPOSITION 3.4. *Given a geometric mesh $\mathcal{T}$ with boundary mesh size h, let $\mathbf{p}$ be a polynomial degree distribution satisfying* (2.12)*, and let the element shape functions be taken as described in Example* 3.1. *Then there exists $C > 0$ independent of h and $\mathbf{p}$ such that the condensed stiffness matrix A^c corresponding to the Dirichlet problem has l^2-condition number*

$$\kappa(A^c) \leq C|\mathbf{p}|(1 + \log|\mathbf{p}|), \tag{3.6}$$

where $|\mathbf{p}| = \max_{K\in\mathcal{T}} p_K$. In the case of Neumann boundary conditions, there holds

$$\kappa(A^c) \leq Ch^{-1}\,|\mathbf{p}|(1 + \log|\mathbf{p}|). \tag{3.7}$$

Proof. For the case of Dirichlet boundary conditions, we refer to [31, Theorem 2.2]. The case of the Neumann boundary conditions is obtained by combining the results of [31] for the p-dependence with those of [7, Theorem 4.1] for the h-dependence.

We mention that the $O(1)$-condition number estimate for $p = 1$ and Dirichlet boundary conditions has already been proved in [44]. □

Applied to the case of linear degree vector $\mathbf{p}$, Proposition 3.4 yields that in the case of a Dirichlet problem the condensed stiffness matrix satisfies

$$\kappa(A^c) \leq C \log N(1 + \log\log N)$$

because $|\mathbf{p}| \leq C \log N = C|\log h|$. Thus, solving Dirichlet problems on geometric meshes can be accomplished efficiently by simple CG-iterations. We remark that the condition number $\kappa(A)$ of the full hp-FEM stiffness matrix can be shown to be $O(|\mathbf{p}|^q)$ for suitable $q \geq 1$; thus, CG-iterations again lead to linear-logarithmic complexity.

Applied to the Neumann problem on geometric meshes with linear degree vectors, Proposition 3.4 yields a bound of the form $\kappa(A^c) \leq CN \log N(1 + \log\log N)$. In this case, preconditioning seems to be desirable for the efficient solution of the resulting linear system. We propose a preconditioner in the following two subsections.

3.4. Neumann problem: Reduction to preconditioning on piecewise linear spaces. The bilinear form B is expressed in (3.3) as a sum of element contributions. The preconditioner C is also constructed elementwise:

$$C(u, v) := \sum_{K\in\mathcal{T}} C_K(u_K, v_K). \tag{3.8}$$

For the construction of the preconditioner, we use the fact that by our discussion in section 3.1.1 the space $\mathcal{P}_{p_K}$ can be written as $\mathcal{P}_{p_K} = \widetilde{\mathcal{V}} \oplus_{i=1}^{3} \widetilde{\mathcal{S}}_i \oplus \widetilde{\mathcal{I}}$, where the polynomial degrees associated with sets of side shape functions $\mathcal{S}_i$ and the internal shape functions $\mathcal{I}$ implicitly depend on K. Correspondingly, a function $u_K \in \mathcal{P}_{p_K}$ can be written as

$$u_K = u_K^{\mathcal{V}} + \sum_{i=1}^{3} u_K^{\mathcal{S}_i} + u_K^{\mathcal{I}}. \tag{3.9}$$

We then define the element contributions C_K of the preconditioner C as

$$C_K(u, v) = B_K(u_K^{\mathcal{V}}, v_K^{\mathcal{V}}) + \sum_{i=1}^{3} B_K(u_K^{\mathcal{S}_i}, v_K^{\mathcal{S}_i}) + B_K(u_K^{\mathcal{I}}, v_K^{\mathcal{I}}). \tag{3.10}$$

We mentioned in section 3.1.1 that the splitting of a basis of $\mathcal{P}_{p_K}$ into vertex, side, and internal shape functions is not unique. The following result, which is due to [3], asserts that for discrete harmonic side shape functions the preconditioner C given by (3.8) is only weakly dependent on $\mathbf{p}$ and independent of the mesh.

PROPOSITION 3.5 (see [3]). *Let $\mathcal{T}$ be a shape-regular mesh consisting of triangles. Assume that the side shape functions are discrete harmonic, i.e.,*

$$B_K(u, v) = 0 \qquad \forall u \in \widetilde{\mathcal{S}} \quad \forall v \in \widetilde{\mathcal{I}}. \tag{3.11}$$

Then there exist c_1, $c_2 > 0$ depending only on the coefficients A, a_0 and the shape-regularity constant γ such that

$$c_1 B(u, u) \le C(u, u) \le c_2 (1 + \ln |\mathbf{p}|)^2 B(u, u) \qquad \forall u \in S^{\mathbf{p}}(\Omega, \mathcal{T}),$$

where $|\mathbf{p}| = \max_{K \in \mathcal{T}} p_K$.

Remark 3.6. The condition (3.11) can be achieved by a process akin to the local static condensation described in section 3.2. By Proposition 3.2, the condition (3.11) can be enforced with work $O(N)$.

Remark 3.7. When solving the system by local static condensation as described in section 3.2, the condition (3.11) should be substituted by

$$B_K(u, v) = 0 \qquad \forall u \in \widetilde{\mathcal{E}} \quad \forall v \in \widetilde{\mathcal{I}}. \tag{3.12}$$

It is easy to see that the energy minimization property of the discrete B-harmonic functions now implies the same condition number for the preconditioner (3.10), (3.12) as in Proposition 3.5. On the other hand, the former preconditioner (3.10), (3.11) can be shown to have the same condition number as the modified one.

We now analyze the cost of applying the preconditioner C, i.e., solving

$$C(u, v) = l(v) \qquad \forall v \in V_N.$$

By the decomposition (3.2) of the basis $\mathcal{B}$ of V_N, the sought function $u \in V_N$ can be written in the form

$$u = u^{\mathcal{V}} + \sum_{\text{edges } e} u^e + \sum_{K \in \mathcal{T}} u_K^{\mathcal{I}}, \tag{3.13}$$

where $u^{\mathcal{V}} \in S^1(\Omega, \mathcal{T})$, $u_e \in \operatorname{span} V_e$, $u_K^{\mathcal{I}} \in \operatorname{span} V_{\mathcal{I}}$. Computing the components $u^{\mathcal{V}}$, u^e, $u_K^{\mathcal{I}}$ amounts to solving a global problem corresponding to a discretization with piecewise linear functions and two sets of local problems:

$$B(u^{\mathcal{V}}, v) = l(v) \qquad \forall v \in V_{\mathcal{V}}, \tag{3.14}$$

$$B(u^e, v) = l(v) \qquad \forall v \in V_e \qquad \text{for all edges } e, \tag{3.15}$$

$$B(u_K^{\mathcal{I}}, v) = l(v) \qquad \forall v \in V_{\mathcal{I}} \qquad \forall K \in \mathcal{T}. \tag{3.16}$$

Solving (3.15) and (3.16) can be accomplished with work $O(N)$ on geometric meshes with linear degree vectors as follows.

PROPOSITION 3.8. *Let $\mathcal{T}$ be a geometric mesh with boundary mesh size h and let $\mathbf{p}$ be a linear degree vector. Then the problems* (3.15) *and* (3.16) *can be solved with work $W = O(N)$, where $N \sim h^{-1}$.*

Proof. First, we note that the stiffness matrices corresponding to (3.15), (3.16) are submatrices of the global stiffness matrix; this guarantees that the problems can be set up with optimal complexity $O(N)$. For the solution step, we observe that the problems decouple into problems associated with single elements or two adjacent elements. Specifically, using Gaussian elimination for each element with cost $O(p_K^6)$, we arrive at the total cost for the solution of (3.16)

$$W \leq C \sum_{K \in \mathcal{T}} p_K^6 = O(N),$$

by a reasoning like that of the proof of Proposition 2.7. The solution of (3.15) decouples into problems associated with each edge of the mesh, and a calculation shows again that the required work is $O(N)$. □

Since $\dim S^1(\Omega, \mathcal{T}) = O(N)$, the cost for solving (3.14) is at least $O(N)$; hence, by Proposition 3.8, the total cost of applying the preconditioner C is controlled by the cost of solving (3.14).

3.5. Efficient solution of piecewise linear discretization. The analysis of the preceding section allowed us to restrict our attention to the case of piecewise linear approximation on geometric meshes $\mathcal{T}$. By the general theory of preconditioning, it suffices to find a spectrally equivalent bilinear form $\widetilde{B}$ on $S^1(\Omega, \mathcal{T}) \times S^1(\Omega, \mathcal{T})$. To that end, the space $S^1(\Omega, \mathcal{T})$ is decomposed further as

$$S^1(\Omega, \mathcal{T}) = V_{\mathcal{H}} \oplus (S^1(\Omega, \mathcal{T}) \cap H_0^1(\Omega)); \tag{3.17}$$

here, $V_{\mathcal{H}}$ is given by

$$V_{\mathcal{H}} := \operatorname{Range} E_{\mathcal{V}},$$

where $E_{\mathcal{V}}$ is the $S^1(\Omega, \mathcal{T})$-discrete harmonic extension operator

$$\begin{aligned} E_{\mathcal{V}} : \gamma_0(S^1(\Omega, \mathcal{T})) &\to S^1(\Omega, \mathcal{T}), \\ u &\mapsto E_{\mathcal{V}} u \quad \text{with} \quad B(E_{\mathcal{V}} u, v) = 0 \qquad \forall v \in S^1(\Omega, \mathcal{T}) \cap H_0^1(\Omega). \end{aligned}$$

The decomposition (3.17) also provides the B-orthogonal splitting (cf. Lemma 2.2)

$$B(u, v) = B(E_{\mathcal{V}} \gamma_0 u, E_{\mathcal{V}} \gamma_0 v) + B(u - E_{\mathcal{V}} \gamma_0 u, v - E_{\mathcal{V}} \gamma_0 v) \tag{3.18}$$

$\forall u, v \in S^1(\Omega, \mathcal{T})$. Due to Proposition 3.4, any spectrally equivalent approximation to the first term on the right-hand side of (3.18) yields the desired bilinear form $\widetilde{B}$. For simplicity of exposition, we now assume that Ω is simply connected. Following the standard construction in the domain decomposition methods, we apply a circulant preconditioning matrix. Let $F_\Omega : \widehat{C} \to \partial\Omega$ be the bi-Lipschitz mapping providing the global parametrization of $\partial\Omega$ by a 2π-periodic function. Assume that our quasi-uniform partitioning of $\mathcal{T}_{|\partial\Omega}$ is the image of the uniform grid $\mathcal{T}_{\widehat{C}}$ on $\widehat{C} := [0, 2\pi]$. Let $\Delta_{\widehat{C},h}$ be the discrete Laplacian defined on the set $\mathcal{T}_{\widehat{C}}$ and associated with the corresponding FE space of periodic piecewise linear functions $V_h(\widehat{C})$:

$$\langle -\Delta_{\widehat{C},h} u, v\rangle_{0,\widehat{C}} = \int_{\widehat{C}} \nabla u \cdot \nabla v \, ds + \int_{\widehat{C}} uv \, ds \qquad \forall\, u, v \in V_h(\widehat{C}). \tag{3.19}$$

The bilinear form $\widetilde{B}$ is then defined on $S^1(\Omega, \mathcal{T}) \times S^1(\Omega, \mathcal{T})$ by

$$\widetilde{B}(u, v) := \langle (-\Delta_{\widehat{C},h})^{1/2} (\gamma_0 u \circ F_\Omega), (\gamma_0 v \circ F_\Omega)\rangle_{0,\widehat{C}} + B(u - E_{\mathcal{V}}\gamma_0 u, v - E_{\mathcal{V}}\gamma_0 v).$$

The symmetric positive definite bilinear form $\widetilde{B}$ is spectrally equivalent to B on $S^1(\Omega, \mathcal{T})$ since

$$\langle (-\Delta)^{1/2} (\gamma_0 u \circ F_\Omega), \gamma_0 u \circ F_\Omega\rangle \sim \|\gamma_0 u\|^2_{1/2,\partial\Omega} \sim \|E_{\mathcal{V}}(\gamma_0 u)\|^2_{1,\Omega}.$$

We are thus left with the efficient solution of the problem

$$\text{Find } u \in S^1(\Omega, \mathcal{T}) \text{ s.t. } \widetilde{B}(u, v) = l(v) \qquad \forall v \in S^1(\Omega, \mathcal{T}). \tag{3.20}$$

Problem (3.20) can be solved in four steps as follows.

Algorithm 3.9 (preconditioner for piecewise (p.w.) linear discretization).

1. *Determine* $\tilde{u} \in S^1(\Omega, \mathcal{T}) \cap H^1_0(\Omega)$ *as the solution of*

$$B(\tilde{u}, v) = l(v) \qquad \forall v \in S^1(\Omega, \mathcal{T}) \cap H^1_0(\Omega). \tag{3.21}$$

This can be done efficiently by CG-iteration, because the stiffness matrix has uniformly bounded condition number.

2. *Determine* $\gamma_0 u$ *as the solution of*

$$\langle (-\Delta_{\widehat{C},h})^{1/2} (\gamma_0 u \circ F_\Omega), (\gamma_0 v \circ F_\Omega)\rangle_{0,\widehat{C}} = l(v) - B(\tilde{u}, v) \;\; \forall v \in S^1(\Omega, \mathcal{T}). \tag{3.22}$$

Note that the right-hand side in (3.22) *depends only on* $\gamma_0 v$, *which makes it possible to compute it with the test functions supported within one near-boundary grid layer. Due to the special structure of the operator* $\Delta_{\widehat{C},h}$ *on uniform meshes, solving* (3.22) *is efficiently accomplished by a forward and a backward FFT.*

3. *Compute (approximately) the discrete* B*-harmonic extension* $E_{\mathcal{V}}\gamma_0 u$. *This can be achieved with the aid of explicit extension operators* [17], *or by CG-iteration of the problem*

$$B(E_{\mathcal{V}}\gamma_0 u, v) = 0 \qquad \forall v \in S^1(\Omega, \mathcal{T}) \cap H^1_0(\Omega),$$

since the stiffness matrix corresponding to this problem has condition number bounded uniformly in N *by Proposition* 3.4.

4. *Set* $u := \tilde{u} + E_{\mathcal{V}}\gamma_0 u$.

Algorithm 3.9 has the following complexity.

PROPOSITION 3.10. *If steps* 1, 3 *in Algorithm* 3.9 *are solved iteratively with a tolerance* $\varepsilon = O(N^{-q})$, *then the solution of* (3.20) *with Algorithm* 3.9 *requires* $O(N \log N)$ *floating point operations.*

Proof. The solution of the problem in step 1 requires $O(N|\log \varepsilon|)$ work. The cost of the FFT is $O(N \log N)$. Finally, the calculation of the harmonic extension by the CG-iteration requires $O(N|\log \varepsilon|)$ arithmetic operations, where $\varepsilon > 0$ is the desired accuracy. Setting $\varepsilon = O(N^{-q})$ completes the proof. □

3.6. Remarks on implementations with static condensation. In computational practice one would likely base an iterative solution fully on local static condensation. One of the advantages of such a procedure is the reduction of the size of the problem that is solved iteratively. We recall the classical static condensation-based scheme as follows.

ALGORITHM 3.11 (solution based on local static condensation).

1. *Compute the local stiffness matrices* A_K *and the local load vectors* l_K.
2. *Compute the condensed local stiffness matrices* A_K^c *(note that this enforces* (3.12)*) and compute the condensed load vectors* l_K^c.
3. *Assemble* $A^c = \mathcal{A}_{K\in\mathcal{T}} A_K^c$ *and assemble the condensed load vectors* l^c *(see, e.g.,* [42]*).*
4. *Solve the linear system* $A^c x = l^c$ *by a preconditioned CG-iteration with the modified preconditioner of Algorithm* 3.12 *below.*
5. *Sweep through the mesh and solve for the internal degrees of freedom.*

Note that, with the exception of solving the condensed system $A^c x = l^c$, all steps of this algorithm can be accomplished with work $O(N)$. For Dirichlet problems, Proposition 3.4 shows that the condition number of the matrix A^c grows polylogarithmically with the problem size and can thus be treated efficiently by CG-iterations. For the Neumann problem, a preconditioner C^c similar to the one introduced above is required.

Static condensation on the element level can be interpreted as choosing new nodal shape functions $V_{\mathcal{V}}^c$ and side shape functions $V_{\mathcal{S}}^c$. Specifically, the mapping Z that accomplishes this transformation of shape functions is given by

$$Z : u \mapsto Zu, \qquad Zu \text{ satisfies} \left\{ \begin{array}{rl} Zu|_e = u|_e & \forall \text{ edges } e, \\ B(Zu, v) = 0 & \forall v \in V_{\mathcal{I}}. \end{array} \right. \tag{3.23}$$

The mapping Z maps the set of piecewise linears $V_{\mathcal{V}}$ and the set of side shape functions $V_{\mathcal{S}}$ onto discrete harmonic nodal shape functions $V_{\mathcal{V}}^c := ZV_{\mathcal{V}}$ and discrete harmonic side shape functions $V_{\mathcal{S}}^c := ZV_{\mathcal{S}}$, respectively. The preconditioner C^c employed for the solution of the condensed linear system can then be realized with the following algorithm.

ALGORITHM 3.12 (preconditioner C^c for a statically condensed stiffness matrix).

Output: solution $u = u_{\mathcal{V}^c} + u_{\mathcal{S}^c} \in \operatorname{span} V_{\mathcal{V}}^c \cup V_{\mathcal{S}}^c$ *such that*

$$C^c(u, v) = l(v) \qquad \forall v \in \operatorname{span} V_{\mathcal{V}}^c \cup V_{\mathcal{S}}^c.$$

1. (a) *Determine* $\widetilde{u} \in \operatorname{span} V_{\mathcal{V}}^c \cap H_0^1(\Omega)$ *as the solution to*

$$B(\tilde{u}, v) = l(v) \qquad \forall v \in \operatorname{span} V_{\mathcal{V}}^c \cap H_0^1(\Omega). \tag{3.24}$$

This can be done efficiently by a CG-iteration, as the stiffness matrix has uniformly bounded condition number (cf. [31]*). Note that the stiffness matrix corresponding to* (3.24) *is a submatrix of* A^c.

(b) *Determine* $\gamma_0 u$ *as the solution of*

$$\langle(-\Delta_{\widehat{C},h})^{1/2}(\gamma_0 u\circ F_\Omega),(\gamma_0 v\circ F_\Omega)\rangle_{0,\widehat{C}} = l(v) - B(\tilde{u},v) \qquad \forall v\in \operatorname{span} V^c_{\mathcal{V}}.$$

(c) *Find* $E^c\gamma_0 u \in \operatorname{span} V^c_{\mathcal{V}}$ *such that*

$$B(E^c\gamma_0 u, v) = 0 \qquad \forall v \in \operatorname{span} V^c_{\mathcal{V}}. \tag{3.25}$$

This could be accomplished by a CG-iteration, as the stiffness matrix corresponding to (3.25) *has uniformly bounded condition number by Proposition* 3.4. *Note again that the stiffness matrix is a submatrix of* A^c.

(d) *Set* $u_{\mathcal{V}^c} := \tilde{u} + E^c\gamma_0 u$.

2. *Introduce* $V^{c,e} := \{v \in V^c_{\mathcal{S}}\,|\, v|_{e'} = 0 \quad$ *for all edges* $e' \neq e\}$. *Then* $u_{\mathcal{S}^c} \in \operatorname{span} V^c_{\mathcal{S}}$ *is given by* $u_{\mathcal{S}^c} = \sum_{\text{edges } e} u^e$, *with* $u^e \in V^{c,e}$ *solving*

$$B(u^e, v) = l(v) \qquad \forall v \in V^{c,e}. \tag{3.26}$$

Note that the stiffness matrices for these edge-based problems are submatrices of A^c. *Solving all edge problems* (3.26) *can be achieved with work* $O(N)$ *by Gaussian elimination.*

We will not analyze the scheme consisting of Algorithms 3.11, 3.12. In view of the following lemma, however, we expect complexity estimates for this scheme as in Proposition 3.10.

LEMMA 3.13. *The mapping* Z *of* (3.23) *is an isomorphism between* $S^1(\Omega,\mathcal{T})$ *and* $ZS^1(\Omega,\mathcal{T})$; *i.e., for some* $C>0$ *there holds*

$$C^{-1}\|Zu\|_{H^1(\Omega)} \le \|u\|_{H^1(\Omega)} \le C\|Zu\|_{H^1(\Omega)} \qquad \forall u \in S^1(\Omega,\mathcal{T}). \tag{3.27}$$

Proof. The lower bound $\|Zu\|_{H^1(\Omega)} \le C\|u\|_{H^1(\Omega)}$ follows from the fact that the energy norm is equivalent to the H^1-norm and the energy minimization properties of the mapping Z. For the upper bound, we write Ω as the union of elements. For a fixed element K, we write $\widehat{u}$, $\widehat{Zu}$ for the pull-backs to the reference element $\hat{K}$ of the functions $u|_K$, $Zu|_K$. We then calculate, using the trace theorem on $\hat{K}$ and exploiting that $\widehat{u}$ is linear,

$$\|\nabla\widehat{u}\|_{L^2(\hat{K})} \sim |\widehat{u}|_{H^{1/2}(\partial\hat{K})} = |\widehat{Z}u|_{H^{1/2}(\partial\hat{K})} \le C\|\nabla\widehat{Z}u\|_{L^2(\hat{K})},$$

where the implied constant in the $\sim$ notation and the constant C are independent of the polynomial degree p_K and the element K. Scaling to the physical element K and summing over all elements yields $\|\nabla u\|_{L^2(\Omega)} \le C\|\nabla Zu\|_{L^2(\Omega)}$. Since $u = Zu$ on $\partial\Omega$, we get the upper bound $\|u\|_{H^1(\Omega)} \le C\|Zu\|_{H^1(\Omega)}$ in (3.27). □

4. Approximation to Poincaré–Steklov operators.

4.1. Poincaré–Steklov operators in elliptic problems. In some applications, the solution u to (1.2) or (1.3) is not the principal quantity of interest, but instead the missing data for a complete set of Cauchy data are sought. The Poincaré–Steklov operator T (also known as the Dirichlet-to-Neumann map) is defined as

$$\begin{array}{rcl} T : H^{1/2}(\partial\Omega) & \to & H^{-1/2}(\partial\Omega),\\ \lambda & \mapsto & \gamma_1 u, \quad u \text{ solves (1.2) with } f=0. \end{array} \tag{4.1}$$

Likewise, we define the Poincaré–Steklov operator S (also called the Neumann-to-Dirichlet map) by

$$\begin{array}{rcl} S : H^{-1/2}(\partial\Omega) & \to & H^{1/2}(\partial\Omega),\\ \psi & \mapsto & \gamma_0 u, \quad u \text{ solves (1.3) with } f=0. \end{array} \tag{4.2}$$

We note that the operators S, T are in fact inverses to each other, i.e., $S^{-1} = T$. Akin to the situations in (1.18), (1.19), the operator T admits a shift theorem. For general Lipschitz domains Ω, [10, Lemma 3.7] asserts for each $s \in [0, 1/2]$ the existence of $C_s > 0$ such that

$$\|T\lambda\|_{H^{-1/2+s}(\partial\Omega)} \leq C_s \|\lambda\|_{H^{1/2+s}(\partial\Omega)} \qquad \forall \lambda \in H^{1/2+s}(\partial\Omega). \tag{4.3}$$

For polygonal domains Ω, this shift theorem holds in a larger range. While this is closely related to (1.19), a precise reference seems to be missing, and we therefore formulate this as the following assumption.

Assumption 4.1. *There exists $\delta_0 > 1/2$ such that for each $\delta \in [0, \delta_0)$ a constant $C_\delta > 0$ can be found with*

$$\|E\lambda\|_{H^{1/2+\delta}(\partial\Omega)} + \|T\lambda\|_{H^{-1/2+\delta}(\partial\Omega)} \leq C_\delta \|\lambda\|_{H^{1/2+\delta}(\partial\Omega)} \qquad \forall \lambda \in H^{1/2+\delta}(\partial\Omega), \tag{4.4}$$

where the extension $E\lambda \in H^1(\Omega)$ solves (1.2) *with $f = 0$.*

Remark 4.2. For the case of Laplace's equation (and thus the case of constant coefficients), Assumption 4.1 can be verified as follows. Let $\delta_0 \in (1/2, 1]$ be defined by (1.19). For general Lipschitz domains Ω, the estimate (4.3) covers the case $\delta \in [0, 1/2]$. For $\delta \in (1/2, 1)$, combining [11, Lemma 2.11] and [11, Lemma 2.7] gives

$$\|T\lambda\|_{H^{-1/2+\delta}(\partial\Omega)} \leq C\|E\lambda\|_{H^{1+\delta}(\Omega)} \leq C_\delta \|\lambda\|_{H^{1/2+\delta}(\partial\Omega)},$$

where the second estimate follows from the regularity result (1.19).

In the case of convex polygons, we have $\delta_0 = 1$.

Remark 4.3. The case $b = 0$, $a_0 = 0$ does not fall directly into our framework because assumption (1.6) is not satisfied. The modification of considering the energy space $V = H^1(\Omega)/\mathbb{R}$, as is standard for Laplace's equation, can be carried out in the case of nonconstant matrix A as well.

4.2. *hp*-FEM approximation of Poincaré–Steklov operators *T* and *S*. We recall that the projector Q_N of (2.6) can be extended to an operator on $H^{-1/2}(\partial\Omega)$ by (2.14).

4.2.1. Approximation of the Poincaré–Steklov operator. Viewing the dual space Y_N' as a subspace of $H^{-1/2}(\partial\Omega)$, we define the approximation S_N to the Poincaré–Steklov operator S as

$$\begin{aligned} S_N : Y_N' &\to Y_N, \\ \Psi &\mapsto S_N\Psi := \gamma_0 u_N, \end{aligned}$$

where $u_N \in V_N$ solves the following discrete Neumann problem:

$$B(u_N, v) = \langle \Psi, v\rangle_{0,\partial\Omega} \qquad \forall v \in V_N.$$

The error analysis for S_N is rather straightforward.

Theorem 4.4. *Under the assumptions of Theorem* 2.13 *(with $f = 0$) there holds*

$$\|S\Psi - S_N\Psi\|_{H^{1/2}\partial\Omega} \leq C\left[h^\delta + h^{b\alpha}\right]. \tag{4.5}$$

Proof. Applying the trace theorem, we obtain

$$\|S\Psi - S_N\Psi\|_{H^{1/2}\partial\Omega} = \|\gamma_0 u - \gamma_0 u_N\|_{H^{1/2}\partial\Omega} \leq C\|u - u_N\|_{1,\Omega}.$$

Now (2.20) yields (4.5). □

4.2.2. Approximation of the Poincaré–Steklov operator. The approximation

$$
\begin{aligned}
T_N : Y_N &\to Y_N', \\
\lambda &\mapsto T_N\lambda
\end{aligned}
\tag{4.6}
$$

to the Poincaré–Steklov operator T defines an element of the dual space Y_N' via

$$\langle T_N\lambda, v\rangle_{0,\partial\Omega} = B(u_N, \widetilde{v}) \qquad \forall v \in Y_N,$$

where $\widetilde{v} \in V_N$ is an arbitrary extension of v, $\gamma_0\widetilde{v} = v$, and $u_N \in V_N$ satisfies

$$\gamma_0 u_N = \lambda \quad \text{ and } \quad B(u_N, v) = 0 \quad \forall\, v \in V_N \cap H_0^1(\Omega).$$

THEOREM 4.5. *Let Ω be a polygon, and let δ_0 be given by Assumption* 4.1. *Then for any $\delta \in [0, \delta_0) \cap [0, 3/2)$ and $\lambda \in H^{1/2+\delta}(\partial\Omega)$ there holds*

$$\|T\lambda - T_N Q_N \lambda\|_{H^{-1/2}(\partial\Omega)} \le C_\delta[h^\delta + h^{b\alpha}], \tag{4.7}$$

where C_δ, $b > 0$ and $\alpha > 0$ is the slope of the degree vector.

Proof. Using $Q_N T_N Q_N = T_N Q_N$, we write

$$T - T_N Q_N = (\mathrm{Id} - Q_N)T + Q_N T(\mathrm{Id} - Q_N) + Q_N(T - T_N)Q_N. \tag{4.8}$$

The first two terms in (4.8) lead to estimates of the form

$$\|(\mathrm{Id} - Q_N)T\lambda\|_{H^{-1/2}(\partial\Omega)} \le C_\delta h^\delta \|T\lambda\|_{H^{-1/2+\delta}(\partial\Omega)} \le C_\delta h^\delta \|\lambda\|_{H^{1/2+\delta}(\partial\Omega)}, \tag{4.9}$$

$$\|Q_N T(\mathrm{Id} - Q_N)\lambda\|_{H^{-1/2}(\partial\Omega)} \le C\|(\mathrm{Id} - Q_N)\lambda\|_{H^{1/2}(\partial\Omega)} \le Ch^\delta \|\lambda\|_{H^{1/2+\delta}(\partial\Omega)}, \tag{4.10}$$

where we exploited the stability and approximation properties of the projector Q_N given in Lemma 2.8 and used the shift theorem for T as detailed in Assumption 4.1. For the third term in (4.8), we first introduce for the elliptic extension $E_N : Y_N \to V_N$. By a reasoning similar to that in the proof of Corollary 2.12, we get for $v \in H^{1/2}(\partial\Omega)$

$$\|E_N(Q_N v)\|_{H^1(\Omega)} \le C\|Q_N v\|_{H^{1/2}(\partial\Omega)} \le C\|v\|_{H^{1/2}(\partial\Omega)}, \tag{4.11}$$

where we used the stability result (2.15) in the last step. For the treatment of the third term in (4.8), we next let $u \in H^1(\Omega)$ be the solution to (1.2) with $f = 0$ and Dirichlet boundary conditions λ; $\tilde{u} \in H^1(\Omega)$ solves (1.2) with $f = 0$ and boundary conditions $Q_N\lambda$; and $u_N \in V_N$ is the *hp*-FEM approximation to u given by (2.3). Reasoning as in [25, Lemma 6.1], the definitions of the operators T, T_N imply

$$
\begin{aligned}
\|Q_N(T - T_N)Q_N\lambda\|_{H^{-1/2}(\partial\Omega)} &= \sup_{v \in H^{1/2}(\partial\Omega)} \frac{\langle (T - T_N)Q_N\lambda, Q_N v\rangle_{0,\partial\Omega}}{\|v\|_{H^{1/2}(\partial\Omega)}} \\
&= \sup_{v \in H^{1/2}(\partial\Omega)} \frac{B(\tilde{u} - u_N, E_N(Q_N v))}{\|v\|_{H^{1/2}(\partial\Omega)}} \\
&\le C\|\tilde{u} - u_N\|_{H^1(\Omega)} \le C\left[\|u - u_N\|_{H^1(\Omega)} + \|u - \tilde{u}\|_{H^1(\partial\Omega)}\right] \\
&\le C\left[\|u - u_N\|_{H^1(\Omega)} + \|\lambda - Q_N\lambda\|_{H^{1/2}(\partial\Omega)}\right].
\end{aligned}
$$

$\|\lambda - Q_N\lambda\|_{H^{1/2}(\partial\Omega)}$ can be bounded as required by (4.10), and Theorem 2.13 allows us to bound $\|u - u_N\|_{H^1(\Omega)}$ in the desired fashion. □

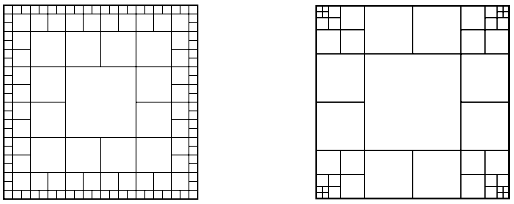


FIG. 4.1. *Refined interface corresponding to Figure* 2.1 *(left) and to Remark* 1.2 *(right).*

Remark 4.6. We employed the $L^2(\partial\Omega)$-projection Q_N in the definition of T_N. Other projections could, in principle, be used as well. Essential is that the projector P_N has the following stability and approximation properties for $\delta \in [0, \delta_0)$:

$$\begin{aligned}\|P_N u\|_{H^{-1/2}(\partial\Omega)} &\leq C\|u\|_{H^{-1/2}(\partial\Omega)},\\ \|P_N u\|_{H^{1/2+\delta}(\partial\Omega)} &\leq C_\delta\|u\|_{H^{1/2+\delta}(\partial\Omega)},\\ \|u - P_N u\|_{H^{1/2}(\partial\Omega)} &\leq C_\delta h^\delta \|u\|_{H^{1/2+\delta}(\partial\Omega)}.\end{aligned}$$

Due to the results of section 3 (see also Table 3.1), the implementation of the finite-dimensional operators S_N and T_N has the linear-logarithmic complexity $O(N \log^q N)$ with respect to $N = O(\dim Y_N)$, except in the case of Neumann boundary conditions, where we arrive at the cost $O(N^{3/2} \log^q N)$.

Remark 4.7. In the case of piecewise constant coefficients in a polygonal domain, an efficient method for matrix-vector product with the discrete Poincaré–Steklov operators was developed in [24, 27, 28, 25]. It is based on a sparse h-FEM approximation to the Schur complement matrix on a rectangle, combined with the reduction of the PDE to the *refined interface*. In this way the Schur complement matrix in each $n \times n$ rectangular subdomain is treated with $O(n \log^2 n)$ arithmetic operations using a truncated Fourier representation. In the case of symmetric and positive definite operators with piecewise constant coefficients, the spectrally equivalent multilevel interface preconditioners (see [27, 28]) lead to the complexity $O(N \log^3 N)$ and the memory requirements $O(N \log N)$ for solving the Schur complement equation on the refined interface.

An example of the refined interface is given in Figure 4.1, where nonconforming decompositions in Figure 4.1 (left) correspond to the geometric meshes in Figure 2.1. Figure 4.1 (right) corresponds to the case of composite meshes refined towards the corner points, related to the situation in Remark 1.2. An extension of this approach to the biharmonic, Stokes, and Lamé equations was discussed in [25].

In this way, the boundary concentrated hp-FEM presented in this paper extends the above-mentioned methods to the case of variable (piecewise analytic) coefficients.

5. Further applications.

5.1. Relation to boundary integral equations. Consider the case of constant coefficients. The results on the sparse approximation to the Poincaré–Steklov operators directly apply to the construction of asymptotically optimal solvers for the classical boundary integral equations involving the weakly singular, hypersingular, and double layer harmonic potential operators V, D, and K, respectively, defined by

$$Vu(x) = \int_\Gamma g(x,y)u(y)dy, \qquad Ku(x) = \int_\Gamma \frac{\partial}{\partial n_y} g(x,y)u(y)dy,$$
$$K'u(x) = \int_\Gamma \frac{\partial}{\partial n_x} g(x,y)u(y)dy, \qquad Du(x) = -\int_\Gamma \frac{\partial}{\partial n_x}\frac{\partial}{\partial n_y} g(x,y)u(y)dy\,,$$

where $g(x,y)$ denotes the fundamental solution for the corresponding elliptic operator $\mathcal{L}$, and $\Gamma = \partial\Omega$ for a Lipschitz domain $\Omega \in \mathbb{R}^2$. The idea is based on the representation of the inverse to the above-mentioned boundary integral operators in terms of interior T_1, S_1 and exterior T_2, S_2 Poincaré–Steklov mappings proposed in [24]. Given a Hilbert space H and an element $g \in H'$, we denote $H_g := \{v \in H : \langle v, g\rangle = 0\}$.

THEOREM 5.1 (see [24]). *The operator* $V^{-1} : H^{1/2}_{g_0}(\Gamma) \to H^{-1/2}_1(\Gamma)$ *has the representation* $V^{-1} = T_1 + T_2$, *where* g_0 *is the Robin potential on* Γ *satisfying* $K'g_0 = -\frac{1}{2}g_0$. *There holds* $(\frac{1}{2}I - K)^{-1}z = (I + S_2 \cdot T_1)z \;\forall\; z \in H^{1/2}(\Gamma)$, *and* $(\frac{1}{2}I + K)^{-1}z = (I + S_1 \cdot T_2)z \;\forall\; z \in H^{1/2}_{g_0}(\Gamma)$. *The operator* $D^{-1} : H^{-1/2}_1(\Gamma) \to H^{1/2}_{g_0}(\Gamma)$ *has the representation* $D^{-1} = S_1 + S_2$.

An important consequence of the above statement is that whenever some efficient discretization (say, with linear-logarithmic cost) is constructed for the operators T_i and S_i, $i = 1, 2$, we immediately obtain the corresponding efficient approximation for the inverse to the classical boundary integral operators in question. We refer to [19, 20, 8] for an alternative approach to data-sparse approximations of T and S based on modern $\mathcal{H}$-matrix arithmetic.

5.2. Application to exterior boundary value problems. BEMs are very often applied to exterior domain problems. In this subsection, we want to briefly show how the boundary concentrated FEM can be adapted to this setting.

In the exterior domain $\Omega_e := \mathbb{R}^d \setminus \Omega$, we consider the Dirichlet problem

$$\mathcal{L}_e u := -\nabla\cdot(A(x)\nabla u) + b(x)\cdot\nabla u + c(x)u = f \quad \text{in } \Omega_e, \tag{5.1a}$$
$$\gamma_0 u = \lambda \quad \text{on } \partial\Omega, \tag{5.1b}$$

with similar assumptions on the data as in section 1.3. In addition, we assume that b, a_0, and f have bounded support Ω_0 such that $\mathcal{L}_e = -\Delta$ in $\mathbb{R}^2 \setminus \Omega_0$, and that u satisfies the "radiation condition" of the form

$$|u(x)| = O(|x|^{-1}), \qquad |\nabla u| = O(|x|^{-2}), \quad |x| \to \infty. \tag{5.2}$$

(This requires a compatibility condition that suppresses the logarithmic growth at infinity typically exhibited by solutions to Laplace's equation on exterior domains.) We approximate (5.2) by imposing homogeneous Neumann conditions on the auxiliary boundary Γ_∞ with $\operatorname{dist}(\Gamma_\infty, \partial\Omega) = R = O(N^{1/2})$. As above, N denotes the number of degrees of freedom on $\partial\Omega_e$. Following [28] (see also the references therein), we use a mesh on $\operatorname{Int}(\Gamma_\infty) \setminus \Omega$ that is a geometric mesh in the sense of Definition 2.3 (see Figure 5.1). The number of levels is again estimated by $\log R = O(\log N)$. We stress that the approximation and solution schemes remain verbatim as for the interior problem.

Our arguments indicate that, in the framework of boundary concentrated *hp*-FEM, there is no essential difference between solving exterior and interior problems in the case of smooth coefficients, and we again expect complexity $O(N \log^q N)$ for the approximation of the exterior Poincaré–Steklov operators.

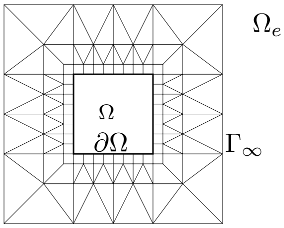


FIG. 5.1. *Geometric mesh on an exterior domain.*

5.3. Application in domain decomposition. The efficient realization of the Poincaré–Steklov operators can also be employed in the context of domain decomposition methods. Let the domain Ω consist of $M_0 \geq 1$ nonoverlapping polygons Ω_i, $\overline{\Omega} = \cup_{i=1}^{M_0}\overline{\Omega}_i$ with $\Gamma := \cup_{i=1}^{M_0}\Gamma_i \setminus \partial\Omega$, and let $f \in H^{-1}(\Omega)$ and $\psi_i \in H^{-1/2}(\Gamma_i)$. Let the bilinear form $B(\cdot,\cdot)$ be written as a sum of subdomain contributions, and consider the problem of finding $u \in H_0^1(\Omega)$ such that

$$\sum_{i=1}^{M_0} B_i(u_{|\Omega_i}, v_{|\Omega_i}) = B(u,v) = F(v) = \int_\Omega f(x)v dx + \sum_{i=1}^{M} \langle \psi_i, v\rangle_{L^2(\Gamma_i)} \qquad \forall v \in H_0^1(\Omega),$$

where the local continuous forms $B_i : V_i \times V_i \to \mathbb{R}$ are supposed to be $H_0^1(\Omega_i)$-elliptic with $V_i := H^1(\Omega_i)$. Letting $u_{0,i} \in H_0^1(\Omega_i)$ be the particular solutions in Ω_i,

$$B_i(u_{0,i}, v) = \int_{\Omega_i} f(x)v dx \qquad \forall v \in H_0^1(\Omega_i), \tag{5.3}$$

and introducing the trace space on Γ,

$$Y_\Gamma = \{u = z_{|\Gamma} : \ z \in H_0^1(\Omega)\}, \qquad \|u\|_{Y_\Gamma} = \inf_{z \in H_0^1(\Omega): z_{|\Gamma} = u} \|z\|_{H^1(\Omega)},$$

we transform the above problem into the interface equation

$$u_{|\Gamma} \in Y_\Gamma : \quad B_\Gamma(u_{|\Gamma}, v) := \sum_{i=1}^{M} \langle T_i u_i, v_i\rangle_{0,\Gamma_i} = \sum_{i=1}^{M} \langle g_i, v\rangle_{\Gamma_i} \quad \forall v \in Y_\Gamma, \tag{5.4}$$

where $g_i = \psi_i - \gamma_{1,i}u_{0,i}$ and $u_i = u_{|\Gamma_i}$, $v_i = v_{|\Gamma_i}$. The local Poincaré–Steklov operators $T_i : H^{1/2}(\Gamma_i) \to H^{-1/2}(\Gamma_i)$ are defined by the B_i-harmonic extensions; see (4.6). Since the bilinear form $B_\Gamma(\cdot,\cdot) : Y_\Gamma \times Y_\Gamma \to \mathbb{R}$ is continuous and coercive, (5.4) is uniquely solvable in Y_Γ for any $g_i \in H^{-1/2}(\Gamma_i)$ providing the trace $u_{|\Gamma}$ (see [25]). Thus, the boundary concentrated FEM can be applied on each subdomain Ω_i separately to numerically realize the operators T_i.

6. Numerics. The main goal of this section is to confirm the principal features of our method (approximation power and conditioning for both full and linear-subspace stiffness matrices) for a simple model problem. Specifically, we consider

$$-\Delta u = 0 \quad \text{on } \Omega = (0,1)^2, \qquad u|_{\partial\Omega} = \begin{cases} \sin \pi x & \text{if } y = 0, \\ 0 & \text{else,} \end{cases} \tag{6.1}$$

with the exact solution $u(x,y) = \sin \pi x \, \frac{\sinh(\pi(1-y))}{\sinh \pi}$. Our calculations are performed with the code CONCEPTS [29]. For quadrilateral elements, this general *hp*-FEM code

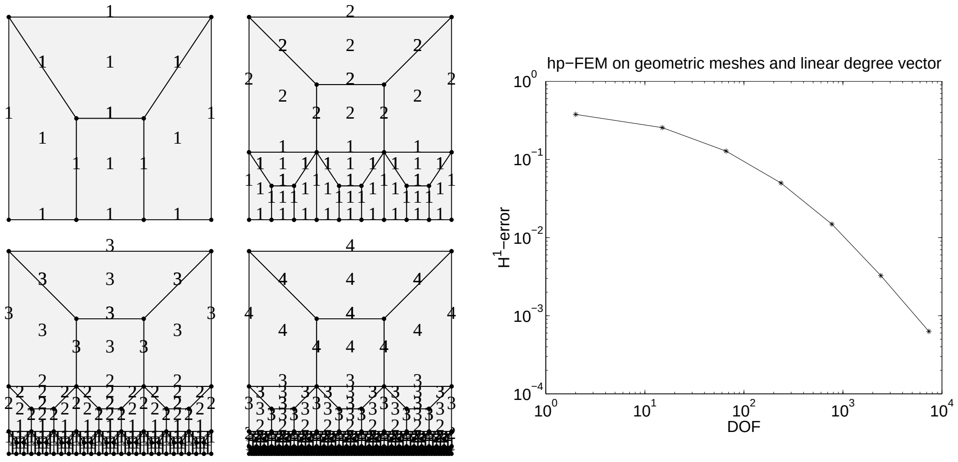


FIG. 6.1. *Geometric meshes for levels* 1–4 *and linear degree vector (left), and* H^1*-error for hp-FEM versus* N *(right).*

TABLE 6.1
Iteration count in the case $p = 1$ *with the mesh in Figure* 2.1.

Level	6	7	8	9	10	11	12
N_{it}	27	30	32	35	37	38	41
N_Γ	376	760	1528	3064	6136	12280	24568
N_Ω	665	1417	2937	5993	12121	24300	48953

employs the so-called Babuška–Szabó shape functions, which are the tensor products of the following one-dimensional shape functions defined on $(-1, 1)$ (we refer to [42, 40] for the details):

$$\varphi_1(x) = \frac{1}{2}(1-x), \quad \varphi_2(x) = \frac{1}{2}(1+x), \quad \varphi_i(x) = \frac{1}{\|L_{i-3}\|_{L^2(-1,1)}} \int_{-1}^{x} L_{i-3}(t)\,dt, \; i \geq 3.$$

Here, the polynomials L_i are the usual Legendre polynomials. First, we check the convergence result of Theorem 2.13, which asserts that for suitable linear degree vector the hp-FEM yields $\|u - u_N\|_{H^1(\Omega)} \leq Ch$. We check this assertion for the meshes and linear degree vectors depicted in Figure 6.1. The mesh size on the portion $y = 0$ of $\partial\Omega$ takes the role of the boundary mesh size h in the statement of Theorem 2.13. The convergence behavior (H^1-error versus number of degrees of freedom DOF $=$ $\dim S^{\mathbf{p}}(\mathcal{T}, \Omega)$) is given in Figure 6.1.

We now turn to our results in Proposition 3.4 concerning the conditioning of the stiffness matrix. For the present Dirichlet problem and $p = 1$, Proposition 3.4 asserts that the condition number is bounded uniformly in h for the meshes of Figure 6.1. The numerical results can be found in Figure 6.2. A second numerical example for the condition number estimates for the stiffness matrix is shown in Table 6.1. For meshes as depicted in Figure 2.1 and polynomial degree $p = 1$, we present in Table 6.1 the number of boundary nodes N_Γ, the number N_Ω of nodes in Ω, as well as the number of CG-iterations N_{it} (with diagonal preconditioning) to reach a residual with l^2-norm below 10^{-6}.

We finally consider the condition number of the full stiffness matrix on the meshes and polynomial degree distributions as depicted in Figure 6.1. From [30] and a rea-

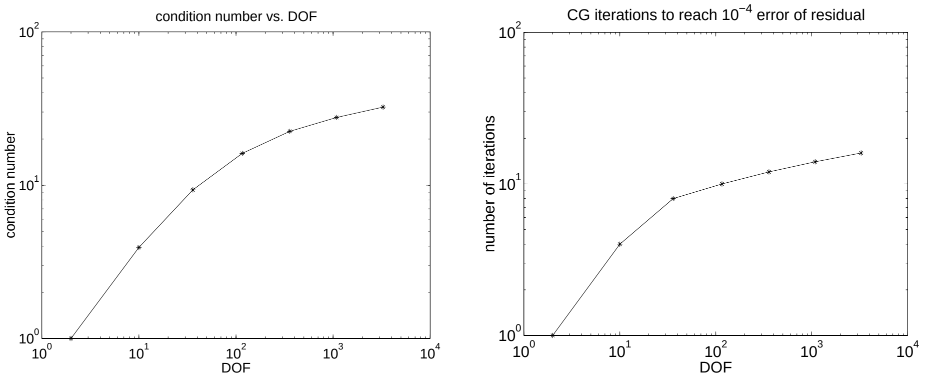


FIG. 6.2. *Condition number of stiffness matrix number of CG-iterations required for a residual of* 10^{-4} *using p.w. linear elements.*

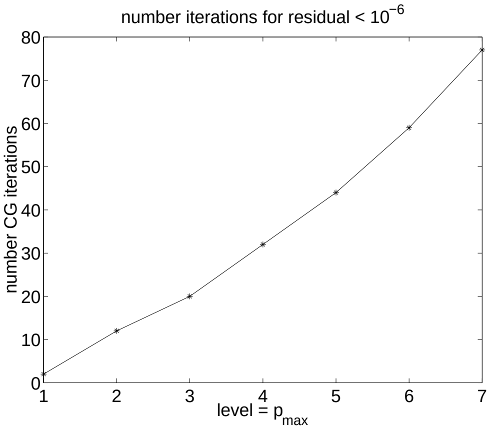


FIG. 6.3. *Iteration count for the full system versus the number of levels.*

soning as in the proof of Proposition 3.4 (see also [31]), the condition number of the full stiffness matrix is bounded by

$$\kappa(A) \leq Cp^4, \tag{6.2}$$

where C is independent of h. Figure 6.3 presents the number of CG-iterations (without preconditioning) to reach a residual of 10^{-6}. The numerical results are slightly better than the growth of $O(p^2)$ expected from (6.2).

Appendix A. Analytic regularity results. In the present section, we are interested in analytic regularity results for solutions to the equation

$$\mathcal{L}u := -\nabla \cdot (A(x)\nabla u) + b(x) \cdot \nabla u + a_0(x)u = f(x) \quad \text{on } \Omega. \tag{A.1}$$

Here, the symmetric matrix A is uniformly positive definite and b is a vector. We furthermore assume the coefficients A, b, c to be analytic; i.e., we stipulate the existence of C_d, $\gamma_d > 0$ such that

$$\|\nabla^p A\|_{L^\infty(\Omega)} + \|\nabla^p b\|_{L^\infty(\Omega)} + \|\nabla^p a_0\|_{L^\infty(\Omega)} \leq C_d \gamma_d^p p! \qquad \forall p \in \mathbb{N}_0. \tag{A.2}$$

The aim of the present section is the proof of the following analytic regularity result for the solution u of (A.1).

THEOREM A.1. *Let $\Omega \subset \mathbb{R}^2$ be a bounded Lipschitz domain with boundary $\partial\Omega$. Let the distance function r be given by* (1.15). *Let f be analytic on Ω and satisfy for some $\delta \in (0,1]$ and C_f, $\gamma_f > 0$*

$$\|r^{p+1-\delta}\nabla^p f\|_{L^2(\Omega)} \leq C_f \gamma_f^p p! \qquad \forall p \in \mathbb{N}_0. \tag{A.3}$$

Finally, let $u \in H^{1+\delta}(\Omega)$ solve (*A.1*) *with data A, b, a_0 satisfying* (*A.2*). *Then there exist C, $\gamma > 0$ depending only on Ω, C_d, C_f, γ_d, γ_f, and δ such that*

$$\|r^{p+1-\delta}\nabla^{p+2}u\|_{L^2(\Omega)} \leq C\gamma^p p! \|u\|_{H^{1+\delta}(\Omega)} \qquad \forall p \in \mathbb{N}_0.$$

Remark A.2. The restrictions on the data A, b, c are not minimal: blow-up akin to that in (A.3) is conceivable. The theorem can also be extended to the case of piecewise analytic data A, b, c.

In order to prove Theorem A.1, we start with the following lemma.

LEMMA A.3. *Let B_R be a ball of radius $R \leq 1$. Assume that A, b, c satisfy* (A.2) *with Ω replaced with B_R. Assume that f satisfies on B_R*

$$\|\nabla^p f\|_{L^2(B_R)} \leq C_f \gamma_f^p p! R^{-p-1+\delta} \qquad \forall p \in \mathbb{N}_0$$

for some C_f, $\gamma_f > 0$, $\delta \in (0,1]$. Let $u \in H^{1+\delta}(B_R)$ solve

$$-\nabla \cdot (A\nabla u) + b \cdot \nabla u + cu = f \qquad \text{on } B_R.$$

Then for every $c \in (0,1)$ there exist constants C, $\gamma > 0$ depending only on C_d, γ_d, C_f, γ_f, δ, and c such that

$$\|\nabla^{p+2}u\|_{L^2(B_{cR})} \leq CR^{-p-1+\delta}\gamma^p p! \left[\|u\|_{H^1(B_R)} + |\nabla u|_{H^\delta(B_R)}\right],$$

where $|\nabla u|_{H^\delta(B_R)}$ stands for the Aronszajn–Slobodeckij norm.

Proof. Using the techniques of [35], we have (see [32, Proposition 5.5.1] for the details)

$$R^p\|\nabla^{p+2}u\|_{L^2(B_{cR})} \leq C\gamma^p p! \left[R^{-1}\|\nabla u\|_{L^2(B_R)} + R^{-2}\|u\|_{L^2(B_R)} + C_f R^{-1+\delta}\right], \tag{A.4}$$

where C, $\gamma > 0$ depend only on γ_f, C_d, and γ_d. For an arbitrary linear function l, the function $u - l$ satisfies

$$L(u-l) = \tilde{f} := f - \mathcal{L}l.$$

From the assumptions on the data, we then conclude that for all $p \in \mathbb{N}_0$

$$\begin{aligned}\|\nabla^p \tilde{f}\|_{L^2(B_R)} &\leq C_f\gamma_f^p p! R^{-p-1+\delta} + C\gamma^p p! \|l\|_{H^1(B_R)} \\ &\leq \tilde{C}\tilde{\gamma}^p R^{-p-1+\delta}\left[C_f + R^{1-\delta}\|l\|_{H^1(B_R)}\right],\end{aligned}$$

where the constants $\tilde{C}$, $\tilde{\gamma} > 0$ depend on C_d, γ_d, and γ_f. Applying (A.4) with u replaced with $u - l$, we obtain

$$\begin{aligned}R^p\|\nabla^{p+2}u\|_{L^2(B_{cR})} \leq C\gamma^p p! [&R^{-1}\|\nabla(u-l)\|_{L^2(B_R)} \\ &+ R^{-2}\|u-l\|_{L^2(B_R)} + R^{-1+\delta}C_f + \|l\|_{H^1(B_R)}].\end{aligned} \tag{A.5}$$

The assumption $u \in H^{1+\delta}(B_R)$ implies the existence of a linear function l such that

$$\|u - l\|_{L^2(B_R)} + R\|\nabla(u - l)\|_{L^2(B_R)} \le CR^{1+\delta}|\nabla u|_{H^\delta(B_R)},$$

where $|\cdot|_{H^\delta(B_R)}$ denotes the Aronszajn–Slobodeckij seminorm. Using $R \le 1$, we get

$$R^{1-\delta+p}\|\nabla^{p+2}u\|_{L^2(B_{cR})} \le C\gamma^p p! \left[|\nabla u|_{H^\delta(B_R)} + \|u\|_{H^1(B_R)} + C_f\right]. \qquad \square$$

Proof of Theorem A.1. Using the Besicovitch covering theorem (see, e.g., [45]), we can construct a covering of Ω by a countable collection $\mathcal{B} = \{B_i \,|\, i \in \mathbb{N}\}$ of closed balls B_i with the following properties:

1. $B_i = B_{r_i}(x_i)$, where $r_i = c' \operatorname{dist}(x_i, \partial\Omega)$ for some fixed $c' \in (0, 1)$;
2. there exists $N \in \mathbb{N}$ such that for all $x \in \Omega$: $|\{i \in \mathbb{N} \,|\, x \in B_i\}| \le N$;
3. there exists $c \in (0, 1)$ such that $\Omega \subset \cup_{i\in\mathbb{N}} B_{cr_i}(x_i)$.

Set

$$C_i^2 := \sum_{p\in\mathbb{N}_0} \frac{1}{(2\gamma_f)^{2p}(p!)^2}\|r^{p+1-\delta}\nabla^p f\|^2_{L^2(B_i)}.$$

The properties of the covering $\mathcal{B}$ and the assumption (A.3) imply

$$r_i^{p+1-\delta}\|\nabla^p f\|_{L^2(B_i)} \le C_i \left(\frac{c'}{1-c'}\right)^{p+1-\delta} (2\gamma_f)^p p! \qquad \forall p \in \mathbb{N}_0,$$

$$\sum_{i\in\mathbb{N}} C_i^2 \le NC_f^2 \sum_{p\in\mathbb{N}_0} \frac{1}{(2\gamma_f)^{2p}(p!)^2}\gamma_f^{2p}(p!)^2 = \frac{4}{3}NC_f^2.$$

From Lemma A.3 we get

$$r_i^{p+1-\delta}\|\nabla^{p+2}u\|_{L^2(B_{cr_i}(x_i))} \le \gamma^p p! \left[C_i + \|u\|_{H^1(B_i)} + |\nabla u|_{H^\delta(B_i)}\right].$$

Using the fact that $\Omega \subset \cup_i B_{cr_i}(x_i)$ and the finite overlap property of the covering $\mathcal{B}$, we get by summation over all balls B_i

$$\begin{aligned}\|r^{p+1-\delta}\nabla^{p+2}u\|^2_{L^2(\Omega)} &\le \left(\frac{1+c'}{c'}\right)^{p+1-\delta} \sum_{i\in\mathbb{N}} r_i^{2(p+1-\delta)}\|\nabla^{p+2}u\|^2_{L^2(B_{cr_i}(x_i))} \\ &\le C(\gamma^p p!)^2 \sum_{i\in\mathbb{N}} C_i^2 + \|u\|^2_{H^1(B_i)} + |\nabla u|^2_{H^\delta(B_i)} \\ &\le C(\gamma^p p!)^2[C_f^2 + \|u\|^2_{H^{1+\delta}(\Omega)}]\end{aligned}$$

for suitable constants C, γ. $\square$

The restriction $\delta \in (0, 1]$ in Theorem A.1 can be removed as follows.

Corollary A.4. *Let $\Omega \subset \mathbb{R}^2$ be a bounded Lipschitz domain and r be defined by* (1.15). *Let $k = \kappa + \delta$ with $\kappa \in \mathbb{N}_0$, $\delta \in [0, 1)$ be given. Assume that $u \in H^{1+k}(\Omega)$ satisfies* (A.1), *with coefficients A, b, c satisfying* (A.2) *and f satisfying*

$$\text{(A.6)} \qquad \|f\|_{H^{\kappa-1}(\Omega)} + \|r^{p+1-\delta}\nabla^{p+\kappa}f\|_{L^2(\Omega)} \le C_f\gamma_f^p p! \qquad \forall p \in \mathbb{N}_0.$$

Then there exist C, $\gamma > 0$ such that

$$\|r^{p+1-\delta}\nabla^{p+2+\kappa}u\|_{L^2(\Omega)} \le C\gamma^p p! \left[C_f + \|u\|_{H^{1+k}(\Omega)}\right].$$

Proof. The corollary is proved by induction on $\kappa \in \mathbb{N}_0$. For $\kappa = 0$, the result holds true by Theorem A.1. Assuming that it holds for all $0 \le \kappa' < \kappa$ for some $\kappa \in \mathbb{N}$, we show that it holds for $\kappa + 1$. The induction hypothesis implies that

$$\|r^{p+1-\delta}\nabla^{p+2+\kappa'}u\|_{L^2(\Omega)} \le C[C_f + \|u\|_{H^{1+\delta+\kappa'}(\Omega)}]\gamma^p p! \qquad \forall p \in \mathbb{N}_0, \quad 0 \le \kappa' < \kappa.$$

Differentiating (A.1) κ times, it is easy to see that $D^\alpha u$ with $|\alpha| = \kappa$ satisfies a differential equation of the form

$$LD^\alpha u = f_\alpha := D^\alpha f + \tilde{u}_\alpha,$$

where $\tilde{u}_\alpha = \sum_{|\beta|\le\kappa+1} \lambda_{\alpha,\beta} D^\beta u$ for some analytic functions $\lambda_{\alpha,\beta}$. The induction hypothesis and the assumptions on f imply that

$$\|r^{p+1-\delta}\nabla^p f_\alpha\|_{L^2(\Omega)} \le C\left[C_f + \|u\|_{H^{1+\delta+\kappa}(\Omega)}\right]\gamma^p p! \qquad \forall p \in \mathbb{N}_0.$$

(See [32, Lemma 4.3.3] for a rigorous proof that products $\lambda_{\alpha,\beta}D^\beta u$ satisfy the desired bounds.) Theorem A.1 therefore allows us to conclude the induction argument. □

Appendix B. Compact embeddings. For $\delta \in (0, 1)$ and a domain $\Omega \subset \mathbb{R}^2$ we define as usual

$$\|u\|^2_{\widetilde{H}^\delta(\Omega)} := \int_\Omega \int_\Omega \frac{|u(x) - u(y)|^2}{|x - y|^{2+2\delta}}\, dxdy + \int_\Omega \frac{|u(x)|^2}{|\mathrm{dist}\,(x, \partial\Omega)|^{2\delta}}\, dx. \tag{B.1}$$

The following lemma is due to von Petersdorff [43] (see also [2] for a proof).

LEMMA B.1. *Let $\mathcal{B} = \{B_i \,|\, i \in \mathbb{N}\}$ be a covering of a domain $\Omega \subset \mathbb{R}^2$ that satisfies a* finite overlap property*; i.e., there exists $N \in \mathbb{N}$ such that*

$$\sup_{x\in\Omega} \mathrm{card}\,\{i \in \mathbb{N} \,|\, x \in B_i\} \le N. \tag{B.2}$$

Then

$$\sum_{i,j\in\mathbb{N}} \int_{B_i}\int_{B_j} \frac{|u(x) - u(y)|^2}{|x - y|^{2+2\delta}}\, dxdy \le N\left(3 + \frac{\pi}{\delta}\right) \sum_{i\in\mathbb{N}} \|u\|^2_{\widetilde{H}^\delta(B_i)}.$$

Let $\Omega \subset \mathbb{R}^2$ be a bounded Lipschitz domain, and let r be defined by (1.15). For $k \in \mathbb{N}$ and $\beta \in (0, 1)$ introduce the norm

$$\|u\|^2_{H^k_\beta(\Omega)} := \|u\|^2_{H^{k-1}(\Omega)} + \|r^\beta \nabla^k u\|^2_{L^2(\Omega)} \tag{B.3}$$

and define the spaces $H^k_\beta(\Omega)$ as the completion of $C^\infty(\overline{\Omega})$ under this norm.

LEMMA B.2. *Let $\Omega \subset \mathbb{R}^2$ be a Lipschitz domain, $\beta \in (0, 1)$. Then for each $0 \le \delta < \min\{1/2, 1 - \beta\}$ there exists $C > 0$ such that*

$$\|u\|_{H^\delta(\Omega)} \le C\|u\|_{H^1_\beta(\Omega)} \qquad \forall u \in H^1_\beta(\Omega).$$

Proof. We start with the following variant of Hardy's inequality in one dimension:

$$\int_0^\infty x^{-2\delta}|u(x)|^2\, dx \le C_\delta \left[\int_0^\infty x^{2(1-\delta)}|u'(x)|^2\, dx + \int_0^\infty |u(x)|^2\, dx\right] \tag{B.4}$$

for all functions u such that the right-hand side is finite (see, e.g., [22, Theorem 330]). In the standard way by locally flattening the boundary (note that Ω is assumed to be Lipschitz), we then obtain from (B.4) the following two-dimensional result: For every $\delta \in [0, 1/2)$ there exists $C_\delta > 0$ such that

$$\|r^{-\delta} u\|_{L^2(\Omega)} \leq C_\delta \left[\|r^{1-\delta}\nabla u\|_{L^2(\Omega)} + \|u\|_{L^2(\Omega)}\right] \qquad \forall u \in H^1_\beta(\Omega). \tag{B.5}$$

We now turn to the proof of the lemma. Let $\mathcal{B} = \{B_i \,|\, i \in \mathbb{N}\}$ be the cover of Ω by balls given in the proof of Theorem A.1. For each i, let $\hat{u}_i = u \circ F_i$, where the affine map $F_i : \hat{B} \to B_i$ maps the unit ball $\hat{B}$ onto B_i. Sobolev's embedding theorem implies

$$\|\hat{u}_i\|^2_{\widetilde{H}^\delta(\hat{B})} \leq C \left[\|\nabla \hat{u}_i\|^2_{L^2(\hat{B})} + \|\hat{u}_i\|^2_{L^2(\hat{B})}\right].$$

Here, we used the assumption $\delta < 1/2$. Scaling back to the balls B_i, we conclude

$$r_i^{-2+2\delta}\|u\|^2_{\widetilde{H}^\delta(B_i)} \leq C \left[\|\nabla u\|^2_{L^2(B_i)} + r_i^{-2}\|u\|^2_{L^2(B_i)}\right].$$

Thus, using the properties of the covering

$$\|u\|^2_{\widetilde{H}^\delta(B_i)} \leq C \left[\|r^{1-\delta}\nabla u\|^2_{L^2(B_i)} + \|r^{-\delta} u\|^2_{L^2(B_i)}\right].$$

Summing over all balls B_i of the covering and using its overlap property together with Lemma B.1, we arrive at

$$\begin{aligned}\|u\|^2_{H^\delta(\Omega)} &\leq C \sum_i \|u\|^2_{\widetilde{H}^\delta(B_i)} \leq C \left[\|r^{1-\delta}\nabla u\|^2_{L^2(\Omega)} + \|r^{-\delta} u\|^2_{L^2(\Omega)}\right] \\ &\leq C \left[\|r^{1-\delta}\nabla u\|^2_{L^2(\Omega)} + \|u\|^2_{L^2(\Omega)}\right],\end{aligned}$$

where, in the last step, we appealed to (B.5). Noting $1 - \delta > \beta$ finishes the proof. □

A consequence of Lemma B.2 is the following.

Theorem B.3. *Let $\Omega \subset \mathbb{R}^2$ be a bounded Lipschitz domain, $\beta \in (0, 1)$. Then for each $0 \leq \delta < \min\{1/2, 1 - \beta\}$ the embedding $H^2_\beta(\Omega) \subset H^{1+\delta}(\Omega)$ is compact. In particular, $H^2_\beta(\Omega) \subset C^0(\overline{\Omega})$.*

Proof. The embedding $H^2_\beta(\Omega) \subset C^0(\overline{\Omega})$ follows from $H^{1+\delta}(\Omega) \subset C^0(\overline{\Omega})$, valid for all $\delta > 0$ by Sobolev's embedding theorem. Because $H^{1+\delta}(\Omega) \subset H^{1+\delta'}(\Omega)$ is compactly embedded for $0 \leq \delta < \delta'$, it suffices to show that for each δ there exists $C > 0$ such that $\|u\|_{H^{1+\delta}(\Omega)} \leq C\|u\|_{H^2_\beta(\Omega)}$, which follows from Lemma B.2 applied to ∇u. □

Theorem B.4. *Let $\hat{K}$ be the reference square or the reference triangle. Let $r(x) = \operatorname{dist}(x, \hat{K})$, $\beta \in (0, 1)$. For $u \in H^2_\beta(\hat{K})$ let Iu be the linear (if $\hat{K} = T$) or the bilinear (if $\hat{K} = S$) interpolant of u. Then*

$$\|u - Iu\|_{H^2_\beta(\hat{K})} \leq C\|r^\beta \nabla^2 u\|_{L^2(\hat{K})}.$$

Proof. Let A_1, A_2, A_3 be three vertices of $\hat{K}$. Exploiting the compactness result of Theorem B.3 in the same way as in the proof of [40, Lemma 4.16], we obtain the existence of $C > 0$ such that

$$\|u\|^2_{H^2_\beta(\hat{K})} \leq C \left[\|r^\beta \nabla^2 u\|^2_{L^2(\hat{K})} + \sum_{i=1}^3 |u(A_i)|^2\right] \qquad \forall u \in H^2_\beta(\hat{K}).$$

As $u(A_i) = Iu(A_i)$ by construction, the result follows. □

Appendix C. Sobolev spaces on the boundary of polygons.

LEMMA C.1. *Let $I = [a, b]$, $I' = [a', b']$ be intervals. Let $\varphi : I \to I'$ be a piecewise smooth bijection. Then for $|s| < 3/2$ the map $u \mapsto u \circ \phi$ is an isomorphism between $H^s(I)$ and $H^s(I')$. The same result holds for the spaces $H^s_{per}(I)$, $H^s_{per}(I')$ of periodic functions if the piecewise smooth function ϕ satisfies additionally $\varphi(a) = \varphi(b)$.*

In particular, for a polygon Ω, let $\varphi : I \to \partial\Omega$ be the parametrization by arc length. Then the map $u \mapsto u \circ \phi$ provides an isomorphism between the spaces $H^s(\partial\Omega)$ and $H^s_{per}(I)$ for $|s| < 3/2$.

Proof. This result is due to Grisvard; see, e.g., [11, Corollary 2.8]. □

For a mesh $\mathcal{T} = \{K\}$, we define piecewise polynomial spaces $S^{p,k}(I, \mathcal{T})$, p, $k \in \mathbb{N}_0$, by

$$S^{p,k}(I, \mathcal{T}) = \{u \in H^k(I) \,|\, u|_K \text{ is a polynomial of degree } p\}.$$

LEMMA C.2. *Let $I \subset \mathbb{R}$ be an interval and $\mathcal{T}$ a quasi-uniform mesh on I with mesh size h; i.e., the nodes $x_0 < x_1 < \cdots < x_N$ of the mesh satisfy $\gamma^{-1}h \le x_{i+1} - x_i \le \gamma h$, $i = 0, \ldots, N-1$, for some $\gamma > 0$. Then for every $\varepsilon \in [0, 1/2)$ and every $p \in \mathbb{N}_0$ there exists $C_{\varepsilon,p} > 0$ such that*

$$\|u\|_{H^\varepsilon(I)} \le C_{\varepsilon,p} h^{-\varepsilon} \|u\|_{L^2(I)} \qquad \forall u \in S^{p,0}(I, \mathcal{T}), \tag{C.1}$$

$$\|u\|_{H^{1+\varepsilon}(I)} \le C_{\varepsilon,p} h^{-(1+\varepsilon)} \|u\|_{L^2(I)} \qquad \forall u \in S^{p,1}(I, \mathcal{T}). \tag{C.2}$$

Proof. We first show (C.1). Equation (C.1) is trivially valid for $\varepsilon = 0$. Let therefore $\varepsilon \in (0, 1/2)$. We characterize the norm $\|\cdot\|_{H^\varepsilon(I)}$ using the K-functional; that is, we have for all $u \in H^\varepsilon(I)$

$$\|u\|^2_{H^\varepsilon(I)} \sim \int_0^\infty t^{-2\varepsilon-1} K^2(u, t)\, dt, \qquad K^2(u, t) := \inf_{v \in H^1(I)} \|u - v\|^2_{L^2(I)} + t^2 |v'|^2_{L^2(I)}.$$

We choose v in the infimum appropriately. For $t \ge h$ we take $v \equiv 0$ and get

$$\int_h^\infty t^{-2\varepsilon-1} K^2(u, t)\, dt \le \int_h^\infty t^{-2\varepsilon-1} \|u\|^2_{L^2(I)}\, dt = \frac{1}{2\varepsilon} h^{-2\varepsilon} \|u\|^2_{L^2(I)}. \tag{C.3}$$

In the range $t \in (0, h)$ we proceed as follows. Let $\varphi \in C_0^\infty(\mathbb{R})$ be a smooth function with $0 \le \varphi(x) \le 1$ that is supported by $[-1, 1]$ and that satisfies $\varphi \equiv 1$ on $[-1/2, 1/2]$. For $\delta > 0$ we set $\varphi_\delta(x) := \varphi(x/\delta)$. For $t \in (0, h)$ we then set

$$\psi_t(x) := 1 - \sum_{i=0}^N \varphi_{t/\gamma}(x - x_i)$$

(γ is the quasiuniformity constant of the mesh $\mathcal{T}$) and choose the function v in the infimum defining $K(t, u)$ as $v(x) := \psi_t(x) u(x)$. Noting the support properties of ψ_t and standard polynomial inverse estimates, we get with the shorthand $I_i := (x_i, x_{i+1})$

$$\begin{aligned}\|u - v\|^2_{L^2(I)} = \|(1 - \psi_t) u\|^2_{L^2(I)} &\le \sum_{i=0}^{N-1} t\gamma^{-1} \|u\|^2_{L^\infty(I_i)} \le C \sum_{i=0}^{N-1} t h^{-1} \|u\|^2_{L^2(I_i)} \\ &\le C t h^{-1} \|u\|^2_{L^2(I)},\end{aligned}$$

$$\begin{aligned}\|v'\|^2_{L^2(I)} &\le \sum_{i=0}^{N-1} \|(u\psi_t)'\|^2_{L^2(I_i)} \le 2\sum_{i=0}^{N-1} \|u'\|^2_{L^2(I_i)} + \|u\|^2_{L^\infty(I_i)}\|\psi_t'\|^2_{L^2(I_i)} \\ &\le C\sum_{i=0}^{N-1} h^{-2}\|u\|^2_{L^2(I_i)} + h^{-1}t^{-1}\|u\|^2_{L^2(I_i)} \le Ch^{-2}\left(1+\frac{h}{t}\right)\|u\|^2_{L^2(I)}.\end{aligned}$$

A straightforward calculation then shows

$$\int_0^h t^{-1-2\varepsilon}K^2(t,u)\,dt \le Ch^{-2\varepsilon}\|u\|^2_{L^2(I)},$$

where the constant C depends on p, ε, I, and γ but is independent of u and h. This proves (C.1). For (C.2), we note again that the case $\varepsilon = 0$ is a standard polynomial inverse estimate. For $\varepsilon \in (0, 1/2)$, we bound

$$\|u\|_{H^{1+\varepsilon}(I)} \le C\left[\|u\|_{L^2(I)} + \|u'\|_{L^2(I)} + \|u'\|_{H^\varepsilon(I)}\right] \le Ch^{-1}\|u\|_{L^2(I)} + C\|u'\|_{H^\varepsilon(I)}.$$

Since $u' \in S^{p-1,0}(I, \mathcal{T})$, we may apply (C.1) to the second term to get

$$\|u\|_{H^{1+\varepsilon}(I)} \le Ch^{-1}\|u\|_{L^2(I)} + C_{\varepsilon,p}h^{-\varepsilon}\|u'\|_{L^2(I)} \le C_{\varepsilon,p}h^{-1-\varepsilon}\|u\|_{L^2(I)}. \qquad \square$$

PROPOSITION C.3. *Let $I \subset \mathbb{R}$ be an interval, $\mathcal{T}$ be a quasi-uniform mesh on I with quasiuniformity constant γ. Let $P : L^2(I) \to S^{p,1}(I, \mathcal{T})$ be a linear operator with*

(i) *$\|Pu\|_{L^2(I)} \le C_{stable}\|u\|_{L^2(I)}$ for all $u \in L^2(I)$;*
(ii) *$Pu = u$ for all $u \in S^{1,1}(I, \mathcal{T})$.*

Then for every $\varepsilon \in [0, 3/2)$ there exists a constant $C_\varepsilon > 0$ depending only on p, C_{stable}, γ, and ε such that

$$\|Pu\|_{H^\varepsilon(I)} \le C_\varepsilon\|u\|_{H^\varepsilon(I)}.$$

Proof. Let $u \in H^\varepsilon(I)$ be arbitrary. By simultaneous approximation in Sobolev spaces (see, e.g., [9]) there exists $C_\varepsilon > 0$ independent of h such that for every $u \in H^\varepsilon(I)$ we can find $q_u \in S^{1,1}(I, \mathcal{T})$ with

$$h^\varepsilon\|u - q_u\|_{H^\varepsilon(I)} + \|u - q_u\|_{L^2(I)} \le C_\varepsilon h^\varepsilon\|u\|_{H^\varepsilon(I)}. \tag{C.4}$$

Exploiting the reproduction assumption (ii) and the stability assumption (i), we get

$$\|u - Pu\|_{L^2(I)} \le \|u - q_u\|_{L^2(I)} + \|P(u - q_u)\|_{L^2(I)} \le (1 + C_{stable})\|u - q_u\|_{L^2(I)}. \tag{C.5}$$

We can therefore estimate with Lemma C.2

$$\begin{aligned}\|Pu\|_{H^\varepsilon(I)} &\le \|u\|_{H^\varepsilon(I)} + \|u - Pu\|_{H^\varepsilon(I)} \le \|u\|_{H^\varepsilon(I)} + \|u - q_u\|_{H^\varepsilon(I)} + \|q_u - Pu\|_{H^\varepsilon(I)} \\ &\le C_\varepsilon\|u\|_{H^\varepsilon(I)} + C_\varepsilon h^{-\varepsilon}\|q_u - Pu\|_{L^2(I)} \\ &\le C_\varepsilon\|u\|_{H^\varepsilon(I)} + C_\varepsilon h^{-\varepsilon}\left\{\|u - q_u\|_{L^2(I)} + \|u - Pu\|_{L^2(I)}\right\} \\ &\le C_\varepsilon\|u\|_{H^\varepsilon(I)} + C_\varepsilon(2 + C_{stable})h^{-\varepsilon}\|u - q_u\|_{L^2(I)} \le C\|u\|_{H^\varepsilon(I)},\end{aligned}$$

where we have used (C.5) and (C.4). This concludes the proof of the proposition. $\square$

Remark C.4. It can be checked that Proposition C.3 also holds if the linear operator P is replaced with an operator $P : H^1(I) \to S^{p,1}(I, \mathcal{T})$ that is stable in $H^1(I)$, i.e., $\|Pu\|_{H^1(I)} \le C_{stable}\|u\|_{H^1(I)}$ for all $u \in H^1(I)$, and that satisfies assumption (ii) of Proposition C.3.

It can also be checked that Proposition C.3 remains valid for periodic functions, i.e., if $P : L^2(I) \to S^{p,1}(I, \mathcal{T}) \cap H^1_{per}(I)$ satisfies $Pu = u$ for all $u \in H^s_{per}(I)$.

Acknowledgment. The authors would like to thank Prof. Dr. W. Hackbusch for valuable comments on the topic of this paper.

REFERENCES

[1] M. Ainsworth and B. Guo, *An additive Schwarz preconditioner for p-version boundary element approximation of the hypersingular operator in three dimensions*, Numer. Math., 85 (2000), pp. 343–366.

[2] M. Ainsworth, W. McLean, and T. Tran, *The conditioning of boundary element equations on locally refined meshes and preconditioning by diagonal scaling*, SIAM J. Numer. Anal., 36 (1999), pp. 1901–1932.

[3] I. Babuška, A. Craig, J. Mandel, and J. Pitkäranta, *Efficient preconditioning for the p-version finite element method in two dimensions*, SIAM J. Numer. Anal., 28 (1991), pp. 624–661.

[4] I. Babuška and B. Q. Guo, *Regularity of the solution of elliptic problems with piecewise analytic data. Part* I. *Boundary value problems for linear elliptic equation of second order*, SIAM J. Math. Anal., 19 (1988), pp. 172–203.

[5] I. Babuška and B. Q. Guo, *Regularity of the solution of elliptic problems with piecewise analytic data,* II. *The trace spaces and application to the boundary value problems with nonhomogeneous boundary conditions*, SIAM J. Math. Anal., 20 (1989), pp. 763–781.

[6] C. Bacuta, *Interpolation between Subspaces of Hilbert Spaces and Application to Shift Theorems for Elliptic Boundary Value Problems and Finite Element Methods*, Ph.D. thesis, Texas A & M University, College Station, TX, 2000.

[7] R. E. Bank and R. L. Scott, *On the conditioning of finite element equations with highly refined meshes*, SIAM J. Numer. Anal., 26 (1989), pp. 1383–1394.

[8] M. Bebendorf and W. Hackbusch, *Existence of $\mathcal{H}$-matrix approximants to the inverse FE-matrix of elliptic operators with L^∞-coefficients*, Numer. Math., to appear.

[9] J. Bramble and R. Scott, *Simultaneous approximation in scales of Banach spaces*, Math. Comput., 32 (1978), pp. 947–954.

[10] M. Costabel, *Boundary integral operators on Lipschitz domains: Elementary results*, SIAM J. Math. Anal., 19 (1988), pp. 613–626.

[11] M. Costabel and E. Stephan, *Boundary integral equations for mixed boundary value problems in polygonal domains and Galerkin approximation*, Math. Models Methods Mechanics, 15 (1985), pp. 175–251.

[12] W. Dahmen, S. Prössdorf, and R. Schneider, *Wavelet approximation methods for pseudodifferential equations* II: *Matrix compression and fast solution*, Adv. Comput. Math., 1 (1993), pp. 259–335.

[13] L. Demkowicz, K. Gerdes, C. Schwab, A. Bajer, and T. Walsh, *A general and flexible Fortran* 90 *hp-FE code*, Computing and Visualization in Science, 1 (1998), pp. 145–163.

[14] L. Greengard and V. Rokhlin, *A new version of the fast multipole method for the Laplace in three dimensions*, in Acta Numerica 6 (1997), A. Iserles, ed., Cambridge University Press, London, 1997, pp. 229–269.

[15] P. Grisvard, *Elliptic Problems in Nonsmooth Domains*, Pitman, Boston, 1985.

[16] B. Guo and W. Cao, *An additive Schwarz method for the h-p version of the finite element method in three dimensions*, SIAM J. Numer. Anal., 35 (1998), pp. 632–654.

[17] G. Haase and S. V. Nepomnyaschikh, *Explicit extension operators on hierarchical grids*, East-West J. Numer. Anal., 5 (1997), pp. 231–248.

[18] W. Hackbusch, *Integral Equations. Theory and Numerical Treatment*, Birkhäuser Boston, Cambridge, MA, 1995.

[19] W. Hackbusch, *A sparse matrix arithmetic based on $\mathcal{H}$-matrices. Part* I: *Introduction to $\mathcal{H}$-matrices*, Computing, 62 (1999), pp. 89–108.

[20] W. Hackbusch and B.N. Khoromskij, *A sparse $\mathcal{H}$-matrix arithmetic. Part* II: *Application to multidimensional problems,* Computing, 64 (2000), pp. 21–47.

[21] W. Hackbusch and Z.P. Nowak, *On the fast matrix multiplication in the boundary element method by panel clustering*, Numer. Math., 54 (1989), pp. 463–491.

[22] G. Hardy, J.E. Littlewood, and G. Pólya, *Inequalities*, Cambridge University Press, Cambridge, UK, 1991.

[23] T.J.R. Hughes, *The Finite Element Method*, Prentice–Hall, Englewood Cliffs, NJ, 1987.

[24] B.N. Khoromskij, *On fast computation with the inverse to harmonic potential operators*, J. Numer. Lin. Alg. Appl., 3 (1996), pp. 91–111.

[25] B.N. Khoromskij, *Lectures on Multilevel Schur-Complement Methods for Elliptic Differential*

Equations, Technical report 98/3, ICA, Universität Stuttgart, Stuttgart, Germany, 1998.

[26] B.N. Khoromskij and J.M. Melenk, *An efficient direct solver for the boundary concentrated FEM in 2D*, Computing, 69 (2002), pp. 91–117.

[27] B.N. Khoromskij and S. Prössdorf, *Multilevel preconditioning on the refined interface and optimal boundary solvers for the Laplace equation*, Adv. Comput. Math., 4 (1995), pp. 331–355.

[28] B.N. Khoromskij and S. Prössdorf, *Fast computation with harmonic Poincaré–Steklov operators on nested refined meshes,* Adv. Comput. Math., 8 (1998), pp. 1–25.

[29] C. Lage, *Concept Oriented Design of Numerical Software*, Technical report 98–07, Seminar für Angewandte Mathematik, ETH Zürich, Zürich, Switzerland, 1998.

[30] J.F. Maitre and O. Pourquier, *Condition number and diagonal preconditioning: Comparison of the p version and the spectral element method*, Numer. Math., 74 (1996), pp. 69–84.

[31] J.M. Melenk, *On condition numbers in hp-FEM with Gauss–Lobatto-based shape functions*, J. Comput. Appl. Math., 139 (2001), pp. 21–48.

[32] J.M. Melenk, *hp Finite Element Methods for Singular Perturbations*, Lecture Notes in Math. 1796, Springer-Verlag, New York, 2002.

[33] J.M. Melenk and C. Schwab, *Fully Discrete hp-Finite Elements: Approximate Element Maps*, in preparation.

[34] J.M. Melenk and B. Wohlmuth, *On residual-based a posteriori error estimation in hp-FEM*, Adv. Comput. Math., 15 (2001), pp. 311–331.

[35] C.B. Morrey, *Multiple Integrals in the Calculus of Variations*, Springer-Verlag, New York, Berlin, 1966.

[36] J. Nečas, *Sur la coercivité des formes sesquilinéaires elliptiques*, Rev. Roumaine Math. Pures Appl., 9 (1964), pp. 47–69.

[37] T.J. Oden, L. Demkowicz, W. Rachowicz, and O. Hardy, *Towards a universal hp finite element strategy. Part* 1*. Constrained approximation and data structure*, Comput. Methods Appl. Mech. Engrg., 77 (1989), pp. 79–112.

[38] L. F. Pavarino and O. B. Widlund, *A polylogarithmic bound for an iterative substructuring method for spectral elements in three dimensions*, SIAM J. Numer. Anal., 33 (1996), pp. 1303–1357.

[39] J. Ruppert, *A Delaunay refinement algorithm for quality* 2*-dimensional mesh generation*, J. Algorithms, 18 (1995), pp. 548–585.

[40] C. Schwab, *p- and hp-Finite Element Methods*, Oxford University Press, London, 1998.

[41] L.R. Scott and S. Zhang, *Finite element interpolation of nonsmooth functions satisfying boundary conditions*, Math. Comput., 54 (1990), pp. 483–493.

[42] B. Szabó and I. Babuška, *Finite Element Analysis*, Wiley, New York, 1991.

[43] T. von Petersdorff, *Randwertprobleme der Elastizitätstheorie für Polyeder—Singularitäten und Approximation mit Randelementmethoden*, Ph.D. thesis, TH Darmstadt, Darmstadt, Germany, 1989.

[44] H. Yserentant, *Coarse grids spaces for domains with a complicated boundary*, Numer. Algorithms, 21 (1999), pp. 387–392.

[45] W.P. Ziemer, *Weakly Differentiable Functions*, Springer-Verlag, New York, Berlin, 1989.

SIAM J. NUMER. ANAL.
Vol. 41, No. 1, pp. 37–63

CONVERGENCE ANALYSIS OF A FINITE VOLUME METHOD FOR MAXWELL'S EQUATIONS IN NONHOMOGENEOUS MEDIA[*]

ERIC T. CHUNG[†], QIANG DU[‡], AND JUN ZOU[§]

Abstract. In this paper, we analyze a recently developed finite volume method for the time-dependent Maxwell's equations in a three-dimensional polyhedral domain composed of two dielectric materials with different parameter values for the electric permittivity and the magnetic permeability. Convergence and error estimates of the numerical scheme are established for general nonuniform tetrahedral triangulations of the physical domain. In the case of nonuniform rectangular grids, the scheme converges with second order accuracy in the discrete L^2-norm, despite the low regularity of the true solution over the entire domain. In particular, the finite volume method is shown to be superconvergent in the discrete $H(\text{curl};\Omega)$-norm. In addition, the explicit dependence of the error estimates on the material parameters is given.

Key words. finite volume method, Maxwell's equations, inhomogeneous medium, stability, convergence

AMS subject classifications. 65M12, 65M15, 78-08

PII. S0036142901398453

1. Introduction.

Let Ω be a general polyhedral domain in $\mathbb{R}^3$, occupied by a material with electric permittivity ε and magnetic permeability μ. Maxwell's equations state that

$$\varepsilon\frac{\partial \mathbf{E}}{\partial t} - \mathbf{curl}\,\mathbf{H} = \mathbf{J} \quad \text{in} \quad \Omega\times(0,T), \tag{1.1}$$

$$\mu\frac{\partial \mathbf{H}}{\partial t} + \mathbf{curl}\,\mathbf{E} = \mathbf{0} \quad \text{in} \quad \Omega\times(0,T), \tag{1.2}$$

$$\text{div}(\varepsilon\mathbf{E}) = \rho \quad \text{in} \quad \Omega\times(0,T), \tag{1.3}$$

$$\text{div}(\mu\mathbf{H}) = 0 \quad \text{in} \quad \Omega\times(0,T), \tag{1.4}$$

where $\mathbf{E} = \mathbf{E}(x,t)$ and $\mathbf{H} = \mathbf{H}(x,t)$ denote the electric and magnetic fields, $\mathbf{J} = \mathbf{J}(x,t)$ denotes the applied current density, and $\rho = \rho(x,t)$ denotes the charge density. This paper is concerned with the case where the domain Ω is composed of two distinct dielectric materials. Let Ω_1 be a polyhedral subdomain strictly lying inside Ω, occupied by a material with electric permittivity ε_1 and magnetic permeability μ_1, and let $\Omega_2 = \Omega\backslash\bar{\Omega}_1$ be occupied by another material with electric permittivity ε_2 and magnetic permeability μ_2. For ease of exposition, we shall consider only the case where the parameters ε_i and μ_i are constant functions in Ω_i, $i = 1,2$, but possibly with great differences in their values. We remark that our subsequent analyses can be

[*] Received by the editors November 16, 2001; accepted for publication (in revised form) June 4, 2002; published electronically February 12, 2003.
http://www.siam.org/journals/sinum/41-1/39845.html

[†] Department of Mathematics, University of California, Los Angeles, CA 90095 (tschung@math.ucla.edu).

[‡] Department of Mathematics, Penn State Univeristy, University Park, PA 16802 and Department of Mathematics, Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong (qdu@math.psu.edu). The research of this author was supported in part by the state major basic research project G199903280 and by NSF DMS-0196522.

[§] Department of Mathematics, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong (zou@math.cuhk.edu.hk). The work of this author was partially supported by Hong Kong RGC grants CUHK4292/00P and CUHK4048/02P.

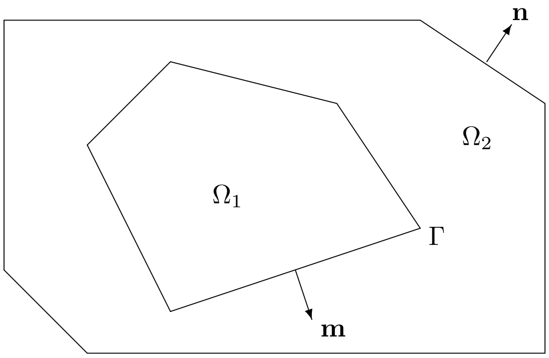


Fig. 1. *Two-dimensional cross-section of dielectric materials Ω_1, Ω_2 and their interface* Γ.

naturally extended to the case with piecewise smooth coefficients as well as multiple subdomains for which our methods have broad applications [3, 11].

Let $\Gamma = \partial\Omega_1$ be the boundary of Ω_1 with a unit outward normal vector $\mathbf{m}$, and let $\partial\Omega$ be the boundary of Ω with a unit outward normal vector $\mathbf{n}$; see Figure 1. We supplement the system (1.1)–(1.4) with the perfect conductor boundary condition and the initial condition given by

$$\mathbf{E} \times \mathbf{n} = \mathbf{0} \quad \text{on} \quad \partial\Omega \times (0,T)\ , \tag{1.5}$$

$$\mathbf{E}(x,0) = \mathbf{E}_0(x) \quad \text{and} \quad \mathbf{H}(x,0) = \mathbf{H}_0(x) \quad \forall x \in \Omega. \tag{1.6}$$

It is well known [3, 19] that the electric and magnetic fields $\mathbf{E}$ and $\mathbf{H}$ satisfy the following physical jump conditions across the interface Γ:

$$[\mathbf{E} \times \mathbf{m}] = \mathbf{0}, \quad [\varepsilon\mathbf{E} \cdot \mathbf{m}] = \rho_\Gamma, \tag{1.7}$$

$$[\mathbf{H} \times \mathbf{m}] = \mathbf{0}, \quad [\mu\mathbf{H} \cdot \mathbf{m}] = 0, \tag{1.8}$$

where $\rho_\Gamma = \rho_\Gamma(x)$ is the surface charge density and, throughout this paper, the jump of any function f across the interface Γ is defined by

$$[f] := f_2|_\Gamma - f_1|_\Gamma,$$

where $f_i = f|_{\Omega_i}$ for $i = 1, 2$.

In addition, we have the following constitutive relations:

$$\mathbf{D} = \varepsilon\mathbf{E}, \quad \mathbf{B} = \mu\mathbf{H}, \tag{1.9}$$

where $\mathbf{D}$ and $\mathbf{B}$ are the electric flux density and the magnetic flux density, respectively.

Over the past few decades, numerical methods for solving Maxwell's equations in homogeneous media have received much attention [11, 20]. The simple and popular Yee's scheme was proposed in 1966 [21], though its convergence analysis was not available until the work by Monk and Süli for nonuniform rectangular grids [14]. In order to handle domains with complicated geometry, both finite element and finite volume methods have been widely studied. For example, some fully discrete finite element methods were used to solve the decoupled time-dependent Maxwell's equations by Monk [13] and Raviart [18]. Second order convergence for the stationary case was established there, while a convergence analysis for the fully discrete time-dependent case was given by Ciarlet and Zou [7]. Chen and Yee proposed a finite volume method to solve Maxwell's equations in [4]. Convergence analyses for both semidiscrete and fully discrete schemes were given by Nicolaides and Wang [16].

For most real applications, however, one is often confronted with the solution of Maxwell's equations in nonhomogeneous media. Many of the aforementioned numerical methods either are not directly applicable or become inefficient (with lower order convergence) for these problems due to different physical characteristics reflected by the electric permittivities and magnetic permeabilities of different media, and due to the extra jump conditions the electric and magnetic fields need to satisfy on the interface; see (1.7)–(1.8). Several attempts have been made to handle the interface Maxwell's problems [4, 5, 20]. For example, Chen and Yee studied a hybrid FDTD/FVTD method for the interface problem [4], assuming that both the tangential components of the electric and magnetic fields are continuous across the interface and the electric field is tangentially piecewise constant on the interface. Chen, Du, and Zou [5] proposed an edge finite element method for solving Maxwell's system with general interface conditions and developed a general framework for its convergence analysis.

Recently, Chung and Zou presented a new finite volume method for Maxwell's equations in nonhomogeneous media [6], together with numerical experiments. In this paper, we will give the convergence analysis of the method for general tetrahedral triangulations. As in many interface problems, the regularity of the analytical solution of Maxwell's system in the entire physical domain is very low, which makes the convergence analysis very difficult. Regardless, we will show that, without making any extra regularity assumptions beyond those that are used for the case of a homogeneous medium [14, 16], the method under consideration is first order convergent for general tetrahedral triangulations and second order convergent for general nonuniform rectangular grids. Furthermore, it is shown that the proposed method has superconvergence in a discrete $H(\mathrm{curl};\Omega)$-norm, and the explicit dependence of the error estimates on the physical material parameters is given. To our knowledge, this seems to be the first rigorous work so far on the convergence of a finite volume method for Maxwell's equations with discontinuous coefficients.

We end this section with some notational conventions to be used in the subsequent analysis. For a nonnegative integer m and $1 \le p < \infty$, we use $W^{m,p}(\Omega)$ to denote the standard Sobolev space equipped with the norm [1]

$$\|u\|_{W^{m,p}(\Omega)} = \left(\sum_{0 \le |\alpha| \le m} \|D^\alpha u\|^p_{L^p(\Omega)} \right)^{1/p}$$

and the seminorm

$$|u|_{W^{m,p}(\Omega)} = \left(\sum_{|\alpha| = m} \|D^\alpha u\|^p_{L^p(\Omega)} \right)^{1/p} .$$

Here $D^\alpha u$ denotes the αth order weak derivative of u. In addition, we define [10]

$$H(\mathrm{curl};\Omega) = \{\mathbf{u} \in L^2(\Omega)^3; \quad \mathbf{curl}\ \mathbf{u} \in L^2(\Omega)^3\},$$

with its seminorm and norm given by

$$|\mathbf{u}|_{H(\mathrm{curl};\Omega)} = \|\mathbf{curl}\ \mathbf{u}\|_{L^2(\Omega)^3}; \quad \|\mathbf{u}\|_{H(\mathrm{curl};\Omega)} = \{\|\mathbf{u}\|^2_{L^2(\Omega)^3} + \|\mathbf{curl}\ \mathbf{u}\|^2_{L^2(\Omega)^3}\}^{\frac{1}{2}} ,$$

respectively. Furthermore, for some $0 < \lambda < 1$, $C^{m,\lambda}(\Omega)$ denotes the standard Hölder spaces of functions whose mth order derivatives are Hölder continuous with exponent λ. The same definitions are adopted on Ω_1 and Ω_2.

We use $L^p(0,T;\mathbf{X})$ to denote the space of all L^p integrable functions $\mathbf{u}(t,\cdot)$ from $[0,T]$ into the Banach space $\mathbf{X}$, and we also define [12]

$$W^{m,p}(0,T;\mathbf{X}) = \left\{\mathbf{u} \in L^p(0,T;\mathbf{X}); \quad \frac{\partial^\alpha \mathbf{u}}{\partial t^\alpha} \in L^p(0,T;\mathbf{X}) \quad \forall |\alpha| \le m\right\},$$

with norm

$$\|\mathbf{u}\|_{W^{m,p}(0,T;\mathbf{X})} = \left\{\sum_{0\le|\alpha|\le m} \left\|\frac{\partial^\alpha \mathbf{u}}{\partial t^\alpha}\right\|_{\mathbf{X}}^p\right\}^{1/p}.$$

When $p = 2$, we set $H^m(\Omega) = W^{m,2}(\Omega)$ and $H^m(0,T;\mathbf{X}) = W^{m,2}(0,T;\mathbf{X})$.

The rest of the paper is organized as follows. Some discrete vector fields and the finite volume method are introduced in sections 2 and 3, respectively. In section 4, we give a discussion of the discrete divergence constraints and stability. The convergence analysis for the general tetrahedral triangulation and the convergence analysis for the case of a nonuniform rectangular grid are given in section 5. Some concluding remarks are given in section 6.

2. Discrete vector fields. We now discuss the triangulation of the domain Ω. We use the Voronoi–Delaunay triangulation [9], which enjoys many elegant geometric properties that allow us to derive the numerical schemes in the subsequent sections. We adopt the notation developed by Nicolaides [15], Nicolaides and Wang [16], and Nicolaides and Wu [17], where a finite volume method was proposed for solving Maxwell's equations with smooth physical coefficients ε and μ.

We first triangulate Ω using the standard tetrahedral elements, which are called the *primal elements*. The triangulation is chosen so that the faces of the primal elements are aligned with the interface Γ. A primal element with at least one face lying on Γ is called an *interface primal element*, and a primal face (edge) lying on Γ is called an interface primal face (edge).

The *dual elements* are the Voronoi polyhedra formed by connecting the circumcenters of adjacent primal elements. Those dual elements (faces and edges) separated by the interface Γ into two parts lying in Ω_1 and Ω_2, respectively, are called the interface dual elements (faces and edges). The definitions and convergence analysis related to dual elements are much more complicated than those related to primal elements, due to the interface. From geometry, it is known that each primal edge is perpendicular to and is in one-to-one correspondence with a dual face, and each dual edge is perpendicular to and in one-to-one correspondence with a primal face.

For the subsequent convergence analysis, we assume that all dihedral angles of each tetrahedron are uniformly acute and the triangulation restricted to each subdomain satisfies

$$K_r \le \frac{h^r_{\max}}{h^r_{\min}} \le \tilde{K}_r, \qquad r = 1, 2\ , \tag{2.1}$$

where $h^r_{\max}$ and $h^r_{\min}$ are, respectively, the local maximum and minimum side lengths of adjacent primal and dual elements in Ω_r, and K_r and $\tilde{K}_r$ are two positive constants.

Let N and L be the numbers of primal and dual elements, respectively, and let F be the number of primal faces (dual edges) and M the number of primal edges (dual faces). Assume that these quantities are numbered sequentially in some order. The individual elements, faces, edges, and nodes of the primal mesh are denoted by τ_i, κ_j, σ_k, and ν_l, respectively. Those quantities related to the dual mesh are denoted by the primed forms such as τ'_i, κ'_j, σ'_k, and ν'_l. The area of κ_j is denoted by s_j, and the length of σ_k is given by h_k. A direction is assigned to each primal and dual edge by the rule that positive direction is from low to high node number. A direction is also assigned to each primal (dual) face so that it is the same as that of the corresponding dual (primal) edge. We denote by F_1 the number of interior primal faces (dual edges) and by M_1 the number of interior primal edges (dual faces). For each dual edge σ'_j of length h'_j, we define a scaled length:

$$\bar{h}'_j = \begin{cases} \frac{1}{\mu_1} h'_j & \text{if} \quad \sigma'_j \in \Omega_1, \\ \frac{1}{\mu_2} h'_j & \text{if} \quad \sigma'_j \in \Omega_2, \\ (\frac{1}{\mu_1} a_j + \frac{1}{\mu_2}(1-a_j))h'_j & \text{otherwise,} \end{cases}$$

where $0 < a_j < 1$ is the ratio of the length of the portion of σ'_j that belongs to Ω_1 over the length of σ'_j. For any u and v in $\mathbb{R}^{F_1}$, we introduce a mesh and parameter dependent inner product defined by

$$(u,v)_W := \sum_{\kappa_j \subset \Omega} u_j v_j s_j \bar{h}'_j = (Su, D'v) = (D'u, Sv), \tag{2.2}$$

where $S := \text{diag}(s_j)$ and $D' := \text{diag}(\bar{h}'_j)$ are $F_1 \times F_1$ diagonal matrices and $(\cdot,\cdot)$ denotes the standard Euclidean inner product. Similarly, for each dual face κ'_j with area s'_j, we define a scaled area:

$$\bar{s}'_j = \begin{cases} \varepsilon_1 s'_j & \text{if} \quad \kappa'_j \in \Omega_1, \\ \varepsilon_2 s'_j & \text{if} \quad \kappa'_j \in \Omega_2, \\ (\varepsilon_1 b_j + \varepsilon_2(1-b_j))s'_j & \text{otherwise,} \end{cases}$$

where $0 < b_j < 1$ is the ratio of the area of the portion of κ'_j that belongs to Ω_1 over the area of κ'_j. Also, we define a mesh and parameter dependent inner product in $\mathbb{R}^{M_1}$ by

$$(u,v)_{W'} := \sum_{\kappa'_j \subset \Omega} u_j v_j \bar{s}'_j h_j = (S'u, Dv) = (Du, S'v), \tag{2.3}$$

where $S' := \text{diag}(\bar{s}'_j)$ and $D := \text{diag}(h_j)$ are $M_1 \times M_1$ diagonal matrices.

For any $\sigma_j \in \partial\kappa_i$, we say that σ_j is oriented positively along $\partial\kappa_i$ if the direction of σ_j agrees with the one of $\partial\kappa_i$ formed by the right-hand rule with the thumb pointing in the direction of σ'_i. Otherwise, we say that σ_j is oriented negatively along $\partial\kappa_i$. For each interior primal face κ_i, we define its discrete circulation by

$$(Cu)_{\kappa_i} := \sum_{\sigma_j \subset \partial\kappa_i} u_j \tilde{h}_j, \tag{2.4}$$

where

$$\tilde{h}_j = \begin{cases} h_j & \text{if } \sigma_j \text{ is oriented positively along } \partial\kappa_i, \\ -h_j & \text{if } \sigma_j \text{ is oriented negatively along } \partial\kappa_i. \end{cases}$$

Similarly, for each interior dual face κ_i' we define its discrete circulation by

$$(C'u)_{\kappa_i'} := \sum_{\sigma_j' \subset \partial\kappa_i'} u_j \tilde{h}_j', \tag{2.5}$$

where

$$\tilde{h}_j' = \begin{cases} \bar{h}_j' & \text{if } \sigma_j' \text{ is oriented positively along } \partial\kappa_i', \\ -\bar{h}_j' & \text{if } \sigma_j' \text{ is oriented negatively along } \partial\kappa_i'. \end{cases}$$

Clearly, C and C' are two linear mappings from $\mathbb{R}^M$ to $\mathbb{R}^{F_1}$ and $\mathbb{R}^{F_1}$ to $\mathbb{R}^{M_1}$, respectively. We remark that (2.4) and (2.5) are the discrete analogues of the integrals

$$\int_{\kappa_i'} \mathbf{E} \cdot \mathbf{n}_i \, d\sigma \quad \text{and} \quad \int_{\kappa_i} \mathbf{H} \cdot \mathbf{n}_i \, d\sigma$$

by Stokes' theorem, where in what follows $\mathbf{n}_i$ represents the unit normal vector for both primal and dual faces.

For each strictly interior dual edge σ_j' with both endpoints of σ_j' lying in Ω and the ith strictly interior dual face κ_i', we define the entries of a $F_1 \times M_1$ matrix G as

$$(G)_{ji} := \begin{cases} 1 & \text{if } \sigma_j' \text{ is oriented positively along } \partial\kappa_i', \\ -1 & \text{if } \sigma_j' \text{ is oriented negatively along } \partial\kappa_i', \\ 0 & \text{if } \sigma_j' \text{ does not meet } \partial\kappa_i'. \end{cases}$$

Let $w \in \mathbb{R}^M$ be a vector whose kth component is the value assigned to the kth primal edge. Let $w_1 \in \mathbb{R}^{M_1}$ be the restriction of w to the interior primal edges. Denote by $w|_{\partial\Omega}$ the components of w that are related to the boundary. Likewise, denote by $v \in \mathbb{R}^{F_1}$ the vector whose jth component represents a value on the jth interior dual edge. Similarly to [15, 16, 17], we have the following result.

LEMMA 2.1. *Let w, w_1, and v be defined as above, and $w|_{\partial\Omega} = 0$; then we have*

$$Cw = GDw_1 \, , \qquad C'v = G^T D'v \, . \tag{2.6}$$

Proof. To see the first relation in (2.6), we note that the ith component of both sides corresponds to the primal face κ_i. By the definition (2.4) and $w|_{\partial\Omega} = 0$, we have

$$(Cw)_{\kappa_i} = \sum_{\sigma_j \subset \partial\kappa_i} w_j \tilde{h}_j = \sum_{j=1}^{M_1} c_j w_j h_j, \qquad (GDw_1)_{\kappa_i} = \sum_{j=1}^{M_1} g_j h_j w_j,$$

where

$$c_j = \begin{cases} 1 & \text{if } \sigma_j \text{ is oriented postively along } \partial\kappa_i, \\ -1 & \text{if } \sigma_j \text{ is oriented negatively along } \partial\kappa_i, \\ 0 & \text{if } \sigma_j \text{ does not meet } \partial\kappa_i \end{cases}$$

for any interior primal edge σ_j, and $g_j = (G)_{ij}$. By the orthogonality between primal and dual meshes, we conclude that c_j and g_j are the same; the first relation in (2.6) is thus proved. The second relation can be proved by a similar technique. □

Using Lemma 2.1, we can show a discrete analogue of the following Green's formula:

$$\int_\Omega \mathbf{curl}\,\mathbf{E}\cdot\mathbf{B}\,dx = \int_\Omega \mathbf{curl}\,\mathbf{B}\cdot\mathbf{E}\,dx,$$

which holds when $\mathbf{E}\times\mathbf{n}=0$ on $\partial\Omega$.

LEMMA 2.2. *With the same definitions as in Lemma* 2.1, *we have*

$$(Cw, D'v) = (C'v, Dw_1). \tag{2.7}$$

Proof. Equation (2.7) follows directly from Lemma 2.1 and (2.6):

$$(C'v, Dw_1) = (G^T D'v, Dw_1) = (D'v, GDw_1) = (D'v, Cw). \qquad \square$$

With the definition of the discrete circulation operator C, we define the following inner product:

$$(u,v)_V := \sum_{\kappa_i\subset\Omega} (Cu)_i (Cv)_i s_i^{-1}\bar{h}'_i = (S^{-1}Cu, D'Cv) = (D'Cu, S^{-1}Cu) \tag{2.8}$$

for any vectors $u, v \in \mathbb{R}^M$, and the induced norm

$$|u|_V := (u,u)_V^{\frac{1}{2}}. \tag{2.9}$$

This norm is equivalent to the discrete seminorm of $H(\text{curl};\Omega)$. We also define

$$\|u\|_V := (\|u\|_{W'}^2 + |u|_V^2)^{\frac{1}{2}}, \tag{2.10}$$

which is a discrete analogue of the norm in $H(\text{curl};\Omega)$.

Let τ_i be a primal element and $\kappa_j \in \partial\tau_i$ be a primal face. We say κ_j is oriented positively along $\partial\tau_i$ if the dual edge σ'_j on κ_j is directed towards the outside of τ_i. Otherwise, we say κ_j is oriented negatively along $\partial\tau_i$. For each primal element τ_i we define a discrete flux by

$$(\mathcal{D}u)_i := \sum_{\kappa_j\subset\partial\tau_i} u_j\tilde{s}_j \qquad \forall u\in\mathbb{R}^{F_1}\ , \tag{2.11}$$

where no components of u on the boundary faces are involved, and $\tilde{s}_j$ is given by

$$\tilde{s}_j = \begin{cases} s_j & \text{if } \kappa_j \text{ is oriented positively along } \partial\tau_i,\\ -s_j & \text{if } \kappa_j \text{ is oriented negatively along } \partial\tau_i.\end{cases}$$

The mapping $\mathcal{D}$ is the discrete version of the divergence operator by noting that

$$\int_{\tau_i} \operatorname{div}\mathbf{u}\,dx = \int_{\partial\tau_i}\mathbf{u}\cdot\mathbf{n}\,ds.$$

Similarly, for each dual element τ'_i, we define a discrete flux by

$$(\mathcal{D}'u)_i := \sum_{\kappa'_j\subset\partial\tau'_i} u_j\tilde{s}'_j \quad \forall u\in\mathbb{R}^{M_1}\ , \tag{2.12}$$

where

$$\tilde{s}'_j = \begin{cases} \bar{s}'_j & \text{if } \kappa'_j \text{ is oriented positively along } \partial\tau'_i, \\ -\bar{s}'_j & \text{if } \kappa'_j \text{ is oriented negatively along } \partial\tau'_i. \end{cases}$$

Next we present a discrete analogue of the identity div(**curl u**) = 0 for the discrete divergence operators $\mathcal{D}$ and $\mathcal{D}'$. To do so, we introduce two matrices B_1 and B'_1. B_1 is a $F_1 \times N$ matrix given by

$$(B_1)_{ji} := \begin{cases} 1 & \text{if } \kappa_j \text{ is oriented positively along } \partial\tau_i, \\ -1 & \text{if } \kappa_j \text{ is oriented negatively along } \partial\tau_i, \\ 0 & \text{if } \kappa_j \text{ does not meet } \partial\tau_i, \end{cases}$$

while B'_1 is a $M_1 \times L$ matrix given by

$$(B'_1)_{ji} := \begin{cases} 1 & \text{if } \kappa'_j \text{ is oriented positively along } \partial\tau'_i, \\ -1 & \text{if } \kappa'_j \text{ is oriented negatively along } \partial\tau'_i, \\ 0 & \text{if } \kappa'_j \text{ does not meet } \partial\tau'_i. \end{cases}$$

Then we have the following relations (cf. [6]).

Lemma 2.3. *We have*

$$\mathcal{D} = B_1^T S, \qquad \mathcal{D}' = (B'_1)^T S', \tag{2.13}$$

$$B_1^T C = \mathbf{0}, \qquad (B'_1)^T C' = \mathbf{0}. \tag{2.14}$$

3. The finite volume method. The finite volume method proposed in Chung and Zou [6] for solving the interface Maxwell's equations (1.1)–(1.8) approximates the edge average of **E** on each primal edge and the face average of **B** on each primal face. The use of the magnetic flux density **B** in the approximation, instead of the magnetic field **H** as in most existing numerical methods, is crucial for maintaining accuracy in interface problems. This observation is supported by the numerical experiments presented in [6].

We now introduce some average quantities. For the magnetic flux density **B**, we define its primal face average $B_f \in \mathbb{R}^{F_1}$ by

$$(B_f)_i := \frac{1}{s_i} \int_{\kappa_i} \mathbf{B} \cdot \mathbf{n}_i \, d\sigma$$

for each primal face κ_i and its dual edge average $B'_e \in \mathbb{R}^{F_1}$ by

$$(B'_e)_i := \frac{1}{h'_i} \int_{\sigma'_i} \mathbf{B} \cdot \mathbf{t}_i \, dl$$

for each noninterface dual edge σ'_i. Further, we let

$$\begin{aligned} (B'_e)_i &:= \alpha_i (B'_{e_1})_i + (1-\alpha_i)(B'_{e_2})_i \\ &:= \alpha_i \frac{1}{h_i^1} \int_{\sigma_i^1} \mathbf{B} \cdot \mathbf{t}_i \, dl + (1-\alpha_i) \frac{1}{h_i^2} \int_{\sigma_i^2} \mathbf{B} \cdot \mathbf{t}_i \, dl \end{aligned} \tag{3.1}$$

for each interface dual edge σ'_i. Here, for $r = 1, 2$, $\sigma_i^r = \sigma'_i \cap \Omega_r$ is the portion of σ'_i in Ω_r and $\alpha_i := \mu_r^{-1} h_i^r (\bar{h}'_i)^{-1}$ with h_i^r being the length of σ_i^r.

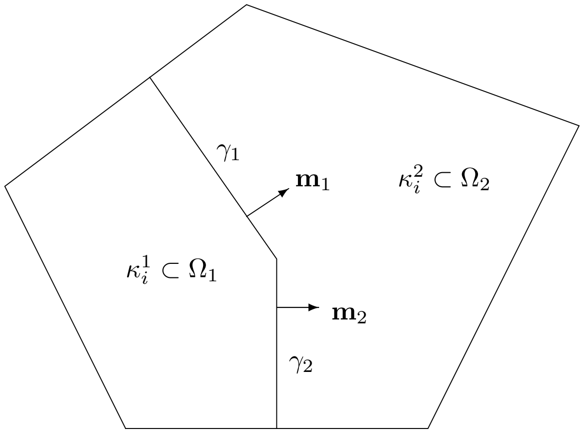


FIG. 2. *A dual face κ_i', divided by the interface into two parts κ_i^1, κ_i^2.*

For the electric field $\mathbf{E}$, we define its primal edge average $E_e \in \mathbb{R}^{M_1}$ by

$$(E_e)_i := \frac{1}{h_i}\int_{\sigma_i} \mathbf{E}\cdot\mathbf{n}_i \, dl$$

for each primal edge σ_i and its dual face average $E_f' \in \mathbb{R}^{M_1}$ by

$$(E_f')_i := \frac{1}{s_i'}\int_{\kappa_i'} \mathbf{E}\cdot\mathbf{n}_i \, d\sigma$$

for each non-interface dual face κ_i', and we let

$$\begin{aligned}(E_f')_i &:= \beta_i(E_{f_1}')_i + (1-\beta_i)(E_{f_2}')_i \\ &:= \beta_i\frac{1}{s_i^1}\int_{\kappa_i^1} \mathbf{E}\cdot\mathbf{n}_i \, d\sigma + (1-\beta_i)\frac{1}{s_i^2}\int_{\kappa_i^2} \mathbf{E}\cdot\mathbf{n}_i \, d\sigma\end{aligned} \tag{3.2}$$

for each interface dual face κ_i'; see Figure 2. Here, for $r = 1, 2$, $\kappa_i^r = \kappa_i' \cap \Omega_r$ is the portion of κ_i' in Ω_i with its area being s_i^r, and $\beta_i := \varepsilon_r s_i^r(\bar{s}_i')^{-1}$.

With the above notation, one can show that for each primal face κ_j and dual face κ_j' the true electric and magnetic fields $\mathbf{E}$ and $\mathbf{B}$ satisfy the equations [6]

$$s_j \frac{d}{dt}(B_f)_j + (CE_e)_{\kappa_j} = 0, \tag{3.3}$$

$$\bar{s}_j' \frac{d}{dt}(E_f')_j - (C'B_e')_{\kappa_j'} = \int_{\kappa_j'} \mathbf{J}\cdot\mathbf{n}_j \, d\sigma\,. \tag{3.4}$$

Let $E \in \mathbb{R}^{M_1}$ and $B \in \mathbb{R}^{F_1}$ be the approximations of the primal edge and face averages of the true solution $\mathbf{E}$ and $\mathbf{B}$ to (1.1)–(1.4), respectively. Note that each dual face (edge) average and the corresponding primal edge (face) average are approximately the same for sufficiently small h. Due to continuity of the tangential component of $\mathbf{E}$ and the normal component of $\mathbf{B}$ across the interface Γ, we naturally come to the following approximations based on (3.3) and (3.4):

Find $E \in \mathbb{R}^{M_1}$ and $B \in \mathbb{R}^{F_1}$ such that $E(0) = E_e(0)$, $B(0) = B_f(0)$, and

$$S' \frac{dE}{dt} - C'B = \tilde{J}, \tag{3.5}$$

$$S \frac{dB}{dt} + CE = \mathbf{0}, \tag{3.6}$$

where $\tilde{J} \in \mathbb{R}^{M_1}$ are defined by the right-hand sides of (3.4), while $E_e(0)$ and $B_f(0)$ are the primal edge average of $\mathbf{E}$ and primal face average of $\mathbf{B}$ at time $t = 0$.

Applying standard results concerning the well-posedness of systems of first order ordinary differential equations, we obtain the following theorem.

THEOREM 3.1. *The semi-discrete scheme* (3.5)–(3.6) *is well-posed.*

4. Discrete divergence constraints and stability. In this section, we show that the solutions E and B of the semidiscrete finite volume scheme (3.5)–(3.6) satisfy the divergence constraint conditions (1.3)–(1.4) at the discrete level.

THEOREM 4.1. *Let E and B be the solutions of* (3.5)–(3.6)*, and let B_f, E_e, and E'_f be the average vectors of* $\mathbf{B}$ *or* $\mathbf{E}$ *as defined in section* 3. *Then*

$$\mathcal{D}B(t) = \mathbf{0}, \qquad \mathcal{D}B_f(t) = \mathbf{0}, \tag{4.1}$$

$$\mathcal{D}'E(t) = \tilde{\rho}(t) + \mathcal{D}'(E_e - E'_f)(0), \qquad \mathcal{D}'E_e(t) = \tilde{\rho}(t) + \mathcal{D}'(E_e - E'_f)(t) \tag{4.2}$$

for any $0 \le t \le T$*, where*

$$\tilde{\rho}_j(t) := \int_{\tau'_j} \rho(x,t)\, dx + \int_{\tau'_j \cap \Gamma} \rho_\Gamma(x,t)\, d\sigma. \tag{4.3}$$

Furthermore, we have the following discrete charge conservation law:

$$(B'_1)^T \tilde{J} = \frac{d\tilde{\rho}(t)}{dt}. \tag{4.4}$$

Proof. Multiplying (3.3) and (3.6) by the matrix B_1^T, and using (2.14), we have

$$\mathcal{D}\frac{dB_f}{dt} = \mathbf{0}, \quad \mathcal{D}\frac{dB}{dt} = \mathbf{0}.$$

So $\mathcal{D}B(t) = \mathcal{D}B_f(t) = \mathcal{D}B_f(0)$. Now (4.1) follows directly from the divergence-free condition (1.4).

To show (4.2) and (4.4), we multiply (3.4) by the matrix $(B'_1)^T$ and then use (2.13) to get

$$\mathcal{D}'\frac{dE'_f}{dt} = (B'_1)^T \tilde{J}. \tag{4.5}$$

Integrating the divergence condition (1.3) on each dual element, we obtain

$$\mathcal{D}'E'_f(t) = \tilde{\rho}(t) \tag{4.6}$$

for any $0 \le t \le T$, which is the second relation in (4.2). Also, the discrete charge conservation law (4.4) follows readily from the two equations above.

Now we multiply (3.5) by the matrix $(B'_1)^T$ and use (4.4) to get

$$\mathcal{D}'\frac{dE}{dt} = (B'_1)^T \tilde{J} = \frac{d\tilde{\rho}(t)}{dt}\ .$$

Integrating in time, we have

$$\mathcal{D}'E(t) = \tilde{\rho}(t) + \mathcal{D}'E(0) - \tilde{\rho}(0),$$

which is the first equation in (4.2) by applying (4.6) at $t = 0$. $\square$

Next we state some stability results for the approximate solutions E and B.

Theorem 4.2. *The solution* (E, B) *to the semidiscrete scheme* (3.5)–(3.6) *satisfies the following stability inequality:*

$$\max_{0\le t\le T}\{\|B(t)\|_W^2 + \|E(t)\|_{W'}^2\} \le 2\|B(0)\|_W^2 + 2\|E(0)\|_{W'}^2 + 4T\int_0^T \|S^{'-1}\tilde{J}(t)\|_{W'}^2\, dt.$$

Proof. Multiplying (3.6) by $D'B$ and (3.5) by DE, and adding up the resulting equations and using (2.7), we obtain

$$\left(S\frac{dB}{dt}, D'B\right) + \left(S'\frac{dE}{dt}, DE\right) = (\tilde{J}, DE),$$

and consequently

$$\frac{1}{2}\frac{d}{dt}\|B(t)\|_W^2 + \frac{1}{2}\frac{d}{dt}\|B(t)\|_{W'}^2 = (\tilde{J}, DE).$$

Integrating with respect to time, we get for any $0 \le s < t$

$$\|B(s)\|_W^2 + \|E(s)\|_{W'}^2 = \|B(0)\|_W^2 + \|E(0)\|_{W'}^2 + 2\int_0^s (\tilde{J}, DE)\, dt.$$

Using the above equation, the desired bound follows from the estimate

$$\begin{aligned}2\int_0^s (\tilde{J}, DE)\, dt &\le 2\int_0^s \|S^{'-1}\tilde{J}(t)\|_{W'}\|E(t)\|_{W'}\, dt\\ &\le 2T\int_0^s \|S^{'-1}\tilde{J}(t)\|_{W'}^2\, dt + \frac{1}{2T}\int_0^s \|E(t)\|_{W'}^2\, dt. \qquad \square\end{aligned}$$

5. Error estimates for the finite volume method. We devote this section to the error analysis of the finite volume scheme (3.5)–(3.6). We will present the discrete L^2-norm error estimates for both a tetrahedral grid and a rectangular grid, where the same convergence orders can be achieved as for noninterface Maxwell's equations. Also, we will show a discrete $H(\text{curl};\Omega)$-norm error estimate, from which one can observe some superconvergence results for the finite volume method.

5.1. Discrete L^2-norm error estimate for tetrahedral grids. The purpose of this section is to develop the error analysis of the numerical scheme (3.5)–(3.6) in the discrete L^2-norms $\|\cdot\|_{W'}$ and $\|\cdot\|_W$. To do so, subtracting (3.3) from the jth component of (3.6), we obtain

$$S\frac{d}{dt}(B - B_f) + C(E - E_e) = 0\ ; \tag{5.1}$$

then subtracting (3.4) from the jth component of (3.5) gives

$$S'\frac{d}{dt}(E - E_f') - C'(B - B_e') = \mathbf{0}\ . \tag{5.2}$$

Now multiplying (5.1) by $D'(B - B'_e)$ and (5.2) by $D(E - E_e)$, and then adding the resulting equalities, we have

$$\begin{aligned}(S(\dot{B} - \dot{B}_f), D'(B - B_e)) + (S'(\dot{E} - \dot{E}'_f), D(E - E_e))\\ = (C'(B - B'_e), D(E - E_e)) - (C(E - E_e), D'(B - B'_e)),\end{aligned} \tag{5.3}$$

where the dot represents the time derivative. By the boundary condition $\mathbf{E} \times \mathbf{n} = \mathbf{0}$ on $\partial\Omega$, we know that all the components of $E - E_e$ on the boundary vanish. So by Lemma 2.2 we see that

$$(C'(B - B'_e), D(E - E_e)) - (C(E - E_e), D'(B - B'_e)) = 0,$$

and consequently we obtain from (5.3) that

$$(\dot{B} - \dot{B}_f, B - B'_e)_W + (\dot{E} - \dot{E}'_f, E - E_e)_{W'} = 0. \tag{5.4}$$

Now we rewrite (5.4) as

$$\begin{aligned}(\dot{B} - \dot{B}'_e, B - B'_e)_W + (\dot{E} - \dot{E}_e, E - E_e)_{W'}\\ = (\dot{E}'_f - \dot{E}_e, E - E_e)_{W'} + (\dot{B}_f - \dot{B}'_e, B - B'_e)_W\end{aligned}$$

or, equivalently, as

$$\frac{1}{2}\frac{d}{dt}(\|B - B'_e\|^2_W + \|E - E_e\|^2_{W'}) = (\dot{E}'_f - \dot{E}_e, E - E_e)_{W'} + (\dot{B}_f - \dot{B}'_e, B - B'_e)_W. \tag{5.5}$$

This enables us to show the following (optimal) first order convergence result for the finite volume scheme (3.5)–(3.6) for solving Maxwell's equations (1.1)–(1.4) on general tetrahedral grids.

THEOREM 5.1. *Assume that* $\mathbf{E}, \mathbf{B} \in W^{1,1}(0, T; W^{1,p}(\Omega_i)^3)$, *for* $i = 1, 2$ *and* $p > 2$, *are the solutions to Maxwell's system* (1.1)–(1.4), *while* E *and* B *are the finite volume solution of* (3.5)–(3.6). *Then the following error estimate holds for some constant* K, *independent of the mesh and the material parameters:*

$$\begin{aligned}&\max_{0\le t\le T}\{\|(E - E_e)(t)\|_{W'} + \|(B - B_f)(t)\|_W\}\\ &\le Kh\sum_{i=1}^{2}\{\|\varepsilon_i^{\frac{1}{2}}\mathbf{E}\|_{W^{1,1}(0,T;W^{1,p}(\Omega_i)^3)} + \|\mu_i^{-\frac{1}{2}}\mathbf{B}\|_{W^{1,1}(0,T;W^{1,p}(\Omega_i)^3)}\}.\end{aligned} \tag{5.6}$$

Proof. We prove this theorem by using (5.5). For each noninterface interior primal edge σ_i, by definition we have

$$(\dot{E}'_f - \dot{E}_e)_i = \frac{1}{s'_i}\int_{\kappa'_i}\dot{\mathbf{E}}\cdot\mathbf{n}_i\,d\sigma - \frac{1}{h_i}\int_{\sigma_i}\dot{\mathbf{E}}\cdot\mathbf{t}_i\,dl,$$

where $\mathbf{n}_i$ is the unit normal vector to the dual face κ'_i. Let τ'_{i_1} and τ'_{i_2} be the two dual elements sharing the same dual face κ'_i; then by the Sobolev embedding theorem we have, for $p > 2$,

$$W^{1,p}(\tau'_{i_1} \cup \tau'_{i_2}) \hookrightarrow L^1(\kappa'_i), \qquad W^{1,p}(\tau'_{i_1} \cup \tau'_{i_2}) \hookrightarrow L^1(\sigma_i)\,.$$

Hence, $(\dot{E}'_f - \dot{E}_e)_i$ is a bounded linear functional on $W^{1,p}(\tau'_{i_1} \cup \tau'_{i_2})^3$ and vanishes for all constant functions. By the Bramble–Hilbert lemma and a standard scaling argument, we obtain

$$|(\dot{E}'_f - \dot{E}_e)_i| \le K h^{1-\frac{3}{p}} |\dot{\mathbf{E}}|_{W^{1,p}(\tau'_{i_1} \cup \tau'_{i_2})^3} \tag{5.7}$$

for some generic constant K.

Next, for each interface primal edge σ_i corresponding to an interface dual face κ'_i, using (3.2) we get

$$\begin{aligned}(\dot{E}'_f - \dot{E}_e)_i &= (\beta_i \dot{E}'_{f_1} + (1-\beta_i)\dot{E}'_{f_2})_i - (\dot{E}_e)_i \\ &= \beta_i (\dot{E}'_{f_1} - \dot{E}_e)_i + (1-\beta_i)(\dot{E}'_{f_2} - \dot{E}_e)_i \,.\end{aligned}$$

Let $O_{i_1} = (\tau'_{i_2} \cup \tau'_{i_1}) \cap \Omega_1$ and $O_{i_2} = (\tau'_{i_2} \cup \tau'_{i_1}) \cap \Omega_2$; then the same reasoning as above shows that $(\dot{E}'_{f_1} - \dot{E}_e)_i$ and $(\dot{E}'_{f_2} - \dot{E}_e)_i$ are bounded linear functionals on $W^{1,p}(O_{i_1})^3$ and $W^{1,p}(O_{i_2})^3$, respectively, and vanish for all constant functions. Again, an application of the Bramble–Hilbert lemma and a scaling argument yield

$$|(\dot{E}'_{f_1} - \dot{E}_e)_i| \le K h^{1-\frac{3}{p}} |\dot{\mathbf{E}}|_{W^{1,p}(O_{i_1})^3}, \tag{5.8}$$

$$|(\dot{E}'_{f_2} - \dot{E}_e)_i| \le K h^{1-\frac{3}{p}} |\dot{\mathbf{E}}|_{W^{1,p}(O_{i_2})^3} \,. \tag{5.9}$$

By the definitions of $\bar{s}'_i$ and β_i, it is easy to see that $\bar{s}'_i \beta_i^2 \le \varepsilon_1 s_i^1$ and $\bar{s}'_i (1-\beta_i)^2 \le \varepsilon_2 s_i^2$. Thus we have

$$\begin{aligned}\bar{s}'_i h_i |(\dot{E}'_f - \dot{E}_e)_i|^2 &\le \bar{s}'_i h_i (2\beta_i^2 |(\dot{E}'_{f_1} - \dot{E}_e)_i|^2 + 2(1-\beta_i)^2 |(\dot{E}'_{f_2} - \dot{E}_e)_i|^2) \\ &\le 2\varepsilon_1 h_i s_i^1 |(\dot{E}'_{f_1} - \dot{E}_e)_i|^2 + 2\varepsilon_2 h_i s_i^2 |(\dot{E}'_{f_2} - \dot{E}_e)_i|^2 \,.\end{aligned}$$

This, along with the estimates (5.7)–(5.9) and the Cauchy–Schwarz inequality, leads to

$$\begin{aligned}\|\dot{E}'_f - \dot{E}_e\|^2_{W'} &= \sum_{\kappa'_i \subset \Omega_1 \cup \Omega_2} \bar{s}'_i h_i |(\dot{E}'_f - \dot{E}_e)_i|^2 + \sum_{\kappa'_i \cap \Gamma \neq \phi} \bar{s}'_i h_i |(\dot{E}'_f - \dot{E}_e)_i|^2, \\ &\le K h^{5-\frac{6}{p}} \sum_{i=1}^{M_1} \left\{ \varepsilon_1 |\dot{\mathbf{E}}|^2_{W^{1,p}(O_{i_1})^3} + \varepsilon_2 |\dot{\mathbf{E}}|^2_{W^{1,p}(O_{i_2})^3} \right\}, \\ &\le K h^{5-\frac{6}{p}} \left\{ \sum_{i=1}^{M_1} \varepsilon_1^{p/2} |\dot{\mathbf{E}}|^p_{W^{1,p}(O_{i_1})^3} + \varepsilon_2^{p/2} |\dot{\mathbf{E}}|^p_{W^{1,p}(O_{i_2})^3} \right\}^{\frac{2}{p}} \left\{ \sum_{i=1}^{M_1} 1 \right\}^{1-\frac{2}{p}} .\end{aligned}$$

Noting the fact that $h^3 \sum_{i=1}^{M_1} 1 \le K$, we conclude that

$$\|\dot{E}'_f - \dot{E}_e\|_{W'} \le K h \sum_{r=1}^{2} |\varepsilon_r^{\frac{1}{2}} \dot{\mathbf{E}}|_{W^{1,p}(\Omega_r)^3}. \tag{5.10}$$

Similarly, we have

$$\|\dot{B}_f - \dot{B}'_e\|_W \le K h \sum_{r=1}^{2} |\mu_r^{-\frac{1}{2}} \dot{\mathbf{B}}|_{W^{1,p}(\Omega_r)^3}. \tag{5.11}$$

By integrating (5.5) over $(0,t)$ and applying the Cauchy–Schwarz inequality, we obtain

$$\begin{aligned}\|(B-B_e')(t)\|_W^2 + \|(E-E_e)(t)\|_{W'}^2 &\le 2\int_0^t (\|(B-B_e')(s)\|_W \|(\dot{B}_f-\dot{B}_e')(s)\|_W \\ &\quad + \|(E-E_e)(s)\|_{W'}\|(\dot{E}_f'-\dot{E}_e)(s)\|_{W'})\, ds, \\ &\le 2\max_{0\le t\le T}(\|(B-B_e')(t)\|_W + \|(E-E_e)(t)\|_{W'}) \\ &\quad \times \int_0^T (\|(\dot{B}_f-\dot{B}_e')(s)\|_W + \|(\dot{E}_f'-\dot{E}_e)(s)\|_{W'})\, ds.\end{aligned}$$

Then, by (5.10) and (5.11), we have

$$\begin{aligned}&\max_{0\le t\le T}(\|(E-E_e)(t)\|_{W'} + \|(B-B_e')(t)\|_W) \\ &\le Kh\sum_{i=1}^{2}(|\varepsilon_i^{\frac{1}{2}}\mathbf{E}|_{W^{1,1}(0,T;W^{1,p}(\Omega_i))^3} + |\mu_i^{-\frac{1}{2}}\mathbf{B}|_{W^{1,1}(0,T;W^{1,p}(\Omega_i))^3}).\end{aligned}$$

In order to complete the proof, we first observe that

$$\|(B-B_f)(t)\|_W \le \|(B-B_e')(t)\|_W + \|(B_e'-B_f)(t)\|_W.$$

So it remains to estimate $\|(B_e'-B_f)(t)\|_W$. Following the same argument as the one that led to (5.11), we have

$$\|B_f - B_e'\|_W \le Kh\sum_{r=1}^{2}|\mu_r^{-\frac{1}{2}}\mathbf{B}|_{W^{1,p}(\Omega_r)^3}.$$

Hence,

$$\begin{aligned}\max_{0\le t\le T}\|(B_f-B_e')(t)\|_W &\le Kh\sum_{r=1}^{2}\max_{0\le t\le T}|\mu_r^{-\frac{1}{2}}\mathbf{B}(t)|_{W^{1,p}(\Omega_r)^3} \\ &\le Kh\sum_{r=1}^{2}\|\mu_r^{-\frac{1}{2}}\mathbf{B}\|_{W^{1,1}(0,T;W^{1,p}(\Omega_r))^3}. \qquad \square\end{aligned}$$

Remark. There are very few studies in the literature concerning the regularity of the solution to the time-dependent Maxwell system (1.1)–(1.4) with discontinuous coefficients. However, for domains with smooth boundaries and interfaces, the regularity $\mathbf{B},\mathbf{E} \in L^2(0,T;W^{1,p}(\Omega_i))$ $(i=1,2)$ can be shown by slightly modifying the proof of Theorem 6.2 [8] in combination with the equivalence between the space $\{\mathbf{w} \in W^{1,p}(\Omega);\ \mathbf{w}\cdot\mathbf{n} = 0 \text{ on } \partial\Omega\}$ and the space

$$\{\mathbf{w} \in L^p(\Omega)^3;\ \mathbf{curl}\,\mathbf{w} \in L^p(\Omega)^3, \mathbf{div}\,\mathbf{w} \in L^p(\Omega)^3,\ \mathbf{w}\cdot\mathbf{n} = 0 \text{ on } \partial\Omega\}.$$

The additional time differentiability $\mathbf{B},\mathbf{E} \in W^{1,1}(0,T;W^{1,p}(\Omega_i))$ can be proved using standard arguments; see, e.g., [2].

5.2. Discrete L^2-norm error estimate for rectangular grids. The first order convergence of the finite volume scheme (3.5)–(3.6) given in the last subsection is generally optimal in terms of the regularities used. In this section, we intend to improve the convergence rate of the scheme (3.5)–(3.6) on rectangular grids by one order; namely, we establish second order convergence by making full use of the local regularities of the fields **E** and **B**. Such a second order convergence result is invalid for general tetrahedral triangulations, even in the case of the noninterface Maxwell's equations [14, 16].

Let Ω be a rectangular cuboid in R^3. Similarly to the case of a polyhedral domain in section 2, we generate the primal and dual triangulations of Ω by using smaller rectangular cuboids. Note that both the primal and dual meshes are now made up of rectangular cuboids. For simplicity, the directions of edges and faces are assigned as follows: a direction is assigned to each primal and dual edge by the rule that positive direction means that it points in the positive axis direction. The directions of primal and dual faces are the same as those of the corresponding dual and primal edges. Below, we adopt the same notations as in section 2.

Clearly, most of the arguments presented in the previous subsection remain valid for the case of rectangular domain Ω considered here. To begin, we rewrite (5.4) as

$$\begin{aligned}&(\dot{B}-\dot{B}_f, B-B_f)_W + (\dot{E}-\dot{E}_e, E-E_e)_{W'}\\ &= (\dot{B}-\dot{B}_f, B'_e-B_f)_W + (\dot{E}'_f-\dot{E}_e, E-E_e)_{W'}\end{aligned}$$

or, equivalently, as

$$\frac{1}{2}\frac{d}{dt}(\|B-B_f\|^2_W + \|E-E_e\|^2_{W'}) \tag{5.12}$$

$$= (\dot{B}-\dot{B}_f, B'_e-B_f)_W + (\dot{E}'_f-\dot{E}_e, E-E_e)_{W'}. \tag{5.13}$$

Next we estimate the terms on the right-hand side of (5.13), and this needs the following two auxiliary lemmas.

LEMMA 5.2. *There exist functions $u(t)$ and $\xi(t) \in \mathbb{R}^{F_1}$ such that all the noninterface components of $\xi(t)$ vanish, all the components of u and ξ are bounded linear functionals of* **B**, *and the following relation holds for all $\phi \in \mathbb{R}^M$ with $\phi|_{\partial\Omega} = 0$:*

$$(C\phi, D'(B_f - B'_e)) = (C\phi, D'u) + (C\phi, \xi)\,. \tag{5.14}$$

Furthermore, the following estimates hold for $u(t)$ and $\xi(t)$:

(5.15)

$$\|u\|_W \le Kh^2 \sum_{i=1}^{2} \|\mu_i^{-\frac{1}{2}}\mathbf{B}\|_{H^3(\Omega_i)^3}\,, \qquad \|D^{'-1}\xi\|_W \le Kh^2 \sum_{i=1}^{2} \|\mu_i^{-\frac{1}{2}}\mathbf{B}\|_{H^3(\Omega_i)^3}.$$

Proof. By definition, for any strictly interior primal face κ_j we have

$$(B_f - B'_e)_j = \frac{1}{s_j}\int_{\kappa_j} \mathbf{B}\cdot\mathbf{n}_j\, d\sigma - \frac{1}{h'_j}\int_{\sigma'_j} \mathbf{B}\cdot\mathbf{t}_j\, dl.$$

Assume that κ_j is parallel to the xy-plane, with P_1 as its center; see Figure 3. We know that the quadrature rule

$$\int_{\kappa_j} \mathbf{B}\cdot\mathbf{n}_j\, d\sigma = s_j\,(\mathbf{B}\cdot\mathbf{n}_j)(P_1)$$

is exact for linear functions.

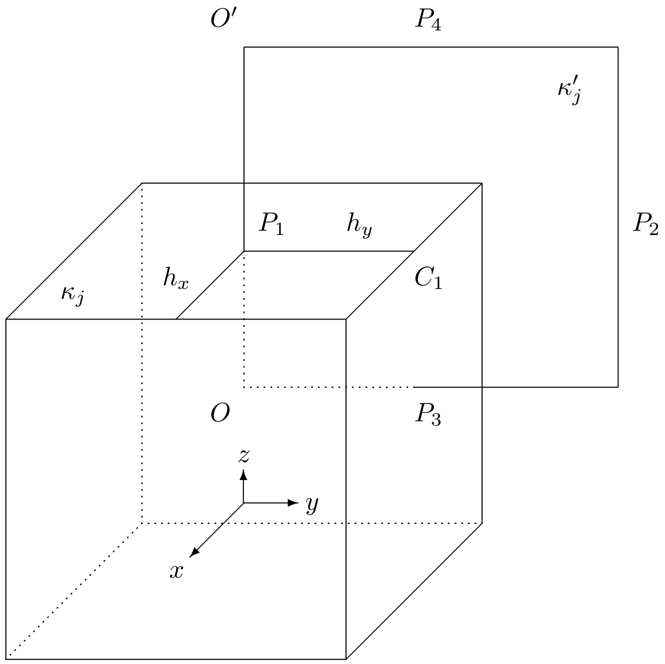


FIG. 3. *Noninterface element.*

Note that P_1 is not the center of the dual edge σ'_j. By adding a first order correction term, it is easy to see that the quadrature rule

$$\int_{\sigma'_j} \mathbf{B}\cdot\mathbf{t}_j\, dl = (\mathbf{B}\cdot\mathbf{t}_j)(P_1)h'_j + \frac{1}{2}(\overline{O'P_1}^2\mathbf{B}_{3z}(O') - \overline{OP_1}^2\mathbf{B}_{3z}(O))$$

is then exact for linear functions. Here $\overline{O'P_1}$ denotes the distance from O' to P_1 and $\mathbf{B}_{3z}$ denotes the derivative of the third component of $\mathbf{B}$ with respect to z. Similar notation will be used below. By the two relations above, we can rewrite $(B_f - B'_e)_j$ as

$$(B_f - B'_e)_j = \frac{1}{h'_j}\tilde{u}_j + u_j, \tag{5.16}$$

where u_j vanishes for linear functions and the first order correction $\tilde{u}_j$ is given by

$$\begin{aligned}\tilde{u}_j := {} & \frac{1}{2\mu_r}(\overline{OP_1}^2\mathbf{B}_{3z} + h_x^2\mathbf{B}_{1x} + h_y^2\mathbf{B}_{2y})(O)\\ & - \frac{1}{2\mu_r}(\overline{O'P_1}^2\mathbf{B}_{3z} + h_x^2\mathbf{B}_{1x} + h_y^2\mathbf{B}_{2y})(O').\end{aligned} \tag{5.17}$$

Here $r = 1$ or 2 is the index corresponding to the subdomain Ω_r in which κ_j lies. Moreover, notice the fact that $\mathbf{B}_{1x}(O) - \mathbf{B}_{1x}(O')$ and $\mathbf{B}_{2y}(O) - \mathbf{B}_{2y}(O')$ vanish for all linear functions, and the terms related to $\mathbf{B}_{1x}$ and $\mathbf{B}_{2y}$ are added to the above equation to make the relation more symmetric.

Next, by (3.1), for an interface primal face κ_i lying on Γ, we have

$$(B_f - B'_e)_i = \alpha_i(B_f - B'_{e_1})_i + (1-\alpha_i)(B_f - B'_{e_2})_i\ .$$

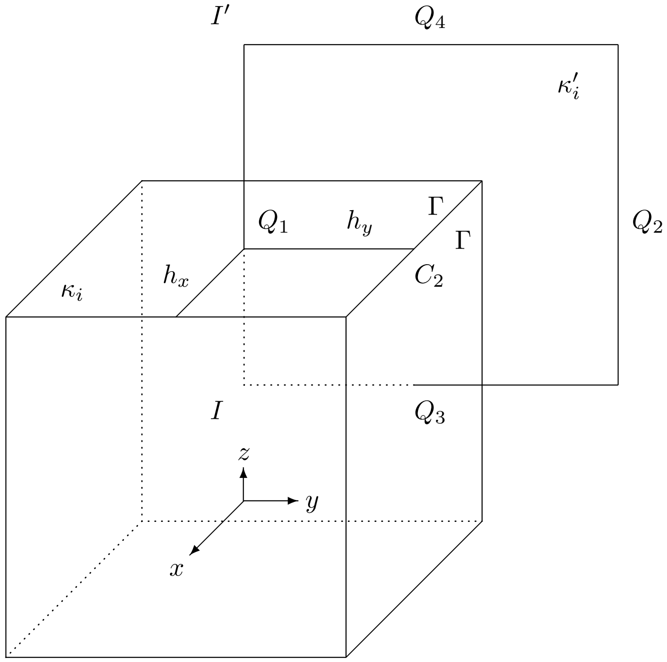


FIG. 4. *Interface element.*

Without loss of generality, we assume that κ_i is parallel to the xy-plane; see Figure 4. It is easy to verify that the quadrature rules

$$\int_{\kappa_i} \mathbf{B}\cdot\mathbf{n}_i \, d\sigma = s_i\,(\mathbf{B}\cdot\mathbf{n}_i)(Q_1)\ ,$$

$$(B'_{e_1})_i = \int_{\sigma_i^1} \mathbf{B}\cdot\mathbf{t}_i \, dl = (\mathbf{B}\cdot\mathbf{t}_i)(Q_1)h_i^1 - \frac{1}{2}\overline{IQ_1}^2\mathbf{B}_{3z}(I),$$

$$(B'_{e_2})_i = \int_{\sigma_i^2} \mathbf{B}\cdot\mathbf{t}_i \, dl = (\mathbf{B}\cdot\mathbf{t}_i)(Q_1)h_i^2 + \frac{1}{2}\overline{I'Q_1}^2\mathbf{B}_{3z}(I')$$

are all exact for linear functions. Using these relations, we can rewrite $(B_f - B'_e)_i$ as

$$(B_f - B'_e)_i = \frac{1}{\bar{h}'_i}\tilde{u}_i + \frac{1}{\bar{h}'_i}\xi_i + u_i, \tag{5.18}$$

where $u_i = \alpha_i u_i^1 + (1-\alpha_i)u_i^2$, u_i^1 and u_i^2 both vanish for linear functions, and the correction terms $\tilde{u}_i$ and ξ_i are given by

$$\begin{aligned}\tilde{u}_i := &\frac{1}{2\mu_1}(\overline{IQ_1}^2\mathbf{B}_{3z} + h_x^2\mathbf{B}_{1x} + h_y^2\mathbf{B}_{2y})(I)\\ &- \frac{1}{2\mu_2}(\overline{I'Q_1}^2\mathbf{B}_{3z} + h_x^2\mathbf{B}_{1x} + h_y^2\mathbf{B}_{2y})(I'),\end{aligned} \tag{5.19}$$

$$\xi_i := \frac{1}{2\mu_2}(h_x^2\mathbf{B}_{1x} + h_y^2\mathbf{B}_{2y})(I') - \frac{1}{2\mu_1}(h_x^2\mathbf{B}_{1x} + h_y^2\mathbf{B}_{2y})(I)\ . \tag{5.20}$$

For the same reason as earlier for the noninterface face κ_i, we have also added some extra terms related to $\mathbf{B}_{1x}$ and $\mathbf{B}_{2y}$ here. Note, however, that due to the jumps across the interface, ξ_i no longer vanishes for linear functions.

By (5.17), (5.19), and the definition of B_1, we can write $\tilde{u} = B_1\tilde{\phi}$ for some $\tilde{\phi} \in \mathbb{R}^N$. Hence for any $\phi \in \mathbb{R}^M$ with $\phi|_{\partial\Omega} = 0$, we get from (5.16) and (5.18) that

$$\begin{aligned}(C\phi, D'(B_f - B_e')) &= (C\phi, \tilde{u}) + (C\phi, D'u) + (C\phi, \xi)\\ &= (C\phi, B_1\tilde{\phi}) + (C\phi, D'u) + (C\phi, \xi)\\ &= (B_1^T C\phi, \tilde{\phi}) + (C\phi, D'u) + (C\phi, \xi)\\ &= (C\phi, D'u) + (C\phi, \xi).\end{aligned}$$

This proves (5.14).

For the estimate (5.15), let u_j be a component of u corresponding to an interior primal face κ_j in Ω_r, $r = 1, 2$. We recall from (5.16) that

$$u_j = (B_f - B_e')_j - \frac{1}{\bar{h}_j'}\tilde{u}_j.$$

By the Sobolev embedding theorem, we have

$$H^3(\tau_{j_1} \cup \tau_{j_2}) \hookrightarrow C^{1,\frac{1}{2}}(\tau_{j_1} \cup \tau_{j_2}),$$

where τ_{j_1} and τ_{j_2} are two elements in Ω_r and share the face κ_j. Hence, u_j is a bounded linear functional of $\mathbf{B}$ in $H^3(\tau_{j_1} \cup \tau_{j_2})^3$ and vanishes for linear fields $\mathbf{B}$. Then, by the Bramble–Hilbert lemma, we have

$$|u_j|^2 \le K(h)\left(|\mathbf{B}|^2_{H^2(\tau_{j_1}\cup\tau_{j_2})^3} + |\mathbf{B}|^2_{H^3(\tau_{j_1}\cup\tau_{j_2})^3}\right).$$

A standard scaling argument yields

$$|u_j|^2 \le Kh\left(|\mathbf{B}|^2_{H^2(\tau_{j_1}\cup\tau_{j_2})^3} + |\mathbf{B}|^2_{H^3(\tau_{j_1}\cup\tau_{j_2})^3}\right) \le Kh\|\mathbf{B}\|^2_{H^3(\tau_{j_1}\cup\tau_{j_2})^3}. \tag{5.21}$$

Now consider a component u_i of u corresponding to an interface face κ_i, which is shared by the element τ_{i_1} in Ω_1 and τ_{i_2} in Ω_2. Recall that

$$u_i = \alpha_i u_i^1 + (1 - \alpha_i)u_i^2,$$

where

$$\begin{aligned}h_i^1 u_i^1 &:= h_i^1(B'_{e_i^1} - (B_f)_i) - \frac{1}{2}\overline{IQ_1}^2\mathbf{B}_{3z}(I)\ ,\\ h_i^2 u_i^2 &:= h_i^2(B'_{e_i^2} - (B_f)_i) + \frac{1}{2}\overline{I'Q_1}^2\mathbf{B}_{3z}(I').\end{aligned}$$

By the Sobolev embedding theorem, u_i^r is a bounded linear functional of $\mathbf{B}$ in $H^3(\tau_{i_r})^3$ and vanishes for all linear fields for $r = 1$ or 2. Hence, again by the Bramble–Hilbert lemma and a scaling argument, we have

$$|u_i^1| \le Kh^{\frac{1}{2}}\|\mathbf{B}\|_{H^3(\tau_{i_1})^3}, \qquad |u_i^2| \le Kh^{\frac{1}{2}}\|\mathbf{B}\|_{H^3(\tau_{i_2})^3}.$$

Similarly to the proof of (5.10), using the above estimates and (5.21) we obtain

$$\begin{aligned}\|u\|_W^2 &= \sum_{\sigma_i'\subset\Omega_1\cap\Omega_2} s_j\bar{h}_j'|u_j|^2 + \sum_{\sigma_i'\cap\Gamma\neq\phi} s_j\bar{h}_j'|u_j|^2\\ &\le \sum_{\sigma_i'\subset\Omega_1\cap\Omega_2} s_j\bar{h}_j'|u_j|^2 + \sum_{\sigma_j'\cap\Gamma\neq\phi} s_j\bar{h}_j'(2\alpha_j^2|u_j^1|^2 + 2(1-\alpha_j)^2|u_j^2|^2)\\ &\le Kh^4\left\{\sum_{\tau_{i_1}\subset\Omega_1}\mu_1^{-1}\|\mathbf{B}\|^2_{H^3(\tau_{i_1})^3} + \sum_{\tau_{i_2}\subset\Omega_2}\mu_2^{-1}\|\mathbf{B}\|^2_{H^3(\tau_{i_2})^3}\right\}\\ &\le Kh^4\left\{\sum_{r=1}^2\|\mu_r^{-\frac{1}{2}}\mathbf{B}\|^2_{H^3(\Omega_r)^3}\right\}^2 .\end{aligned}$$

We are now ready to estimate ξ. For each interface primal face κ_i shared by the element τ_{i_1} in Ω_1 and τ_{i_2} in Ω_2, we rewrite ξ_i using the interface condition (1.8) as

$$\begin{aligned}\xi_i := &\left\{\frac{1}{2}(h_x^2\mathbf{H}_{1x} + h_y^2\mathbf{H}_{2y})(I') - \frac{1}{2}(h_x^2\mathbf{H}_{1x} + h_y^2\mathbf{H}_{2y})(Q_1)\right\}\\ &+\left\{\frac{1}{2}(h_x^2\mathbf{H}_{1x} + h_y^2\mathbf{H}_{2y})(Q_1) - \frac{1}{2}(h_x^2\mathbf{H}_{1x} + h_y^2\mathbf{H}_{2y})(I)\right\}.\end{aligned}\tag{5.22}$$

By the Hölder continuity of $\mathbf{H}_{1x}$, we have

$$|\mathbf{H}_{1x}(I') - \mathbf{H}_{1x}(Q_1)| \le Kh^{\frac{1}{2}}\|\mathbf{H}\|_{C^{1,\frac{1}{2}}(\tau_{i_2})^3}.$$

Similar estimates hold for the other pairs in (5.22). This leads to

$$|\xi_i| \le Kh^{\frac{5}{2}}\left\{\|\mathbf{H}\|_{C^{1,\frac{1}{2}}(\tau_{i_1})^3} + \|\mathbf{H}\|_{C^{1,\frac{1}{2}}(\tau_{i_2})^3}\right\}.$$

Consequently, by the fact that $\xi_i = 0$ for any noninterface primal face, we get

$$\begin{aligned}\|D^{'-1}\xi\|_W^2 &= \sum_{i=1}^{F_1} s_i\bar{h}_i'|(\bar{h}_j')^{-1}\xi_i|^2\\ &\le Kh^6\sum_{\kappa_i\subset\Gamma}\left\{\mu_1\|\mathbf{H}\|^2_{C^{1,\frac{1}{2}}(\tau_{i_1})^3} + \mu_2\|\mathbf{H}\|^2_{C^{1,\frac{1}{2}}(\tau_{i_2})^3}\right\}\\ &\le Kh^4\sum_{r=1}^2\|\mu_r^{\frac{1}{2}}\mathbf{H}\|^2_{C^{1,\frac{1}{2}}(\Omega_r)^3}. \qquad \square\end{aligned}$$

LEMMA 5.3. *There exist functions $v(t)$, $\lambda(t) \in \mathbb{R}^{M_1}$, and $w(t) \in \mathbb{R}^{F_1}$, such that all the noninterface components of $\lambda(t)$ vanish and all the components of v, w, and λ are bounded linear functionals of $\mathbf{E}$, and the following relation holds for all $\phi \in \mathbb{R}^M$ with $\phi|_{\partial\Omega} = 0$:*

$$(\dot{E}_f' - \dot{E}_e, \phi)_{W'} = (\dot{v}, \phi)_{W'} + (D'\dot{w}, C\phi) + (S^{'-1}\dot{\lambda}, \phi)_{W'} .\tag{5.23}$$

Furthermore, we have the following estimates for $v(t)$, $\lambda(t)$, $w(t)$, and $p > 3$:

$$\|\dot{v}\|_{W'} \le Kh^2\sum_{i=1}^2\|\epsilon_i^{\frac{1}{2}}\dot{\mathbf{E}}\|_{H^3(\Omega_i)^3}\ , \qquad \|\dot{w}\|_W \le Kh^2\sum_{i=1}^2\|\epsilon_i^{\frac{1}{2}}\dot{\mathbf{E}}\|_{W^{2,p}(\Omega_i)^3},\tag{5.24}$$

$$\|S^{'-1}\dot{\lambda}\|_{W'} \ \le\ Kh^2\sum_{i=1}^2\|\epsilon_i^{\frac{1}{2}}\dot{\mathbf{E}}\|_{H^3(\Omega_i)^3}\ .\tag{5.25}$$

Proof. The proof is similar to that of Lemma 5.2. First, we consider a noninterface dual face κ'_j lying in Ω_r $(r = 1, 2)$. Recall that

$$(E'_f - E_e)_j = \frac{1}{s'_j} \int_{\kappa'_j} \mathbf{E} \cdot \mathbf{n}_j \, d\sigma - \frac{1}{h_j} \int_{\sigma_j} \mathbf{E} \cdot \mathbf{t}_j \, dl.$$

We see from Figure 3 that C_1 is the center of the primal edge σ_j, so the quadrature rule

$$\int_{\sigma_j} \mathbf{E} \cdot \mathbf{t}_j \, dl = \mathbf{E}_1(C_1) h_j$$

is exact for all linear functions. However, C_1 is not the center of the dual face κ'_j. By adding a first order correction term $\tilde{w}_j$, the quadrature rule

$$\int_{\kappa'_j} \mathbf{E} \cdot \mathbf{n}_j \, d\sigma = \mathbf{E}_1(C_1) s'_j + \tilde{w}_j$$

is then exact for all linear functions, where $\tilde{w}_j$ is given by

$$2\tilde{w}_j = [\mathbf{E}_{1y}(P_2)\overline{P_2C_1}^2 - \mathbf{E}_{1y}(P_1)\overline{P_1C_1}^2]\overline{P_3P_4} + [\mathbf{E}_{1z}(P_4)\overline{P_4C_1}^2 - \mathbf{E}_{1z}(P_3)\overline{P_3C_1}^2]\overline{P_1P_2}.$$

By direct computations, $\tilde{w}_j$ can be represented by the discrete circulation as follows:

$$\tilde{w}_j := \frac{1}{\varepsilon_r}(C'w)_j \ , \tag{5.26}$$

where the components of w corresponding to the four edges of κ'_j containing the points P_1, P_2, P_3, and P_4 are assigned, respectively, the following values:

$$\begin{aligned}
w(P_1) &:= \frac{1}{2}\varepsilon_r\mu_r(h_y^2\mathbf{E}_{1y}(P_1) - h_x^2\mathbf{E}_{2x}(P_1)),\\
w(P_2) &:= \frac{1}{2}\varepsilon_r\mu_r((\overline{P_2C_1}^2\mathbf{E}_{1y}(P_2) - h_x^2\mathbf{E}_{2x}(P_2)),\\
w(P_3) &:= \frac{1}{2}\varepsilon_r\mu_r(h_x^2\mathbf{E}_{3x}(P_3) - \overline{P_3C_1}^2\mathbf{E}_{1z}(P_3)),\\
w(P_4) &:= \frac{1}{2}\varepsilon_r\mu_r(h_x^2\mathbf{E}_{3x}(P_4) - \overline{P_4P_1}^2\mathbf{E}_{1z}(P_4)).
\end{aligned}$$

We remark that for the verification of (5.26) we have used the simple fact that $\mathbf{E}_{2x}(P_1)$ and $\mathbf{E}_{2x}(P_2)$, as well as $\mathbf{E}_{3x}(P_1)$ and $\mathbf{E}_{3x}(P_2)$, are equal, respectively, for all linear functions. Using (5.26), we can rewrite $\dot{E}'_f - \dot{E}_e$ as

$$(\dot{E}'_f - \dot{E}_e)_j = \frac{1}{\bar{s}'_j}(C'\dot{w})_j + \dot{v}_j, \tag{5.27}$$

where $\dot{v}_j$ is a functional which vanishes for all linear functions.

Now consider an interface dual face κ'_i. By the definition (3.2) we have

$$\bar{s}'_i(\dot{E}'_f - \dot{E}_e)_i = \frac{d}{dt}\{\varepsilon_1 s_i^1(E'_{f_1} - E_e)_i + \varepsilon_2 s_i^2(E'_{f_2} - E_e)_i\}.$$

Without loss of generality, we assume that κ_i' is parallel to the zy-plane and perpendicular to the interface primal face κ_i; see Figure 4. It is straightforward to verify that the three quadrature rules

$$\begin{aligned}
(E_e)_i &= \frac{1}{h_i}\int_{\sigma_i} \mathbf{E}\cdot\mathbf{t}_i\,dl = \mathbf{E}_1(C_2),\\
(E'_{f_1})_i &= \frac{1}{s_i^1}\int_{\kappa_i^1} \mathbf{E}\cdot\mathbf{n}_i\,d\sigma = \frac{1}{s_i^1}\{\mathbf{E}_1(C_2)s_i^1 + \tilde{w}_i^1\},\\
(E'_{f_2})_i &= \frac{1}{s_i^2}\int_{\kappa_i^2} \mathbf{E}\cdot\mathbf{n}_i\,d\sigma = \frac{1}{s_i^2}\{\mathbf{E}_1(C_2)s_i^2 + \tilde{w}_i^2\}
\end{aligned}$$

are all exact for linear functions, where

$$\tilde{w}_i^1 := \frac{1}{2}[-\mathbf{E}_{1y}(Q_1)\overline{Q_1C_2}^2]\overline{Q_3C_2} + \frac{1}{2}[-\mathbf{E}_{1z}(Q_3)\overline{Q_3C_2}^2]\overline{Q_1C_2}$$

and

$$\begin{aligned}
\tilde{w}_i^2 :=& \frac{1}{2}[\mathbf{E}_{1y}(Q_2)\overline{Q_2C_2}^2]\overline{Q_3Q_4} + \frac{1}{2}[-\mathbf{E}_{1y}(Q_1)\overline{Q_1C_2}^2]\overline{Q_4C_2}\\
&+ \frac{1}{2}[\mathbf{E}_{1z}(Q_4)\overline{Q_4C_2}^2]\overline{Q_1Q_2} + \frac{1}{2}[-\mathbf{E}_{1z}(Q_3)\overline{Q_3C_2}^2]\overline{Q_2C_2}.
\end{aligned}$$

Then we have

$$\bar{s}_i'(\dot{E}_f' - \dot{E}_e)_i = \frac{d}{dt}(\varepsilon_1\tilde{w}_i^1 + \varepsilon_2\tilde{w}_i^2) + \frac{d}{dt}(\varepsilon_1 s_i^1 v_i^1 + \varepsilon_2 s_i^2 v_i^2),$$

where v_i^1 and v_i^2 are linear functionals which vanish for all linear functions. We further write

$$\frac{d}{dt}(\varepsilon_1\tilde{w}_i^1 + \varepsilon_2\tilde{w}_i^2) = (C'\dot{w})_i + \dot{\lambda}_i, \tag{5.28}$$

where the components of w on the four edges of κ_i' containing the points Q_1, Q_2, Q_3, and Q_4 are assigned, respectively, the following values:

$$\begin{aligned}
w(Q_1) :=& \frac{1}{2\bar{h}'_{i_1}}(\varepsilon_1\overline{Q_3C_2}h_y^2\mathbf{E}_{1y}(Q_1) + \varepsilon_2\overline{Q_4C_2}h_y^2\mathbf{E}_{1y}(Q_1))\\
&+ \frac{1}{2\bar{h}'_{i_1}}(-\varepsilon_1\overline{Q_3C_2}h_x^2\mathbf{E}_{2x}(Q_1) - \varepsilon_2\overline{C_2Q_4}h_x^2\mathbf{E}_{2x}(Q_1)),\\
w(Q_2) :=& \frac{1}{2\bar{h}'_{i_2}}(\varepsilon_2\overline{Q_3Q_4}\,\overline{C_2Q_2}^2\mathbf{E}_{1y}(Q_2) - \varepsilon_2\overline{Q_3Q_4}h_x^2\mathbf{E}_{2x}(Q_2)),\\
w(Q_3) :=& \frac{1}{2\bar{h}'_{i_3}}(-\varepsilon_1\overline{Q_1C_2}\,\overline{Q_3C_2}^2\mathbf{E}_{1z}(Q_3) - \varepsilon_2\overline{C_2Q_2}\,\overline{Q_3C_2}^2\mathbf{E}_{1z}(Q_3))\\
&+ \frac{1}{2\bar{h}'_{i_3}}(\varepsilon_1\overline{Q_1C_2}h_x^2\mathbf{E}_{3x}(Q_3) + \varepsilon_2\overline{C_2Q_2}h_x^2\mathbf{E}_{3x}(Q_3)),\\
w(Q_4) :=& \frac{1}{2\bar{h}'_{i_4}}(-\varepsilon_2\overline{Q_1Q_2}\,\overline{C_2Q_4}^2\mathbf{E}_{1z}(Q_4) + \varepsilon_2\overline{Q_1Q_2}h_x^2\mathbf{E}_{3x}(Q_4)),
\end{aligned}$$

and λ_i is a term due to the jump in the coefficients across the interface:

$$\begin{aligned}\lambda_i =& \frac{1}{2}\{-\varepsilon_1\overline{Q_3C_2}h_x^2\mathbf{E}_{2x}(Q_1) - \varepsilon_2\overline{C_2Q_4}h_x^2\mathbf{E}_{2x}(Q_1) + \varepsilon_2\overline{Q_3Q_4}h_x^2\mathbf{E}_{2x}(Q_2)\}\\ &+ \frac{1}{2}\{-\varepsilon_1\overline{Q_1C_2}h_x^2\mathbf{E}_{3x}(Q_3) - \varepsilon_2\overline{C_2Q_2}h_x^2\mathbf{E}_{3x}(Q_3) + \varepsilon_2\overline{Q_1Q_2}h_x^2\mathbf{E}_{3x}(Q_4)\}\\ :\equiv& \frac{1}{2}\,\mathrm{I} + \frac{1}{2}\,\mathrm{II}\,.\end{aligned}$$

As above, we can write

$$(\dot{E}'_f - \dot{E}_e)_i = \frac{1}{\bar{s}'_i}(C'\dot{w})_i + \frac{1}{\bar{s}'_i}\dot{\lambda}_i + \dot{v}_i, \tag{5.29}$$

where $\dot{v}_i = (\varepsilon_1 s_i^2 \dot{v}_i^1 + \varepsilon_2 s_i^2 \dot{v}_i^2)/\bar{s}'_i$. It is easy to see by using (5.27) and (5.29) that

$$\begin{aligned}(\dot{E}'_f - \dot{E}_e, \phi)_{W'} &= (C'\dot{w}, D\phi) + (\dot{v}, \phi)_{W'} + (S^{'-1}\dot{\lambda}, \phi)_{W'}\\ &= (D'\dot{w}, C\phi) + (\dot{v}, \phi)_{W'} + (S^{'-1}\dot{\lambda}, \phi)_{W'}.\end{aligned}$$

The estimates in (5.24) can be proved similarly to those in Lemma 5.2. We show only (5.25).

First, we rewrite $\dot{\mathrm{I}}$ as $\dot{\mathrm{I}} = \dot{\delta}_1 + \dot{\delta}_2$ with

$$\begin{aligned}\dot{\delta}_1 &= -\varepsilon_2\overline{C_2Q_4}h_x^2\dot{\mathbf{E}}_{2x}(Q_1) + \varepsilon_2\overline{C_2Q_4}h_x^2\dot{\mathbf{E}}_{2x}(Q_2),\\ \dot{\delta}_2 &= -\varepsilon_1\overline{Q_3C_2}h_x^2\dot{\mathbf{E}}_{2x}(Q_1) + \varepsilon_2\overline{Q_3C_2}h_x^2\dot{\mathbf{E}}_{2x}(Q_2).\end{aligned}$$

Note that the term $\dot{\delta}_1$ clearly vanishes for any linear field $\mathbf{E}$, so it can be absorbed into the term $\dot{v}_i$. The remaining term $\dot{\delta}_2$ can be written as

$$\dot{\delta}_2 = \varepsilon_1\overline{Q_3C_2}h_x^2\{-\dot{\mathbf{E}}_{2x}(Q_1) + \dot{\mathbf{E}}_{2x}(C_2)\} - \varepsilon_2\overline{Q_3C_2}h_x^2\{\dot{\mathbf{E}}_{2x}(C_2) - \dot{\mathbf{E}}_{2x}(Q_2)\}$$

by using the interface condition (1.7) and the fact that the function ρ_Γ depends only on the spatial variables. Then, by the Hölder continuity of $\dot{\mathbf{E}}_{2x}$, we have

$$\begin{aligned}|\dot{\mathbf{E}}_{2x}(Q_1) - \dot{\mathbf{E}}_{2x}(C_2)| &\le Kh^{\frac{1}{2}}\|\dot{\mathbf{E}}\|_{C^{1,\frac{1}{2}}(\tau'_{i_1})},\\ |\dot{\mathbf{E}}_{2x}(Q_2) - \dot{\mathbf{E}}_{2x}(C_2)| &\le Kh^{\frac{1}{2}}\|\dot{\mathbf{E}}\|_{C^{1,\frac{1}{2}}(\tau'_{i_2})},\end{aligned}$$

where τ'_{i_r} is the intersection of Ω_r with the union of all dual elements sharing the dual face κ'_i $(r = 1, 2)$. Hence,

$$|\delta_2| \le Kh^{\frac{7}{2}}\left\{\epsilon_1^{\frac{1}{2}}\|\dot{\mathbf{E}}\|_{C^{1,\frac{1}{2}}(\tau'_{i_1})} + \epsilon_2^{\frac{1}{2}}\|\dot{\mathbf{E}}\|_{C^{1,\frac{1}{2}}(\tau'_{i_2})}\right\}.$$

The term II can be estimated in the same manner. The rest of the proof is the same as the proof for ξ in (5.15). □

We are now ready to give the main result of this section.

Theorem 5.4. *Assume that the following regularity hypotheses hold for the solution of the interface Maxwell system* (1.1)–(1.8):

$$\mathbf{E} \in W^{1,1}(0,T;H^3(\Omega_i)^3) \cap W^{2,1}(0,T;W^{2,p}(\Omega_i)^3), \qquad \mathbf{B} \in W^{1,1}(0,T;H^3(\Omega_i)^3)$$

for $i = 1, 2$ *and* $p > 3$*, and* $(E,\ B)$ *is the solution of* (3.5)–(3.6) *on a nonuniform rectangular grid of size* h*. Then we have*

$$\max_{0\le t\le T}\{\|(E-E_e)(t)\|_{W'} + \|(B-B_f)(t)\|_W\}$$

$$\le Kh^2\sum_{i=1}^{2}\{\|\epsilon_i^{\frac{1}{2}}\mathbf{E}\|_{W^{1,1}(0,T;H^3(\Omega_i)^3)} \tag{5.30}$$

$$+ \|\epsilon_i^{\frac{1}{2}}\mathbf{E}\|_{W^{2,1}(0,T;W^{2,p}(\Omega_i)^3)} + \|\mu_i^{-\frac{1}{2}}\mathbf{B}\|_{W^{1,1}(0,T;H^3(\Omega_i)^3)}\}.$$

Proof. It follows from (5.1), (5.13), (5.14), and (5.23) that

$$\begin{aligned}
&\frac{1}{2}\frac{d}{dt}(\|B-B_f\|_W^2 + \|E-E_e\|_{W'}^2)\\
&\quad= (C(E-E_e), D'(B_f - B_e')) + (\dot v, E-E_e)_{W'}\\
&\qquad+ (D'\dot w, C(E-E_e)) + (S^{'-1}\dot\lambda, E-E_e)_{W'}\\
&\quad= (C(E-E_e), D'u) + (C(E-E_e), \xi) + (\dot v, E-E_e)_{W'}\\
&\qquad- (\dot w, \dot B - \dot B_f)_W + (S^{'-1}\dot\lambda, E-E_e)_{W'}\\
&\quad= -\,(\dot B - \dot B_f, u)_W - (\dot B - \dot B_f, D^{'-1}\xi)_W + (\dot v, E-E_e)_{W'}\\
&\qquad- (\dot w, \dot B - \dot B_f)_W + (S^{'-1}\dot\lambda, E-E_e)_{W'}\,.
\end{aligned}$$

Integrating over $(0, t_1)$, we have

$$\begin{aligned}
&\frac{1}{2}(\|B-B_f\|_W^2 + \|E-E_e\|_{W'}^2)(t_1)\\
&\quad= \int_0^{t_1}[-(\dot B - \dot B_f, u)_W - (\dot B - \dot B_f, D^{'-1}\xi)_W + (\dot v, E-E_e)_{W'}\\
&\qquad\qquad -(\dot w, \dot B - \dot B_f)_W + (S^{'-1}\dot\lambda, E-E_e)_{W'}]dt\,.
\end{aligned}$$

Then, by integration by parts,

$$\begin{aligned}
&\frac{1}{2}(\|B-B_f\|_W^2 + \|E-E_e\|_{W'}^2)(t_1)\\
&\quad= \int_0^{t_1}[(\dot v, E-E_e)_{W'} + (S^{'-1}\dot\lambda, E-E_e)_{W'}]\,dt\\
&\qquad+ \int_0^{t_1}(B-B_f, \dot u + \ddot w)_W\,dt + \int_0^{t_1}(B-B_f, D^{'-1}\dot\xi)_W\,dt\\
&\qquad- (B-B_f, \dot w + u)_W(t_1) - (B-B_f, D^{'-1}\xi)_W(t_1).
\end{aligned}$$

Now the desired estimate follows from the Cauchy–Schwarz inequality and the estimates in Lemmas 5.2 and 5.3. □

5.3. Superconvergence in the discrete $H(\mathbf{curl};\Omega)$-norm. We now show that the finite volume scheme (3.5)–(3.6) has certain superconvergence property; namely, the errors $E - E_e$ and $B - B_f$ are also second order convergent in a discrete $H(\mathrm{curl};\Omega)$-norm. To do so, we first differentiate (3.5) with respect to t to obtain

$$S'\frac{d^2E}{dt^2} - C'\frac{dB}{dt} = \frac{d\tilde J}{dt},$$

and then by (3.6) we obtain

$$S'\frac{d^2E}{dt^2}+C'S^{-1}CE=\frac{d\tilde{J}}{dt}. \tag{5.31}$$

We supplement (5.31) with the following initial conditions:

$$E(0)=E_e(0), \qquad \dot{E}(0)=\dot{E}_e(0). \tag{5.32}$$

Upon rewriting (5.31) as

$$S'\frac{d^2}{dt^2}(E-E_e)+C'S^{-1}C(E-E_e)=\frac{d\tilde{J}}{dt}-S'\frac{d^2E_e}{dt^2}-C'S^{-1}CE_e,$$

and by (3.3), we then have

$$S'\frac{d^2}{dt^2}(E-E_e)+C'S^{-1}C(E-E_e)=S'\frac{d^2}{dt^2}(E_f'-E_e)+\frac{d}{dt}(C'(B_f-B_e')). \tag{5.33}$$

This indicates that $E-E_e$ satisfies the ordinary differential equation (5.33) with the homogeneous initial conditions

$$(E-E_e)(0)=0, \qquad (\dot{E}-\dot{E}_e)(0)=0. \tag{5.34}$$

Multiplying (5.33) by $D(\dot{E}-\dot{E}_e)$, we obtain

$$\begin{aligned}&(S'(\ddot{E}-\ddot{E}_e),D(\dot{E}-\dot{E}_e))+(C'S^{-1}C(E-E_e),D(\dot{E}-\dot{E}_e))\\ &=(S'(\ddot{E}_f'-\ddot{E}_e),D(\dot{E}-\dot{E}_e))+(C'(\dot{B}_f-\dot{B}_e'),D(\dot{E}-\dot{E}_e)).\end{aligned}$$

Then, using (2.7), we get

$$\begin{aligned}&(S'(\ddot{E}-\ddot{E}_e),D(\dot{E}-\dot{E}_e))+(D'S^{-1}C(E-E_e),C(\dot{E}-\dot{E}_e))\\ &=(S'(\ddot{E}_f'-\ddot{E}_e),D(\dot{E}-\dot{E}_e))+(D'(\dot{B}_f-\dot{B}_e'),C(\dot{E}-\dot{E}_e)),\end{aligned}$$

which can be written as

$$\begin{aligned}&\frac{1}{2}\frac{d}{dt}\|\dot{E}-\dot{E}_e\|_{W'}^2+\frac{1}{2}\frac{d}{dt}\|E-E_e\|_V^2\\ &=(\ddot{E}_f'-\ddot{E}_e,\dot{E}-\dot{E}_e)_{W'}+(D'(\dot{B}_f-\dot{B}_e'),C(\dot{E}-\dot{E}_e)).\end{aligned} \tag{5.35}$$

The following theorem gives a superconvergence result for $E-E_e$.

Theorem 5.5. *Assume that*

$$\mathbf{E}\in W^{2,1}(0,T;H^3(\Omega_i)^3)\cap W^{3,1}(0,T;W^{2,p}(\Omega_i)^3), \qquad \mathbf{B}\in W^{2,1}(0,T;H^3(\Omega_i)^3)$$

satisfy the interface Maxwell system (1.1)–(1.8) *for* $i=1,2$ *and* $p>3$, *and* $(E,\,B)$ *is the solution of* (3.5)–(3.6) *on a nonuniform rectangular grid of size* h. *Then we have*

$$\begin{aligned}&\max_{0\le t\le T}\{\|(\dot{E}-\dot{E}_e)(t)\|_{W'}+\|(E-E_e)(t)\|_V\}\\ &\le Kh^2\sum_{i=1}^{2}\{\|\epsilon_i^{\frac{1}{2}}\mathbf{E}\|_{W^{2,1}(0,T;H^3(\Omega_i)^3)}\\ &+\|\epsilon_i^{\frac{1}{2}}\mathbf{E}\|_{W^{3,1}(0,T;W^{2,p}(\Omega_i)^3)}+\|\mu_i^{-\frac{1}{2}}\mathbf{B}\|_{W^{2,1}(0,T;H^3(\Omega_i)^3)}\}.\end{aligned} \tag{5.36}$$

Proof. By Lemma 5.3 we have

$$(\dot{E}'_f - \dot{E}_e, E - E_e)_{W'} = (\dot{v}, E - E_e)_{W'} + (D'\dot{w}, C(E - E_e)) + (S'^{-1}\dot{\lambda}, E - E_e)_{W'}.$$

A proof similar to the one for (5.23) leads to the following relations:

$$(\ddot{E}'_f - \ddot{E}_e, E - E_e)_{W'} = (\ddot{v}, E - E_e)_{W'} + (D'\ddot{w}, C(E - E_e)) + (S'^{-1}\ddot{\lambda}, E - E_e)_{W'},$$
$$(\dddot{E}'_f - \dddot{E}_e, E - E_e)_{W'} = (\dddot{v}, E - E_e)_{W'} + (D'\dddot{w}, C(E - E_e)) + (S'^{-1}\dddot{\lambda}, E - E_e)_{W'},$$

with $\ddot{v}, \dddot{v}, \ddot{w}, \dddot{w}, \ddot{\lambda}, \ \dddot{\lambda}$ obeying the same estimates as those stated in Lemma 5.3. In addition, by Lemma 5.2 and (5.1), we have

$$(C(E - E_e), D'(B_f - B'_e)) = (C(E - E_e), u) + (C(E - E_e), \xi).$$

Again, by a proof similar to the one of Lemma 5.2 we deduce that

$$(C(E - E_e), D'(\dot{B}_f - \dot{B}'_e)) = (C(E - E_e), \dot{u}) + (C(E - E_e), \dot{\xi}),$$
$$(C(E - E_e), D'(\ddot{B}_f - \ddot{B}'_e)) = (C(E - E_e), \ddot{u}) + (C(E - E_e), \ddot{\xi}),$$

with the corresponding estimates for $\dot{u}, \ddot{u}, \dot{\xi}$, and $\ddot{\xi}$ as those stated in Lemma 5.2. Now, integrating (5.35) over $[0, t_1]$, and by (5.34), we obtain

$$\begin{aligned}
&\|(\dot{E} - \dot{E}_e)(t_1)\|^2_{W'} + \|(E - E_e)(t_1)\|^2_V \\
&= 2\int_0^{t_1} (\ddot{E}'_f - \ddot{E}_e, \dot{E} - \dot{E}_e)_{W'}\, ds + 2\int_0^{t_1} (D'(\dot{B}_f - \dot{B}'_e), C(\dot{E} - \dot{E}_e))\, ds.
\end{aligned}$$

An application of integration by parts yields

$$\begin{aligned}
&\|(\dot{E} - \dot{E}_e)(t_1)\|^2_{W'} + \|(E - E_e)(t_1)\|^2_V \\
&= 2\int_0^{t_1} (\ddot{E}'_f - \ddot{E}_e, \dot{E} - \dot{E}_e)_{W'}\, ds \\
&\quad + 2(D'(\dot{B}_f - \dot{B}'_e), C(E - E_e))(t_1) - 2\int_0^{t_1} (D'(\ddot{B}_f - \ddot{B}'_e), C(E - E_e))\, ds.
\end{aligned}$$

Substituting the relations given in the beginning of the proof into the above equation, and using the Cauchy–Schwarz inequality together with the estimates in Lemmas 5.2 and 5.3, we obtain the desired estimate. $\square$

The following theorem gives a superconvergence result for $B - B_f$.

THEOREM 5.6. *Under the same assumptions as in Theorem* 5.5, *we have*

$$\begin{aligned}
&\max_{0\le t\le T}\left\{ \|(\dot{B} - \dot{B}_f)(t)\|_W + \sup_{\phi\in\mathbb{R}^{M_1}} \frac{|(C'(B - B_f), D\phi)|}{\|\phi\|_V} \right\} \\
&\quad \le Kh^2 \sum_{i=1}^{2} \{ \|\epsilon_i^{\frac{1}{2}}\mathbf{E}\|_{W^{2,1}(0,T;H^3(\Omega_i)^3)} + \|\epsilon_i^{\frac{1}{2}}\mathbf{E}\|_{W^{3,1}(0,T;W^{2,p}(\Omega_i)^3)} \\
&\qquad + \|\mu_i^{-\frac{1}{2}}\mathbf{B}\|_{W^{2,1}(0,T;H^3(\Omega_i)^3)} \}\ .
\end{aligned}$$

Proof. By (5.1) and (5.36), we obtain

$$\begin{aligned}
&\max_{0\le t\le T} \|(\dot{B} - \dot{B}_f)(t)\|_W \\
&\quad \le\ Kh^2 \sum_{i=1}^{2} \{ \|\epsilon_i^{\frac{1}{2}}\mathbf{E}\|_{W^{2,1}(0,T;H^3(\Omega_i)^3)} +\ \|\epsilon_i^{\frac{1}{2}}\mathbf{E}\|_{W^{3,1}(0,T;W^{2,p}(\Omega_i)^3)} \\
&\qquad + \|\mu_i^{-\frac{1}{2}}\mathbf{B}\|_{W^{2,1}(0,T;H^3(\Omega_i)^3)} \}.
\end{aligned}$$

By (5.2), we have

$$C'(B-B_f) = S'\frac{d}{dt}(E-E_e) - S'\frac{d}{dt}(E'_f - E_e) - C'(B_f - B'_e). \tag{5.37}$$

For any $\phi \in \mathbb{R}^{M_1}$, multiplying (5.37) by $D\phi$ and using (2.7), we obtain

$$(C'(B-B_f), D\phi) = (\dot{E} - \dot{E}_e, \phi)_{W'} - (\dot{E}'_f - \dot{E}_e, \phi)_{W'} - (D'(B_f - B'_e), C\phi).$$

First, by (5.36) we have

$$\begin{aligned} &|(\dot{E} - \dot{E}_e, \phi)_{W'}| \\ &\quad \le\ Kh^2 \|\phi\|_{W'} \sum_{i=1}^{2} \{ \|\epsilon_i^{\frac{1}{2}} \mathbf{E}\|_{W^{2,1}(0,T;H^3(\Omega_i)^3)} +\ \|\epsilon_i^{\frac{1}{2}} \mathbf{E}\|_{W^{3,1}(0,T;W^{2,p}(\Omega_i)^3)} \\ &\qquad\qquad + \|\mu_i^{-\frac{1}{2}} \mathbf{B}\|_{W^{2,1}(0,T;H^3(\Omega_i)^3)} \}. \end{aligned}$$

Then, using (5.23) and (5.25), we easily derive

$$|(\dot{E}'_f - \dot{E}_e, \phi)_{W'}| \le Kh^2 \|\phi\|_V \sum_{i=1}^{2} \{ \|\epsilon_i^{\frac{1}{2}} \mathbf{E}\|_{W^{2,1}(0,T;H^3(\Omega_i)^3)} + \|\epsilon_i^{\frac{1}{2}} \mathbf{E}\|_{W^{3,1}(0,T;W^{2,p}(\Omega_i)^3)} \},$$

while using (5.14) and (5.15) we have

$$|(D'(B_f - B'_e), C\phi)| \le Kh^2 \|\phi\|_V \sum_{i=1}^{2} \|\mu_i^{-\frac{1}{2}} \mathbf{B}\|_{W^{2,1}(0,T;H^3(\Omega_i)^3)}.$$

Collecting the above results leads to

$$\begin{aligned} \frac{|(C'(B-B_f), C\phi)|}{\|\phi\|_V} &\le K_1 h^2 \sum_{i=1}^{2} (\|\epsilon_i^{\frac{1}{2}} \mathbf{E}\|_{W^{2,1}(0,T;H^3(\Omega_i)^3)} + \|\epsilon_i^{\frac{1}{2}} \mathbf{E}\|_{W^{3,1}(0,T;W^{2,p}(\Omega_i)^3)}) \\ &\quad + K_2 h^2 \sum_{i=1}^{2} \|\mu_i^{-\frac{1}{2}} \mathbf{B}\|_{W^{2,1}(0,T;H^3(\Omega_i)^3)} \end{aligned}$$

for any $\phi \in \mathbb{R}^{M_1}$. □

6. Conclusion. Through a detailed analysis, we have established some rigorous convergence results for a finite volume method for the time-dependent Maxwell's equations in a three-dimensional polyhedral domain. Different materials are allowed to occupy portions of the domain, and interface conditions are imposed. Our analysis does not require extra regularity assumptions on the solutions of the interface problem beyond those for the analogous convergence results for noninterface Maxwell's equations, and our estimates also exhibit the detailed dependence on the material parameters. For brevity, we have chosen the case of two subdomains in our derivations, though much of our theory can be generalized to cases involving multiple subdomains. Implementations and applications of the methods discussed here are currently underway, and the results will be reported elsewhere.

Acknowledgments. The authors would like to thank the editor and the referees whose constructive suggestions and comments significantly improved the presentation of the paper.

REFERENCES

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

[2] H. T. Banks and J. Zou, *Regularity and approximation of systems arising in electromagnetic interrogation of dielectric materials*. Numer. Funct. Anal. Optim., 20 (1999), pp. 609–627.

[3] A. Chatterjee, L. C. Kempel, and J. L. Volakis, *Finite Element Method for Electromagnetics: Antennas, Microwave Circuits, and Scattering Applications*, IEEE Press, New York, 1998.

[4] J. S. Chen and K. S. Yee, *The finite-difference time-domain and the finite-volume time-domain methods in solving Maxwell's equations*. IEEE Trans. Antennas and Propagation, 45 (1997), pp. 354–363.

[5] Z. Chen, Q. Du, and J. Zou, *Finite element methods with matching and non-matching meshes for Maxwell equations with discontinuous coefficients*, SIAM J. Numer. Anal., 37 (2000), pp. 1542–1570.

[6] T. S. Chung and J. Zou, *A finite volume method for Maxwell's equations with discontinuous physical coefficients*, Int. J. Appl. Math., 7 (2001), pp. 201–223.

[7] P. Ciarlet, Jr., and J. Zou, *Fully discrete finite element approaches for time-dependent Maxwell equations*, Numer. Math., 82 (1999), pp. 193–219.

[8] G. Duvaut and J. Lions, *Inequalities in Mechanics and Physics*, Springer-Verlag, New York, 1976.

[9] S. Fortune, *Voronoi diagrams and Delaunay triangulations*, in Computing in Euclidean Geometry, World Scientific, Singapore, 1992, pp. 193–233.

[10] V. Girault and P. A. Raviart, *Finite Element Approximation of the Navier-Stokes Equations*, Springer-Verlag, New York, 1979.

[11] J. Jin, *The Finite Element Method in Electromagnetics*, John Wiley and Sons, New York, 1993.

[12] J. L. Lions and E. Magenes, *Non-homogeneous Boundary Value Problems and Applications* I, Springer-Verlag, Berlin, Heidelberg, 1972.

[13] P. Monk, *Analysis of a finite element method for Maxwell's equations*, SIAM J. Numer. Anal., 29 (1992), pp. 714–729.

[14] P. Monk and E. Süli, *A convergence analysis of Yee's scheme on nonuniform grids*, SIAM J. Numer. Anal., 31 (1994), pp. 393–412.

[15] R. A. Nicolaides, *Direct discretization of planer div-curl problems*, SIAM J. Numer. Anal., 29 (1992), pp. 32–56.

[16] R. A. Nicolaides and D. Q. Wang, *Convergence analysis of a covolume scheme for Maxwell's equations in three dimensions*, Math. Comp., 67 (1998), pp. 947–963.

[17] R. A. Nicolaides and X. Wu, *Covolume solutions of three-dimensional div-curl equations*, SIAM J. Numer. Anal., 34 (1997), pp. 2195–2203.

[18] P. A. Raviart, *Finite Element Approximation of the Time Dependent Maxwell Equations*, Technical Report GdR SPARCH #6, Ecole Polytechnique, Palaiseau Cedex, France, 1993.

[19] N. O. Sadiku, *Elements of Electromagnetics*, Oxford University Press, Oxford, UK, 2001.

[20] A. Taflove, *Computational Electrodynamics*, Artech House, Boston, MA, 1995.

[21] K. S. Yee, *Numerical solution of initial boundary value problems involving Maxwell's equations in isotropic media*, IEEE Trans. Antennas and Propagation, 14 (1966), pp. 302–307.

SIAM J. NUMER. ANAL.
Vol. 41, No. 1, pp. 64–89

A MIXED SPECTRAL-DIFFERENCE METHOD FOR THE STEADY STATE BOLTZMANN–POISSON SYSTEM[*]

CHRISTIAN RINGHOFER[†]

Abstract. The approximate solution of the Boltzmann transport equation via Galerkin-type series expansion methods leads to a system of conservation laws in space and time for the expansion coefficients. In this paper, we derive discretization methods for these equations in the mean field approximation, which are based on the entropy principles of the underlying Boltzmann equation, and discuss the performance of these discretizations and the series expansion approach in nonequilibrium regimes.

Key words. Boltzmann equation, Galerkin methods, finite differences

AMS subject classifications. 65N35, 65N05

PII. S003614290138958X

1. Introduction.

This paper is concerned with the numerical solution of the steady state Boltzmann–Poisson system, describing the transport of an ensemble of electrons or holes in a crystal interacting with a phonon background under the mean field approximation. The Boltzmann–Poisson system in steady state is given by (see [14])

$$(1)\qquad \text{(a)}\quad \nabla_x \cdot [\nabla_k \varepsilon(k) f] - q\nabla_k \cdot [\nabla_x V(x) f] + \frac{1}{\lambda} Q(f) = 0,$$

$$\text{(b)}\quad -\sigma \Delta_x V + q[D^{dop}(x) - \rho] = 0, \quad \rho(x) := \int f(x,k)dk,$$

where the phase space density function $f(x,k)$ is a function of position $x \in R_x^d$ ($d =$ 1, 2, or 3) and wave vector $k \in R_k^3$. The function $\varepsilon(k)$ describes the energy band under consideration. So $\nabla_k \varepsilon$ denotes the velocity with which a particle (electron or hole) with wave vector k travels. (If more than one energy band is considered, one density function f per band would have to be computed.) The Boltzmann equation (1) arises from a many body problem under the mean field approximation. So the function $V(x)$ denotes the mean field potential and q denotes the charge of the particle ($q = -1$ for electrons, $q = 1$ for holes). $\rho(x)$, as defined in (1)(b), denotes the density of particles in physical space and σ stands for the dielectricity constant of the material. The function $D^{dop}(x)$ in (1)(b) models the doping concentration, the background density of ions due to the implantation of donor and acceptor atoms into the crystal. Equation (1) has already been brought into a scaled dimensionless form, where the parameter λ denotes scaled mean-free path, i.e., the average distance a particle travels before it undergoes a collision event. (See [14, Chap. 1] for details of the scaling.) Finally, collisions of electrons or holes with the phonon background (the vibrations of the crystal lattice)

[*] Received by the editors May 20, 2001; accepted for publication (in revised form) June 5, 2002; published electronically February 12, 2003. This research was supported by NSF grants DMS 970-6792 and INT 960-3253.

http://www.siam.org/journals/sinum/41-1/38958.html

[†] Department of Mathematics, Arizona State University, Tempe, AZ 85287-1804 (ringhofer@asu.edu).

are modeled by the integral operator Q in (1)(a). For the semiconductor Boltzmann–Poisson problem, Q is of the form

$$Q[f](x,k) = \int_{R_k^3} S(k,k')[f(x,k)e^{\varepsilon(k)} - f(x,k')e^{\varepsilon(k')}]dk', \tag{2}$$

where $S(k,k')$ denotes the scaled scattering cross section. Because of the principle of detailed balance [3], the scattering cross section S is a symmetric function ($S(k,k') = S(k',k)$), which guarantees that the charge ρ is conserved by the collision operator Q. Moreover, the above form of the collision operator guarantees that the Maxwellians of the form $f(x,k) = c(x)e^{-\varepsilon(k)}$ are in the kernel of Q. The collision operator Q models the generation and annihilation of phonons, resulting in a specified gain or loss of energy of the particles. Consequently, the function S is a distribution of the form

$$S(k,k') = \sum_{\nu=-1}^{1} s_\nu(k,k')\delta(\varepsilon(k) - \varepsilon(k') + \nu\omega), \tag{3}$$

where ω is the amount of energy gained or lost due to the collisions. Because of the symmetry of the scattering cross section,

$$s_\nu(k,k') = s_{-\nu}(k',k)$$

has to hold. The assumption of a linear collision operator of the form (2) implies that we neglect electron-electron interactions. The steady state Boltzmann–Poisson system (1) is subject to a mixed set of Dirichlet and Neumann boundary conditions which will be discussed in the context of the discretization in section 3.

The outline of this paper is as follows. We will discretize the Boltzmann equation (1)(a) by a spectral Galerkin method in the wave vector direction, obtaining a system of conservation laws in x, which is solved by a difference method. This difference method is the actual topic of this paper. The spectral Galerkin method is chosen such as to preserve the entropy properties of the Boltzmann equation. So in section 2 we briefly review the general concepts involved in entropy based series expansion methods for the Boltzmann equation. In section 3, we derive the actual difference scheme which, in more than one spatial dimension, is based on a staggered grid approach. The resulting difference equations are nonlinear due to the coupling to the Poisson equation (1)(b). In section 4, we prove the stability of the linearization of these equations. Section 5 is devoted to the actual implementation of the method in one spatial dimension and to a numerical test example. The example consists of a standard $n^+ - n - n^+$ semiconductor diode with a $50nm$ channel. We demonstrate that in this example the solution to the Boltzmann equation exhibits truly kinetic features, which cannot be modeled by fluid dynamical approximations.

2. Entropy based Galerkin methods.

This paper is concerned with the spatial discretization of first order systems of partial differential equations in physical (x-) space, which arise from employing a certain type of series expansion method for the Boltzmann equation (1) in the wave vector direction. Series expansion methods for the Boltzmann–Poisson system in the context of modeling semiconductor devices were first used effectively in [8]. The expansion methods considered in this paper make use of the entropy property of the Boltzmann equation. In this section, we briefly present the use of such entropy principles in a somewhat more general framework than is actually used in this paper. It is generally more convenient to consider entropy principles for the case of the time dependent Boltzmann equation in the absence of boundary conditions ($x \in R_x^d$, $k \in R_k^3$), which we will do in this section. So the following is a brief review of general concepts described in detail in [7] or [11]. It should be pointed out here that, for the purpose of this paper, we approach the subject from a somewhat different angle. The entropy concept is usually relevant for the transient kinetic problem in the absence of boundary conditions. As a matter of fact, it does not hold in the form described here in the presence of boundary conditions. (See [5].) However, when discretizing the steady state Boltzmann equation with boundary conditions, one still has to deal with a rather stiff system of partial differential equations due to rapid potential variation; i.e., the same problems which make the discretization of the drift-diffusion or the energy transport system non-trivial arise again. The subject of this paper is the development of discretization methods for these systems, and we will only make use of the structure of the free-space operator induced by the free-space entropy property.

Generally, an entropy is given by a function $H(f,x,k)$, satisfying the following properties.

P1. The entropy is dissipated by the collision operator, so

$$\int_{R_x^d \times R_k^3} \partial_f H(f,x,k) Q[f](x,k) dxdk > 0$$

holds.

P2. The entropy is preserved by the free streaming operator, so

$$\int_{R_x^d \times R_k^3} \partial_f H(f,x,k)[\nabla_x \cdot (\nabla_k \varepsilon(k) f) - q \nabla_k \cdot (\nabla_x V(x) f)] dxdk = 0$$

holds.

P3. For fixed x and k the function H is a strictly convex function of the variable f and the functional

$$\eta(f) = \int_{R_x^d \times R_k^3} H(f(x,k),x,k) dxdk \tag{4}$$

is a convex functional of f.

The above properties guarantee that, for the time dependent Boltzmann equation

$$\partial_t f + \nabla_x \cdot [\nabla_k \varepsilon(k) f] - q \nabla_k \cdot [\nabla_x V(x) f] + \frac{1}{\lambda} Q(f) = 0, \tag{5}$$

the functional η is nonincreasing in time. This can be easily verified by integrating (5) against $\partial_f H(f,x,k)$ over $R_x^3 \times R_k^3$. However, since this paper is concerned with the steady state problem, we will use the structure induced by the entropy functional for a stability estimate for the steady state problem. Entropy based Galerkin methods are truncated series expansions of the solution f of (1) which preserve this structure. To this end, we introduce the inverse of $\partial_f H$ with respect to the variable f by

$$g = \partial_f H(f,x,k) \quad \Longleftrightarrow \quad f = \mu(g,x,k),$$

which exists since $\partial_f H$ is a monotone function of f. To obtain an approximation which preserves the entropy property, we apply a Galerkin procedure to the Boltzmann equation (5) in the entropy variable g. Thus, after choosing some basis functions $\{\phi_m(k),\, m = 0, 1, \dots, M\}$, we set

$$g(x,k,t) = \sum_{n=0}^{M} f_n(x,t)\phi_n(k), \quad f(x,k,t) = \mu\Big(\sum_{n=0}^{M} f_n(x,t)\phi_n(k), x, k\Big)$$

and integrate the Boltzmann equation for the entropy variable g, which is of the form

$$\partial_t \mu(g,x,k) + \nabla_x \cdot [\nabla_k \varepsilon(k)\mu(g,x,k)] - q\nabla_k \cdot [\nabla_x V(x)\mu(g,x,k)] + \frac{1}{\lambda}Q(\mu(g,x,k)) = 0$$

against each of the basis functions with respect to the wave vector k. This yields a first order system of partial differential equations of the form

$$\text{(6) (a)} \quad \partial_t G(F) + \sum_{j=1}^{d}[\partial_{x_j} A^j(F) - q(\partial_{x_j} V)B^j(F)] + \frac{1}{\lambda}C(F) = 0, \quad F = (f_0, \dots, f_M)^T,$$

$$\text{(b)} \quad G_m(F,x) = \int \phi_m(k) f(x,k,t)dk, \quad A^j_m(F,x) = \int \phi_m(k)(\partial_{k_j}\varepsilon(k)) f(x,k,t)dk,$$

$$\text{(c)} \quad B^j_m(F,x) = \int \phi_m(k)\partial_{k_j} f(x,k,t)dk, \quad C_m(F,x) = \int \phi_m(k) Q[f](x,k,t)dk,$$

which, by virtue of construction, automatically satisfies the entropy estimate

$$\partial_t \eta(F) \le 0, \quad \eta(F) = \int F^T(x,t)G(F,x)dx,$$

where, for simplicity, we denote the entropy function $\eta(F)$ of the coefficient vector F with the same symbol as the entropy $\eta(f)$ in (4) evaluated at the corresponding linear combination of the basis functions. The existence of the above estimate automatically guarantees that the generally nonlinear first order system (6) is hyperbolic [11]. The actual form of the system is determined by the choice of basis function ϕ_m and the form of the entropy function H.

Choice of entropies. The classical physical entropy, which arises in the kinetic description of fluids, where the collision operator Q models particle-particle interactions and is nonlinear, is given by the natural logarithm $\partial_f H = \ln f$, $H = f(\ln f - 1)$. Since we are dealing with a greatly simplified linear operator, and also are considering a plasma driven by the field $\nabla_x V$, there is a much larger degree of freedom in the choice of entropy. A straightforward calculation gives

$$\text{(7)} \qquad \int \partial_f H(f,x,k)Q[f](x,k)dk$$

$$= \frac{1}{2}\int_{R^3_k \times R^3_k} S(k,k')[\partial_f H(f,x,k) - \partial_f H(f',x,k')][fe^{\varepsilon} - f'e^{\varepsilon'}]dkdk',$$

where the prime denotes evaluation of the corresponding function at k'. Thus, any entropy H whose derivative is a nondecreasing function of fe^{ε} (meaning $\partial_f H(f,x,k) = h_1(fe^{\varepsilon(k)}, x)$ with $\partial_u h_1(u,x) \ge 0$) will satisfy the first requirement on the entropy since the term $[h_1(fe^{\varepsilon(k)},x)-h_1(f'e^{\varepsilon(k')},x)][f(x,k)e^{\varepsilon(k)}-f(x,k')e^{\varepsilon(k')}]$ in the integrand of (7) will always be nonnegative. To satisfy the second requirement, namely, that the entropy is preserved by the free streaming operator, we note that

$$\partial_f H(f,x,k)[\nabla_x\cdot(\nabla_k\varepsilon(k)f)-q\nabla_k\cdot(\nabla_x V(x)f)] \tag{8}$$

$$=[\nabla_x\cdot(\nabla_k\varepsilon(k)H)-q\nabla_k\cdot(\nabla_x V(x)H)]-[\nabla_k\varepsilon\cdot\nabla_2 H-q\nabla_x V(x)\cdot\nabla_3 H]$$

holds, where ∇_2,∇_3 denote the derivatives with respect to the second and third variables of $H(f,x,k)$. The integral over the whole phase space of the first term on the right-hand side of (8) is zero since it represents a total derivative. So the second condition on the entropy H is satisfied if the second term on the right-hand side of (8) vanishes. The general solution of $[\nabla_k\varepsilon\cdot\nabla_2-q\nabla_x V(x)\cdot\nabla_3]H(f,x,k)=0$ is given by all functions of the total energy $\varepsilon(k)+qV(x)$. So $H(f,x,k)$, and therefore also $\partial_f H(f,x,k)$, should depend on x and k only through the total energy $\varepsilon(k)+qV(x)$. Combining this with the first requirement $\partial_f H(f,x,k)=h_1(fe^{\varepsilon(k)},x)$, we see that any function H satisfying

$$\partial_f H(f,x,k)=h(f\exp(\varepsilon(k)+qV(x)),\quad h'(u)\geq 0,$$

can be chosen as an entropy for the Boltzmann equation. The physical entropy, the logarithm, now corresponds to the choice

$$h(u)=\ln u,\quad H(f,x,k)=f[\ln f-1+\varepsilon(k)+qV(x)],\quad \mu(g,x,k)=\exp[g-\varepsilon(k)-qV(x)].$$

If this entropy is used with the basis functions $1,k,\varepsilon(k)$, we obtain the classical hydrodynamic model for semiconductors. If more terms are used in the Galerkin approximation (6), we face the problem that, because of the exponential function μ, the involved integrals will in general become infinite. This problem can be remedied by using special restricted sets of basis functions developed by Levermore [11]. However, we face the additional problem that we have to evaluate the quite complicated integrals of the collision operator Q. In the Levermore approach, as well as in the hydrodynamic models, the collision operator is usually replaced by some form of BGK approximation with fitted relaxation times. Indeed, the Galerkin approach (6) can be philosophically viewed in two different ways. One is to regard (6) as deriving extensions of the Euler equations. The other one is to regard (6) as an actual numerical method for the Boltzmann equation. The difference is that in the first approach only relatively few terms are taken which have some physical interpretation. In the second approach, arbitrarily many terms in the expansion (6) have to be generated automatically, and the physical interpretation of the high order terms is not important. In this paper, we follow the second approach. This means that we use the simplest possible entropy, namely a quadratic function.

$$\eta(u)=u,\quad H(f,x,k)=\frac{1}{2}f^2\exp[\varepsilon(k)+qV(x)],\quad \mu(g,x,k)=g\exp[-\varepsilon(k)-qV(x)]. \tag{9}$$

This choice makes the Galerkin equations (6) linear, as long as we do not couple the Boltzmann equation to the Poisson equation (1)(b), and the entropy estimate becomes an L^2 estimate with the weight function $\exp[\varepsilon+qV]$. Methods based on these basis functions have been analyzed and implemented in [17], [18], and suitable time discretization methods have been developed in [15]. All of this work was concerned with the linear time dependent Boltzmann equation for a given potential V. In this paper, we treat the steady state Boltzmann–Poisson problem instead, using the same linear entropy. So the role of the entropy will not be to provide an a priori estimate on the time dependent solution. Instead, we will make use of the effect of the entropy on the structure of the involved nonlinear operators. The convergence of Galerkin methods using the linear entropy in the linear case (i.e., for a given potential) has been analyzed in [16], [18].

3. Difference schemes. In this section, we present the spatial discretization of the Galerkin equations arising from (6) together with the Poisson equation (1)(b). When designing the spatial discretization one is confronted with two somewhat contradictory priorities. On one hand, the discretization should reflect the entropy preservation property of the free streaming operator; that is, a discrete equivalent of P3 in section 2 should hold. In the case of the quadratic entropy used in this paper, this means that the spatially discretized free streaming operator should be antisymmetric with respect to the weight function e^{qV}. This results in obviously desirable stability properties of the discretization. On the other hand, we wish to locally conserve charge. Building the zero order moment of the Boltzmann equation (1)(a) with respect to the wave vector k yields the continuity equation

$$(10) \qquad \nabla_x \cdot \langle \nabla_k \varepsilon \rangle = 0, \quad \langle \nabla_k \varepsilon \rangle := \int \nabla_k \varepsilon(k) f(x,k) dk.$$

This continuity equation should hold locally; that is, we should be able to apply a discrete version of Gauss's theorem locally over any submesh. In the context of a finite element discretization in the spatial direction, the first priority would suggest a straightforward finite element discretization using a weighted scalar product, which would automatically not be locally conservative. A similar problem exists when using difference schemes. Our approach is the following: We will split the balance equations (6)(a) into those governing the even and odd order moments of the kinetic density function f. The difference scheme for the even order moment equations will be designed in such a way that it locally conserves the appropriate momenta. The difference scheme for the odd order moment equations will then be chosen such that the whole scheme has the appropriate entropy properties outlined in the previous section. This represents a compromise between the two priorities of local conservation and entropy dissipation. This compromise is acceptable, since the collisions modeled by the operator Q are nonelastic and the odd order moments are not conserved anyway. In the steady state case, the odd order moments have to be viewed rather as constitutive relations.

First we observe that the free streaming operator L, defined by

$$L[f] = \nabla_x \cdot [\nabla_k \varepsilon(k) f] - q \nabla_k \cdot [\nabla_x V(x) f],$$

induces a natural decomposition of the function space for f. If the band energy $\varepsilon(k)$ is an even function of the wave vector k (which always can be assumed), the operator L maps even functions of k into odd functions and vice versa. So we split the density function f, as well as the Boltzmann equation, into its even and odd parts

$$(11) \qquad (a) \quad f = f^e + f^o, \quad f^e(x,k) = f^e(x,-k), \quad f^o(x,k) = -f^o(x,-k),$$

$$(b) \quad L[f^o] + Q^e[f^e + f^o] = 0, \quad (c) \quad L[f^e] + Q^o[f^e + f^o] = 0,$$

where Q^e, Q^o denote the even and odd parts of the collision operator. Next we observe that the important conservation laws (for charge and energy) are given by the even part of the Boltzmann equation. Since the given collision operator does not preserve momentum anyway, the moments of the odd part of the Boltzmann equation gives rise to constitutive relations rather than conservation laws. Consequently, we will dicretize (11)(b) conservatively and choose the discretization of (11)(c) to satisfy the entropy conservation property P2 of the operator L. After splitting into even and odd parts, the entropy property of the free streaming operator L reads

$$\int_{R^3_x \times R^3_k} e^{\varepsilon+V} f^e L[f^o] dx dk + \int_{R^3_x \times R^3_k} e^{\varepsilon+V} f^o L[f^e] dx dk = 0. \tag{12}$$

We will use (12) as a weak definition of $L[f^e]$ in terms of $L[f^o]$. Moreover, we can also assume that in general the scattering cross section $S(k,k')$ (3) is an even function of its arguments. This implies that the collision operator Q maps even functions of k into even functions and odd functions into odd functions, giving

$$Q^e[f^e + f^o] = Q[f^e], \quad Q^o[f^e + f^o] = Q[f^o].$$

In keeping with the spirit of the entropy based Galerkin approach outlined in the previous section, we expand f^e and f^o into

$$f^e(x,k) = \sum_{n=0}^{N_e} f^e_n(x) e^{-\varepsilon} \phi^e_n(k), \quad f^o(x,k) = \sum_{n=0}^{N_o} f^o_n(x) e^{-\varepsilon} \phi^o_n(k),$$

where we have absorbed the factor e^{-qV} in (9) into the coefficients $f^{e,o}_n$ for notational convenience. Integrating (11)(b) against the even basis functions ϕ^e_m and (11)(c) against the odd basis functions ϕ^o_m yields the system

$$\text{(13)} \qquad \text{(a)} \quad \sum_{\nu=1}^{d} [A^{eo}_\nu \partial_{x_\nu} F^o - q(\partial_{x_\nu} V) B^{eo}_\nu F^o] + \frac{1}{\lambda} C^e F^e = 0,$$

$$\text{(b)} \quad \sum_{\nu=1}^{d} [A^{oe}_\nu \partial_{x_\nu} F^e - q(\partial_{x_\nu} V) B^{oe}_\nu F^e] + \frac{1}{\lambda} C^o F^o = 0,$$

where $F^{e,o}(x)$ denotes the coefficient vector $(f^{e,o}_0, \dots, f^{e,o}_{N_{e,o}})$ and the matrices are defined in the obvious way by

$$\text{(14) (a)} \quad A^{eo}_\nu(m,n) = \int [\phi^e_m (\partial_{k_\nu} \varepsilon) e^{-\varepsilon} \phi^o_n] dk, \quad A^{oe}_\nu(m,n) = \int [\phi^o_m (\partial_{k_\nu} \varepsilon) e^{-\varepsilon} \phi^e_n] dk,$$

$$\text{(b)} \quad B^{eo}_\nu(m,n) = \int \phi^e_m \partial_{k_\nu} [e^{-\varepsilon} \phi^o_n] dk, \quad B^{oe}_\nu(m,n) = \int \phi^o_m \partial_{k_\nu} [e^{-\varepsilon} \phi^e_n] dk,$$

$$\text{(c)} \quad C^e(m,n) = \int \phi^e_m Q[e^{-\varepsilon} \phi^e_n] dk, \quad C^o(m,n) = \int \phi^o_m Q[e^{-\varepsilon} \phi^o_n] dk.$$

As mentioned, we will discretize the even part (13)(a) by a conservative difference method and the odd part (13)(b) in its weak form defined by the relation (12). A simple calculation gives that the relations

$$B^{eo}_\nu + (B^{oe}_\nu)^T + A^{eo}_\nu = 0, \quad \nu = 1, \dots, d, \tag{15}$$

hold. Furthermore, because of the symmetry of the scattering cross section S, the odd collision operator is given by

$$C^o(m,n) = \int \phi^o_m \sigma(k) \phi^o_n dk, \quad \sigma(k) = \int S(k,k')dk'.$$

Therefore the matrix C^o, corresponding to the odd part of the collision operator, is invertible as a consequence of the fact that the kernel of Q contains only even functions. This suggests eliminating F^o locally by inverting C^o in (13)(b) and inserting the resulting expression into (13)(a). To discretize (13) in the spatial direction, we will need difference operators which are conservative and satisfy a discrete version of Gauss's theorem. This is nontrivial in more than one spatial dimension if a general nonrectangular mesh is used, unless they do not act on the same grids. We therefore assume two separate grids for F^e and F^o and difference operators acting between them. We define the meshes

$$M^e = \{\mathbf{x}_j,\ j = 0,\dots,J_e\}, \quad M^o = \{\mathbf{y}_j,\ j = 0,\dots,J_o\}$$

and difference operators D^{eo}_ν, D^{oe}_ν, both approximating ∂_{x_ν}, acting between them

$$(D^{eo}_\nu u^o)(\mathbf{x}_j) = \sum_{s=0}^{J_o} d^{eo}_\nu(j,s) u^o(\mathbf{y}_s), \quad (D^{oe}_\nu u^e)(\mathbf{y}_j) = \sum_{s=0}^{J_e} d^{oe}_\nu(j,s) u^e(\mathbf{x}_s)$$

for grid functions u^e, u^o defined on the respective meshes. Using appropriate discrete integration operators I_e, I_o, the discrete integration by parts formula we assume takes the form

$$(16)\ \text{(a)} \quad I_e[(u^e)^T(D^{eo}_\nu u^o)] = -I_o[(u^o)^T(D^{oe}_\nu u^e)] + b^e(u^e)^T b^o(u^o), \quad \nu = 1,\dots,d,$$

$$\text{(b)} \quad I_e[u^e] = \sum_{s=0}^{J_e} \gamma_e(s) u^e(\mathbf{x}_s), \quad I_o[u^o] = \sum_{s=0}^{J_o} \gamma_o(s) u^o(\mathbf{y}_s)$$

for vector grid functions u^e, u^o with appropriate integration weights γ_e, γ_o. The boundary operators b^e, b^o correspond to evaluation of the grid function at the boundary points of the respective meshes M^e, M^o. There is one minor problem caused by the introduction of the dual meshes M^e, M^o, which is that the free streaming operator L contains derivatives of the density function as well as zero order terms, which couples the meshes. This is remedied by using an interpolation formula which is in some sense generic for the free streaming operator. In order to write the term $q(\partial_{x_\nu}V)(\partial_{k_\nu}f)$ solely in terms of spatial derivatives of f, we replace it by $\partial_{x_\nu}[qV\partial_{k_\nu}f] - qV\partial_{x_\nu}\partial_{k_\nu}f$. Thus (13)(a) is discretized by

$$(17) \quad L^{eo}F^o + \frac{1}{\lambda}C^e F^e = 0, \quad L^{eo}F^o = \sum_{\nu=1}^{d}[D^{eo}_\nu(A^{eo}_\nu F^o - qVB^{eo}_\nu F^o) + qVB^{eo}_\nu D^{eo}_\nu F^o].$$

The odd part (13)(b) of the Boltzmann equation is now discretized using the dual operator to L^{eo} as defined by (12). This gives

$$(18) \qquad \text{(a)} \quad L^{oe}F^e + \frac{1}{\lambda}C^o F^o = 0,$$

$$\text{(b)} \quad L^{oe}F^e = \sum_{\nu=1}^{d} e^{-qV}[(A^{eo}_\nu - qVB^{eo}_\nu)^T D^{oe}_\nu(e^{qV}F^e) + (B^{eo}_\nu)^T D^{oe}_\nu(qVe^{qV}F^e)].$$

This represents a consistent discretization of (13)(b) because of the relation (15) for the coefficient matrices, as can be easily verified. By virtue of construction, the discrete system (17), (18) now has the desired symmetry properties and accurately represents a discrete version of the conservation law (10). We summarize these statements in the following proposition.

PROPOSITION 1. *The discretized operators* L^{eo}, L^{oe} *in* (17), (18) *satisfy the equality*

$$I_e[e^{qV}(F^e)^T L^{eo} F^o] + I_o[e^{qV}(F^o)^T L^{oe} F^e] \tag{19}$$

$$= \sum_{\nu=1}^{d} b^e(e^{qV}F^e)^T b^o(A_\nu^{eo}F^o - qVB_\nu^{eo}F^o) + b^e(qVe^{qV}F^e)b^o(B_\nu^{eo}D_\nu^{eo}F^o)$$

for any choice of grid functions F^e, F^o, V. *Furthermore, the discrete conservation law*

$$I_e[\mathbf{e}^T L^{eo} F^o] = \sum_{\nu=1}^{d} b^e(\mathbf{e})^T b^o(A_\nu^{eo}F^o) \tag{20}$$

holds, where $\mathbf{e}$ *denotes the first unit vector in* R^{N_e}.

Proof. Equation (19) holds by virtue of construction, using the discrete integration by parts formulae (16). Equation (20) is obtained from (19) by setting $F^e = \mathbf{e}$ and observing that because of (14)(b) the first row of B_ν^{eo} vanishes. □

We now turn to coupling (17), (18) to the Poisson equation (1)(b). Using the obvious discretization induced by the difference operators D_ν^{eo}, D_ν^{oe}, the discretization of the Poisson equation reads

$$-\sigma \sum_{\nu=1}^{d} D_\nu^{eo} D_\nu^{oe} V^e + q[D^{dop} - \rho^T F^e] = 0, \quad \rho^T = (\rho_1, \dots, \rho_{N_e}), \quad \rho_n = \int e^{-\varepsilon}\phi_n^e dk.$$

Therefore, if the even coefficient vector F^e is only defined on the mesh M^e, so is the potential. The discretization of the Boltzmann equation (17), (18) requires, however, the potential on both meshes M^e and M^o, and therefore some interpolation formula is needed which computes the potential V on the mesh M^o from V^e. So, in summary, the discretized Boltzmann–Poisson system is given by

$$\text{(a)} \quad L^{eo}(F^e, F^o) + \frac{1}{\lambda}C^e F^e = 0, \quad \text{(b)} \quad L^{oe}(F^e) + \frac{1}{\lambda}C^o F^o = 0, \tag{21}$$

$$\text{(c)} \quad -\sigma \sum_{\nu=1}^{d} D_\nu^{eo} D_\nu^{oe} V^e + q[D^{dop} - \rho^T F^e] = 0, \quad V^o = S(V^e),$$

$$\text{(d)} \quad L^{eo}(F^e, F^o) = \sum_{\nu=1}^{d} [D_\nu^{eo}(A_\nu^{eo}F^o - qV^o B_\nu^{eo}F^o) + qV^e B_\nu^{eo} D_\nu^{eo} F^o],$$

$$\text{(e)} \quad L^{oe}(F^e) = \sum_{\nu=1}^{d} e^{-qV^o}[(A_\nu^{eo} - qVB_\nu^{eo})^T D_\nu^{oe}(e^{qV^e}F^e) + (B_\nu^{eo})^T D_\nu^{oe}(qV^e e^{qV^e}F^e)],$$

where V^o is somehow interpolated from V^e using the interpolation operator S. Notice that L^{eo} and L^{oe} are now nonlinear operators because the potentials V^e, V^o are given in terms of F^e through the solution of the Poisson equation (21)(c). In principle, any interpolation formula could be used which is of a high enough order as to keep the total order of accuracy of (17), (18), and the choice of interpolation formula cannot be discussed further without being more specific about the structure of the mesh and the form of the difference operators. In practice, we will take the following approach: The Galerkin approximation to the Boltzmann–Poisson system will reduce to the drift-diffusion system if only the basis functions $1, k_1, k_2, k_3$ are used with $N_e = 0$, $N_o = 2$. In this case, the matrices B^{eo}_ν and C^e vanish and the matrices A^{eo}_ν will form the rows of a multiple of the identity matrix; i.e., there will be no mixed derivatives appearing in the equations. Given a certain mesh structure and a set of difference operators, we will choose the interpolation formula to compute V^o in such a way that the resulting scheme reduces to the well-known Scharfetter–Gummel scheme on this mesh.

Boundary conditions. For practical applications, the boundary will be decomposed in two types of boundary segments, namely insulating parts at which the particle fluxes normal to the boundary vanish and parts at which particles are injected according to a Maxwellian distribution in such a way that charge neutrality in the Poisson equation is preserved (meaning $\rho = D$ in (1)(b)). The insulation property simply translates into the odd component of the density function being equal to zero. The injection of particles results in a quite complicated mixed boundary condition involving the diagonalization of the matrices A^{eo}_ν, A^{oe}_ν. (See [15] for more details.) For the solution of the transient problem, it is quite essential to capture the precise structure of the boundary conditions in order to avoid artificial reflections. For the steady state problem, it turns out the structure of the boundary conditions has actually relatively little impact on the solution as long as charge neutrality is guaranteed. As mentioned before, we will actually eliminate the odd component vector F^o for the practical implementation of the scheme and solve (21) in the form

$$\lambda L^{eo}(F^e, -\lambda(C^o)^{-1}L^{oe}(F^e)) + C^e F^e = 0.$$

In this formulation, dealing with the mixed boundary condition resulting from the particle injection is actually quite cumbersome. We therefore use the simplified set of boundary conditions

$$(22) \qquad \text{(a)} \quad b^e(F^e) = D^{dop}\mathbf{e}, \quad V^e = V^e_b \quad \text{for } \mathbf{x}_j \in \partial M^e_{Dir},$$

$$\text{(b)} \quad b^o(F^o) = 0, \quad b^o(D^{eo}V^e) = 0 \quad \text{for } \mathbf{y}_j \in \partial M^o_{Neu},$$

where the subscripts Dir, Neu denote the union of Dirichlet and Neumann segments of the boundary and b^e, b^o are the boundary operators in (16). Note that b^o is actually the discretization of the outward normal component of a vector. Also, (22)(a) tacitly assumes that $\phi^e_0 = 1$ holds and that the basis functions form an orthonormal system, which implies that the physical space density ρ is actually given by F^e_0 and therefore Maxwellians of the form $e^{-\varepsilon}$ correspond to the first unit vector $\mathbf{e}$.

4. Stability of the linearization. In this section, we prove that the linearization of the scheme defined in the previous section around the equilibrium solution is stable. This implies, among other things, that close to equilibrium Newton's method will be locally quadratically convergent when applied to the nonlinear system (21). To this end, we first reformulate (21) slightly by essentially writing it as a correction to a discretization of the drift-diffusion–Poisson equations. We assume that the Galerkin approximation (13) to the Boltzmann equation contains the balance equations for charge and momentum, i.e., that the functions $1, k$ are contained in the set of basis functions. So we set

$$\phi_0^e = 1, \quad \phi_n^o = k_n, \quad n = 1, \dots, d.$$

Next we split the system (21) into the balance equations for charge and momentum and the rest by setting

$$F^e = \begin{pmatrix} F_0^e \\ F_1^e \end{pmatrix}, \quad F^o = \begin{pmatrix} F_0^o \\ F_1^o \end{pmatrix}, \quad L^{eo} = \begin{pmatrix} L_0^{eo} \\ L_1^{eo} \end{pmatrix}, \quad L^{oe} = \begin{pmatrix} L_0^{oe} \\ L_1^{oe} \end{pmatrix},$$

where F_0^e and L_0^{eo} denote the first component of F^e, L^{eo} and F_1^e and L_1^{eo} denote the other N_e components. In the same way, F_0^o, L_0^{oe} denote the first d components of F^o, L^{oe} and F_1^o, L_1^{oe} denote the other $N_o + 1 - d$ components. Accordingly, we partition the matrices A_ν^{eo}, B_ν^{eo}, C^e, and C^o into

$$A_\nu^{eo} = \begin{pmatrix} A_\nu^{00} & A_\nu^{01} \\ A_\nu^{10} & A_\nu^{11} \end{pmatrix}, \quad B_\nu^{eo} = \begin{pmatrix} 0 & 0 \\ B_\nu^{10} & B_\nu^{11} \end{pmatrix},$$

$$C^e = \begin{pmatrix} 0 & 0 \\ 0 & C_e^{11} \end{pmatrix}, \quad C^o = \begin{pmatrix} C_o^{00} & C_o^{01} \\ C_o^{10} & C_o^{11} \end{pmatrix}.$$

Note that the first row of B^{eo} and the first row and column of C^e vanish because of the conservation properties. In this partition, the system (21) now becomes

$$\text{(23)} \qquad \text{(a)} \quad L_0^{eo}(F_0^o, F_1^o) = 0, \quad \text{(b)} \quad L_0^{oe}(F_0^e, F_1^e) + \frac{1}{\lambda}[C_o^{00} F_0^o + C_o^{01} F_1^o] = 0,$$

$$\text{(c)} \quad -\sigma \sum_{\nu=1}^{d} D_\nu^{eo} D_\nu^{oe} V^e + q[D^{dop} - F_0^e] = 0, \quad V^o = S(V^e),$$

$$\text{(d)} \quad b^e(F_0^e) = D^{dop}, \quad b^e(V^e) = V_b^e, \quad \mathbf{x}_j \in \partial M_{Dir}^e,$$

$$b^o(F_0^o) = 0, \quad b^o(D^{eo} V^e) = 0 \quad \mathbf{y}_j \in \partial M_{Neu}^o,$$

(24)

$$\text{(a)} \quad L_1^{eo}(F_0^e, F_1^e, F_0^o, F_1^o) + \frac{1}{\lambda} C_e^{11} F_1^e = 0, \quad \text{(b)} \quad L_1^{oe}(F_0^e, F_1^e) + \frac{1}{\lambda}[C_o^{10} F_0^o + C_o^{11} F_1^o] = 0,$$

$$\text{(c)} \quad b^e(F_1^e) = 0, \quad \mathbf{x}_j \in \partial M_{Dir}^e, \quad b^o(F_1^o) = 0, \quad \mathbf{y}_j \in \partial M_{Neu}^o,$$

where the involved operators are given by

$$\text{(25)} \qquad \text{(a)} \quad L_0^{eo}(F_0^o, F_1^o) = \sum_{\nu=1}^{d} D_\nu^{eo}(A_\nu^{00} F_0^o + A_\nu^{01} F_1^o),$$

$$\text{(b)} \quad L_0^{oe}(F_0^e, F_1^e)$$

$$= \sum_{\nu=1}^{d} e^{-qV^o}[(A_\nu^{00})^T D_\nu^{oe}(e^{qV^e}F_0^e) + (A_\nu^{10} - qVB_\nu^{10})^T D_\nu^{oe}(e^{qV^e}F_1^e) + (B_\nu^{10})^T D_\nu^{oe}(qV^e e^{qV^e}F_1^e)],$$

$$\text{(c)} \quad L_1^{eo}(F_0^e, F_1^e, F_0^o, F_1^o)$$

$$= \sum_{\nu=1}^{d}[D_\nu^{eo}(A_\nu^{10}F_0^o + A_\nu^{11}F_1^o - qV^o B_\nu^{10}F_0^o) - qV^o B_\nu^{11}F_1^o) + qV^e B_\nu^{10} D_\nu^{eo}F_0^o + qV^e B_\nu^{11} D_\nu^{eo}F_1^o],$$

$$\text{(d)} \quad L_1^{oe}(F_0^e, F_1^e)$$

$$= \sum_{\nu=1}^{d} e^{-qV^o}[(A_\nu^{01})^T D_\nu^{oe}(e^{qV^e}F_0^e)(A_\nu^{11} - qVB_\nu^{11})^T D_\nu^{oe}(e^{qV^e}F_1^e) + (B_\nu^{11})^T D_\nu^{oe}(qV^e e^{qV^e}F_1^e)].$$

We write the system (23), (24) as a nonlinear equation for F_1^e and F_1^o, where the remaining variables are given implicitly by (23). So write (24) as

$$(26) \qquad \text{(a)} \quad K^e(F_1^e, F_1^o) = 0, \quad K^o(F_1^e, F_1^o) = 0,$$

$$\text{(b)} \quad K^e(F_1^e, F_1^o) = L_1^{eo}(F_1^e, F_1^e, F_0^o, F_1^o) + \frac{1}{\lambda}C_e^{11}F_1^e,$$

$$\text{(c)} \quad K^o(F_1^e, F_1^o) = L_1^{oe}(F_0^e, F_1^e) + \frac{1}{\lambda}[C_o^{10}F_0^o + C_o^{11}F_1^o]$$

with V^e, V^o, F_0^o, F_0^e given in terms of F_1^e, F_1^o as the solution of the equations (23). So, in order to evaluate K^e, K^o a drift-diffusion–Poisson problem has to be solved where F_1^e, F_1^o appear as source terms. We will show that the linearization of (26) is stable around the equilibrium solution. The equilibrium solution is given by a Maxwellian in the wave vector direction multiplied by the function e^{iqV}, where the potential V is such that it satisfies the resulting nonlinear Poisson equation (1). We first confirm that this solution is an exact solution of our difference scheme by proving the following theorem.

THEOREM 1. *If the boundary potential V_b^e in* (23)(b) *is given by $V_b^e = -\frac{1}{q}\ln D^{dop}$, and the corresponding discrete Poisson problem*

$$-\sigma \sum_{\nu=1}^{d} D_\nu^{eo} D_\nu^{oe} V^e + q[D^{dop} - e^{-qV^e}] = 0$$

has a solution, the system

$$K^e(F_1^e, F_1^o) = 0, \quad K^o(F_1^e, F_1^o) = 0$$

has a solution $F_1^e = 0$, $F_1^o = 0$.

The proof is deferred to the appendix.

If we now consider the linearization around this equilibrium solution, we have the following theorem.

THEOREM 2. *Let $\delta F_1^e, \delta F_1^o$ be the solution of the linearized problem*

$$\text{(27)} \qquad \text{(a)} \quad dK^e(0,0)(\delta F_1^e, \delta F_1^o) = R_1^e, \quad dK^o(0,0)(\delta F_1^e, \delta F_1^o) = R_1^o,$$

$$\text{(b)} \quad b^e(\delta F_1^e) = 0, \quad \mathbf{x}_j \in \partial M_{Dir}^e, \quad b^o(\delta F_1^o) = 0, \quad \mathbf{y}_j \in \partial M_{Neu}^o.$$

Then there exists a constant c, dependent only on the collision matrices C^e, C^o, such that $\delta F_1^e, \delta F_1^o$ satisfy $||(\delta F_1^e, \delta F_1^o)|| \le c\lambda ||(R_1^e, R_1^o)||$, where the norm is defined by

$$||(\delta F_1^e, \delta F_1^o)|| = I_e[e^{qV^e}|\delta F_1^e|^2] + I_o[e^{qV^o}|\delta F_1^o|^2],$$

and V^e, V^o are the potentials corresponding to the equilibrium solution.

The proof is deferred to the appendix.

Two remarks should be made at this point. First, while the reformulation and partitioning of the problem seems like a formality at first sight, it does have a practical implication. The stability result of Theorem 2 guarantees, among other things, that the solution can be computed by a locally quadratically convergent Newton method of the form

$$dK^e(F_1^e, F_1^o)(\delta F_1^e, \delta F_1^o) = -K^e(F_1^e, F_1^o), \quad dK^o(F_1^e, F_1^o)(\delta F_1^e, \delta F_1^o) = -K^o(F_1^e, F_1^o).$$

However, to compute the right-hand side of the linearized equations, we have to evaluate the terms K^e, K^o, which involves solving the nonlinear problem (23) exactly, which corresponds to solving a nonlinear drift-diffusion–Poisson problem at each Newton step. While this is not really necessary in practice, the convergence of Newton's method can be improved dramatically if a few extra iterations are performed on the low-dimensional system (23) within each Newton step. Second, the above formulation tacitly assumes that K^e, K^o can be evaluated, that is, that the drift-diffusion–Poisson system can be solved by the discretization (23) for any source terms arising from F_1^e, F_1^o. This is the main reason why the interpolation operator S in (21) is chosen such that the resulting scheme reduces to the Scharfetter–Gummel scheme on the given mesh, which is a well tested and incredibly robust discretization. (See [19] for an overview.)

5. Implementation and numerical test example. In this section we present a numerical test example in one spatial dimension and outline how to actually compute the coefficient matrices C^e, C^o, corresponding to the collision operator. The computation of these matrices is not completely trivial because of the presence of the $\delta-$ functions in the integral kernel (3). (See [12], [13].) In one spatial dimension, the choice of meshes M^e and M^o and the choice of the corresponding difference operators D_1^{eo}, D_1^{oe} is quite obvious. We define the grids by

$$M^e = \{\mathbf{x}_0 < \cdots < \mathbf{x}_J\}, \quad M^o = \left\{\mathbf{y}_j : \mathbf{y}_j = \frac{1}{2}(\mathbf{x}_j + \mathbf{x}_{j+1}),\ j = 0, \ldots, J-1\right\}$$

and the difference operators by

$$(D^{oe}u^e)(\mathbf{y}_j) = \frac{u^e(\mathbf{x}_{j+1}) - u^e(\mathbf{x}_j)}{\mathbf{x}_{j+1} - \mathbf{x}_j}, \quad (D^{eo}u^o)(\mathbf{x}_j) = \frac{u^o(\mathbf{y}_j) - u^o(\mathbf{y}_{j-1})}{\mathbf{y}_j - \mathbf{y}_{j-1}}.$$

If we define the discrete integral operators I_e, I_o by

$$I_e(u^e) = \sum_{j=1}^{J-1} (\mathbf{y}_j - \mathbf{y}_{j-1}) u^e(\mathbf{x}_j), \quad I_o(u^o) = \sum_{j=0}^{J-1} (\mathbf{x}_{j+1} - \mathbf{x}_j) u^o(\mathbf{y}_j),$$

then the discrete integration by parts formula (16) holds with the boundary terms b^e, b^o given by

$$b^e(u^e) = \begin{pmatrix} -u^e(\mathbf{x}_0) \\ u^e(\mathbf{x}_J) \end{pmatrix}, \quad b^o(u^o) = \begin{pmatrix} -u^o(\mathbf{y}_0) \\ u^o(\mathbf{y}_{J-1}) \end{pmatrix}.$$

In order for the scheme to reduce to the exponentially fitted Scharfetter–Gummel scheme in one dimension, the interpolation operator S for the potential V^o is chosen such that e^{qV} is expressed as $\frac{d}{dV} e^{qV}$. So S is chosen as

$$S(V^e)(\mathbf{y}_j) = \frac{1}{q} \ln\left(\frac{\exp[qV(\mathbf{x}_{j+1})] - \exp[qV(\mathbf{x}_j)]}{qV(\mathbf{x}_{j+1}) - qV(\mathbf{x}_j)} \right).$$

For a general band energy function $\varepsilon(k)$, the matrices A^{eo}, B^{eo}, C^e, C^o will have to be computed numerically. This can be done at a considerable computational expense, since it has to be done only once for a given set of basis functions. In particular, the collision matrices C^e, C^o have to be computed exercising some care so that the conservation and dissipation properties are not destroyed by the involved numerical integration. Generally, we have to compute integrals of the form $\int \phi_m(k) Q[e^{-\varepsilon} \phi_n(k)] dk$, where ϕ_m, ϕ_n denote either the even or the odd basis functions. Using the symmetry of the collision operator, we will compute the matrix elements of the collision operator as

$$C(m,n)$$

$$= \frac{1}{2} \sum_{\nu=-1}^{1} \int_{R_k^3 \times R_{k'}^3} s_\nu(k,k') \delta(\varepsilon(k) - \varepsilon(k') + \nu\omega)[\phi_m(k) - \phi_m(k')][\phi_n(k) - \phi_n(k')] dk dk'$$

and use numerical integration formulas for the integral in this form. Note that, in this form, it is apparent that the matrix C will be symmetric and that, if either ϕ_m or ϕ_n are constants, the matrix element will be zero, corresponding to conservation of charge and the Maxwellian kernel. Thus, we just have to make sure that numerical integration is applied in the same way in the k and k' variables to preserve these properties. The presence of the $\delta-$ functions in the integration kernel suggests writing the wave vector k in polar coordinates $k = r(\cos\alpha, \sin\alpha \sin\beta, \sin\alpha \cos\beta)$ giving

$$C(m,n)$$

$$= \frac{1}{2} \sum_{\nu=-1}^{1} \int s_\nu \delta(\varepsilon - \varepsilon' + \nu\omega)[\phi_m - \phi'_m][\phi_n - \phi'_n] r^2 \sin(\alpha)(r')^2 \sin(\alpha') dr d\alpha d\beta dr' d\alpha' d\beta',$$

where the $'$ denotes evaluation at r', α', β'. We now choose as basis functions spherical harmonic functions in the angular directions and polynomials in the radial direction $r = |k|$,

$$\phi_m(r,\alpha,\beta) = P_{m_1}(r) \Gamma_{m_2}(\alpha,\beta), \quad m = (m_1, m_2).$$

To take care of the $\delta-$ function in the integral kernel, we change variables in the integral from r to ε. Thus, we need to invert the band energy function ε as a function of the radius r for fixed angles α, β:

$$\varepsilon(r,\alpha,\beta) = u \quad \Longleftrightarrow \quad r = g(\varepsilon,\alpha,\beta). \tag{28}$$

In the transformed variables, the integration with respect to ε' can now be carried out exactly using the δ function. So the matrix element $C(m,n)$ now becomes

$$C(m,n) = \frac{1}{2}\sum_{\nu=-1}^{1}\int_0^\infty d\varepsilon \int d\alpha d\beta d\alpha' d\beta' [g^2(g')^2 \frac{dg}{d\varepsilon}\frac{dg'}{d\varepsilon'}\sin(\beta)\sin(\beta')] \tag{29}$$

$$\times\, \theta(\varepsilon+\nu\omega)s_\nu[P_{m_1}\Gamma_{m_2} - P'_{m_1}\Gamma'_{m_2}][P_{n_1}\Gamma_{n_2} - P'_{n_1}\Gamma'_{n_2}],$$

where the $'$ now denotes evaluation at $r = g(\varepsilon+\nu\omega,\alpha',\beta')$ and α',β', and θ is the Heaviside function. The structure of the scattering cross sections s_ν is usually sufficiently simple, such that the integrals with respect to the angular variables can be carried out exactly, leaving only the one-dimensional integral with respect to the energy to be evaluated numerically. The procedure outlined above is applicable in principle to an arbitrary band structure. It becomes significantly simpler if the band energy is actually radially symmetric, i.e., $\varepsilon = \varepsilon(|k|)$, since in this case the inverse band energy function g depends only on the energy. A popular choice is the dispersion relation due to Kane which is of the form (see [9])

$$\varepsilon(k) = \frac{|k|^2}{1+\sqrt{1+\kappa|k|^2}}.$$

We should point out that, if the band energy is actually not radially symmetric, the solution will actually depend on how the sample is aligned with the crystal direction, which might pose a considerable technological problem.

We now present results of a numerical test on the standard $n^+ - n - n^+$ silicon diode with a channel of $50nm$ length. This means that the doping concentration D^{dop} in (1) is given by a step function of the form

$$D^{dop}(x) = \begin{pmatrix} 10^{24}m^{-3} & 0 < x < 50nm \\ 10^{21}m^{-3} & 50nm < x < 100nm \\ 10^{24}m^{-3} & 100nm < x < 150nm \end{pmatrix}.$$

For the sake of simplicity, we restrict ourselves to parabolic band structures. So (in scaled variables) $\varepsilon(k) = \frac{|k|^2}{2}$ holds, which makes the computation of the inverse function $g(\varepsilon,\alpha,\beta)$ in (28) trivial. ($g(\varepsilon,\alpha,\beta) = \sqrt{2\varepsilon}$ holds and all integrations in (29) can be carried out exactly.) Also, since the problem in one spatial dimension admits solutions which are cylindrically symmetric around the k_1 direction (i.e., $f(x,k) = f(x_1,k_1,k_2^2+k_3^2)$), we choose only spherical harmonics with this symmetry as basis functions. For scattering cross sections we choose Fermi's golden rule formula

(30)

$$\text{(a)}\quad Q[f](x,k) = \frac{\hbar 2 V_L F(\xi)^2\pi}{\Omega}\sum_{\nu=\pm1}\left(N_\xi + \frac{1+\nu}{2}\right)\delta(\varepsilon(k)-\varepsilon(k')+\nu\hbar\omega)f(x,k')dk' - \frac{f(x,k)}{\tau(k)},$$

$$\text{(b)} \quad \frac{1}{\tau(k)} = \frac{2V_L F(\xi)^2 m_*^{3/2}}{\hbar^2 \Omega} 4\pi^2 \sum_{\nu=\pm 1} \left(N_\xi + \frac{1+\nu}{2} \right) \sqrt{2(\varepsilon(k) - \nu\hbar\omega)}^+,$$

$$\text{(c)} \quad \varepsilon(k) = \frac{\hbar^2 |k|^2}{2m_*}, \quad \Omega = \ln\left(\frac{N_\xi + 1}{N_\xi} \right),$$

where $F(\xi)$ denotes the frequency of the lattice state with momentum ξ and N_ξ denotes its occupation number. V_L denotes the volume of one lattice cell. $F(\xi)$ is given according to the formula

$$F(\xi) = \sqrt{\frac{q^2 \hbar\omega}{2V_L |\xi|^2 \varepsilon_0} \left(\frac{1}{\varepsilon_\infty} - \frac{1}{\varepsilon_s} \right)}.$$

The constants are given by the following expressions:

Symbol	Value	Unit	Meaning
q	$1.602 * 10^{-19}$	C	electron charge
$\hbar$	$6.626196 * 10^{-34}$	kgm^2/sec	Planck constant
m_*	$0.063 * 0.109 * 10^{-31}$	kg	effective electron mass
$\hbar\omega$	0.036	eV	emission/absorption energy
ε_0	$8.85 * 10^{-12}$	$\frac{C}{Vm}$	dielectricity constant (vacuum)
ε_∞	10.92	1	
ε_s	12.9	1	

We take into account only one single phonon momentum vector (corresponding to a $\delta-$ function collision potential; see [2], [4], [10]), which is evaluated at equilibrium, meaning $|\xi|^2 = m_* KT$ holds at room temperature. This gives, for an occupation number N_ξ corresponding to room temperature, numerical values of

$$\frac{\hbar 2V_L F(\xi)^2 \pi}{\Omega} N_\xi = 5.9356e * 10^5 \frac{m^3}{sec^2}, \quad \frac{2V_L F(\xi)^2 m_*^{3/2}}{\hbar^2 \Omega} N_\xi 4\pi^2 = 9.6044 * 10^{23} \frac{1}{m\sqrt{kg}}$$

for the constants necessary to evaluate (30). Finally, choosing computational units of $10^{-7} m$ for x and $10^7 m^{-1}$ for k, this gives a value of $\lambda = 0.3974$ for the dimensionless parameter λ in (1) with a scaled band energy $\varepsilon = \frac{|k|^2}{2}$.

The results below were obtained by varying the applied bias (the difference of the boundary potential V_b^e between the right and the left endpoint) form $-0.1V$ to $-0.5V$. A uniform mesh with 151 gridpoints and 128 expansion terms (16 in the energy direction and 8 in the angular direction) was used. The obtained figures did not change significantly when either doubling the number of expansion terms or halving the mesh size.

Figures 1–4 depict the potential as well as the physical densities for electrons, electron velocity, and electron energy as functions of the spatial variable x. Figure 1 shows the potential distribution $V(x)$. Figure 2 shows the electron density $\langle 1 \rangle(x) = \int f(x,k)dk = \rho^T F^e$. Figure 3 shows the velocity distribution $u(x) = \frac{\langle \partial_{k_1} \varepsilon \rangle(x)}{\langle 1 \rangle(x)}$ with $\langle \partial_{k_1} \varepsilon \rangle = \int \partial_{k_1} \varepsilon(k) f(x,k)dk$. Figure 4 shows the corresponding energy densities given by $w(x) = \frac{\langle \varepsilon \rangle(x)}{\langle 1 \rangle(x)}$.

Figures 5–8 depict the kinetic density $f(x,k)$ for the bias value of $0.4V$ for various values of x. Using spherical harmonic basis functions $\phi_m(k)$ gives the kinetic density function f in spherical coordinates as $f = f^{sph}(x, r, \cos(\alpha), \beta)$ with

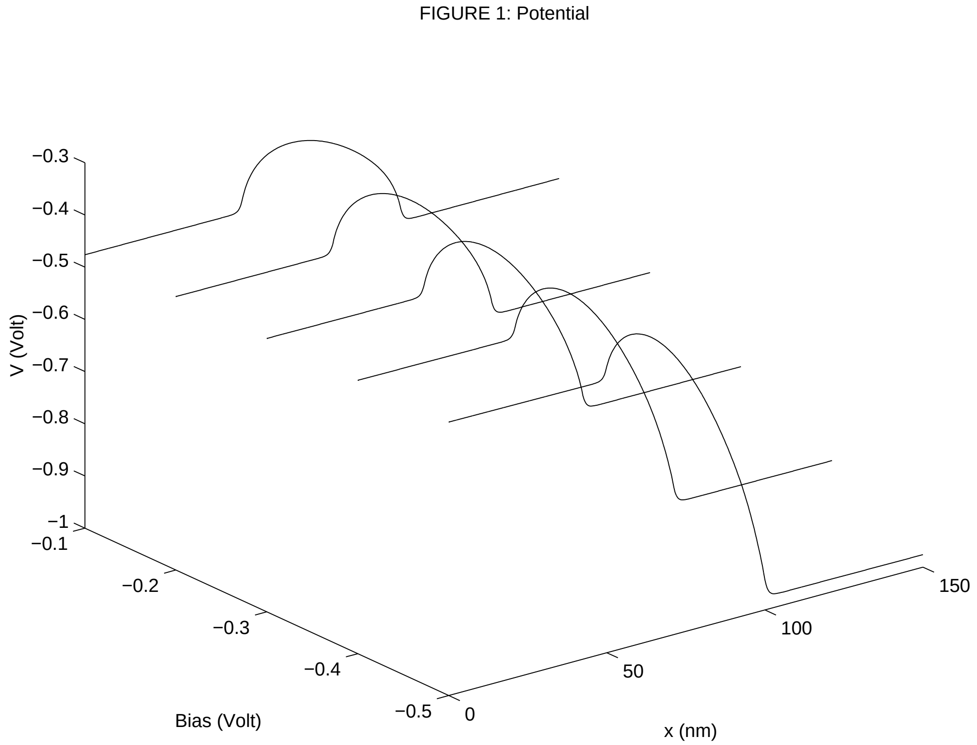


Fig. 1. *Potential.*

$k = (r\cos(\alpha), r\sin(\alpha)\sin(\beta), r\sin(\alpha)\cos(\beta))$, $\alpha \in [0,\pi]$, $\beta \in [-\pi,\pi]$. It is, however, more instructive to look at the kinetic density $f = f^{car}$ in Cartesian coordinates. In the case of one spatial dimension, the kinetic energy density will be cylindrically symmetrical around the k_1-axis. So $f^{car}(x,k) = f(x, k_1, \sqrt{k_2^2 + k_3^2})$ and $f^{sph}(x,k) = f(x, r, \cos(\alpha))$ hold. f^{car} and f^{sph} are related through $f^{car}(x, r\cos(\alpha), r\sin(\alpha)) = r^2 f^{sph}(x, r, \cos(\alpha))$, where the factor r^2 arises from the infinitesimal volume element due to the coordinate transformation. So Figures 5–8 show the function $r^2 f^{sph}(x, r, \cos(\alpha))$ for fixed values of x as a function of $k_1 = r\cos(\alpha)$ and $\sqrt{k_2^2 + k_3^2} = r\sin(\alpha)$, which is the density against which any function of Cartesian coordinates has to be integrated to compute expectations.

The purpose of these computations is to investigate how far away from the fluid dynamic regime we are. (For such a short channel and the given applied bias, we expect to see some distinctly nonequilibrium phenomena.) Figure 5 shows the distribution to the left of the channel. It is essentially given by a forward and a backward traveling Maxwellian of roughly the same amplitude. Figures 6 and 7 show the distribution at the beginning and the end of the channel. First we notice that the wave has developed a second peak in the forward as well as in the backward traveling component. Moreover, it has become definitely asymmetric in the k_1 direction at the end of the channel (in Figure 7). After the electron has left the channel in Figure 8, remnants of the second peak are still visible, but the solution has become symmetric in k_1 again. Figure 5 could have been produced by a hydrodynamic model or even by drift-diffusion equations, considering the low values of the group velocity in Figure 3 at this point. The second peak in Figure 8 (after leaving the channel) could have

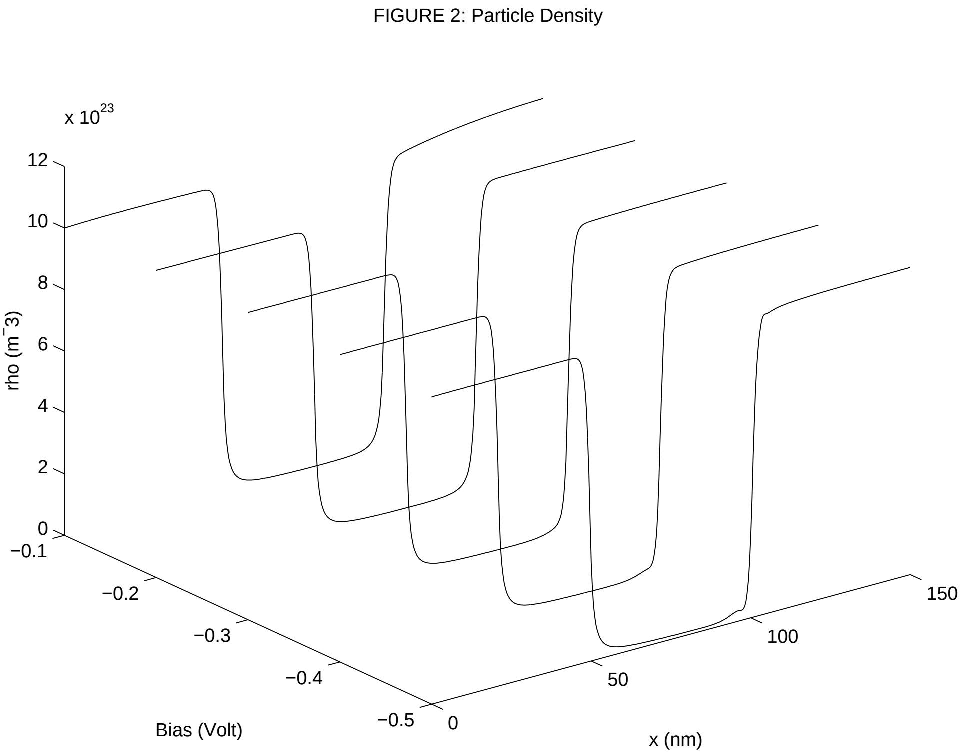


Fig. 2. *Particle density.*

been produced by a SHE-model. However, the strongly asymmetric distribution at the channel end (Figure 7) is a truly kinetic phenomenon and could not have been produced by either of the above approximations.

Concluding remarks. We conclude this section with some remarks concerning future work and the extension of this approach to two spatial dimensions. First, going to two spatial dimensions means a dramatic increase in the computational cost, since in addition to the added spatial dimension the cylindrical symmetry of the density function is lost. In the two-dimensional case, the density has to be taken of the form $f(x,k) = f(x_1, x_2, k_1, k_2, |k_3|)$, thus going effectively from a three- to a five-dimensional problem. The computations above were carried out using 16 expansion terms in the energy direction and 8 terms in the angular directions, so a system of 128 one-dimensional conservation laws was solved. Roughly estimating the cost in the two-dimensional case, allowing for a little less resolution and still exploiting the symmetry in the k_3 direction, this would translate into solving anywhere between 256 and 1024 two-dimensional conservation laws. On the other hand, we do not expect it to be necessary to solve the full kinetic problem in the whole device; i.e., for the simulation of a transistor this amount of resolution would only be necessary in a rather narrow region around the channel. Indeed, the promise of the expansion approach in two spatial dimensions is that it allows for a model hierarchy ranging from a Scharfetter–Gummel solution of the drift-diffusion equation (i.e., taking only three terms) to the full Boltzmann solution. Taking this into account, we estimate that a two-dimensional solution would involve around half a million variables. This raises, of course, the issue of the iterative solution of the involved linear system. In the one-

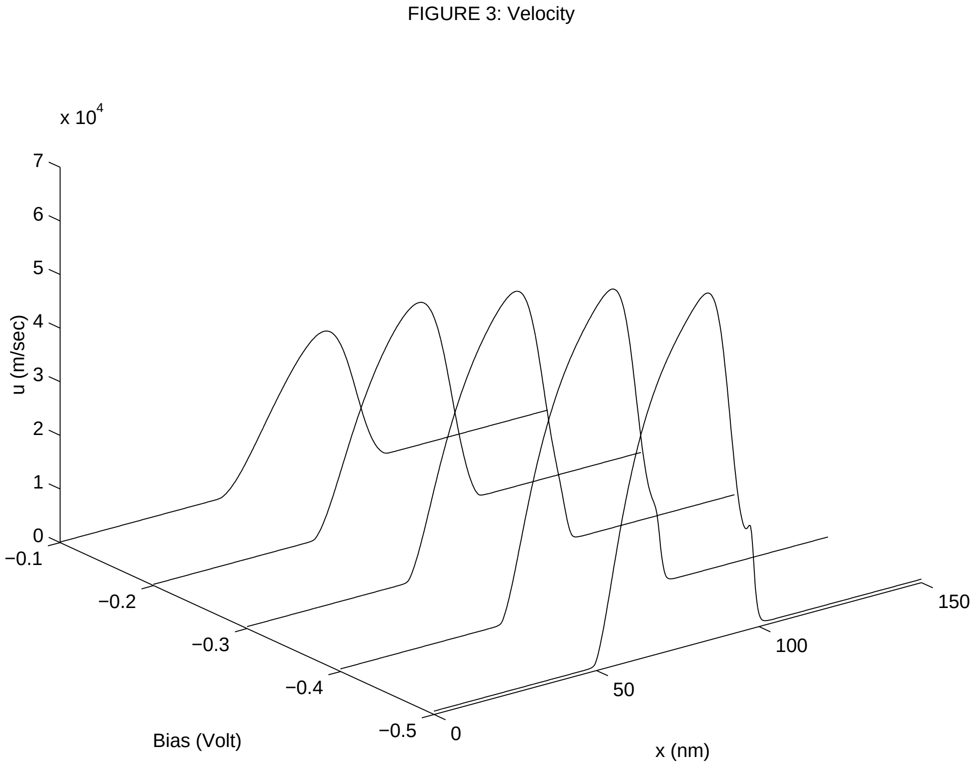


Fig. 3. *Velocity.*

dimensional case, a direct solution of the linear system was combined with a standard Gummel iteration procedure for the coupling to Poisson's equation. (See [19].) One possibility is to precondition the system by a solution of the drift-diffusion equations (i.e., by the the equations corresponding to the first three terms only). This would result in a pseudo time marching algorithm where the diffusive part of the operator is discretized implicitly in time and the hyperbolic part explicitly, as first proposed in [17] for the solution of the transient problem.

Finally, we would like to comment on the differences in the approaches to solving the steady state problem and the transient Boltzmann equation. There have been several approaches to deterministic solutions of the transient Boltzmann equation using methods designed for hyperbolic conservation laws. (See [1], [6], and references therein.) The transient solution (as well as the conservation law resulting from its Galerkin approximation in the wave vector direction) will exhibit a variety of hyperbolic and dispersive transient phenomena which would be unduly damped by the pseudo time marching algorithm suggested above. (See [15].) On the other hand, because of this fact, a pseudo time marching algorithm, based on the discretization presented here, will arrive at the correct steady state much faster, albeit with the wrong transient response, than a method designed for hyperbolic equations. So the method presented in this paper is really designed for the computation of steady states and for accurate deterministic simulations of the transient behavior it is probably preferable to use methods which take into account the hyperbolic and dispersive nature of the transient problem as given in [6] and [15].

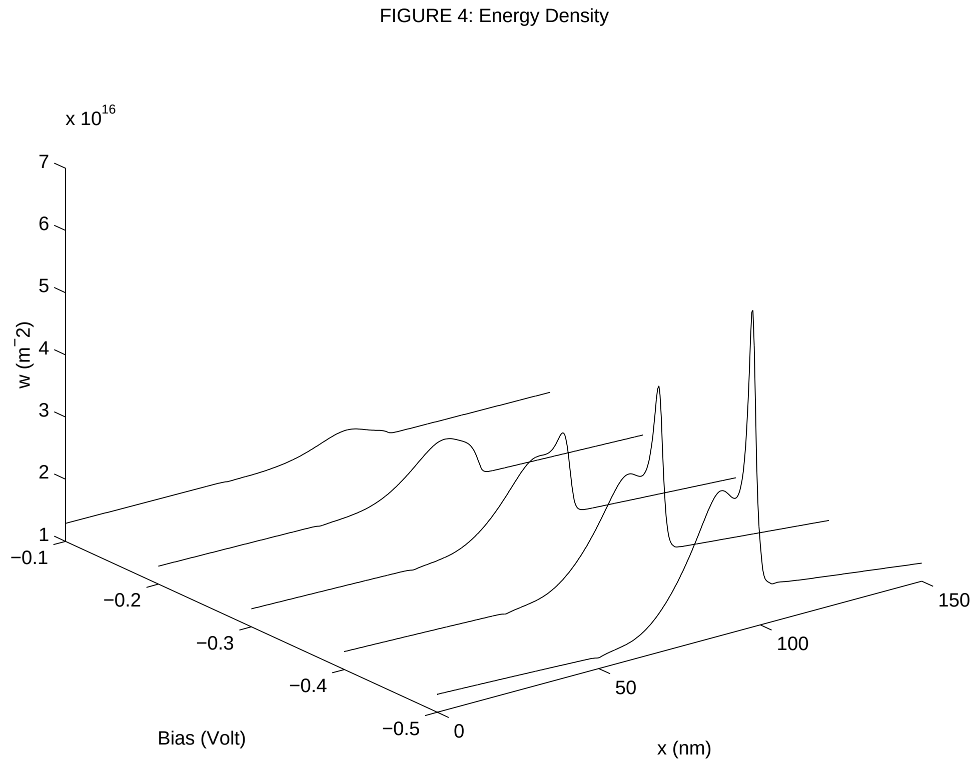


Fig. 4. *Energy density.*

6. Appendix.

Proof of Theorem 1. Setting $F_1^e = 0$, $F_1^o = 0$ in (26) gives the equations

$$\sum_{\nu=1}^{d}[D_\nu^{eo}(A_\nu^{10}F_0^o - qV^o B_\nu^{10}F_0^o)) + qV^e B_\nu^{10} D_\nu^{eo} F_0^o] = 0,$$

$$\sum_{\nu=1}^{d} e^{-qV^o}[(A_\nu^{01})^T D_\nu^{oe}(e^{qV^e}F_0^e)] + \frac{1}{\lambda}C_o^{10}F_0^o = 0,$$

which depend only on F_0^o and discrete derivatives of $e^{qV^e}F_0^e$. For $F_1^e = 0$, $F_1^o = 0$ these terms are given by (23) as the solution of

$$\sum_{\nu=1}^{d} D_\nu^{eo}(A_\nu^{00}F_0^o) = 0,$$

$$\sum_{\nu=1}^{d} e^{-qV^o}[(A_\nu^{00})^T D_\nu^{oe}(e^{qV^e}F_0^e)] + \frac{1}{\lambda}C_o^{00}F_0^o = 0,$$

$$-\sigma\sum_{\nu=1}^{d} D_\nu^{eo}D_\nu^{oe}V^e + q[D^{dop} - F_0^e] = 0,$$

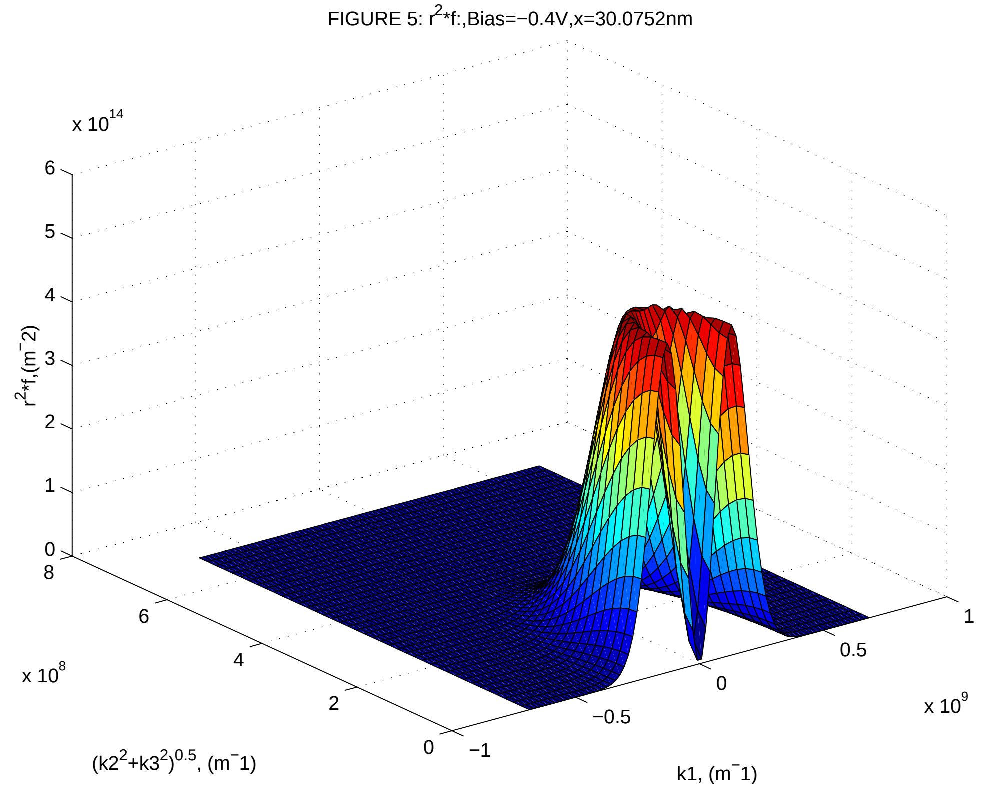


FIG. 5. $r^2 * f$: *Bias* $= -0.4V$, $x = 30.0752nm$.

and setting $F_0^o = 0$ and $F_0^e = e^{-qV}$ satisfies all of the above equations and the boundary conditions, except for the Poisson equation. Therefore V^e has to be chosen as the solution of

$$-\sigma \sum_{\nu=1}^{d} D_\nu^{eo} D_\nu^{oe} V^e + q[D^{dop} - e^{-qV^e}] = 0$$

together with the corresponding boundary conditions (23)(d) which corresponds to the equilibrium solution. $\square$

Proof of Theorem 2. The linearized problem (27) is given by the equations

$$\text{(31)} \qquad \text{(a)} \quad dL_1^{eo}(F_0^e, 0, 0, 0)(\delta F_0^e, \delta F_1^o, \delta F_0^o, \delta F_1^o) + \frac{1}{\lambda} C_e^{11} \delta F_1^e = R_1^e,$$

$$\text{(b)} \quad dL_1^{oe}(F_0^e, 0)(\delta F_0^e, \delta F_1^e) + \frac{1}{\lambda}[C_o^{10} \delta F_0^o + C_o^{11} \delta F_1^o] = R_1^o,$$

where the terms $\delta F_0^e, \delta F_0^o, \delta V^e, \delta V^o$ are given in terms of $\delta F_1^e, \delta F_1^o$ by linearizing the drift-diffusion–Poisson system (23):

$$\text{(32)} \qquad \text{(a)} \quad dL_0^{eo}(0, 0)(\delta F_0^o, \delta F_1^o) = 0,$$

$$\text{(b)} \quad dL_0^{oe}(F_0^e, 0)(\delta F_0^e, \delta F_1^e) + \frac{1}{\lambda}[C_o^{00} \delta F_0^o + C_o^{01} \delta F_1^o] = 0,$$

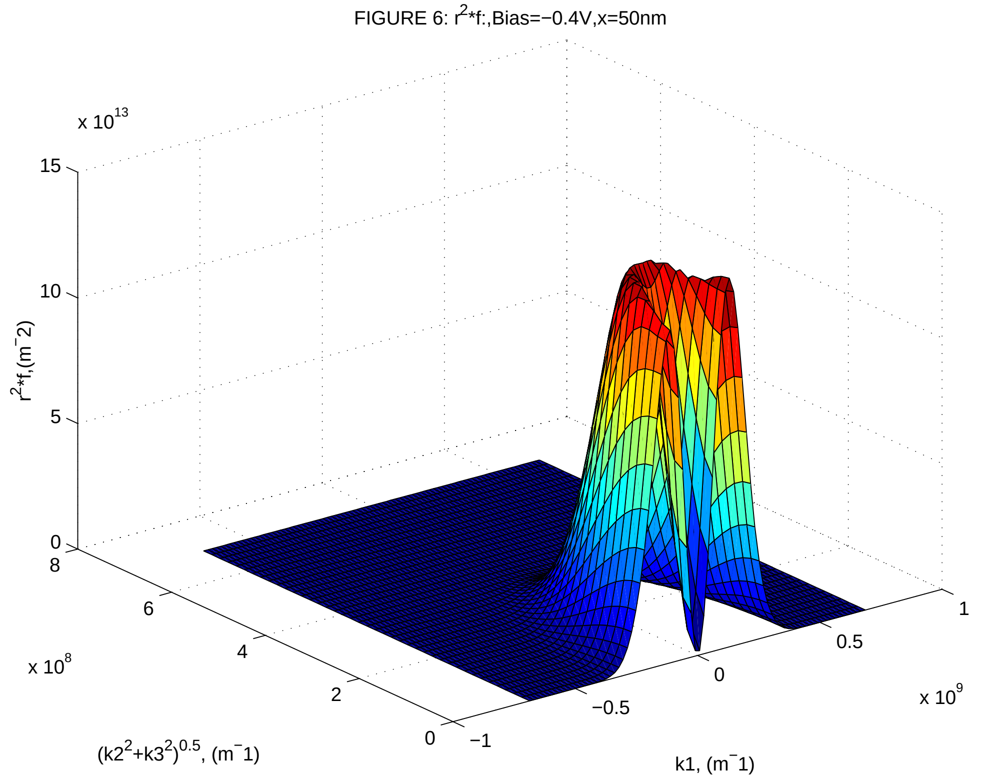


Fig. 6. $r^2 * f$: *Bias* $= -0.4V$, $x = 50nm$.

$$\text{(c)} \quad -\sigma \sum_{\nu=1}^{d} D_\nu^{eo} D_\nu^{oe} \delta V^e + q[D^{dop} - \delta F_0^e] = 0$$

together with homogeneous boundary conditions of the form (23)(d). Combining (31)(a–b) with (32)(a–b) we obtain the linearized system for the full equations

$$(33) \qquad \text{(a)} \quad dL^{eo}(F^e, 0)(\delta F^e, \delta F^o) + \frac{1}{\lambda} C^e \delta F^e = R^e,$$

$$\text{(b)} \quad dL^{oe}(F^e)(\delta F^e) + \frac{1}{\lambda} C^o \delta F^o = R^o,$$

with

$$\delta F^e = \begin{pmatrix} \delta F_0^e \\ \delta F_1^e \end{pmatrix}, \quad \delta F^o = \begin{pmatrix} \delta F_0^o \\ \delta F_1^o \end{pmatrix}, \quad F^e = \begin{pmatrix} F_0^e \\ 0 \end{pmatrix},$$

$$R^e = \begin{pmatrix} 0 \\ R_1^e \end{pmatrix}, \quad R^o = \begin{pmatrix} 0 \\ R_1^o \end{pmatrix}.$$

So we basically obtain the same linearized system as if we had linearized (21) directly, except for the crucial fact that the first component of R^e and the first three components of R^o are zero. Now, by virtue of construction of the discrete operators L^{eo}, L^{oe}

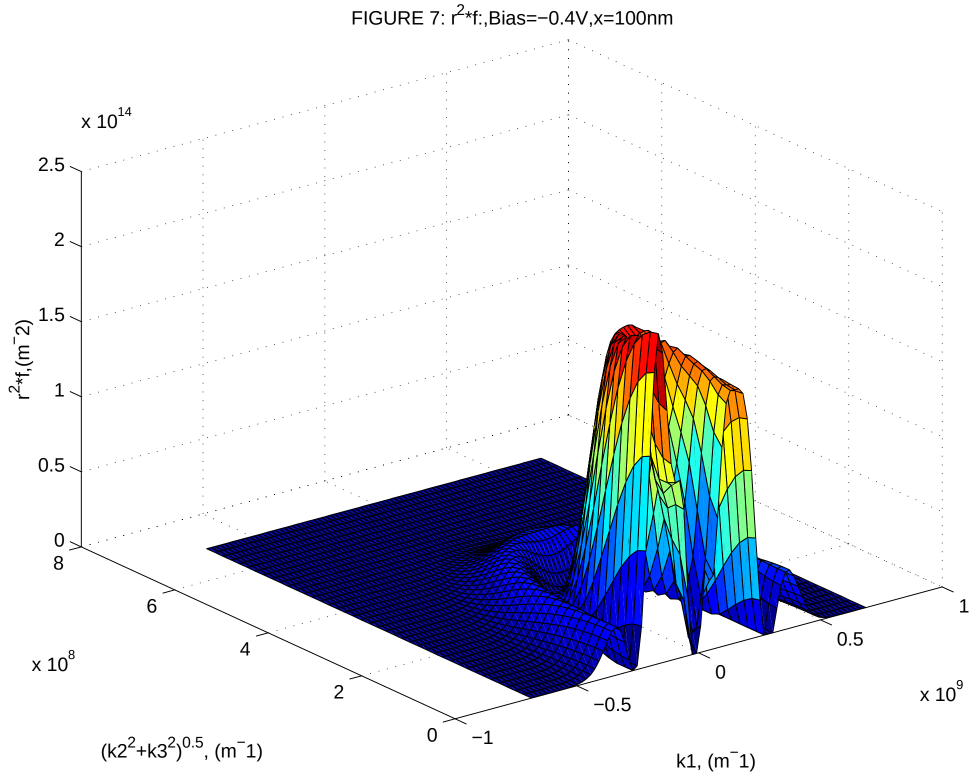


FIG. 7. $r^2 * f$: *Bias* $= -0.4V$, $x = 100nm$.

(see Proposition 1) and because of the structure of the boundary conditions, we have that

$$I_e[h^e(F^e)^T L^{eo}(F^e, F^o)] + I_o[h^o(F^e, F^o)^T L^{oe}(F^e)] = 0 \tag{34}$$

holds for any choice of F^e, F^o, where I_e, I_o denote the discrete integral operators defined in (16) and the entropies h^e, h^o are given by $h^e(F^e) = e^{qV^e} F^e$, $h^o(F^e, F^o) = e^{qV^o} F^o$ and, of course, the potentials V^e, V^o depend on F^e through the Poisson equation. Differentiating (34) twice functionally with respect to all of its arguments yields

$$\begin{aligned}
&I_e[(d^2h^e(F^e)(\delta F^e)^2)^T L^{eo}(F^e, F^o)] \\
&+ 2I_e[(dh^e(F^e)(\delta F^e))^T (dL^{eo}(F^e, F^o)(\delta F^e, \delta F^o))] \\
&+ I_e[h^e(F^e)(d^2L^{eo}(F^e, F^o)^T (\delta F^e, \delta F^o)^2)] \\
&+ I_o[(d^2h^o(F^e, F^o)(\delta F^e, \delta F^o)^2)^T L^{oe}(F^e)] \\
&+ 2I_o[(dh^o(F^e, F^o)(\delta F^e, \delta F^o))^T (dL^{oe}(F^e)(\delta F^e))] \\
&+ I_o[h^o(F^e, F^o)(d^2L^{oe}(F^e)^T (\delta F^e)^2)] = 0.
\end{aligned}$$

Inserting the equilibrium solution $F^o = 0$, $F^e = e^{-qV^e}\mathbf{e}$ into the above equations, we

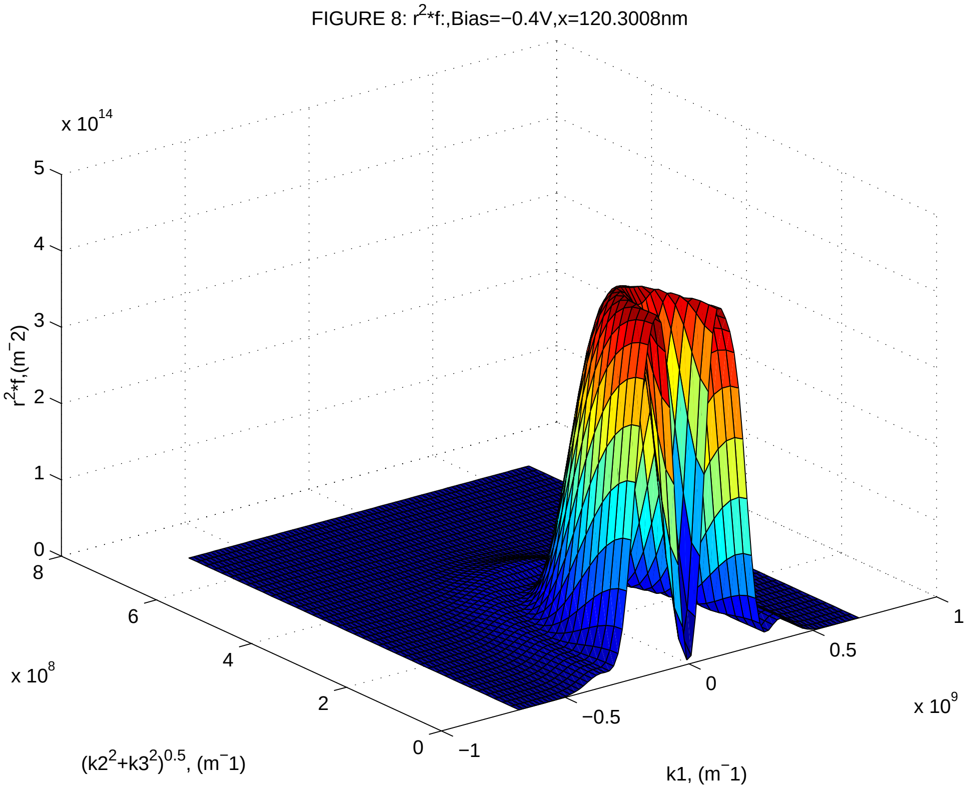


Fig. 8. $r^2 * f$: *Bias* $= -0.4V$, $x = 120.3008nm$.

see that

$$L^{eo}(F^e, F^o) = 0, \quad L^{oe}(F^e) = 0, \quad h^o(F^e, F^o) = 0, \quad h^e(F^e) = \mathbf{e}$$

holds, which eliminates three of the six terms in the above relations. Thus, we obtain

$$
\begin{aligned}
(35) \qquad & 2I_e[(dh^e(F^e)(\delta F^e))^T(dL^{eo}(F^e,F^o)(\delta F^e,\delta F^o))] \\
& + I_e[h^e(F^e)(d^2L^{eo}(F^e,F^o)^T(\delta F^e,\delta F^o)^2)] \\
& + 2I_o[(dh^o(F^e,F^o)(\delta F^e,\delta F^o))^T(dL^{oe}(F^e)(\delta F^e))] = 0,
\end{aligned}
$$

where in the second term of (35) only the first component of L^{eo} is left because of the form of h^e. However, from (25) it can be seen that L^{eo}_0 is a pure divergence operator, and therefore, because of the structure of the boundary conditions, the discrete integral over L^{eo}_0 vanishes for all arguments. Therefore the same is true for its second functional derivative, and we obtain the formula

$$
\begin{aligned}
& 2I_e[(dh^e(F^e)(\delta F^e))^T(dL^{eo}(F^e,F^o)(\delta F^e,\delta F^o))] \\
& + 2I_o[(dh^o(F^e,F^o)(\delta F^e,\delta F^o))^T(dL^{oe}(F^e)(\delta F^e))] = 0
\end{aligned}
$$

which we use to estimate the linearized equations (33). Multiplying (33)(a) by $(dh^e(F^e)(\delta F^e))^T$ and (33)(b) by $(dh^o(F^e,0)(\delta F^e,\delta F^o))^T$, and applying the discrete integral operators I_e, I_o, we obtain

$$\begin{aligned} &I_e[(dh^e(F^e)(\delta F^e))^T C^e \delta F^e] + I_o[(dh^o(F^e,0)(\delta F^e,\delta F^o))^T C^o \delta F^o] \\ &= \lambda I_e[(dh^e(F^e)(\delta F^e))^T R^e] + \lambda I_o[(dh^o(F^e,0)(\delta F^e,\delta F^o))^T R^o]. \end{aligned} \tag{36}$$

Computing the first Frechet derivatives of h^e, h^o at the equilibrium solution gives

$$dh^e(F^e)(\delta F^e) = q\delta V^e \mathbf{e} + e^{qV^e}\delta F^e, \quad dh^o(F^e,0)(\delta F^o,\delta V^o) = e^{qV^o}\delta F^o,$$

with $\delta V^e, \delta V^0$ the solutions of the linearized Poisson equation with homogeneous boundary conditions. Inserting this into (36), and using the fact that the first row and column of C^e vanishes, and that the first component of R^e, R^o is also zero, we get

$$\begin{aligned} &I_e[e^{qV^e}(\delta F_1^e)^T C_e^{11} \delta F_1^e] + I_o[e^{qV^o}(\delta F^o)^T C^o \delta F^o] \\ &= \lambda I_e[e^{qV^e}(\delta F_1^e)^T R_1^e] + \lambda I_o[e^{qV^o}(\delta F_1^o)^T R_1^o]. \end{aligned}$$

Now the matrices C_e^{11} and C^o are strictly positive definite. Using the Cauchy–Schwarz inequality gives the result. □

REFERENCES

[1] M. Anile, J. Carrillo, I. Gamba, and C. Shu, *Approximation of the BTE by a relaxation-time operator: Simulations for a* 50*nm-channel Si diode*, VLSI Design, to appear.

[2] P. Argyres, *Quantum kinetic equations for electrons in high electric and phonon fields*, Phys. Lett. A, 171 (1992).

[3] N. Ashcroft and M. Mermin, *Solid State Physics*, Holt-Saunders, New York, 1976.

[4] J. Barker and D. Ferry, Phys. Rev. Lett., 42 (1997).

[5] N. Ben Abdallah and J. Dolbeault, *Entropies relatives pour le système de Vlasov-Poisson dans des domaines bornés*, preprint archive, TMR-Project: Asymptotic methods in kinetic theory; also available online from http://www.math.tu-berlin.de/~tmr/(1999).

[6] J. Carrillo, I. Gamba, and C. Shu, *Computational macroscopic approximations to the* 1*-D relaxation-time kinetic system for semiconductors*, Phys. D, 146 (2000), pp. 289–306.

[7] C. Cercignani, *The Boltzmann Equation and Its Applications*, Appl. Math. Sci. 67, Springer-Verlag, New York, 1988.

[8] E. Fatemi and F. Odeh, *Upwind finite difference solution of Boltzmann equation applied to electron transport in semiconductor devices*, J. Comput. Phys., 108 (1993), pp. 209–217.

[9] M. Fischetti and S. Laux, *Monte Carlo analysis of electron transport in small semiconductor devices including band structure effects*, Phys. Rev. B, 38 (1988), pp. 9721–9745.

[10] F. Fromlet, P. Markowich, and C. Ringhofer, *A Wignerfunction approach to phonon scattering*, VLSI Design, 9 (1999), pp. 339–350.

[11] D. Levermore, *Moment closure hierarchies for kinetic theories*, J. Statist. Phys., 83 (1996), pp. 1021–1065.

[12] A. Majorana and R. Pidatella, *A finite difference scheme for solving the Boltzmann-Poisson system for semiconductor devices,* J. Comput. Phys., to appear.

[13] A. Majorana, *Spherical harmonics type expansion for the Boltzmann equation in semiconductor devices*, Le Matematice LIII, (1998), pp. 331–344.

[14] P. Markowich, C. Ringhofer, and C. Schmeiser, *Semiconductor Equations*, Springer-Verlag, Vienna, 1990.

[15] C. Ringhofer, *Space-time discretization of series expansion methods for the Boltzmann transport equation*, SIAM J. Numer. Anal., 38 (2000), pp. 442–465.

[16] C. Ringhofer, C. Schmeiser, and A. Zwirchmayr, *Moment methods for the semiconductor Boltzmann equation on bounded position domains*, SIAM J. Numer. Anal., 39 (2001), pp. 1078–1095.

[17] C. Schmeiser and A. Zwirchmayr, *Galerkin methods for the semiconductor Boltzmann equation*, in Proceedings of the Third International Congress on Industrial and Applied Mathematics, Hamburg, Germany, 1995.

[18] C. Schmeiser and A. Zwirchmayr, *Convergence of moment methods for linear kinetic equations*, SIAM J. Numer. Anal., 36 (1998), pp. 74–88.

[19] S. Selberherr, *Analysis of Semiconductor Devices*, 2nd ed., Wiley, New York, 1981.

SIAM J. NUMER. ANAL.
Vol. 41, No. 1, pp. 90–111

KRYLOV SUBSPACE ACCELERATION OF WAVEFORM RELAXATION[*]

ANDREW LUMSDAINE[†] AND DEYUN WU[‡]

Abstract. In this paper we describe and analyze Krylov subspace techniques for accelerating the convergence of waveform relaxation for solving time-dependent problems. A new class of accelerated waveform methods, convolution Krylov subspace methods, is presented. In particular, we give convolution variants of the CG algorithm and the GMRES algorithm and analyze their convergence behavior. We prove that the convolution Krylov subspace algorithms for initial value problems have the same convergence bounds as their linear algebra counterparts. Analytical examples are given to illustrate the operation of convolution Krylov subspace methods. Experimental results are presented which show the convergence behavior of traditional and convolution waveform methods applied to solving a linear initial value problem as well as the convergence behavior of static Krylov subspace methods applied to solving the associated linear algebraic equation.

Key words. convolution, dynamic iteration, Galerkin method, Krylov subspace methods, waveform relaxation

AMS subject classifications. 65L60, 65L05, 65R20, 65J10

PII. S0036142996313142

1. Introduction.

Dynamic iteration methods for initial value problems were first studied by Picard (1893) and Lindelöf (1894) in the context of existence and uniqueness of solutions to ODEs. In the early 1980's, dynamic iteration was reintroduced (with the name "waveform relaxation") as an efficient method for solving the large sparsely coupled differential equation systems generated by the simulation of integrated circuits [15, 43]. Since then, this method has been extended and applied to various other application areas [20, 24, 40]. Waveform relaxation continues to attract interest because of its natural medium-scale parallelism.

With the waveform approach, a dynamic system of equations is first decomposed spatially (i.e., at the equation level). Individual equations, or sets of equations taken together, are then solved iteratively by using values from previous iterates of other equations as input. Thus, the iterates are functions ("waveforms") rather than vectors.

Unfortunately, the convergence rate of standard waveform relaxation can be prohibitively slow for many problems of interest. As with relaxation-based approaches for linear algebra (e.g., Jacobi), application of appropriate acceleration is necessary to make the waveform approach practical. Previous approaches for accelerating the convergence of waveform relaxation include the shifted Picard iteration [34], multigrid [18, 41], SOR [23], Chebyshev acceleration [17], convolution SOR [30], $\mathbb{L}^2$ Krylov subspace methods [19], and adaptive window size selection [14].

Many of these waveform acceleration techniques are analogous to acceleration methods for iteratively solving linear systems of equations. However, in most cases, the generalizations of those approaches to waveform relaxation do not accelerate con-

[*] Received by the editors December 5, 1996; accepted for publication (in revised form) April 17, 2001; published electronically February 12, 2003. This work was supported in part by National Science Foundation grants CCR92-09815 and CCR95-02710.

http://www.siam.org/journals/sinum/41-1/31314.html

[†] Computer Science Department, Indiana University, Bloomington, IN 47405 (lums@cs.indiana.edu).

[‡] IdeaDavinci Lab, 1550 Alameda, Suite 303, San Jose, CA 95126 (deyun@ideadavinci.com).

vergence to the same degree as their linear algebra counterparts [23]. An analysis of why linear acceleration of waveform relaxation can, in general, be expected to be limited is given in [28].

One acceleration method for waveform relaxation that does, in fact, provide the same degree of acceleration as the analogous linear algebra method is convolution SOR, developed in [30, 31]. Inspired by convolution SOR, we use a convolution-based approach to develop an entirely new class of algorithms for accelerating the convergence of waveform relaxation, namely, convolution Krylov subspace methods. As particular exemplars of this new class of algorithms, we develop and analyze convolution GMRES (CGMRES) and biconvolution CG (BiCCG). Analysis of these methods shows that the convolution algorithms for linear differential equations and the corresponding algorithms for the associated linear algebraic equations have the same convergence rate bounds. In other words, the convolution Krylov subspace methods accelerate the convergence of waveform relaxation to the same degree as their linear algebra counterparts.

In the next two sections, we first review waveform relaxation and the $\mathbb{L}^2$ Krylov subspace techniques presented in [19]. The convolution Krylov subspace methods are then developed and analyzed. Experimental results comparing various waveform approaches are presented, and we conclude with a discussion and suggestions for further work.

2. Waveform relaxation. The mathematical description of waveform methods that we will be use throughout this paper is based on the model initial value problem:

$$\left\{ \begin{aligned} \tfrac{d}{dt}\boldsymbol{x}(t) + \boldsymbol{A}\boldsymbol{x}(t) &= \boldsymbol{f}(t), \\ \boldsymbol{x}(0) &= \boldsymbol{x}_0, \end{aligned} \right. \tag{2.1}$$

where $\boldsymbol{A} \in \mathbb{R}^{n\times n}$, $\boldsymbol{f}(t) \in \mathbb{R}^n$ is a given input, and $\boldsymbol{x}(t) \in \mathbb{R}^n$ is the unknown vector to be computed over an interval of interest $[0, T]$.

In (2.1), let $\boldsymbol{A} = \boldsymbol{M} - \boldsymbol{N}$ be a splitting of $\boldsymbol{A}$. The waveform relaxation algorithm based on this splitting is expressed in matrix form as follows.

ALGORITHM 1 (waveform relaxation for linear systems).

1. *Initialize:* Pick $\boldsymbol{x}^0$.
2. *Iterate:* For waveform iteration $k = 0, 1, \ldots$

   Solve

$$\left\{ \begin{aligned} \tfrac{d}{dt}\boldsymbol{x}^{k+1}(t) + \boldsymbol{M}\boldsymbol{x}^{k+1}(t) &= \boldsymbol{N}\boldsymbol{x}^k(t) + \boldsymbol{f}(t), \\ \boldsymbol{x}^{k+1}(0) &= \boldsymbol{x}_0 \end{aligned} \right.$$

   for $\boldsymbol{x}^{k+1}(t)$ on $[0, T]$.

Using operator notation, the waveform relaxation iteration can be expressed as

$$\boldsymbol{x}^{k+1} = \mathcal{K}\boldsymbol{x}^k + \boldsymbol{\psi}, \tag{2.2}$$

where the variables are defined on $\mathbb{L}^2([0, T], \mathbb{R}^n)$. The operator

$$\mathcal{K} : \mathbb{L}^2([0, T], \mathbb{R}^n) \to \mathbb{L}^2([0, T], \mathbb{R}^n)$$

is defined by

$$(\mathcal{K}\boldsymbol{x})(t) = \int_0^t e^{-\mathbf{M}(t-s)}\boldsymbol{N}\boldsymbol{x}(s)ds, \tag{2.3}$$

and $\boldsymbol{\psi} \in \mathbb{L}^2([0,T],\mathbb{R}^n)$ is given by

$$\boldsymbol{\psi}(t) = e^{-\mathbf{M}t}\boldsymbol{x}_0 + \int_0^t e^{-\mathbf{M}(t-s)}\boldsymbol{f}(s)ds.$$

It is obvious then that $\boldsymbol{x}$ will satisfy

$$(\boldsymbol{I} - \mathcal{K})\boldsymbol{x} = \boldsymbol{\psi}, \tag{2.4}$$

where $\boldsymbol{I}$ is the identity operator.

2.1. Properties. In this section we briefly review some relevant properties of the operator $\mathcal{K}$.

Remark. Associated with the initial value problem (2.1) is a linear algebraic problem

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}. \tag{2.5}$$

Similarly, associated with the waveform relaxation equation (2.4) is a preconditioned linear system of equations

$$(\boldsymbol{I} - \boldsymbol{M}^{-1}\boldsymbol{N})\boldsymbol{x} = \boldsymbol{M}^{-1}\boldsymbol{b}. \tag{2.6}$$

In what follows, we relate properties (in particular, spectral properties) of $\boldsymbol{M}^{-1}\boldsymbol{N}$ to properties of $\mathcal{K}$ and relate the behavior of algorithms applied to (2.4) to the behavior of algorithms applied to (2.6).

LEMMA 2.1. *The operator $\mathcal{K}$ as defined in* (2.3) *is compact, has zero spectral radius, and has adjoint operator $\mathcal{K}^*$ given by*

$$(\mathcal{K}^*\boldsymbol{x})(t) = \int_t^T \left[e^{-\mathbf{M}(s-t)}\boldsymbol{N}\right]^T \boldsymbol{x}(s)ds.$$

Remark. In general, $\mathcal{K}$ is not self-adjoint with respect to the $\mathbb{L}^2$ inner product, even when $\boldsymbol{M}^{-1}\boldsymbol{N}$, the matrix for the corresponding linear system, is symmetric in $\mathbb{R}^n$ (or Hermitian in $\mathbb{C}^n$).

Since $\mathcal{K}$ is compact with zero spectral radius, a straightforward convergence result can be stated.

THEOREM 2.2. *The waveform relaxation algorithm* (2.2) *generates a sequence of iterates $\{\boldsymbol{x}^k\}$ such that $\boldsymbol{x}^k \to \boldsymbol{x}$ as $k \to \infty$.*

Although $\mathcal{K}$ has zero spectral radius, it is highly nonnormal and thus the characteristics of the operator are far from trivial. That is, the spectrum itself provides very little insight into the behavior of iterative methods involving the operator $\mathcal{K}$. One approach to understanding iterative methods involving $\mathcal{K}$ is to consider the case for $T \to \infty$, in which case spectral properties of the operator apparently do provide information about the behavior of iterative methods involving $\mathcal{K}$. A detailed analysis of waveform relaxation for the $T \to \infty$ case is given in [23].

Unfortunately, the spectrum of $\mathcal{K}$ is discontinuous as a function of T—for any finite T, the spectral radius of $\mathcal{K}$ is zero. Thus, the degree to which the results for infinite T apply to real problems (which necessarily use finite T) is problem-dependent. One tool for understanding the behavior of $\mathcal{K}$ for finite T, and one that in some sense unifies the two cases of finite and infinite T, is pseudospectral analysis [38].

Definition. Let $\mathbb{X}$ be a Banach space with norm $\|\cdot\|$. The ϵ-pseudospectrum of a densely defined closed linear operator $\mathcal{A}:\mathbb{X}\to\mathbb{X}$ is defined as

$$\Lambda_\epsilon(\mathcal{A}) \equiv \left\{\lambda\in\mathbb{C} : \|(\lambda\boldsymbol{I}-\mathcal{A})^{-1}\| \geq \epsilon^{-1}\right\},$$

where it is understood that $\|(\lambda\boldsymbol{I}-\mathcal{A})^{-1}\| = \infty$ for $\lambda\in\Lambda(\mathcal{A})$. Here $\Lambda(\mathcal{A})$ is the spectrum of $\mathcal{A}$.

The following result (the proof is given in [22]) shows that the pseudospectrum is continuous as $T\to\infty$.

THEOREM 2.3. *Let $\mathcal{K}_T$ and $\mathcal{K}_\infty$ denote the operator $\mathcal{K}$ on $\mathbb{L}^2([0,T],\mathbb{R}^n)$ and $\mathbb{L}^2([0,\infty),\mathbb{R}^n)$, respectively. Then, for $\varepsilon>0$,*

$$cl \lim_{T\to\infty}\Lambda_\varepsilon(\mathcal{K}_T) = \Lambda_\varepsilon(\mathcal{K}_\infty).$$

2.2. Example. To aid our subsequent discussion, we provide a graphical illustration of the spectrum and pseudospectrum of $\mathcal{K}$ for the model problem (2.1). We take $\boldsymbol{A}$ to be a symmetric positive definite matrix:

$$\boldsymbol{A} = \frac{1}{\Delta x^2}\begin{bmatrix} 2 & -1 & & & 0 \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ 0 & & & -1 & 2 \end{bmatrix}_{n\times n}, \tag{2.7}$$

which is obtained, e.g., from discretizing the one-dimensional heat equation with spatial discretization Δx. For this example, we use a Jacobi splitting to obtain $\boldsymbol{M}$ and $\boldsymbol{N}$ from $\boldsymbol{A}$, and we take $\Delta x = 1/16$, $n=17$, and $\varepsilon = 10^{-3}$.

Figure 2.1 shows the spectra and pseudospectra of $\mathcal{K}_T$ and $\mathcal{K}_\infty$ for $\boldsymbol{A}$ given in (2.7), where $\mathcal{K}_T$ and $\mathcal{K}_\infty$, respectively, denote the operator $\mathcal{K}$ on $\mathbb{L}^2([0,T],\mathbb{R}^n)$ and $\mathbb{L}^2([0,\infty),\mathbb{R}^n)$. Since the matrix $\boldsymbol{M}^{-1}\boldsymbol{N}$ is normal, the pseudospectrum of $\mathcal{K}_\infty$

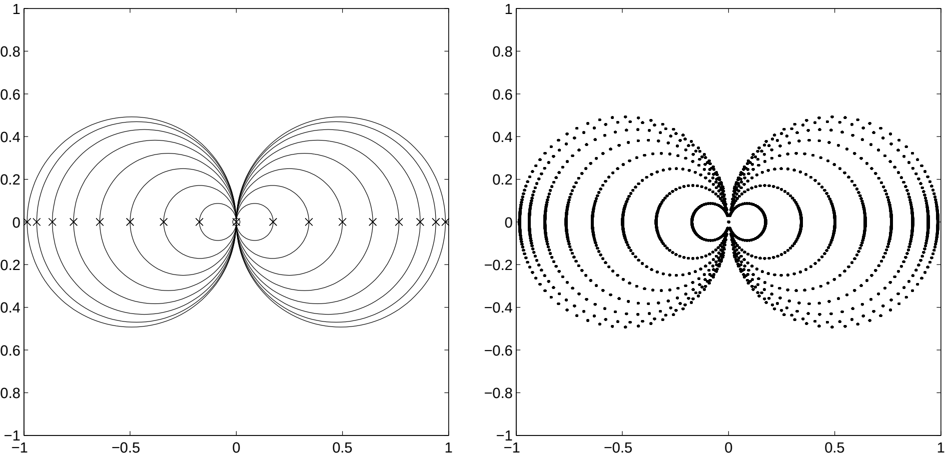


FIG. 2.1. *Spectrum and pseudospectrum of $\mathcal{K}_T$ and $\mathcal{K}_\infty$ for $\boldsymbol{A}$ given in* (2.7)*. In the left figure, $\Lambda(\mathcal{K}_\infty)$ is the union of the interiors of the circles shown, $\Lambda(\mathcal{K}_T)$ is the origin (indicated with* o*), and the eigenvalues of $\boldsymbol{M}^{-1}\boldsymbol{N}$ are indicated with* x*. In the right figure, $\Lambda_\varepsilon(\mathcal{K}_\infty)$ is the union of the interiors of the dotted circles.*

is also very close to the spectrum. The spectra and pseudospectra were plotted using the formulae given in [22]. Note that $\Lambda_\varepsilon(\mathcal{K}_T)$ does not have a known formula (only bounds are known), and so it is not plotted.

3. Hilbert space acceleration methods. For solving linear algebra problems, Krylov subspace algorithms form sequences of approximate solutions $\{\boldsymbol{x}^k\}$ with

$$\boldsymbol{x}^k = \boldsymbol{x}^0 + \sum_{i=0}^{k-1} \alpha_i \boldsymbol{A}^i \boldsymbol{r}^0,$$

where $\boldsymbol{x}^0$ is the initial estimate for $\boldsymbol{x}$ and $\boldsymbol{r}^k = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}^k$ is the residual associated with the kth iterate. That is, each $\boldsymbol{x}^k$ is a member of the affine Krylov subspace

$$\boldsymbol{x}^k \in \boldsymbol{x}^0 + \mathbb{K}^k(\boldsymbol{A}, \boldsymbol{r}^0) = \boldsymbol{x}^0 + \operatorname{span}\{\boldsymbol{r}^0, \boldsymbol{A}\boldsymbol{r}^0, \ldots, \boldsymbol{A}^{k-1}\boldsymbol{r}^0\} \subset \mathbb{R}^n.$$

Algorithms for generating $\{\boldsymbol{x}^k\}$ typically do so by enforcing some type of Galerkin or minimal residual condition on the iterates. To enforce these conditions, it is only necessary that the underlying space has certain geometric properties, namely, that a notion of orthogonality exists. This is usually taken to mean Hilbert space, but, as we will see, it seems that Hilbert space geometry may be too strong and that weaker geometric conditions can yield effective methods (see the discussion in section 7). Thus, Krylov subspace algorithms can readily be extended from $\mathbb{R}^n$ to Hilbert space (a fact that has been known since the early development of Krylov subspace iterative methods for linear algebra [10]). By embedding (2.4) into an appropriate Hilbert space, it is a rather straightforward matter to develop Krylov subspace acceleration techniques for waveform relaxation. A natural Hilbert space for this problem is $\mathbb{L}^2([0,T], \mathbb{R}^n)$.

3.1. Waveform GMRES. By Lemma 2.1, it is obvious that $\mathcal{K}$ is not self-adjoint with respect to the $\mathbb{L}^2$ inner product. Thus, in order to accelerate waveform relaxation, we must restrict our attention to those Krylov subspace algorithms suitable for non-Hermitian linear systems. At present, the premier such algorithm is the GMRES algorithm of Saad and Schultz [33].

The waveform GMRES (WGMRES) algorithm is as follows.

```
ALGORITHM 2 (WGMRES).
  1. Start: Set $\boldsymbol{r}^0 = \boldsymbol{\psi} - (\boldsymbol{I} - \mathcal{K})\boldsymbol{x}^0$, $\boldsymbol{v}^1 = \boldsymbol{r}^0/\|\boldsymbol{r}^0\|$, $\beta = \|\boldsymbol{r}^0\|$.
  2. Iterate: For $k = 1, 2, \ldots,$ until satisfied do:
             $h_{jk} = \langle(\boldsymbol{I} - \mathcal{K})\boldsymbol{v}^k, \boldsymbol{v}^j\rangle, \ j = 1, 2, \ldots, k$
       $\widetilde{\boldsymbol{v}}^{k+1} = (\boldsymbol{I} - \mathcal{K})\boldsymbol{v}^k - \sum_{j=1}^{k} h_{jk}\boldsymbol{v}^j$
         $h_{k+1,k} = \|\widetilde{\boldsymbol{v}}^{k+1}\|$
          $\boldsymbol{v}^{k+1} = \widetilde{\boldsymbol{v}}^{k+1}/h_{k+1,k}$.
  3. Form approximate solution:
     $\boldsymbol{x}^k = \boldsymbol{x}^0 + \boldsymbol{V}^k\boldsymbol{y}^k$, where $\boldsymbol{y}^k$ minimizes $\|\beta\boldsymbol{e}_1 - \bar{\boldsymbol{H}}^k\boldsymbol{y}^k\|$,
     $\boldsymbol{V}^k = (\boldsymbol{v}^1, \boldsymbol{v}^2, \ldots, \boldsymbol{v}^k)$, $\boldsymbol{e}_1 = (1, 0, \ldots, 0)^T$, and $\bar{\boldsymbol{H}}^k = (h_{ij})_{(k+1)\times k}$.
```

Remark. Symbolically, this algorithm is identical to the GMRES algorithm on $\mathbb{R}^n$. The difference is that the vectors and vector-space operations are understood to be defined on the Hilbert space $\mathbb{L}^2([0,T], \mathbb{R}^n)$ rather than on $\mathbb{R}^n$. That is, $\boldsymbol{x}^k$, $\boldsymbol{r}^k$, $\boldsymbol{\psi}$, $\boldsymbol{v}^k$, $\widetilde{\boldsymbol{v}}^k \in \mathbb{L}^2([0,T], \mathbb{R}^n)$, $h_{jk} \in \mathbb{C}$, and $\langle\cdot,\cdot\rangle$ is the $\mathbb{L}^2$ inner product.

3.2. Analysis of WGMRES. In order to analyze WGMRES, it is useful to recall that GMRES is a Galerkin method. The use of a Galerkin method over a Krylov space generated by $(\boldsymbol{I}-\mathcal{K})$ is discussed in [25] and [29], where the approach is called the method of moments (see also [42]). The following two results can be found in [19, 21].

THEOREM 3.1. *Let $\boldsymbol{X}$ be a Hilbert space and let $\mathcal{A} : \boldsymbol{X} \to \boldsymbol{X}$ be a bounded bijective linear operator. Let $\boldsymbol{X}^k \subset \boldsymbol{X}$ be a k-dimensional subspace with $\boldsymbol{X}^k \subset \boldsymbol{X}^{k+1}$ for all $k \in \mathbb{N}$. If $\boldsymbol{x}$ is in the closure of $\mathbb{S} = \bigcup_{k=1}^{\infty} \boldsymbol{X}^k$, then the Galerkin method for the operator equation $\mathcal{A}\boldsymbol{x} = \boldsymbol{f}$ converges.*

COROLLARY 3.2. *The Galerkin method for $(\boldsymbol{I}-\mathcal{K})\boldsymbol{x} = \boldsymbol{\psi}$ converges in the space $\mathbb{L}^2([0,T],\mathbb{R}^n)$, with finite dimensional subspaces $\mathbb{K}^m(\mathcal{K},\boldsymbol{\psi}) = \{\boldsymbol{\psi},\mathcal{K}\boldsymbol{\psi},\dots,\mathcal{K}^{m-1}\boldsymbol{\psi}\}$ for all $m \in \mathbb{N}$.*

As an immediate consequence, we have that WGMRES converges.

Unfortunately, because of the nonnormality of the operator $\mathcal{K}$, nothing much can be said about the rate of convergence, particularly in relationship to the spectrum of $\boldsymbol{M}^{-1}\boldsymbol{N}$ (or, more to the point, in relationship to the behavior of GMRES applied to solving a linear system of equations based on $\boldsymbol{M}^{-1}\boldsymbol{N}$).

Although precise statements about the convergence behavior of GMRES cannot be made, certain (rather pessimistic) qualitative statements can be made. Nevanlinna considered the general case of linear acceleration of waveform relaxation in [28], with the conclusion that significant speedups (of the sort one sees for linear algebra problems) would not be achievable in general.

Intuitively, we can see that it is much more difficult for GMRES to be effective when applied to waveform relaxation. Figure 2.1 shows the spectrum and pseudospectrum of $\mathcal{K}$. For the matrix problem, the spectrum of $\boldsymbol{M}^{-1}\boldsymbol{N}$ consists of a set of distinct eigenvalues on the real axis. For the waveform problem, the pseudospectrum of $\mathcal{K}$ fills a two-dimensional region in the complex plane. Clearly, it is much more difficult to find a good minimizing polynomial for the waveform case than for the linear algebra case. In fact, by using conformal mapping techniques, it is possible to show that there is, in fact, no essential speedup possible for WGMRES for a large class of problems [37] (on the infinite interval).

However, all is not lost. Similar pessimistic results hold for waveform relaxation accelerated with SOR [23]. However, with the use of convolution techniques, Reichelt developed a variant of SOR for waveform relaxation that does for (2.4) what algebraic SOR does for (2.6) [31]. That convolution techniques can provide the desired rate of acceleration for SOR gives us hope that similar results can be achieved for Krylov subspace techniques. We develop such a class of algorithms and prove their rates of convergence in the next section.

4. Convolution methods.

The principle behind convolution SOR (CSOR), and, indeed, the convolution methods developed in this paper, is that rather than simply taking linear combinations of waveform iterates, the methods take sums weighted by a convolution kernel. The resulting algorithms thus circumvent the limitations of linear acceleration as described in [23] and [28]. In fact, CSOR and the convolution Krylov subspace algorithms developed here exhibit speedup precisely comparable to that in the associated linear algebra problem.

4.1. CSOR.

A waveform relaxation algorithm using CSOR for solving (2.1) is shown in Algorithm 3. The algorithm takes an ordinary Gauss–Seidel waveform relaxation step to obtain a value for the intermediate variable $\widehat{x}_i^{k+1}$. The iterate x_i^{k+1} is obtained by adding a correction obtained by convolving $\widehat{x}_i^{k+1} - x_i^{k+1}$ with a kernel function $\omega(t)$. This is in contrast to simple waveform SOR in which x_i^{k+1} is obtained by multiplying $\widehat{x}_i^{k+1} - x_i^{k+1}$ with a scalar parameter ω. With the convolution, the CSOR method correctly accounts for the temporal frequency-dependence of the spectrum of the Jacobi waveform relaxation operator (e.g., Jacobi waveform relaxation smoothes high frequency components of the error waveform more rapidly than low frequency components) by, in effect, using a different SOR parameter for each frequency [31].

Algorithm 3 (Gauss–Seidel waveform relaxation with CSOR acceleration).

1. *Initialize:* Pick vector waveform $\boldsymbol{x}^0(t) \in C^1([0,T],\mathbb{R}^n)$ with $\boldsymbol{x}^0(0) = \boldsymbol{x}_0$.
2. *Iterate:* For $k = 0, 1, \ldots,$ until converged,
   - *Solve* for scalar waveform $\widehat{x}_i^{k+1}(t) \in C^1([0,T],\mathbb{R})$ with $\widehat{x}_i^{k+1}(0) = x_{0i}$,

$$\left(\tfrac{d}{dt} + a_{ii}\right)\widehat{x}_i^{k+1}(t) = f_i(t) - \sum_{j=1}^{i-1} a_{ij} x_i^{k+1}(t) - \sum_{j=i+1}^{n} a_{ij} x_i^{k}(t).$$

   - *Overrelax* to generate $x_i^{k+1}(t) \in C^1([0,T],\mathbb{R})$,

$$x_i^{k+1}(t) = x_i^k(t) + \int_0^t \omega(\tau) \cdot \left[\widehat{x}_i^{k+1}(t-\tau) - x_i^k(t-\tau)\right] d\tau. \tag{4.1}$$

4.2. Convolution Krylov subspace algorithms. In this section and the next, we incorporate convolution into the Krylov subspace approach for accelerating waveform relaxation. We begin by identifying some key spaces and associated operations to be used in what follows.

Assume $f, g \in \mathbb{L}^2(\mathbb{R},\mathbb{R})$ are functions, and $\boldsymbol{x} = (x_1, \ldots, x_n)^T,\ \boldsymbol{y} = (y_1, \ldots, y_n)^T \in \mathbb{L}^2(\mathbb{R},\mathbb{R}^n)$ are vectors of functions. Define

$$\begin{aligned}
(f \star g)(t) &= \int_{-\infty}^{\infty} f(s) g(t-s) ds \in \mathbb{L}^1(\mathbb{R},\mathbb{R}),\\
(f \star \boldsymbol{x})(t) &= ((f \star x_1)(t) \ldots (f \star x_n)(t))^T \in \mathbb{L}^1(\mathbb{R},\mathbb{R}^n),\\
\langle \boldsymbol{x}, \boldsymbol{y} \rangle_\star(t) &= \sum_{i=1}^{n} (x_i \star \widetilde{y_i})(t) \in \mathbb{L}^1(\mathbb{R},\mathbb{R}),
\end{aligned}$$

where $\widetilde{f}(t) = f(-t)$, and $\widetilde{\boldsymbol{x}}(t) = \boldsymbol{x}(-t)$.

Remark. Although the above formulae are defined in general for $\mathbb{L}^2$ functions, one important subset of $\mathbb{L}^2$ that will figure prominently is $C_0([0,\infty),\mathbb{R})$, the compactly supported continuous functions, and we will be working with this space in what follows. The following technique is purely for the analysis of the algorithms, although the implementational computations are not related to it. For brevity we will indicate $\mathcal{K}_T$ by $\mathcal{K}$.

From the basic operational calculus [26, 27, 36], it is known that convolution induces a ring structure on $C_0([0,\infty),\mathbb{R})$. As with any ring structure, this ring structure, can be algebraically extended to a quotient field (in a manner similar to extending the ring of integers to the field of rational numbers). Define $\mathbb{Q}$ to be the set of ordered pairs ("fractions")

$$\mathbb{Q} = \{f/g : f, g \in C_0([0,\infty),\mathbb{R})\}.$$

By a result of Titchmarsh [36], there are no zero divisors; i.e., $f \star g = 0$ implies that either $f = 0$ or $g = 0$. The axiomatic operations required for field structure are then readily defined. An operational calculus has been developed on this basis, adding a theoretical underpinning to the operational calculus of Heaviside [26, 27].

A result of Foiaş shows that the range of convolution is dense in $\mathbb{L}^1$, a result that was later extended to the continuous case [7, 35]. More recently, a constructive proof of these results has been given by Bäumer [2]. As a direct consequence of the injectivity and dense range properties of convolution, we have the following.

LEMMA 4.1. *Let $\mathbb{Q}$ be the quotient field induced by the convolution ring $\{C_0([0,\infty),\mathbb{R}),\star\}$. For any element $f/g \in \mathbb{Q}$, there is a sequence $\{\phi_i\}$ with $\phi_i \in C_0([0,\infty),\mathbb{R})$ such that*

$$\lim_{i\to\infty} \phi_i \star g = f.$$

Remark. $\mathbb{Q}$ is, in fact, a space of generalized functions; the limit $\lim_{i\to\infty}\phi_i$ may not necessarily exist in $C_0([0,\infty),\mathbb{R})$. For example, the element in $\mathbb{Q}$ identified with f/f is the Dirac δ-distribution. In this regard, $\mathbb{Q}$ is a completion of $C_0([0,\infty),\mathbb{R})$ with respect to the convolution operator. Notice that a function $f \in C_0([0,\infty),\mathbb{R})$ can be identified with any quotient of the form $(f\star g)/g$, $g \neq 0$; i.e., $f \in C_0([0,\infty),\mathbb{R})$ implies $f \in \mathbb{Q}$.

Using this definition of convolution between elements of $\mathbb{Q}$ and continuous functions, we can readily create a vector space over the field $\mathbb{Q}$ using convolution.

PROPOSITION 4.2. *Let $\mathbb{Q}$ be the quotient field induced by the convolution ring $\{C_0([0,\infty),\mathbb{R}),\star\}$. For $q \in \mathbb{Q}$ and $\boldsymbol{x} \in C_0([0,\infty),\mathbb{R}^n)$, let $\phi_i \to q$ in the sense of Lemma* 4.1. *With the convolution operation defined by*

$$q \star \boldsymbol{x} = \lim_{i\to\infty} \phi_i \star \boldsymbol{x},$$

$C_0([0,\infty),\mathbb{R}^n)$ forms a vector space over $\mathbb{Q}$.

Definition. Assume $\mathcal{A}$ is a bounded linear operator defined on $C_0([0,\infty),\mathbb{R}^n)$ and $\boldsymbol{r}^0 \in C_0([0,\infty),\mathbb{R}^n)$ is fixed. The m-dimensional convolution Krylov subspace generated by $\mathcal{A}$ and $\boldsymbol{r}^0$ is defined to be

$$\begin{aligned}\mathbb{K}_\star^m(\mathcal{A},\boldsymbol{r}^0) &= \operatorname{span}_\star\{\boldsymbol{r}^0, \mathcal{A}\boldsymbol{r}^0, \ldots, \mathcal{A}^{m-1}\boldsymbol{r}^0\}\\ &= \left\{\sum_{i=0}^{m-1} \alpha_i \star \mathcal{A}^i\boldsymbol{r}^0 : \alpha_i \in \mathbb{Q}\right\} \subset C_0([0,\infty),\mathbb{R}^n).\end{aligned}$$

Notice that this definition of the convolution Krylov subspace differs from the usual definition of a Krylov subspace due to the convolution operator; by $\operatorname{span}_\star$ we mean combinations under convolution. However, as with methods that use the traditional definition of a Krylov subspace, we seek to find the element of $\mathbb{K}_\star^m(\mathcal{A},\boldsymbol{r}^0)$ that best satisfies (2.4).

4.3. Examples.

Using convolutions in Krylov subspace methods may seem counterintuitive. To demonstrate the general operation of convolution Krylov subspace methods, we present two examples.

Example. Consider the following initial value problem:

$$\left\{\begin{aligned} \tfrac{d}{dt}\boldsymbol{x}(t) + \boldsymbol{A}\boldsymbol{x}(t) &= \boldsymbol{0},\\ \boldsymbol{x}(0) &= (0,1)^T, \end{aligned}\right. \tag{4.2}$$

where $\boldsymbol{A}$ is a 2×2 matrix

$$\boldsymbol{A} = \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}.$$

For splitting $\boldsymbol{A} = \boldsymbol{M} - \boldsymbol{N}$ with $\boldsymbol{M} = 2\boldsymbol{I}$, in order to solve $(\boldsymbol{I} - \mathcal{K})\boldsymbol{x} = \boldsymbol{\psi}$, we choose

$$\boldsymbol{x} = \alpha_0 \star \boldsymbol{r}^0 + \alpha_1 \star (\boldsymbol{I} - \mathcal{K})\boldsymbol{r}^0.$$

By the Galerkin conditions,

$$\langle(\boldsymbol{I} - \mathcal{K})\boldsymbol{x}, \boldsymbol{r}^0\rangle_\star = \langle\boldsymbol{\psi}, \boldsymbol{r}^0\rangle_\star,$$

$$\langle(\boldsymbol{I} - \mathcal{K})\boldsymbol{x}, (\boldsymbol{I} - \mathcal{K})\boldsymbol{r}^0\rangle_\star = \langle\boldsymbol{\psi}, (\boldsymbol{I} - \mathcal{K})\boldsymbol{r}^0\rangle_\star,$$

we can find

$$\begin{cases} \alpha_0 = \frac{2}{3}[e^{-2t} \star (-2 + e^{-2t})]/[(e^{-2t} - 1) \star (e^{-2t} - 1)], \\ \alpha_1 = -\frac{2}{3}[-1 + 2e^{-2t})]/[(e^{-2t} - 1) \star (e^{-2t} - 1)] \end{cases}$$

and

$$\begin{cases} x_1 = -(te^{-2t})/(te^{-2t} - 1) = e^{-2t}\sinh t, \\ x_2 = -(e^{-2t})/(te^{-2t} - 1) = e^{-2t}\cosh t, \end{cases}$$

which is the analytic solution for (4.2).

Example. Let us consider another initial value problem:

$$\begin{cases} \frac{d}{dt}\boldsymbol{x}(t) + \boldsymbol{A}\boldsymbol{x}(t) = \boldsymbol{0}, \\ \qquad\qquad \boldsymbol{x}(0) = (0, 1)^T, \end{cases} \tag{4.3}$$

where $\boldsymbol{A}$ is a 2×2 matrix

$$\boldsymbol{A} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}.$$

In this case, we take splitting $\boldsymbol{A} = \boldsymbol{M} - \boldsymbol{N}$ with $\boldsymbol{M} = 0$. Again choose

$$\boldsymbol{x} = \alpha_0 \star \boldsymbol{r}^0 + \alpha_1 \star (\boldsymbol{I} - \mathcal{K})\boldsymbol{r}^0.$$

As in the last example, by using the Galerkin conditions, we can find

$$\begin{cases} \alpha_0 = -(6t - t^3)/(6t + t^3), \\ \alpha_1 = 6/(6t + t^3) \end{cases}$$

and

$$\begin{cases} x_1 = t^3/(6t + t^3) = \sin t, \\ x_2 = (3t^2)/(6t + t^3) = \cos t, \end{cases}$$

which is the analytic solution for (4.3).

4.4. Fourier transform. Define the Fourier–Laplace transform on $\mathbb{L}^2(\mathbb{R},\mathbb{R})$ to be

$$\widehat{f}(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-izx} f(x)dx \quad \text{for } z \in \mathbb{C}.$$

It is well known that $\|\widehat{f}\|_{\mathbb{L}^2(\mathbb{R},\mathbb{R})} = \|f\|_{\mathbb{L}^2(\mathbb{R},\mathbb{R})}$.

Remarks.

1. $\langle \boldsymbol{x}, \boldsymbol{y}\rangle_\star\widehat{\ }(\xi) = \langle \widehat{\boldsymbol{x}}(\xi), \widehat{\boldsymbol{y}}(\xi)\rangle$, where $\langle\cdot,\cdot\rangle$ is the Hermitian inner product in $\mathbb{C}^n$, and $\xi \in \mathbb{R}$.
2. $\langle \boldsymbol{x}, \boldsymbol{y}\rangle_\star = \widetilde{\langle \boldsymbol{y}, \boldsymbol{x}\rangle_\star} = \langle \widetilde{\boldsymbol{y}}, \widetilde{\boldsymbol{x}}\rangle_\star$.
3. For $f, g \in C_0([0,\infty),\mathbb{R})$ $(f/g)\widehat{\ } = \frac{\widehat{f}}{\widehat{g}}$ is a meromorphic function with discrete poles [1]. (Here $\frac{\cdot}{\cdot}$ indicates normal division.)
4. Under the convolution inner product "$\langle\cdot,\cdot\rangle_\star$," the convolution adjoint operator of $\mathcal{K}$ is the same as the $\mathbb{L}^2$ adjoint operator $\mathcal{K}^*$ of $\mathcal{K}$. Thus, even with real symmetric $\boldsymbol{A}$, $\mathcal{K}$ is not self-adjoint with respect to $\langle\cdot,\cdot\rangle_\star$.

Since convolution is closely related to the Fourier transform (which is isometric on the $\mathbb{L}^2(\mathbb{R},\mathbb{R})$ space), we restrict our analysis to $\mathbb{L}^2$ spaces. In this context, we view $C_0([0,\infty),\mathbb{R}^n)$ as a subspace of $\mathbb{L}^2(\mathbb{R},\mathbb{R}^n)$ by extending elements in $C_0([0,\infty),\mathbb{R}^n)$ trivially on $(-\infty,0)$. Hence, for any $f \in C_0([0,\infty),\mathbb{R})$ we have $\|\widehat{f}\| = \|f\|$ in the $\mathbb{L}^2(\mathbb{R},\mathbb{R})$ norm.

4.5. The convolution GMRES algorithm. In this section, we introduce the convolution GMRES (CGMRES) algorithm. Analogous to GMRES for linear systems of equations, CGMRES is appropriate for operator systems where $\mathcal{A}$ is not self-adjoint with respect to the convolution inner product.

ALGORITHM 4 (CGMRES). Let $\mathcal{A} : C_0([0,\infty),\mathbb{R}^n) \to \overline{C_0([0,\infty),\mathbb{R}^n)}$ be a bounded linear operator. By Bäumer [2], $\mathcal{A}$ is extendable to $\overline{C_0([0,\infty),\mathbb{R}^n)}$, the vector-valued generalized function space, i.e., the completion of $C_0([0,\infty),\mathbb{R}^n)$ under operator $\mathcal{A}$, which is a vector space over $\mathbb{Q}$ due to the commutative of $\mathcal{A}$ and $\star$. Let $\boldsymbol{f} \in C_0([0,\infty),\mathbb{R}^n)$.

1. Pick $\boldsymbol{x}^0 \in C_0([0,\infty),\mathbb{R}^n)$ and compute $\boldsymbol{r}^0 = \boldsymbol{f} - \mathcal{A}\boldsymbol{x}^0,\ \beta = |\widehat{\boldsymbol{r}^0}|^\vee, \boldsymbol{v}^1 = \boldsymbol{r}^0/\beta$.
2. For $j = 1, \ldots,$ until converged,

$$\begin{aligned}
\boldsymbol{w}^j &= \mathcal{A}\boldsymbol{v}^j \\
h_{ij} &= \langle \boldsymbol{w}^j, \boldsymbol{v}^i\rangle_\star, i = 1,\ldots,j \\
\boldsymbol{w}^j &= \boldsymbol{w}^j - \sum_{i\le j} h_{ij} \star \boldsymbol{v}^i \\
h_{j+1,j} &= |\widehat{\boldsymbol{w}^j}|^\vee, \ \text{ if } h_{j+1,j} \equiv 0, \text{ set } m = j \text{ and go to step 3} \\
\boldsymbol{v}^{j+1} &= \boldsymbol{w}^j/h_{j+1,j}.
\end{aligned}$$

3. Compute $\boldsymbol{y}^m$, the minimizer of $\langle(\beta\star(\delta_0,0,\ldots,0)^T - \bar{\boldsymbol{H}}^m\star\boldsymbol{y}), (\beta\star(\delta_0,0,\ldots,0)^T - \bar{\boldsymbol{H}}^m \star \boldsymbol{y})\rangle_\star$ and $\boldsymbol{x}^m = \boldsymbol{x}^0 + \boldsymbol{V}^m \star \boldsymbol{y}^m$, where $\bar{\boldsymbol{H}}^m = (h_{ij})_{(m+1)\times m}$, $\boldsymbol{V}^m = (\boldsymbol{v}^1, \boldsymbol{v}^2, \ldots, \boldsymbol{v}^m)$, and δ_0 is the Dirac δ-distribution.

Remarks.

1. In the above algorithms, "$|\cdot|$" is the Euclidean norm defined in $\mathbb{C}^n$, and "$\vee$" means the inverse Fourier transform.
2. By $\boldsymbol{x}/f$ for $\boldsymbol{x} \in C_0([0,\infty),\mathbb{R}^n)$ and $f \in C_0([0,\infty),\mathbb{R})$ we mean the vector $(x_1/f, x_2/f, \ldots, x_n/f)^T \in \mathbb{Q}^n$.

3. It is not hard to see that there is a compactly supported sequence $\beta_i \in C_0(-\infty,\infty)$ such that $\beta_i \to \beta \in \mathbb{L}^2(\mathbb{R})$. Rewrite $\boldsymbol{v}^1 = \boldsymbol{r}^0/\beta = (\boldsymbol{r}^0 \star h)/(\beta \star h)$, where $h \neq 0$ is any function in $C_0([0,\infty),\mathbb{R})$. Now $\beta_i \star h \in C_0([0,\infty),\mathbb{R})$. We view $\mathcal{A}\boldsymbol{v}^1 = \lim_{i\to\infty}(\mathcal{A}\boldsymbol{r}^0 \star h)/(\beta_i \star h)$ which is well-defined on the completion vector space $\overline{C_0([0,\infty),\mathbb{R}^n)}$ over $\mathbb{Q}$.

4.6. Convergence of CGMRES. To analyze the convergence of the CGMRES algorithm we begin with the convolution Petrov–Galerkin conditions

$$\langle \boldsymbol{f} - \mathcal{A}\boldsymbol{x}, \boldsymbol{w}\rangle_\star = 0 \quad \forall \boldsymbol{w} \in \mathcal{A}\mathbb{K},$$

where for brevity $\mathbb{K}$ indicates $\mathbb{K}_\star^m(\mathcal{A}, \boldsymbol{r}^0)$. Based on these conditions, we say that $\boldsymbol{f} - \mathcal{A}\boldsymbol{x}$ is *convolutionally perpendicular* to $\mathcal{A}\mathbb{K}$. Note that $\langle \boldsymbol{x}, \boldsymbol{y}\rangle_\star = 0$ if and only if $\langle \hat{\boldsymbol{x}}(\xi), \hat{\boldsymbol{y}}(\xi)\rangle_\star = 0$ for all $\xi \in \mathbb{R}$.

Therefore, after taking Fourier transforms, the CGMRES algorithm becomes the GMRES algorithm for the linear algebraic equation

$$\boldsymbol{A}(\xi)\boldsymbol{u} = \widehat{\boldsymbol{f}}(\xi)$$

at each fixed $\xi \in \mathbb{R}$, where $\boldsymbol{A}(\xi) = (i\xi + d)^{-1}(i\xi\boldsymbol{I} + \boldsymbol{A})$.

Notice that Fourier transform is isometric on $\mathbb{L}^2$. Therefore, by Proposition 6.15 in [32], we have the following result.

THEOREM 4.3. *Assume $\boldsymbol{A} = \boldsymbol{X}\,\mathrm{diag}(\lambda_1,\dots,\lambda_n)\boldsymbol{X}^{-1}$ has a splitting $\boldsymbol{M} - \boldsymbol{N}$ with $\boldsymbol{M} = d\boldsymbol{I}$, $d > 0$. Define*

$$\epsilon^{(k)} = \max_{\xi\in\mathbb{R}} \min_{\substack{p\,\in\,\mathbb{P}_k\\ p(0)\,=\,1}} \max_{j=1,\dots,n} \left| p\left(\frac{i\xi + \lambda_j}{i\xi + d}\right)\right|.$$

Then the residual $\boldsymbol{r}^k = \boldsymbol{\psi} - (\boldsymbol{I} - \mathcal{K})\boldsymbol{x}^k$ obtained by the CGMRES algorithm satisfies

$$\|\boldsymbol{r}^k\|_{\mathbb{L}^2} \le \kappa_2(\boldsymbol{X})\epsilon^{(k)}\|\boldsymbol{r}^0\|_{\mathbb{L}^2},$$

where $\kappa_2(\boldsymbol{X}) = \|\boldsymbol{X}\|_2\|\boldsymbol{X}^{-1}\|_2$ is the condition number of $\boldsymbol{X}$ under the 2-norm.

Proof. For each fixed $\xi \in \mathbb{R}$, it is known that

$$\left|\widehat{\boldsymbol{r}^k}(\xi)\right|_2 \le \kappa_2(\boldsymbol{X}) \min_{\substack{p\,\in\,\mathbb{P}_k\\ p(0)\,=\,1}} \max_{j=1,\dots,n} \left| p\left(\frac{i\xi + \lambda_j}{i\xi + d}\right)\right| \cdot \left|\widehat{\boldsymbol{r}^0}(\xi)\right|_2.$$

The right-hand side of the above inequality is bounded by $\kappa_2(\boldsymbol{X})\epsilon^{(k)}|\hat{\boldsymbol{r}}^0(\xi)|_2$. Integrating both sides in ξ, we get the desired inequality. □

In order to estimate $\epsilon^{(k)}$, we assume that eigenvalues of $\boldsymbol{A}$ are included in an ellipse $E(c,e,a)$, centered at c, with focal distance e and semimajor axis a, which excludes the origin.

Assume a kth order polynomial p satisfies $p(0) = 1$. Then, since $d > 0$, $\xi \in \mathbb{R}$, $i\xi + d \neq 0$,

$$q_\xi(z) = p\left(\frac{i\xi + z}{i\xi + d}\right)$$

is of degree k in z and $q_\xi(-i\xi) = 1$. For fixed ξ,

$$\{q_\xi \in \mathbb{P}_k : \; q_\xi(-i\xi) = 1\} = \{q \in \mathbb{P}_k : \; q(-i\xi) = 1\}.$$

In fact, assume $q(z) = 1 + a_1(z + i\xi) + \cdots + a_k(z + i\xi)^k$. Define kth order polynomials

$$\begin{aligned} p(z) &= 1 + a_1(d + i\xi)z + \cdots + a_k(d + i\xi)^k z^k, \\ q_\xi(z) &= p\left(\frac{i\xi + z}{i\xi + d}\right), \end{aligned}$$

and then $q(z) = q_\xi(z)$. Therefore, the above two sets are equal. Hence,

$$\begin{aligned} \epsilon^{(k)} &\le \max_{\xi\in\mathbb{R}} \min_{\substack{p\,\in\,\mathbb{P}_k \\ p(0)\,=\,1}} \max_{z\in E(c,e,a)} \left| p\left(\frac{i\xi + z}{i\xi + d}\right)\right| \\ &= \max_{\xi\in\mathbb{R}} \min_{\substack{q_\xi\,\in\,\mathbb{P}_k \\ q_\xi(-i\xi)\,=\,1}} \max_{z\in E(c,e,a)} |q_\xi(z)| \\ &= \max_{\xi\in\mathbb{R}} \min_{\substack{q\,\in\,\mathbb{P}_k \\ q(-i\xi)\,=\,1}} \max_{z\in E(c,e,a)} |q(z)|. \end{aligned}$$

By an estimate in [32, p. 192],

$$\min_{q\in\mathbb{P}_k,\, q(-i\xi)=1} \max_{z\in E(c,e,a)} |q(z)| = \frac{C_k\left(\frac{a}{e}\right)}{\left|C_k\left(\frac{c+i\xi}{e}\right)\right|},$$

where C_k is the kth order Chebyshev polynomial. Hence,

$$\epsilon^{(k)} \le \max_{\xi\in\mathbb{R}} \frac{C_k\left(\frac{a}{e}\right)}{\left|C_k\left(\frac{c+i\xi}{e}\right)\right|}.$$

The following is a key lemma for analyzing the convergence of CGMRES.

LEMMA 4.4. *If c, e are real and positive, then*

$$\min_{\xi\in\mathbb{R}} \left|C_k\left(\frac{c + i\xi}{e}\right)\right| = C_k\left(\frac{c}{e}\right).$$

Proof. By the definition of the Chebyshev polynomial $C_k(z)$, or by noticing that $C_k(z)$ has real coefficients, we have $C_k(\overline{z}) = \overline{C_k(z)}$. As an immediate consequence,

$$\min_{\xi\in\mathbb{R}} \left|C_k\left(\frac{c + i\xi}{e}\right)\right| = \min_{\xi\ge 0} \left|C_k\left(\frac{c + i\xi}{e}\right)\right|.$$

In order to find the minimum, define

$$f(\xi) = \left|C_k\left(\frac{c + i\xi}{e}\right)\right|^2.$$

Let $w = \rho e^{i\theta}$ such that

$$\frac{1}{e}(c + i\xi) = \frac{1}{2}(w + w^{-1}).$$

Then

$$(\rho + \rho^{-1})\cos\theta = \frac{2c}{e}, \quad (\rho - \rho^{-1})\sin\theta = \frac{2\xi}{e}.$$

By a computation, we get

$$f'(\xi) = \frac{k}{2e\beta^2}[(\rho^{2k} - \rho^{-2k})(\rho + \rho^{-1})\sin\theta - 2(\rho - \rho^{-1})\cos\theta\sin(2k\theta)],$$

where $\beta^2 = |\rho e^{i\theta} - \rho^{-1}e^{-i\theta}|^2$.

If $\rho = 1$ or $\theta = 0$, then $\xi = 0$ and $f'(\xi) = 0$. However, in our case, since $c > e$, $\rho \neq 1$. Otherwise, $\cos\theta = \frac{c}{e} > 1$, which is impossible.

By the relation $(\rho - \rho^{-1})\sin\theta = \frac{2\xi}{e}$, if $0 < \theta < \frac{\pi}{2}$, then $\rho > 1$; if $-\frac{\pi}{2} < \theta < 0$, then $\rho < 1$. By symmetry, in order to prove that $f'(\xi) > 0$ for $\xi > 0$, it is enough to prove that

$$g(\rho, \theta) = (\rho^{2k} - \rho^{-2k})(\rho + \rho^{-1})\sin\theta - 2(\rho - \rho^{-1})\cos\theta\sin(2k\theta) > 0$$

for $0 < \theta < \frac{\pi}{2}$ and $\rho > 1$. By introduction, one can prove that $\sin(m\theta) \leq m\sin\theta$ for $0 < \theta < \frac{\pi}{2}$ and $m \in \mathbb{N}$. Also, notice that $\rho + \rho^{-1} > 2$. Hence, in order to prove $g(\rho, \theta) > 0$, it is sufficient to prove that

$$g(\rho) = \rho^{2k} - \rho^{-2k} - k(\rho^2 - \rho^{-2}) > 0$$

for $\rho > 1$. However, $g(\rho)$ is a strictly increasing function for $\rho > 1$, which implies $g(\rho) > g(1) = 0$. This proves the lemma. □

Notice that the Chebyshev polynomial $C_k(z)$ has properties $|C_k(z)| = |C_k(\overline{z})| = |C_k(-z)| = |C_k(-\overline{z})|$. Therefore, by the last lemma, we have the following consequence.

COROLLARY 4.5. *If c is a complex number and $e \neq 0$ is real, then*

$$\min_{\xi \in \mathbb{R}} \left| C_k\left(\frac{c + i\xi}{e}\right) \right| = C_k\left(\frac{|Re\, c|}{|e|}\right).$$

Finally, we have the following convergence result for CGMRES.

THEOREM 4.6. *Assume $\boldsymbol{A} = \boldsymbol{X}\Lambda\boldsymbol{X}^{-1} = \boldsymbol{M} - \boldsymbol{N}$, $\boldsymbol{M} = d\boldsymbol{I}$, $d > 0$. The spectrum of $\boldsymbol{A}$ is included in ellipse $E(c, e, a)$, centered at c, with focal distance e and semimajor axis a, which excludes origin. Then the residual $\boldsymbol{r}^k = \boldsymbol{\psi} - (\boldsymbol{I} - \mathcal{K})\boldsymbol{x}^k$ obtained by the CGMRES algorithm satisfies the estimate*

$$\|\boldsymbol{r}^k\|_{\mathbb{L}^2} \leq \kappa_2(\boldsymbol{X}) \frac{C_k\left(\frac{a}{e}\right)}{C_k\left(\frac{c}{e}\right)} \|\boldsymbol{r}^0\|_{\mathbb{L}^2}.$$

Remark. Thus, CGMRES applied to (2.4) is bounded by the same rate of convergence as GMRES applied to the associated problem (2.6).

5. Biconvolution acceleration methods. The CGMRES algorithm demonstrates that the use of convolution can accelerate waveform relaxation in a manner similar to GMRES applied to the associated linear algebra problem. However, there is still something unsatisfying about the algorithm in that even for Hermitian $\boldsymbol{A}$ one must use CGMRES. In this section, we turn our attention to a method that exploits Hermitian properties of the matrix $\boldsymbol{A}$, namely, biconvolution CG (BiCCG). The development of BiCCG will use a convolution bilinear form in place of the convolution inner product used by CGMRES.

We begin by defining the following bilinear form:

$$[\boldsymbol{x}, \boldsymbol{y}]_\star(t) = \sum_{i=1}^{n}(x_i \star y_i)(t) \in \mathbb{L}^1(\mathbb{R}, \mathbb{R}).$$

Remarks.

1. Convolution between a function and a vector of functions is the same as defined in section 4. However, the convolution bilinear form $[\cdot,\cdot]_\star$ is different from the previously defined inner product $\langle\cdot,\cdot\rangle_\star$.
2. Notice that $[\boldsymbol{x},\boldsymbol{y}]_\star\widehat{\;}(\xi) = [\widehat{\boldsymbol{x}}(\xi),\widehat{\boldsymbol{y}}(\xi)]$. Here $[\cdot,\cdot]$ is the bilinear form in $\mathbb{C}^n$ defined by
$$[\boldsymbol{z},\boldsymbol{w}] = \sum_{i=1}^{n} z_i w_i$$
for $\boldsymbol{w},\boldsymbol{z}\in\mathbb{C}^n$. Note that this is *not* the typical inner product in $\mathbb{C}^n$. It is, however, the bilinear form used in the BiCG algorithm [8], hence the name "biconvolution CG."
3. Since, by assumption, functions and vectors are compactly supported, $f\star\boldsymbol{x}$ and $[\boldsymbol{x},\boldsymbol{y}]_\star$ are in fact in $\mathbb{L}^2$.

5.1. The BiCCG algorithm. The CG algorithm is a popular and effective iterative method for solving symmetric positive definite systems of equations [11, 13]. Waveform extensions (using scalar parameters) of the CG algorithm are not well-defined, even for symmetric positive definite $\boldsymbol{A}$, since, in general, the operator $\mathcal{K}$ is not self-adjoint with respect to the $\mathbb{L}^2$ inner product. On the other hand, as we will see below, it is possible to develop a well-defined waveform extension to CG using convolution, i.e., the BiCCG algorithm.

Definition. An operator $\mathcal{A}:\ \mathbb{L}^2([0,T],\mathbb{R}^n)\to\mathbb{L}^2([0,T],\mathbb{R}^n)$ is called *convolution self-adjoint* if for any $\boldsymbol{u},\ \boldsymbol{v}\in\mathbb{L}^2([0,T],\mathbb{R}^n),\ [\mathcal{A}\boldsymbol{u},\boldsymbol{v}]_\star = [\boldsymbol{u},\mathcal{A}\boldsymbol{v}]_\star$.

Remark. If $\boldsymbol{A}$ is Hermitian, the operator $\mathcal{K}$ is convolution self-adjoint with respect to the convolution bilinear form "$[\cdot,\cdot]_\star$."

Definition. An operator $\mathcal{A}:\ \mathbb{L}^2([0,T],\mathbb{R}^n)\to\mathbb{L}^2([0,T],\mathbb{R}^n)$ is called *convolution definite* if for any nonzero $\boldsymbol{u}\in\mathbb{L}^2([0,T],\mathbb{R}^n),\ [\mathcal{A}\boldsymbol{u},\boldsymbol{u}]_\star\neq 0$ in the $\mathbb{L}^2$ sense.

Since $(s+d)^{-1}(s\boldsymbol{I}+\boldsymbol{A})$ is symmetric positive definite for positive real s and symmetric positive definite $\boldsymbol{A}$, the following lemma follows immediately from the above definitions and the equivalence (via the Fourier transform) between $(\boldsymbol{I}-\mathcal{K})$ and $(s+d)^{-1}(s\boldsymbol{I}+\boldsymbol{A})$.

LEMMA 5.1. *If $\boldsymbol{A}$ is real symmetric with splitting $\boldsymbol{M}-\boldsymbol{N},\ \boldsymbol{M}=d\boldsymbol{I},\ d>0$, then $(\boldsymbol{I}-\mathcal{K})$ is convolution self-adjoint. Furthermore, if $\boldsymbol{A}$ is positive definite, then $(\boldsymbol{I}-\mathcal{K})$ is also convolution definite.*

If $\mathcal{A}$ is convolution self-adjoint and convolution definite on $\mathbb{L}^2([0,T],\mathbb{R}^n)$, then we can define the following BiCCG algorithm (analogous to CG).

ALGORITHM 5 (BiCCG). Let $\mathcal{A}: C_0([0,\infty),\mathbb{R}^n)\to C_0([0,\infty),\mathbb{R}^n)$ be a bounded linear operator. As before, $\mathcal{A}$ is extendable to $\overline{C_0([0,\infty),\mathbb{R}^n)}$, the vector-valued generalized function space, which is again a vector space over $\mathbb{Q}$. Let $\boldsymbol{f}\in C_0([0,\infty),\mathbb{R}^n)$.

1. Pick $\boldsymbol{x}^0\in C_0([0,\infty),\mathbb{R}^n)$ and compute $\boldsymbol{r}^0=\boldsymbol{f}-\mathcal{A}\boldsymbol{x}^0,\ \boldsymbol{p}^0=\boldsymbol{r}^0$.
2. For $j=0,1,\ldots$ until converged,
$$\begin{aligned}\alpha_j &= [\boldsymbol{r}^j,\boldsymbol{r}^j]_\star/[\mathcal{A}\boldsymbol{p}^j,\boldsymbol{p}^j]_\star\\ \boldsymbol{x}^{j+1} &= \boldsymbol{x}^j+\alpha_j\star\boldsymbol{p}^j\\ \boldsymbol{r}^{j+1} &= \boldsymbol{r}^j-\alpha_j\star\mathcal{A}\boldsymbol{p}^j,\\ \beta_j &= [\boldsymbol{r}^{j+1},\boldsymbol{r}^{j+1}]_\star/[\boldsymbol{r}^j,\boldsymbol{r}^j]_\star\\ \boldsymbol{p}^{j+1} &= \boldsymbol{r}^{j+1}+\beta_j\star\boldsymbol{p}^j.\end{aligned}$$

Note that the Fourier transform of the BiCCG algorithm for $\mathcal{A}=\boldsymbol{I}-\mathcal{K}$ is the BiCG algorithm (see [6, 8, 13]) for matrix $\boldsymbol{A}(\xi)=(i\xi\boldsymbol{I}+\boldsymbol{M})^{-1}(i\xi\boldsymbol{I}+\boldsymbol{A})$. Therefore, we can view the BiCG algorithm (for a complex system) as a continuation of the CG algorithm (for a real system). Also, notice that by [5] there is a CG algorithm for matrix $\boldsymbol{A}(\xi) = (i\xi + d)^{-1}(i\xi\boldsymbol{I} + \boldsymbol{A})$.

5.2. Laplace and Fourier transforms. To analyze the convergence of the BiCCG algorithm, we first give some definitions for weighted Sobolev spaces (see [16]) and introduce some results related to Laplace and Fourier transforms.

Definition. For a real number $\lambda > 0$, $\alpha \geq 0$, define the weighted Sobolev space $\mathbb{H}^\alpha_\lambda(\mathbb{R}, \mathbb{R})$ according to

$$\mathbb{H}^\alpha_\lambda(\mathbb{R}, \mathbb{R}) = \{u \in \mathbb{L}^2(\mathbb{R}, \mathbb{R}) : \ (\lambda^2 + |\xi|^2)^{\alpha/2}|\widehat{u}(\xi)| \in \mathbb{L}^2(\mathbb{R}, \mathbb{R})\}.$$

In particular, if $\lambda = 1$, then the weighted $\mathbb{H}^\alpha_1$ space is the regular $\mathbb{H}^\alpha$ space [16]. Also notice that $\mathbb{H}^\alpha_\lambda$ can be defined by another equivalent norm as follows:

$$\mathbb{H}^\alpha_\lambda(\mathbb{R}, \mathbb{R}) = \{u \in \mathbb{L}^2(\mathbb{R}, \mathbb{R}) : \ (|\lambda| + |\xi|)^{\alpha}|\widehat{u}(\xi)| \in \mathbb{L}^2(\mathbb{R}, \mathbb{R})\}.$$

Now assume $\boldsymbol{A} = \boldsymbol{U}\,\mathrm{diag}(\lambda_1, \ldots, \lambda_n)\boldsymbol{U}^T = \boldsymbol{U}\Lambda\boldsymbol{U}^T$ is real symmetric positive definite, where $\boldsymbol{U}\boldsymbol{U}^T = \boldsymbol{I}$, and $0 < \lambda_1 \leq \cdots \leq \lambda_n$. Define a weighted $\mathbb{H}^\alpha_{\mathbf{A}}$ as follows:

$$\mathbb{H}^\alpha_{\mathbf{A}}(\mathbb{R}, \mathbb{R}^n) = \{\boldsymbol{u} = (u_1, \ldots, u_n)^T : (\lambda_j + |\xi|)^\alpha|(\boldsymbol{U}^T\widehat{u})_j(\xi)| \in \mathbb{L}^2(\mathbb{R}, \mathbb{R})\}.$$

By the definition, we can see that $\mathbb{H}^\alpha_{\mathbf{A}} = \mathbb{H}^\alpha_{\lambda_1} \times \mathbb{H}^\alpha_{\lambda_2} \times \cdots \times \mathbb{H}^\alpha_{\lambda_n}$. Also, on $\mathbb{H}^\alpha_{\mathbf{A}}$, we can define an inner product

$$\begin{aligned}(\boldsymbol{u}, \boldsymbol{v})_\alpha &= \int_{\mathbb{R}} \widehat{\boldsymbol{u}}^T(\xi)(|\xi|\boldsymbol{I} + \boldsymbol{A})^{2\alpha}\overline{\widehat{\boldsymbol{v}}(\xi)}d\xi \\ &= \int_{\mathbb{R}} \langle \widehat{\boldsymbol{u}}(\xi), \widehat{\boldsymbol{v}}(\xi)\rangle_{(|\xi|\mathbf{I}+\mathbf{A})^{\mathbf{2}\alpha}} d\xi,\end{aligned}$$

which makes $\mathbb{H}^\alpha_{\mathbf{A}}$ a Hilbert space with norm

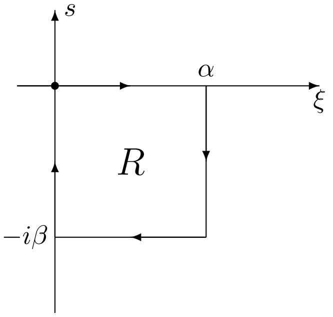
$$\|\boldsymbol{u}\|_{\mathbb{H}^\alpha_{\mathbf{A}}} = \left(\sum_{j=1}^n \int_{\mathbb{R}} (\lambda_j + |\xi|)^{2\alpha}|(\boldsymbol{U}^T\widehat{u})_j(\xi)|^2 d\xi\right)^{1/2}.$$

Now we want to study the relationship between the Laplace transform and the Fourier transform.

Since $f \in \mathbb{L}^2([0, T], \mathbb{R})$, and the trivial extension of f to $\mathbb{L}^2(\mathbb{R}, \mathbb{R})$ is compactly supported, the Fourier transform of f is therefore entire in $\mathbb{C}$. By Cauchy's theorem,

$$\int_R (\widehat{f}(z))^2 dz = 0,$$

where R is a rectangle as shown in the figure at the right. Therefore,



$$\begin{aligned}&\int_0^\alpha (\widehat{f}(\xi))^2 d\xi + \int_0^{-\beta} (\widehat{f}(\alpha + is))^2 d(is) \\ &+ \int_\alpha^0 (\widehat{f}(\xi - i\beta))^2 d\xi + \int_{-\beta}^0 (\widehat{f}(is))^2 d(is) = 0.\end{aligned}$$

For fixed β, let $\alpha \to \infty$. By a property of the Fourier transform (see [12]), $\lim_{\alpha\to\infty} \widehat{f}(\alpha + is) = 0$. Also, since $(\widehat{f}(\alpha + is))^2$ is absolutely integrable,

$$\lim_{\alpha\to\infty} \int_0^{-\beta} (\widehat{f}(\alpha + is))^2 d(is) = i \int_0^{-\beta} \lim_{\alpha\to\infty} (\widehat{f}(\alpha + is))^2 ds = 0.$$

The third term becomes $-\int_0^\infty (\widehat{f}(\xi - i\beta))^2 d\xi$. Let $\beta \to \infty$, and then by

$$\widehat{f}(\xi - i\beta) = \frac{1}{\sqrt{2\pi}} \int_0^\infty e^{-\beta x - i\xi x} f(x)dx \to 0$$

we get that

$$\lim_{\beta\to\infty} \int_0^\infty (\widehat{f}(\xi - i\beta))^2 d\xi = 0.$$

Hence, we conclude that

$$i \int_0^\infty (\widehat{f}(\xi))^2 d\xi = \int_0^\infty (\widehat{f}(-is))^2 ds \geq 0, \tag{5.1}$$

since $\widehat{f}(-is)$ is the Laplace transform and is real. We denote the Laplace transform by $\mathcal{L}f(s) = \widehat{f}(-is)$. This relationship between the Laplace and Fourier transforms plays an important role in what follows.

For subsequent analysis, we require the following definition (see [16]).

Definition.

$$\mathbb{H}_0^\alpha\left([0,\infty),\mathbb{R}\right) = \text{closure of } C_0^\infty[0,\infty) \text{ in } \mathbb{H}^\alpha\left([0,\infty),\mathbb{R}\right),$$

where $C_0^\infty[0,\infty) = \{f : f \in C^\infty[0,\infty) \text{ with compact support }\}$.

If we assume that $f \in \mathbb{H}_0^{1/2}\left([0,\infty),\mathbb{R}\right)$, then by Cauchy's theorem again,

$$\int_R z(\widehat{f}(z))^2 dz = 0,$$

so that

$$-\int_0^\infty \xi(\widehat{f}(\xi))^2 d\xi = \int_0^\infty s(\mathcal{L}f(s))^2 ds \geq 0. \tag{5.2}$$

Combining (5.1) and (5.2) yields a key lemma.

LEMMA 5.2. *For $f \in \mathbb{H}_0^{1/2}([0,\infty),\mathbb{R})$, $\lambda \in \mathbb{C}$,*

$$i \int_0^\infty (i\xi + \lambda)(\widehat{f}(\xi))^2 d\xi = \int_0^\infty (s + \lambda)(\mathcal{L}f(s))^2 ds.$$

Because of this key equality, we can give the following definitions.

Definition. Assume $\boldsymbol{f} = (f_1, \ldots, f_n)^T$, $\boldsymbol{g} = (g_1, \ldots, g_n)^T \in \mathbb{H}_0^{1/2}([0,\infty),\mathbb{R}^n)$, and real symmetric and positive definite $\boldsymbol{A} = \boldsymbol{U}\,\mathrm{diag}(\lambda_1, \ldots, \lambda_n)\boldsymbol{U}^T$ with $\boldsymbol{U}\boldsymbol{U}^T = \boldsymbol{I}$ and $0 < \lambda_1 \leq \cdots \leq \lambda_n$. Define an $\boldsymbol{A}$-weighted inner product by

$$\begin{aligned}\langle \boldsymbol{f}, \boldsymbol{g}\rangle_{\mathbf{A}} &= i \int_0^\infty \langle \hat{\boldsymbol{f}}, \hat{\boldsymbol{g}}\rangle_{i\xi\mathbf{I}+\mathbf{A}} d\xi \\ &= \sum_{j=1}^n i \int_0^\infty (i\xi + \lambda_j) \left(\widehat{\boldsymbol{U}\boldsymbol{f}}\right)_j (\xi) \left(\widehat{\boldsymbol{U}\boldsymbol{g}}\right)_j (\xi) d\xi.\end{aligned}$$

Definition. For $\boldsymbol{f} \in \mathbb{H}_0^{1/2}([0,\infty),\mathbb{R}^n)$, define a norm by

$$\|\boldsymbol{f}\|_{\mathbb{K}_{\mathbf{A}}^{1/2}} = \langle \boldsymbol{f}, \boldsymbol{f}\rangle_{\mathbf{A}}^{1/2}.$$

The new normed space is again a Hilbert space and is denoted by $\mathbb{K}_{\mathbf{A}}^{1/2}([0,\infty),\mathbb{R}^n)$.

Remarks.

1. By integration over a wedge instead of over a rectangle as in Lemma 5.2, one can see that

$$ie^{i\theta}\int_0^\infty (ire^{i\theta}+\lambda)(\widehat{f}(re^{i\theta}))^2 dr = \int_0^\infty (s+\lambda)(\mathcal{L}f(s))^2 ds.$$

   This means the integral

$$i\int_{R(\theta)} (iz+\lambda)(\widehat{f}(z))^2 dz = \int_{R(\theta+\pi/2)} (z+\lambda)(\mathcal{L}f(z))^2 dz$$

   is invariant in θ, where $R(\theta) = \{re^{i\theta} : 0 \le r < +\infty\}$ is a ray starting at the origin.
2. The $\mathbb{K}_{\mathbf{A}}^{1/2}$ norm is bounded by the $\mathbb{H}_{\mathbf{A}}^{1/2}$ norm.
3. By Theorem 11.1, Chapter 1 in [16], $\mathbb{H}_0^{1/2}(\mathbb{R}^+,\mathbb{R}^n) = \mathbb{H}^{1/2}(\mathbb{R}^+,\mathbb{R}^n)$, where $\mathbb{R}^+ = \{x \in \mathbb{R} : x > 0\}$ is the positive half ray of $\mathbb{R}$.

5.3. Convergence of the BiCCG algorithm. By using the projection properties and Chebyshev polynomials, one can prove the following well-known theorem [32] for CG applied to the linear system $\boldsymbol{Ax} = \boldsymbol{b}$.

THEOREM 5.3. *Assume matrix $\boldsymbol{A}$ is real symmetric and positive definite. Let $\{\boldsymbol{x}_m\}$ be the sequence of approximate solutions obtained by the CG algorithm and let $\boldsymbol{x}_*$ be the exact solution. Then the iterates satisfy*

$$\|\boldsymbol{x}_* - \boldsymbol{x}_m\|_{\mathbf{A}} \le 2\left[\frac{\sqrt{\kappa(\boldsymbol{A})}-1}{\sqrt{\kappa(\boldsymbol{A})}+1}\right]^m \|\boldsymbol{x}_* - \boldsymbol{x}_0\|_{\mathbf{A}},$$

where $\kappa(\boldsymbol{A}) = \frac{\lambda_{\max}}{\lambda_{\min}}$ is the condition number of the matrix $\boldsymbol{A}$.

For the BiCCG algorithm we can prove a similar convergence theorem.

THEOREM 5.4. *Let $\boldsymbol{M}-\boldsymbol{N}$ be a splitting of a real symmetric positive definite matrix $\boldsymbol{A}$ with $\boldsymbol{M} = d\boldsymbol{I}$, $d > 0$. Then the BiCCG algorithm applied to* (2.4) *generates a sequence of iterates $\{\boldsymbol{x}^m\}$ that satisfy the weighted $\frac{1}{2}$-estimates*

$$\|\boldsymbol{x}^* - \boldsymbol{x}^m\|_{\mathbb{K}_{\mathbf{A}}^{1/2}} \le 2\left[\frac{\sqrt{\kappa(\boldsymbol{A})}-1}{\sqrt{\kappa(\boldsymbol{A})}+1}\right]^m \|\boldsymbol{x}^* - \boldsymbol{x}^0\|_{\mathbb{K}_{\mathbf{A}}^{1/2}}, \tag{5.3}$$

where $\boldsymbol{x}^*$ is the exact solution.

Proof. By taking Fourier transform on the BiCCG algorithm, we can see that, for each m, $\boldsymbol{x}^* - \boldsymbol{x}^m \in \mathbb{L}^2([0,\infty),\mathbb{R}^n)$. Since $\boldsymbol{x}^* - \boldsymbol{x}^m$ is differentiable on $[0,T]$ and $(\boldsymbol{x}^* - \boldsymbol{x}^m)(0) = 0$, by Urysohn's lemma, we can assume that $\boldsymbol{x}^* - \boldsymbol{x}^m$ is also compactly supported on $[0,\infty)$. Therefore $\boldsymbol{x}^* - \boldsymbol{x}^m$ is in fact in $\mathbb{H}_0^{1/2}([0,\infty),\mathbb{R}^n)$. By taking Laplace transform on the BiCCG algorithm, we can see that $\mathcal{L}(\boldsymbol{x}^m)(s)$ is an approximate solution obtained by the CG algorithm applied to real matrix $(s+d)^{-1}(s\boldsymbol{I}+\boldsymbol{A})$. By Theorem 5.3, for each fixed $s \ge 0$,

$$\|\mathcal{L}(\boldsymbol{x}^* - \boldsymbol{x}^m)(s)\|_{\frac{s\mathbf{I}+\mathbf{A}}{s+d}} \le 2\left[\frac{\sqrt{\kappa(s\boldsymbol{I}+\boldsymbol{A})}-1}{\sqrt{\kappa(s\boldsymbol{I}+\boldsymbol{A})}+1}\right]^m \|\mathcal{L}(\boldsymbol{x}^* - \boldsymbol{x}^0)(s)\|_{\frac{s\mathbf{I}+\mathbf{A}}{s+d}}. \tag{5.4}$$

Since $\boldsymbol{A}$ is positive definite, there exists an unitary matrix $\boldsymbol{U}$ such that

$$\boldsymbol{A} = \boldsymbol{U}\Lambda\boldsymbol{U}^T,$$

where $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$, and $0 < \lambda_1 \leq \cdots \leq \lambda_n$. It is easy to see that

$$\kappa(s\boldsymbol{I} + \boldsymbol{A}) = \frac{s + \lambda_n}{s + \lambda_1}.$$

Notice that

$$\max_{s \in \mathbb{R}^+ \cup \{0\}} \kappa(s\boldsymbol{I} + \boldsymbol{A}) = \max_{s \in \mathbb{R}^+ \cup \{0\}} \left[\frac{s + \lambda_n}{s + \lambda_1}\right] = \kappa(\boldsymbol{A}).$$

Since $f(x) = \frac{\sqrt{x}-1}{\sqrt{x}+1}$ is an increasing function on $[0, +\infty)$, we have

$$\begin{aligned} \frac{\sqrt{\kappa(s\boldsymbol{I} + \boldsymbol{A})} - 1}{\sqrt{\kappa(s\boldsymbol{I} + \boldsymbol{A})} + 1} &\leq \frac{\sqrt{\max_{s \in \mathbb{R}^+ \cup \{0\}} \kappa(s\boldsymbol{I} + \boldsymbol{A})} - 1}{\sqrt{\max_{s \in \mathbb{R}^+ \cup \{0\}} \kappa(s\boldsymbol{I} + \boldsymbol{A})} + 1} \\ &= \frac{\sqrt{\kappa(\boldsymbol{A})} - 1}{\sqrt{\kappa(\boldsymbol{A})} + 1}. \end{aligned}$$

By (5.4) we get

$$(5.5) \qquad \|\mathcal{L}(\boldsymbol{x}^* - \boldsymbol{x}^m)(s)\|^2_{\frac{s\mathbf{I}+\mathbf{A}}{s+d}} \leq 4 \left[\frac{\sqrt{\kappa(\boldsymbol{A})} - 1}{\sqrt{\kappa(\boldsymbol{A})} + 1}\right]^{2m} \|\mathcal{L}(\boldsymbol{x}^* - \boldsymbol{x}^0)(s)\|^2_{\frac{s\mathbf{I}+\mathbf{A}}{s+d}}.$$

Multiplying both sides of (5.5) by $(s + d)$, we obtain

$$(5.6) \qquad \|\mathcal{L}(\boldsymbol{x}^* - \boldsymbol{x}^m)(s)\|^2_{s\mathbf{I}+\mathbf{A}} \leq 4 \left[\frac{\sqrt{\kappa(\boldsymbol{A})} - 1}{\sqrt{\kappa(\boldsymbol{A})} + 1}\right]^{2m} \|\mathcal{L}(\boldsymbol{x}^* - \boldsymbol{x}^0)(s)\|^2_{s\mathbf{I}+\mathbf{A}}.$$

Integrating both sides of (5.6) with respect to s and taking the square root gives the desired inequality (5.3). □

Remarks.

1. Thus, under the weighted $\mathbb{K}_{\mathbf{A}}^{1/2}$ norm, BiCCG applied to (2.4) is bounded by the same rate of convergence as CG applied to the associated problem (2.6).
2. Another important point to note is that, for the Laplace transform, we have the pointwise inequality (5.4). For the Fourier transform, it is not known whether a similar pointwise inequality is true or not. Such an inequality with the Fourier transform is unnecessary, however, because, by Theorem 5.4, it is the integral of the Fourier transforms of $\boldsymbol{x}^* - \boldsymbol{x}^m$ and $\boldsymbol{x}^* - \boldsymbol{x}^0$ which must satisfy a similar inequality.

Notice that, since $((\boldsymbol{I} - \mathcal{K})^n)\hat{} = ((\boldsymbol{I} - \mathcal{K})\hat{})^n$, and the CG algorithm terminates in finite steps, we have the following result regarding finite termination.

COROLLARY 5.5. *The BiCCG algorithm applied to* (2.4) *terminates in finite steps.*

Remark. For the algebraic equation (2.5), the CG algorithm terminates in finite steps. For the differential equation (2.1), we should expect that finite termination is possible [34]. Although the finite termination property is typically not important in practice, the fact that convolution Krylov subspace methods exhibit this property is another indication that they are the "right" generalization from linear algebra problems to waveform problems. The Hilbert space methods in section 3 do not exhibit finite termination.

6. Numerical experiments. In this section, we present preliminary experimental results using convolution Krylov subspace methods. For our model problem, we take the one-dimensional heat equation with unit spatial dimension and $T = 64$ for the temporal dimension. The problem is discretized with 64 spatial points and 32 temporal points and is integrated using backward-Euler.

Two convolution Krylov subspace methods are examined: BiCCG and the convolution variant of the generalized conjugate residual algorithm [4], CGCR. The CGCR algorithm is included rather than convolution GMRES because, although it is theoretically equivalent to GMRES, it is much simpler to implement. Thus, the CGCR results should be taken to be indicative of CGMRES.

The experimental code for BiCCG was written in C++, using the CG module from the IML++ class library [3]. Although IML++ was developed for solving linear systems of equations, by using it with a waveform class and by overloading the appropriate operators, the same CG code was able to be used for both linear algebra problems and waveform problems. The code for CGCR was similarly based on a GCR module (auxiliary to the IML++ library distribution). The experimental code for CSOR and WGMRES was written in C. The convolution kernel for CSOR was obtained according to the algorithms given in [30, 31].

Elements in $\mathbb{Q}$ were applied to functions by first convolving with the numerator and then deconvolving with the denominator. The deconvolution was implemented in a variety of ways, including computation of $(\frac{\hat{f}}{\hat{g}})^{\vee}$, inversion of the convolution based recurrence, and a Newton iteration. All approaches gave similar results.

Figure 6.1 compares the convergence rates of waveform relaxation, WGMRES, CSOR, and BiCCG applied to solving the model problem. We also show the performance of CG applied to solving the corresponding linear algebra problem. For this experiment, BiCCG and CGCR have remarkably better convergence behavior than the other waveform methods and is, in fact, much better than CG itself. This be-

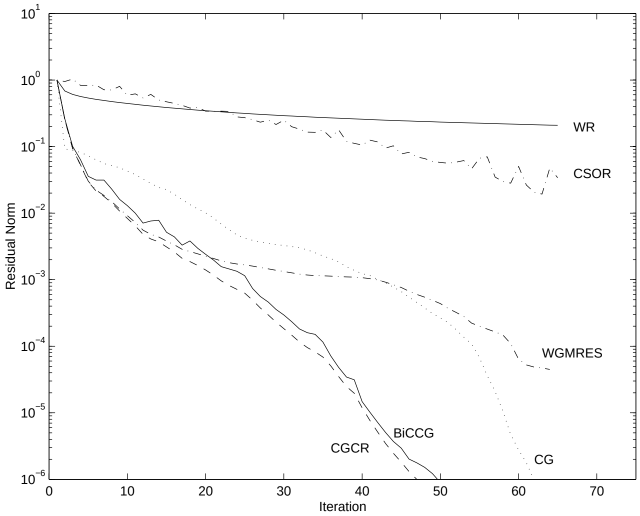


FIG. 6.1. *Convergence of waveform relaxation, WGMRES, CSOR, BiCCG, and CGCR applied to solving a model initial value problem. Also shown is the convergence of CG applied to the corresponding linear algebra problem.*

havior is typical of BiCCG and CGCR over a wide variety of experiments that we conducted (with a particular *caveat*, which we discuss below).

7. Discussion.

7.1. Deconvolution.

It is well known that, in general, deconvolution is an ill-posed problem and that therefore numerical deconvolution will be ill-conditioned. This ill-posedness arises because the "divisor" in the deconvolution may have zero values at particular frequencies.

Theoretically, the convolution Krylov subspace algorithms avoid this ill-posedness naturally because of our choice of functions (i.e., compactly supported $\mathbb{L}^2$ functions). That is, the deconvolution is well-defined in the $\mathbb{L}^2$ sense.

Practically, however, the issue becomes somewhat more delicate, because we need to be concerned not simply with possible zero divisors, but with divisors that are small in a relative numerical sense. Since the convolution Krylov algorithms can be interpreted as being simultaneous iterative processes in the frequency domain, one at each frequency, small divisors can occur during the solution process if the residual values at particular frequencies are small relative to others.

The numerical ill-conditioning can be avoided in part by ensuring that each temporal frequency is present in the initial waveform $\boldsymbol{x}^0$. In our experiments, we effected this by setting $\boldsymbol{x}^0$ to have random values as a function of $t \neq 0$. Unfortunately, for problems having a large number of timepoints, higher temporal frequencies will tend to converge at a much higher rate than the lower frequencies, and numerical instabilities due to deconvolution may appear. Practical implementations of convolution Krylov subspace algorithms (if there turn out to be such things) should be able to circumvent this difficulty via windowing (which may be attractive for memory conservation reasons at any rate) or perhaps by restarting. Alternatively, it may be possible to modify the algorithms in such a way as to equalize the rates of convergence at all frequencies, or through the incorporation of some kind of regularization procedure.

7.2. Conclusion.

As should be evident from this paper, convolution Krylov subspace methods are tremendously interesting, and we have scratched only the surface here. These methods appear to be the "right" generalization of linear algebra acceleration techniques to waveform relaxation. Moreover, they open some entirely new lines of inquiry about Krylov subspace iterations. For instance, the vector space defined by convolution with generalized function is seemingly more abstract than $\mathbb{R}^n$ or $\mathbb{L}^2$, where Krylov subspace algorithms are normally thought to be appropriate. The geometry of Hilbert space is explicitly present only in the transform domain. Finally, there have been a number of algorithms developed recently for the efficient iterative solution, large nonsymmetric linear systems of equations—QMR [9] and Bi-CGSTAB [39] to name just two. Adaptation of these and other methods to the convolution case should be relatively straightforward in terms of description and implementation (although, as with BiCCG and CGMRES, analysis may be somewhat less straightforward). However, a comprehensive experimental study of an assortment of convolution Krylov subspace methods, particularly if applied to practical application problems, would help to shed light on whether or not these methods will be practical in real life.

Acknowledgments. Convolution Krylov subspace algorithms have tantalized the authors for several years. During that time, we have had helpful discussions on this topic with Boris Bäumer, Roland Freund, Alex Himonas, Ken Jackson, Tom Manteuffel, Mark Reichelt, Stefan Vandewalle, and Jacob White and we wish to thank them. The authors also wish to thank the referees for their careful reading and helpful suggestions for strengthening this paper.

REFERENCES

[1] L. V. Ahlfors, *Complex Analysis*, 3rd ed., McGraw-Hill, New York, 1979.

[2] B. Bäumer, *A Vector-Valued Operational Calculus and Abstract Cauchy Problems*, Ph.D. thesis, Louisiana State University, Baton Rouge, LA, 1997.

[3] J. Dongarra, A. Lumsdaine, X. Niu, R. Pozo, and K. Remington, *A sparse matrix library in C++ for high performance architectures*, in Proceedings of the Object Oriented Numerics Conference, Sun River, OR, 1994.

[4] H. C. Elman, *Iterative Methods for Large Sparse Nonsymmetric Systems of Linear Equations*, Ph.D. thesis, Computer Science Department, Yale University, New Haven, CT, 1982.

[5] V. Faber and T. Manteuffel, *Necessary and sufficient conditions for the existence of a conjugate gradient method*, SIAM J. Numer. Anal., 21 (1984), pp. 352–362.

[6] R. Fletcher, *Conjugate gradient methods for indefinite systems*, in Numerical Analysis, G. Watson, ed., Springer-Verlag, Berlin, New York, 1975.

[7] C. Foiaş, *Approximation des opérateurs de J. Mikusinski par des fonctions continues*, Studia Math., 21 (1961), pp. 73–74.

[8] R. W. Freund, *Conjugate gradient-type methods for linear systems with complex symmetric coefficient matrices*, SIAM J. Sci. Statist. Comput., 13 (1992), pp. 425–448.

[9] R. W. Freund and N. M. Nachtigal, *A quasi-minimal residual method for non-Hermition linear systems*, Numer. Math., 60 (91), pp. 315–339.

[10] R. M. Hayes, *Iterative methods of solving linear problems on Hilbert space*, in Contributions to the Solution of Systems of Linear Equations and the Determination of Eigenvalues, Nat. Bur. Standards Appl. Math. 39, O. Taussky, ed., U.S. Government Printing Office, Washington, D.C., 1954, pp. 71–103.

[11] M. R. Hestenes and E. Stiefel, *Methods of conjugate gradients for solving linear systems*, J. Research Nat. Bur. Standards, 49 (1952), pp. 409–436.

[12] L. Hörmander, *The Analysis of Linear Partial Differential Operators*. I., 2nd ed., Springer-Verlag, Berlin, New York, 1990.

[13] C. Lanczos, *Solution of systems of linear equations by minimized iterations*, J. Research Nat. Bur. Standards, 49 (1952), pp. 33–53.

[14] B. Leimkuhler, *Estimating waveform relaxation convergence*, SIAM J. Sci. Comput., 14 (1993), pp. 872–889.

[15] E. Lelarasmee, A. E. Ruehli, and A. L. Sangiovanni-Vincentelli, *The waveform relaxation method for time domain analysis of large scale integrated circuits*, IEEE Trans. CAD, 1 (1982), pp. 131–145.

[16] J. L. Lions and E. Magenes, *Non-Homogeneous Boundary Value Problems and Applications*, Vol. 1, Springer-Verlag, New York, 1972.

[17] C. Lubich, *Chebyshev acceleration of Picard-Lindelöf iteration*, BIT, 32 (1992), pp. 535–538.

[18] C. Lubich and A. Osterman, *Multigrid dynamic iteration for parabolic problems*, BIT, 27 (1987), pp. 216–234.

[19] A. Lumsdaine, *Theoretical and Practical Aspects of Parallel Numerical Algorithms for Initial Value Problems, with Applications*, Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA, 1992.

[20] A. Lumsdaine, M. W. Reichelt, J. M. Squyres, and J. K. White, *Accelerated waveform methods for parallel transient simulation of semiconductor devices*, IEEE Trans. CAD, 15 (1996), pp. 716–726.

[21] A. Lumsdaine and J. K. White, *Accelerating dynamic iteration methods with application to parallel semiconductor device simulation*, Numer. Funct. Anal. Optim., 16 (1995), pp. 395–414.

[22] A. Lumsdaine and D. Wu, *Spectra and pseudospectra of waveform relaxation operators*, SIAM J. Sci. Comput., 18 (1997), pp. 286–304.

[23] U. Miekkala and O. Nevanlinna, *Convergence of dynamic iteration methods for initial value problems*, SIAM J. Sci. Statist. Comput., 8 (1987), pp. 459–482.

[24] U. Miekkala and O. Nevanlinna, *Iterative Solution of Systems of Linear Differential Equations*, Acta Numer. 5, Cambridge University Press, Cambridge, UK, 1996, pp. 259–307.

[25] G. Miel, *Iterative refinement of the method of moments*, Numer. Funct. Anal. Optim., 9 (1987/1988), pp. 1193–1200.

[26] J. Mikusiński, *Operational Calculus*, Vol. 1, 2nd ed., Pergamon Press, Oxford, UK, 1983.

[27] J. Mikusiński, *Operational Calculus*, Vol. 2, 2nd ed., Pergamon Press, Oxford, UK, 1987.

[28] O. Nevanlinna, *Linear acceleration of Picard-Lindelöf iteration*, Numer. Math., 57 (1990), pp. 147–156.

[29] P. Omari, *On the fast convergence of a Galerkin-like method for equations of the second kind*, Math. Z., 201 (1989), pp. 529–539.

[30] M. Reichelt, *Accelerated Waveform Relaxation Techniques for the Parallel Transient Simulation of Semiconductor Devices*, Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA, 1993.

[31] M. W. Reichelt, J. K. White, and J. Allen, *Optimal convolution SOR acceleration of waveform relaxation with application to parallel simulation of semiconductor devices*, SIAM J. Sci. Comput., 16 (1995), pp. 1137–1158.

[32] Y. Saad, *Iterative Methods for Sparse Linear Systems*, PWS, Boston, MA, 1996.

[33] Y. Saad and M. H. Schultz, *GMRES: A generalized minimum residual algorithm for solving nonsymmetric linear systems*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 856–869.

[34] R. D. Skeel, *Waveform iteration and the shifted Picard splitting*, SIAM J. Sci. Statist. Comput., 10 (1989), pp. 756–776.

[35] K. Skornik, *On the Foiaş theorem on convolution of continuous functions*, in Complex Analysis and Applications '85, Publ. House Bulgar. Acad. Sci., Sofia, Bulgaria, 1986.

[36] E. C. Titchmarsh, *The zeros of certain integral functions*, Proc. London Math. Soc., 25 (1926), pp. 283–302.

[37] L. N. Trefethen, *private communication*, 1992.

[38] L. N. Trefethen, *Pseudospectra of matrices*, in Proceedings of 14th Dundee Biennial Conference on Numerical Analysis, 1991.

[39] H. A. van der Vorst, *Bi-CGSTAB: A fast and smoothly converging variant of Bi-CG for the solution of nonsymmetric linear systems*, SIAM J. Sci. Statist. Comput., 13 (1992), pp. 631–644.

[40] S. Vandewalle, *Parallel Multigrid Waveform Relaxation for Parabolic Problems*, Teubner-Skripten zur Numerik, B. G. Teubner, Stuttgart, Germany, 1993.

[41] S. Vandewalle and R. Piessens, *Efficient parallel algorithms for solving initial-boundary value and time-periodic parabolic partial differential equations*, SIAM J. Sci. Statist. Comput., 13 (1992), pp. 1330–1346.

[42] Y. V. Vorobyev, *Method of Moments in Applied Mathematics*, Gordon and Breach, New York, 1965.

[43] J. K. White and A. Sangiovanni-Vincentelli, *Relaxation Techniques for the Simulation of VLSI Circuits*, Engineering and Computer Science Series, Kluwer Academic Publishers, Norwell, MA, 1986.

SIAM J. Numer. Anal.
Vol. 41, No. 1, pp. 112–134

VELOCITY-CORRECTION PROJECTION METHODS FOR INCOMPRESSIBLE FLOWS*

J. L. GUERMOND† AND JIE SHEN‡

Abstract. We introduce and study a new class of projection methods—namely, the velocity-correction methods in standard form and in rotational form—for solving the unsteady incompressible Navier–Stokes equations. We show that the rotational form provides improved error estimates in terms of the H^1-norm for the velocity and of the L^2-norm for the pressure. We also show that the class of fractional-step methods introduced in [S. A. Orsag, M. Israeli, and M. Deville, *J. Sci. Comput.*, 1 (1986), pp. 75–111] and [K. E. Karniadakis, M. Israeli, and S. A. Orsag, *J. Comput. Phys.*, 97 (1991), pp. 414–443] can be interpreted as the rotational form of our velocity-correction methods. Thus, to the best of our knowledge, our results provide the first rigorous proof of stability and convergence of the methods in those papers. We also emphasize that, contrary to those of the above groups, our formulations are set in the standard L^2 setting, and consequently they can be easily implemented by means of any variational approximation techniques, in particular the finite element methods.

Key words. Navier–Stokes equations, projection methods, fractional-step methods, incompressibility, finite elements, spectral approximations

AMS subject classifications. 65M12, 35Q30, 35J05, 76D05

PII. S0036142901395400

1. Introduction.

We consider in this paper the time discretization of the unsteady incompressible Navier–Stokes equations in primitive variables. For a given body force $f(t)$ and an initial solenoidal vector field v_0, we look for u and p such that

$$(1.1)\qquad \begin{cases} \partial_t \mathsf{u} - \nu\nabla^2\mathsf{u} + \mathsf{u}\cdot\nabla\mathsf{u} + \nabla\mathsf{p} = f \qquad \text{in } \Omega\times[0,T], \\ \nabla\cdot\mathsf{u} = 0 \qquad \text{in } \Omega\times[0,T], \\ \mathsf{u}|_\Gamma = 0, \\ \mathsf{u}|_{t=0} = v_0 \qquad \text{in } \Omega. \end{cases}$$

The boundary condition on the velocity is set to zero for the sake of simplicity. The fluid domain Ω is open and bounded in $\mathbb{R}^d$ ($d = 2$ or 3 in practical situations). The domain boundary Γ is assumed to be smooth; e.g., Γ is Lipschitzian and Ω is locally on one side of its boundary.

The goal of this paper is to present a new class of fractional-step projection methods. The original projection method was introduced by Chorin [3] and Temam [15] in the late 60s. An important class of projection methods is the so-called pressure-correction methods introduced in [5, 8]. These schemes consist of two substeps per time step: the pressure is treated explicitly in the first substep and corrected in the second substep by projecting the intermediate velocity onto the space of divergence-free

*Received by the editors September 20, 2001; accepted for publication (in revised form) August 12, 2002; published electronically March 13, 2003.
http://www.siam.org/journals/sinum/41-1/39540.html

†LIMSI (CNRS-UPR 3152), BP 133, 91403 Orsay, France (guermond@limsi.fr). Part of this work was completed while this author was visiting Texas Institute of Computational and Applied Mathematics, Austin, TX, during the period of August, 2001, to July, 2002, under a TICAM Visiting Faculty Fellowship.

‡Department of Mathematics, Purdue University, West Lafayette, IN 47907 (shen@math.purdue.edu). The work of this author was partially supported by NSF grants DMS-9706951 and DMS-0074283.

fields. These schemes are widely used in practice and have been rigorously analyzed in [4, 14, 7].

The new class of projection methods that we introduce in this paper also consist of two substeps per time step: here the viscous (velocity) term is treated explicitly in the first substep and corrected in the second one. Two versions of the method are presented: a standard form and a rotational form. We prove stability and $\mathcal{O}(\delta t^2)$ convergence in the L^2-norm of the velocity for both versions. We also prove improved error estimates for the rotational form, namely, $\mathcal{O}(\delta t^{3/2})$ convergence in the H^1-norm of the velocity and the L^2-norm of the pressure. Such estimates are new and, as indicated by our numerical results, appear to be the best possible under the general context considered in this paper. Since this class of projection methods can be viewed as the dual class of pressure-correction methods, we shall hereafter refer to them as velocity-correction methods.

An interesting aspect of the new class of projection methods is its relation to a set of schemes introduced in [10] and [9]. These schemes have never been analyzed rigorously, partly because they do not fit into any standard weak setting. As originally presented in [10] and [9], these schemes use normal traces of second derivatives of the velocity at the boundary, introducing formidable difficulties in analysis as well as in implementation. In contrast, the new schemes are set in the standard L^2 weak setting and consequently can be naturally implemented and analyzed by means of finite elements or spectral methods. In fact, the schemes in [10] and [9] are formally equivalent, in the spatial continuous case, to the rotational forms of our velocity-correction methods. Thus, to the best of our knowledge, this paper provides the first rigorous proof of stability and convergence of the methods introduced in [10] and [9].

The paper is organized as follows. In the next section, we introduce the velocity-correction method in standard form and prove error estimates for both the time continuous and the time discrete versions of the method. Then, in section 3, we introduce the rotational form of the velocity-correction method and show that this version yields better convergence rates than its standard counterpart. In section 4, we present numerical results, using a finite element method and a Legendre spectral method, which are consistent with our theoretical analysis. In section 5, we examine the relation between the splitting schemes introduced in [10] and [9] and our velocity-correction methods in rotational form. In section 6, we indicate how nonlinear terms can be treated in the velocity-correction schemes. We present concluding remarks in section 7.

We now introduce some notation. We shall use the standard Sobolev spaces $L^2(\Omega)^d$, $H^1(\Omega)^d$, $H^{-1}(\Omega)^d$, etc., whose norm will be denoted by $\|\cdot\|_{0,\Omega}$, $\|\cdot\|_{1,\Omega}$, $\|\cdot\|_{-1,\Omega}$, etc. The L^2 scalar product for scalar and vector valued functions is denoted by $(\cdot,\cdot)$. We equip $H^1_0(\Omega)^d$ with the following norm:

$$\forall\beta\in H^1_0(\Omega)^d,\qquad \|\beta\|_{1,\Omega} := (\|\nabla\cdot\beta\|^2_{0,\Omega} + \|\nabla\times\beta\|^2_{0,\Omega})^{1/2}. \tag{1.2}$$

We introduce two spaces of solenoidal vector fields

$$H = \{v\in L^2(\Omega)^d;\ \nabla\cdot v = 0;\ v\cdot n|_\Gamma = 0\}, \tag{1.3}$$

$$V = \{v\in H^1(\Omega)^d;\ \nabla\cdot v = 0;\ v|_\Gamma = 0\}, \tag{1.4}$$

and we define P_H to be the L^2 projection onto H.

We denote by d_t and ∂_t the time derivative and the partial derivative with respect to time, respectively. Let $\delta t > 0$ be a time step and set $t^k = k\delta t$ for $0 \le k \le K =$

$[T/\delta t]$. Let $\phi^0, \phi^1, \dots, \phi^K$ be some sequence of functions in some Banach space E. We shall use the following discrete norms:

$$\|\phi\|_{l^2(E)} := \left(\delta t \sum_{k=0}^{K} \|\phi^k\|_E^2\right)^{1/2}, \qquad \|\phi\|_{l^\infty(E)} := \max_{0\le k\le K}\left(\|\phi^k\|_E^2\right). \tag{1.5}$$

We denote by c a generic constant which is independent of ε and δt but possibly depends on the data and the solution, and the value of which may vary at each occurrence.

Since the nonlinear term does not contribute in any essential way to the error analysis of projection methods, we shall carry out our analysis for the linearized equations only, so as to avoid technicalities which may obscure the essential ideas in the proof. Our proofs can be adapted to account for the nonlinearity using standard techniques (cf. [16, 14, 7]).

2. Velocity-correction methods: Standard form.

2.1. Introduction of the scheme. Before introducing velocity-correction methods, let us recall the second-order pressure-correction scheme. Hereafter, we take $\nu = 1$ for simplicity and drop the nonlinear term. A second-order pressure-correction scheme is written in the following form: set $u^0 = \mathsf{u}(0)$, $p^0 = \mathsf{p}(0)$, and choose u^1 and p^1 to be reasonable approximations of $\mathsf{u}(\delta t)$ and $\mathsf{p}(\delta t)$; then for $k \ge 1$ we look for $(\tilde u^{k+1}; u^{k+1}, p^{k+1})$ such that

$$\begin{cases} \frac{1}{2\delta t}(3\tilde u^{k+1} - 4u^k + u^{k-1}) - \nabla^2 \tilde u^{k+1} + \nabla p^k = f(t^{k+1}), \\ \tilde u^{k+1}|_\Gamma = 0 \end{cases} \tag{2.1}$$

and

$$\begin{cases} \frac{3}{2\delta t}(u^{k+1} - \tilde u^{k+1}) + \nabla(p^{k+1} - p^k) = 0, \\ \nabla\cdot u^{k+1} = 0, \\ u^{k+1}\cdot n|_\Gamma = 0, \end{cases} \tag{2.2}$$

where n is the outward normal of Ω. For a rigorous analysis of this scheme and its variants, we refer to [4, 14, 7].

Now we propose to adopt a new point of view by switching the role of pressure and velocity. We first treat the viscous (velocity) term explicitly in the first substep and then correct it in the second substep. The corresponding scheme is as follows: set $\tilde u^0 = v_0$ and choose $\tilde u^1$ to be a good approximation of $\mathsf{u}(\delta t)$; then for $k \ge 1$ we look for $(u^{k+1}, p^{k+1}; \tilde u^{k+1})$ such that

$$\begin{cases} \frac{1}{2\delta t}(3u^{k+1} - 4\tilde u^k + \tilde u^{k-1}) - \nabla^2 \tilde u^k + \nabla p^{k+1} = f(t^{k+1}), \\ \nabla\cdot u^{k+1} = 0, \\ u^{k+1}\cdot n|_\Gamma = 0 \end{cases} \tag{2.3}$$

and

$$\begin{cases} \frac{3}{2\delta t}(\tilde u^{k+1} - u^{k+1}) - \nabla^2(\tilde u^{k+1} - \tilde u^k) = 0, \\ \tilde u^{k+1}|_\Gamma = 0. \end{cases} \tag{2.4}$$

We shall hereafter refer to this scheme as the standard form of the velocity-correction method. Note that there is a strong similarity between the velocity-correction method and the pressure-correction method. In fact, our velocity-correction scheme can be regarded as the dual of the pressure-correction scheme.

Note also that (2.3) can be written as

$$u^{k+1} = P_H\left(\frac{4}{3}\tilde{u}^k - \frac{1}{3}\tilde{u}^{k-1} + \frac{2\delta t}{3}(\nabla^2\tilde{u}^k + f(t^{k+1}))\right),$$

where P_H is the L^2 projection onto H. Hence, the method defined by (2.3)–(2.4) falls into the class of the projection methods as introduced by Chorin and Temam. Since the projection step precedes the viscous step, one could also refer to these methods as "projection–diffusion" methods as in [1].

We observe from (2.4) that $\nabla^2(\tilde{u}^{k+1} - \tilde{u}^k)\cdot n|_\Gamma = 0$, which implies that

$$\nabla^2\tilde{u}^{k+1}\cdot n\big|_\Gamma = \nabla^2\tilde{u}^k\cdot n|_\Gamma = \cdots = \nabla^2\tilde{u}^0\cdot n|_\Gamma. \tag{2.5}$$

We then derive from the above and (2.3) that

$$\left.\frac{\partial p^{k+1}}{\partial n}\right|_\Gamma = (f(t^{k+1}) + \nabla^2\tilde{u}^0)\cdot n|_\Gamma. \tag{2.6}$$

This is obviously an artificial Neumann boundary condition for the pressure, which will introduce a numerical boundary layer on the pressure and limit the accuracy of the scheme, just as in the case of pressure-correction schemes.

2.2. Implementation of the standard form. When working with H^1-conformal finite elements, it is difficult to solve (2.3) as a weak Poisson problem for the pressure, for there is a second derivative in the right-hand side which cannot be tested against gradients. To avoid this difficulty, we rewrite the algorithm in an equivalent form by making algebraic substitutions.

By subtracting step (2.3) at time t^k from step (2.3) at time t^{k+1} and by substituting step (2.4) at time t^k into the resulting equation, one obtains a new equivalent form of the projection step:

$$\begin{cases} \frac{1}{2\delta t}(3u^{k+1} - 7\tilde{u}^k + 5\tilde{u}^{k-1} - \tilde{u}^{k-2}) + \nabla(p^{k+1} - p^k) = f(t^{k+1}) - f(t^k), \\ \nabla\cdot u^{k+1} = 0, \\ u^{k+1}\cdot n|_\Gamma = 0. \end{cases} \tag{2.7}$$

Note that in this form the projection step can be solved easily as a weak Poisson problem as follows:

$$\begin{cases} \text{Find } p^{k+1} \text{ in } H^1(\Omega)/\mathbb{R} \text{ such that } \forall q \text{ in } H^1(\Omega)/\mathbb{R} \\ (\nabla(p^{k+1} - p^k), \nabla q) = (f(t^{k+1}) - f(t^k) + \frac{1}{2\delta t}(7\tilde{u}^k - 5\tilde{u}^{k-1} + \tilde{u}^{k-2}), \nabla q). \end{cases} \tag{2.8}$$

Once the pressure is known, the new viscous velocity $\tilde{u}^{k+1}$ is evaluated by solving

$$\begin{cases} \frac{1}{2\delta t}(3\tilde{u}^{k+1} - 4\tilde{u}^k + \tilde{u}^{k-1}) - \nabla^2\tilde{u}^{k+1} + \nabla p^{k+1} = f(t^{k+1}), \\ \tilde{u}^{k+1}|_\Gamma = 0. \end{cases} \tag{2.9}$$

Note that the projected velocity u^{k+1} has been completely eliminated from the algorithm (2.8)–(2.9); hence, it is not necessary to evaluate this quantity, i.e., $\tilde{u}^{k+1}$ is the approximate velocity to be considered in practice.

2.3. The time continuous version: A singularly perturbed PDE. Just as in the pressure-correction case (cf., e.g., [11, 14]), the behavior of the error for the velocity-correction scheme (2.7)–(2.9) is determined by the corresponding singularly perturbed system:

$$\partial_t u^\varepsilon - \nabla^2 u^\varepsilon + \nabla p^\varepsilon = f, \qquad u^\varepsilon|_\Gamma = 0, \tag{2.10}$$

$$\nabla\cdot(u^\varepsilon - \varepsilon(\nabla p_t^\varepsilon - f_t)) = 0, \qquad \left(\frac{\partial p_t^\varepsilon}{\partial n} - f_t\cdot n\right)\Big|_\Gamma = 0, \tag{2.11}$$

$$u^\varepsilon|_{t=0} = \mathsf{u}(0),\ p^\varepsilon|_{t=0} = \mathsf{p}(0). \tag{2.12}$$

Note that (2.11) is obtained by taking the divergence of (2.7) and letting $\delta t \to 0$. Its singular nature comes from the nonphysical boundary condition $(\frac{\partial p_t^\varepsilon}{\partial n} - f_t\cdot n)|_\Gamma = 0$, which introduces a numerical boundary layer for the pressure.

The following theorem characterizes the errors $\mathsf{u} - u^\varepsilon$ and $\mathsf{p} - p^\varepsilon$.

Theorem 2.1. *If the solution of* (1.1) *is sufficiently smooth, we have*

$$\|\mathsf{u} - u^\varepsilon\|_{L^2(L^2)} + \varepsilon^{\frac14}\|\mathsf{u} - u^\varepsilon\|_{L^\infty(L^2)} + \varepsilon^{\frac12}(\|\mathsf{u} - u^\varepsilon\|_{L^\infty(H^1)} + \|\mathsf{p} - p^\varepsilon\|_{L^\infty(L^2)}) \le c\varepsilon.$$

Proof. We shall first derive some a priori estimates.

We denote $e = \mathsf{u} - u^\varepsilon$ and $q = \mathsf{p} - p^\varepsilon$. Subtracting (2.10) from (1.1), we find

$$\begin{cases} e_t - \nabla^2 e + \nabla q = 0, \quad e|_\Gamma = 0, \\ \nabla\cdot e = -\varepsilon\nabla\cdot(\nabla p_t^\varepsilon - f_t) = \epsilon\nabla\cdot\nabla q_t - \epsilon\nabla\cdot(\nabla\partial_t\mathsf{p} - f_t), \quad (\frac{\partial p_t^\varepsilon}{\partial n} - f_t\cdot n)|_\Gamma = 0, \end{cases} \tag{2.13}$$

with $e(0) = 0$ and $q(0) = 0$. Taking the inner product of (2.13) with (e, q), we find

$$\begin{aligned} \frac12 d_t\|e\|_{0,\Omega}^2 + \|\nabla e\|_{0,\Omega}^2 + \frac{\varepsilon}{2} d_t\|\nabla q\|_{0,\Omega}^2 &= \varepsilon(\nabla\partial_t\mathsf{p} - f_t, \nabla q) \\ &\le \frac{\varepsilon}{2}\|\nabla\partial_t\mathsf{p} - f_t\|_{0,\Omega}^2 + \frac{\varepsilon}{2}\|\nabla q\|_{0,\Omega}^2. \end{aligned}$$

Thus, an application of the Gronwall lemma leads to

$$\|e(t)\|_{0,\Omega}^2 + \varepsilon\|\nabla q(t)\|_{0,\Omega}^2 + \int_0^t \|\nabla e\|_{0,\Omega}^2 ds \le c\varepsilon. \tag{2.14}$$

Now, noticing that $e(0) = 0$ and $q(0) = 0$ imply that $e_t(0) = 0$, we can repeat the computation above to obtain

$$\|e_t(t)\|_{0,\Omega}^2 + \varepsilon\|\nabla q_t(t)\|_{0,\Omega}^2 + \int_0^t \|\nabla e_t\|_{0,\Omega}^2 ds \le c\varepsilon, \tag{2.15}$$

which implies, in particular, that

$$\|u_t^\varepsilon\|_{L^\infty(L^2)} + \|\nabla p_t^\varepsilon\|_{L^\infty(L^2)} \le c. \tag{2.16}$$

We are now in position to derive the desired error estimates. Consider the following parabolic dual problem:

$$\begin{cases} w_t + \nabla^2 w + \nabla r = e(s), \quad s \in (0, t), \\ \nabla\cdot w = 0, \\ w|_\Gamma = 0, \quad w(t) = 0. \end{cases} \tag{2.17}$$

It is well known (and an easy matter to show) that

$$\int_0^t (\|w\|_{2,\Omega}^2 + \|\nabla r\|_{0,\Omega}^2) ds \le c \int_0^t \|e(s)\|_{0,\Omega}^2 ds. \tag{2.18}$$

Taking the inner product of (2.17) with $e(s)$ and using the error equation (2.13) and the fact that $\nabla \cdot w = 0$, we infer

$$\begin{aligned} \|e\|_{0,\Omega}^2 &= (e, w_t) + (e, \nabla^2 w) + (\nabla r, e) \\ &= d_t(e,w) - (e_t, w) + (\nabla^2 e, w) - (r, \nabla \cdot e) \\ &= d_t(e,w) - \varepsilon(\nabla r, \nabla(p_t^\varepsilon - f_t)). \end{aligned} \tag{2.19}$$

Integrating the equation above on the interval $[0,t]$, we find

$$\int_0^t \|e\|_{0,\Omega}^2 ds \le \varepsilon \left(\int_0^t \|\nabla r\|_{0,\Omega}^2 ds \right)^{\frac{1}{2}} \left(\int_0^t \|\nabla (p_t^\varepsilon - f_t)\|_{0,\Omega}^2 ds \right)^{\frac{1}{2}}.$$

Using this bound together with (2.16) and (2.18), we finally obtain

$$\|e\|_{L^2(L^2)} \le c\varepsilon.$$

Next, we consider a second parabolic dual problem:

$$\begin{cases} w_t + \nabla^2 w + \nabla r = e_t(s), \quad s \in (0,t), \\ \nabla \cdot w = 0, \\ w|_\Gamma = 0, \quad w(t) = 0. \end{cases} \tag{2.20}$$

Owing to (2.15), we have

$$\int_0^t (\|w\|_{2,\Omega}^2 + \|\nabla r\|_{0,\Omega}^2) ds \le c \int_0^t \|e_t(s)\|_{0,\Omega}^2 ds \le c\varepsilon. \tag{2.21}$$

Taking the inner product of (2.20) with $e(s)$, and proceeding in the same fashion as above, we find

$$\frac{1}{2} d_t \|e\|_{0,\Omega}^2 = d_t(e,w) - (r, \nabla \cdot e) = d_t(e,w) - \varepsilon(\nabla r, \nabla(p_t^\varepsilon - f_t)). \tag{2.22}$$

Integrating this equation in time and using (2.21), we deduce

$$\|e(t)\|_{0,\Omega}^2 \le 2\varepsilon \left(\int_0^t \|\nabla r\|_{0,\Omega}^2 ds \right)^{\frac{1}{2}} \left(\int_0^t \|\nabla (p_t^\varepsilon - f_t)\|_{0,\Omega}^2 ds \right)^{\frac{1}{2}} \le c\varepsilon^{\frac{3}{2}}.$$

To estimate $\|e\|_{L^\infty(H^1)}$, we take the inner product of the first equation in (2.13) with e_t as follows:

$$\begin{aligned} \|e_t\|_{0,\Omega}^2 + \frac{1}{2} d_t \|\nabla e\|_{0,\Omega}^2 &= (q, \nabla \cdot e_t) = \varepsilon(\nabla q, \partial_t \nabla (p_t^\varepsilon - f_t)) \\ &= \varepsilon d_t (\nabla q, \nabla(p_t^\varepsilon - f_t)) - \varepsilon(\nabla q_t, \nabla(p_t^\varepsilon - f_t)). \end{aligned} \tag{2.23}$$

Integrating this equation in time and using the a priori estimates in (2.16), we obtain

$$\|e_t\|_{L^2(L^2)}^2 + \|\nabla e\|_{L^\infty(L^2)}^2 \le C\varepsilon(\|\nabla q\|_{L^\infty(L^2)} + \|\nabla q_t\|_{L^2(L^2)}) \|\nabla(p_t^\varepsilon - f_t)\|_{L^\infty(L^2)} \le c\varepsilon.$$

Finally, using the estimate above and (2.15), we derive

$$\|q\|_{L^\infty(L^2)} \le c\varepsilon^{\frac{1}{2}}.$$

The proof is now complete. □

2.4. Error estimates for the standard velocity-correction scheme. In this section we derive error estimates for the standard velocity-correction scheme (2.3)–(2.4). Hereafter we assume that the following nonessential initialization hypothesis holds:

$$\text{(H)} \quad \begin{cases} \|\mathsf{u}(\delta t) - \tilde{u}^1\|_{0,\Omega} \le c\delta t^2, \\ \|\mathsf{u}(\delta t) - \tilde{u}^1\|_{1,\Omega} \le c\delta t^{3/2}, \\ \|\mathsf{u}(\delta t) - \tilde{u}^1\|_{2,\Omega} \le c\delta t. \end{cases}$$

Remark 2.1. We point out that this hypothesis would hold, in particular, if $(\tilde{u}^1, u^1, p^1)$ were obtained by using a first-order velocity-correction projection scheme. This would amount to replacing the second-order BDF (backward difference formula) time stepping in (2.3) with the backward Euler time stepping at the very first time step.

THEOREM 2.2. *Under the initialization hypothesis* (H) *and provided that the solution to* (1.1) *is smooth enough in time and space, the solution* $(u^k, \tilde{u}^k, p^k)$ *to* (2.3)–(2.4) *is such that*

$$\begin{aligned} \|\mathsf{u} - u\|_{l^2(L^2(\Omega)^d)} + \|\mathsf{u} - \tilde{u}\|_{l^2(L^2(\Omega)^d)} &\le c(\mathsf{u}, \mathsf{p}, T)\, \delta t^2, \\ \|\mathsf{u} - u\|_{l^\infty(L^2(\Omega)^d)} + \|\mathsf{u} - \tilde{u}\|_{l^\infty(L^2(\Omega)^d)} &\le c(\mathsf{u}, \mathsf{p}, T)\, \delta t^{\frac{3}{2}}, \\ \|\mathsf{u} - \tilde{u}\|_{l^\infty(H^1(\Omega)^d)} + \|\mathsf{p} - p\|_{l^\infty(L^2(\Omega))} &\le c(\mathsf{u}, \mathsf{p}, T)\, \delta t. \end{aligned}$$

Note that the discrete norms in the theorem above, and subsequently in later sections, are defined in (1.5). By comparing the time discrete version (2.3)–(2.4) and the time continuous version (2.10)–(2.12), one observes that ε in (2.10)–(2.12) corresponds to δt^2 in (2.3)–(2.4). Thus, the results of Theorem 2.2 are fully consistent with those of Theorem 2.1.

The proof of Theorem 2.2 follows exactly the same procedure as the proof of Theorem 2.1 but for the time discretization. The main technical difficulty comes from the treatment of the second-order BDF term, which will be treated in detail in the proof of Theorem 3.1. Thus, we omit the proof here to avoid unnecessary repetition.

3. Velocity-correction method: Rotational form.

3.1. Introduction of the scheme. The main obstacle in proving error estimates better than first-order on the velocity in the H^1-norm and the pressure in the L^2-norm comes from the fact that the algorithm enforces the nonphysical pressure Neumann boundary condition (2.6). This phenomenon is reminiscent of the numerical boundary layer induced by the nonphysical boundary condition $\partial_n p^{k+1}|_\Gamma = \cdots = \partial_n p^0|_\Gamma$ enforced by the pressure-correction method in its standard form; cf. [14, 7]. Thus, in order to obtain better approximation for the pressure, we need to correct this nonphysical boundary condition. Considering the identity $\nabla^2\tilde{u}^k = \nabla\nabla\cdot\tilde{u}^k - \nabla\times\nabla\times\tilde{u}^k$ and the fact that we are searching for divergence-free solutions, we are led to replace $-\nabla^2\tilde{u}^k$ in (2.3)–(2.4) with $\nabla\times\nabla\times\tilde{u}^k$. The new scheme is as follows:

$$(3.1) \qquad \begin{cases} \frac{1}{2\delta t}(3u^{k+1} - 4\tilde{u}^k + \tilde{u}^{k-1}) + \nabla\times\nabla\times\tilde{u}^k + \nabla p^{k+1} = f(t^{k+1}), \\ \nabla\cdot u^{k+1} = 0, \\ u^{k+1}\cdot n|_\Gamma = 0 \end{cases}$$

and

$$\left\{\begin{array}{l}\frac{3}{2\delta t}(\tilde{u}^{k+1}-u^{k+1})-\nabla^2\tilde{u}^{k+1}-\nabla\times\nabla\times\tilde{u}^k=0,\\ \tilde{u}^{k+1}|_\Gamma=0.\end{array}\right. \tag{3.2}$$

This scheme is hereafter referred to as the rotational form of the velocity-correction algorithm.

Observing from (3.2) that $(\nabla^2\tilde{u}^{k+1}+\nabla\times\nabla\times\tilde{u}^k)\cdot n|_\Gamma=0$, we derive from (3.1) that

$$\left.\frac{\partial p^{k+1}}{\partial n}\right|_\Gamma=(f(t^{k+1})+\nabla^2\tilde{u}^{k+1})\cdot n|_\Gamma, \tag{3.3}$$

which, unlike (2.6), is a consistent Neumann boundary condition for the pressure. This is the main reason why the rotational form yields a much better pressure approximation than the standard form.

3.2. Implementation of the rotational form. As in the standard form of the method, the projection step (3.1) cannot be solved as a weak Poisson problem when working with H^1-conformal finite elements, for there is a second derivative in the right-hand side. This difficulty can be solved once more by making algebraic substitutions.

By subtracting step (3.1) at time t^k from step (3.1) at time t^{k+1} and by substituting step (3.2) at time t^k into the resulting equation, a more adequate form of the projection step is obtained:

$$\left\{\begin{array}{l}\frac{1}{2\delta t}(3u^{k+1}-7\tilde{u}^k+5\tilde{u}^{k-1}-\tilde{u}^{k-2})+\nabla(p^{k+1}-p^k+\nabla\cdot\tilde{u}^k)\\ \qquad=f(t^{k+1})-f(t^k),\\ \nabla\cdot u^{k+1}=0,\\ u^{k+1}\cdot n|_\Gamma=0.\end{array}\right. \tag{3.4}$$

The new viscous velocity $\tilde{u}^{k+1}$ is evaluated by solving

$$\left\{\begin{array}{l}\frac{1}{2\delta t}(3\tilde{u}^{k+1}-4\tilde{u}^k+\tilde{u}^{k-1})-\nabla^2\tilde{u}^{k+1}+\nabla p^{k+1}=f(t^{k+1}),\\ \tilde{u}^{k+1}|_\Gamma=0.\end{array}\right. \tag{3.5}$$

Note that the new form of the projection step is still not satisfactory since a second derivative remains in the form of the term $\nabla\nabla\cdot\tilde{u}^k$. To remove this final difficulty, we introduce an auxiliary pressure $\phi^{k+1}=p^{k+1}-p^k+\nabla\cdot\tilde{u}^k$. The final algorithm is as follows:

$$\left\{\begin{array}{l}\frac{1}{2\delta t}(3u^{k+1}-7\tilde{u}^k+5\tilde{u}^{k-1}-\tilde{u}^{k-2})+\nabla\phi^{k+1}=f(t^{k+1})-f(t^k),\\ \nabla\cdot u^{k+1}=0,\\ u^{k+1}\cdot n|_\Gamma=0,\end{array}\right. \tag{3.6}$$

$$p^{k+1}=\phi^{k+1}+p^k-\nabla\cdot\tilde{u}^k, \tag{3.7}$$

$$\left\{\begin{array}{l}\frac{1}{2\delta t}(3\tilde{u}^{k+1}-4\tilde{u}^k+\tilde{u}^{k-1})-\nabla^2\tilde{u}^{k+1}=f(t^{k+1})-\nabla p^{k+1},\\ \tilde{u}^{k+1}|_\Gamma=0.\end{array}\right. \tag{3.8}$$

In practice, the projection step is processed as follows:

$$\left\{\begin{array}{l}\text{Find } \phi^{k+1} \text{ in } H^1(\Omega)/\mathbb{R} \text{ such that } \forall q \text{ in } H^1(\Omega)/\mathbb{R},\\ (\nabla\phi^{k+1},\nabla q)=(f(t^{k+1})-f(t^k)+\frac{1}{2\delta t}(7\tilde{u}^k-5\tilde{u}^{k-1}+\tilde{u}^{k-2}),\nabla q).\end{array}\right. \tag{3.9}$$

Note once again that the projected velocity u^{k+1} has been eliminated from the algorithm.

3.3. A time continuous version. We emphasize that it is informative to study the time continuous version of the scheme, since it both reveals the behavior of the splitting error and indicates the procedure to follow for obtaining stability and convergence results on the discrete system.

By neglecting some small terms, the following can be considered an "approximate" time continuous version of the scheme (3.6)–(3.8):

$$\partial_t u^\varepsilon - \nabla^2 u^\varepsilon + \nabla p^\varepsilon = f, \qquad u^\varepsilon|_\Gamma = 0, \tag{3.10}$$

$$\nabla\cdot u^\varepsilon - \varepsilon\nabla\cdot(\nabla\phi - \varepsilon f_t) = 0, \qquad (\nabla\phi - \varepsilon f_t)\cdot n|_\Gamma = 0, \tag{3.11}$$

$$\phi = \varepsilon p_t^\varepsilon + \nabla\cdot u^\varepsilon, \tag{3.12}$$

with $u^\varepsilon(0) = \mathsf{u}(0)$ and $p^\varepsilon(0) = \mathsf{p}(0)$. Note that (3.10) and (3.12) correspond, respectively, to (3.8) and (3.7), while (3.11) corresponds to the divergence of (3.6), and $\varepsilon \sim \Delta t$.

Without going into the full details of proving the well-posedness of (3.10)–(3.12) and providing a detailed error analysis as we did for (2.10)–(2.12), we just indicate how to derive the first a priori estimate. This will guide us to prove the stability of the discrete scheme and will show that this scheme provides a better control on the divergence of the approximate velocity.

Taking the inner product of $\varepsilon u_t^\varepsilon$ with the time derivative of (3.10), we find

$$\begin{aligned}\frac{\varepsilon}{2}d_t\|u_t^\varepsilon\|_{0,\Omega}^2 + \varepsilon\|\nabla u_t^\varepsilon\|_{0,\Omega}^2 &= \varepsilon(u_t^\varepsilon, f_t) + \varepsilon(p_t^\varepsilon, \nabla\cdot u_t^\varepsilon)\\ &= \varepsilon(u_t^\varepsilon, f_t) - (\nabla\cdot u^\varepsilon - \phi, \nabla\cdot u_t^\varepsilon)\\ &= \varepsilon(u_t^\varepsilon, f_t) - \frac{1}{2}d_t\|\nabla\cdot u^\varepsilon\|_{0,\Omega}^2 + (\phi, \nabla\cdot u_t^\varepsilon).\end{aligned}$$

Noting that

$$(\phi, \nabla\cdot u_t^\varepsilon) = (\phi, \varepsilon\nabla^2\phi_t - \varepsilon^2\nabla\cdot f_t) = -\frac{\varepsilon}{2}d_t\|\nabla\phi\|_{0,\Omega}^2 + \varepsilon^2(\nabla\phi, f_t),$$

we obtain

$$\frac{\varepsilon}{2}d_t\|u_t^\varepsilon\|_{0,\Omega}^2 + \varepsilon\|\nabla u_t^\varepsilon\|_{0,\Omega}^2 + \frac{1}{2}d_t\|\nabla\cdot u^\varepsilon\|_{0,\Omega}^2 + \frac{\varepsilon}{2}d_t\|\nabla\phi\|_{0,\Omega}^2 = \varepsilon(u_t^\varepsilon, f_t) + \varepsilon^2(\nabla\phi, f_t).$$

Using the fact that the initial data are such that $u_t^\varepsilon(0) = f(0) + \nabla^2 u^\varepsilon(0) - \nabla p^\varepsilon(0) = f(0) + \nabla^2\mathsf{u}(0) - \nabla\mathsf{p}(0)$, the Gronwall lemma yields

$$\|u_t^\varepsilon(t)\|_{0,\Omega}^2 + \|\nabla\phi(t)\|_{0,\Omega}^2 + \frac{1}{\varepsilon}\|\nabla\cdot u^\varepsilon(t)\|_{0,\Omega}^2 + \int_0^t \|\nabla u_t^\varepsilon\|_{0,\Omega}^2 ds \le c, \qquad t \in [0,T].$$

Let us define $e = \mathsf{u} - u^\varepsilon$ and $\psi = \varepsilon\partial_t(\mathsf{p} - p^\varepsilon) + \nabla\cdot u^\varepsilon$. By working with the error equation, the above results become

$$\|e_t(t)\|_{0,\Omega}^2 + \|\nabla\psi(t)\|_{0,\Omega}^2 + \frac{1}{\varepsilon}\|\nabla\cdot u^\varepsilon(t)\|_{0,\Omega}^2 + \int_0^t \|\nabla e_t\|_{0,\Omega}^2 ds \le c\varepsilon^2, \qquad t \in [0,T].$$

A remarkable consequence, which is essential for obtaining improved error estimates, is that we have

$$\|\nabla\cdot u^\varepsilon\|_{L^\infty(L^2)} \leq c\varepsilon^{\frac{3}{2}}. \tag{3.13}$$

3.4. Error analysis. We now turn our attention to the error analysis of the discrete scheme (3.1)–(3.2). The main result in this section is the following.

THEOREM 3.1. *Under the initialization hypothesis* (H)*, if* (u,p)*, the solution to* (1.1)*, is smooth enough in time and space, the solution* $(u^k,\tilde{u}^k,p^k)$ *to* (3.1)–(3.2) *satisfies the estimates*

$$\|\mathsf{u}-u\|_{l^2(L^2(\Omega)^d)} + \|\mathsf{u}-\tilde{u}\|_{l^2(L^2(\Omega)^d)} \leq c(\mathsf{u},\mathsf{p},T)\,\delta t^2,$$
$$\|\mathsf{u}-\tilde{u}\|_{l^2(H^1(\Omega)^d)} + \|\mathsf{p}-p\|_{l^2(L^2(\Omega))} \leq c(\mathsf{u},\mathsf{p},T)\,\delta t^{3/2}.$$

The remainder of this section is devoted to the proof of the above theorem. Let us introduce some notation. For any sequence $\phi^0,\phi^1,\dots$, we set

$$\delta_t\phi^k = \phi^k - \phi^{k-1}, \quad \delta_{tt}\phi^k = \delta_t(\delta_t\phi^k), \quad \delta_{ttt}\phi^k = \delta_t(\delta_{tt}\phi^k).$$

For any sequence of functions in $H_0^1(\Omega)^d \cap H^2(\Omega)^d$, say $\phi^0,\phi^1,\dots$, we set

$$D_t\phi^k = -\nabla^2\phi^k - \nabla\times\nabla\times\phi^{k-1}.$$

We shall make use of the following identity:

$$\forall\beta\in H_0^1(\Omega)^d, \qquad (D_t\phi^{k+1},\beta) = (\nabla\cdot\phi^{k+1},\nabla\cdot\beta) + (\nabla\times\delta_t\phi^{k+1},\nabla\times\beta). \tag{3.14}$$

Hereafter we shall make use of the following notation:

$$\left\{\begin{array}{rcl} e^k &=& \mathsf{u}(t^k) - u^k, \\ \tilde{e}^k &=& \mathsf{u}(t^k) - \tilde{u}^k, \\ \tilde{\psi}^k &=& \mathsf{u}(t^{k+1}) - \tilde{u}^k, \\ \epsilon^k &=& \mathsf{p}(t^k) - p^k. \end{array}\right. \tag{3.15}$$

The proof of Theorem 3.1 will be carried out through a sequence of estimates presented below.

3.4.1. Stability and the improved estimate on $\|\nabla\cdot\tilde{u}^k\|_{0,\Omega}$.

LEMMA 3.1. *Provided that the solution of* (1.1) *is smooth enough in space and time and satisfies the initialization hypothesis* (H)*, then we have the following error estimates:*

$$\|\nabla\cdot\tilde{u}\|_{l^\infty(L^2(\Omega)^d)} \leq c(\mathsf{u},\mathsf{p},T)\,\delta t^{3/2},$$
$$\|\tilde{e}-e\|_{l^\infty(L^2(\Omega)^d)} \leq c(\mathsf{u},\mathsf{p},T)\,\delta t^2,$$
$$\|\delta_t\tilde{e}-\delta_t e\|_{l^2(L^2(\Omega)^d)} \leq c(\mathsf{u},\mathsf{p},T)\,\delta t^{5/2}.$$

Proof. The proof of this lemma follows the procedure set out in section 3.3 for the time continuous counterpart of the scheme. The critical step here consists of working with the time increments $\delta_t e^{k+1}$ and $\delta_t\tilde{e}^{k+1}$, which corresponds to taking the inner product of $\varepsilon\partial_t u^\varepsilon$ with the time derivative of (3.10).

Step 1: Let us first write the equations that control the time increments of the errors. By defining $R^k = \partial_t \mathsf{u}(t^k) - (3\mathsf{u}(t^k) - 4\mathsf{u}(t^{k-1}) + \mathsf{u}(t^{k-2}))/2\delta t$, we have for $k \geq 2$

$$\left\{\begin{array}{l} \frac{1}{2\delta t}(3\delta_t e^{k+1} - 4\delta_t \tilde{e}^k + \delta_t \tilde{e}^{k-1}) + D_t\tilde{\psi}^k + \nabla(\delta_t \epsilon^{k+1} + \nabla\cdot\tilde{u}^k) = \delta_t R^{k+1}, \\ \nabla\cdot\delta_t e^{k+1} = 0, \\ \delta_t e^{k+1}\cdot n|_\Gamma = 0, \end{array}\right. \tag{3.16}$$

$$\left\{\begin{array}{l} \frac{3}{2\delta t}\delta_t \tilde{e}^{k+1} + D_t\tilde{e}^{k+1} = \frac{3}{2\delta t}\delta_t e^{k+1} + D_t\tilde{\psi}^k, \\ \tilde{e}^{k+1}|_\Gamma = 0. \end{array}\right. \tag{3.17}$$

Step 2: Let us multiply (3.16) by $4\delta t\delta_t e^{k+1}$ and integrate over Ω. We obtain

$$\begin{aligned} 2(\delta_t e^{k+1}, 3\delta_t e^{k+1} - 4\delta_t \tilde{e}^k + \delta_t \tilde{e}^{k-1}) + 4\delta t(\delta_t e^{k+1}, D_t\tilde{\psi}^k) &= 4\delta t(\delta_t e^{k+1}, \delta_t R^{k+1}) \\ &\leq 4\delta t(\|\delta_t e^{k+1} - \delta_t\tilde{e}^{k+1}\|_{0,\Omega} + \|\delta_t\tilde{e}^{k+1}\|_{0,\Omega})\|\delta_t R^{k+1}\|_{0,\Omega} \\ &\leq \delta t\|\delta_t\tilde{e}^{k+1}\|^2_{1,\Omega} + \delta t\|\delta_t e^{k+1} - \delta_t\tilde{e}^{k+1}\|^2_{0,\Omega} + c\delta t^7, \end{aligned}$$

where we have used the Poincaré inequality and the fact that $\|\delta_t R^{k+1}\|_{0,\Omega} \leq c\delta t^3$. Note also that we have used the inequality $2ab \leq \gamma a^2 + b^2/\gamma$, which holds for all $\gamma > 0$. We shall repeatedly use this standard trick hereafter without mentioning it anymore.

Since the treatment of the approximate time derivative is quite involved, we show the details. Let us define

$$\begin{aligned} I &= 2(\delta_t e^{k+1}, 3\delta_t e^{k+1} - 4\delta_t\tilde{e}^k + \delta_t\tilde{e}^{k-1}) \\ &= 6(\delta_t e^{k+1}, \delta_t e^{k+1} - \delta_t\tilde{e}^{k+1}) + 2(\delta_t e^{k+1} - \delta_t\tilde{e}^{k+1}, 3\delta_t\tilde{e}^{k+1} - 4\delta_t\tilde{e}^k + \delta_t\tilde{e}^{k-1}) \\ &\quad + 2(\delta_t\tilde{e}^{k+1}, 3\delta_t\tilde{e}^{k+1} - 4\delta_t\tilde{e}^k + \delta_t\tilde{e}^{k-1}) \end{aligned}$$

and denote by I_1, I_2, and I_3 the three terms in the right-hand side. Owing to the standard identities

$$\begin{aligned} 2(a^{k+1}, a^{k+1} - a^k) &= |a^{k+1}|^2 + |a^{k+1} - a^k|^2 - |a^k|^2, \\ 2(a^{k+1}, 3a^{k+1} - 4a^k + a^{k-1}) &= |a^{k+1}|^2 + |2a^{k+1} - a^k|^2 + |\delta_{tt}a_{k+1}|^2 \\ &\quad - |a^k|^2 - |2a^k - a^{k-1}|^2, \end{aligned} \tag{3.18}$$

we deduce

$$\begin{aligned} I_1 &= 3\|\delta_t e^{k+1}\|^2_{0,\Omega} + 3\|\delta_t e^{k+1} - \delta_t\tilde{e}^{k+1}\|^2_{0,\Omega} - 3\|\delta_t\tilde{e}^{k+1}\|^2_{0,\Omega}, \\ I_3 &= \|\delta_t\tilde{e}^{k+1}\|^2_{0,\Omega} + \|2\delta_t\tilde{e}^{k+1} - \delta_t\tilde{e}^k\|^2_{0,\Omega} + \|\delta_{ttt}\tilde{e}^{k+1}\|^2_{0,\Omega} \\ &\quad - \|\delta_t\tilde{e}^k\|^2_{0,\Omega} - \|2\delta_t\tilde{e}^k - \delta_t\tilde{e}^{k-1}\|^2_{0,\Omega}. \end{aligned}$$

For the remaining term I_2, we make use of (3.17) as follows:

$$\frac{3}{2\delta t}I_2 = 2(D_t\tilde{e}^{k+1} - D_t\tilde{\psi}^k, 3\delta_t\tilde{e}^{k+1} - 4\delta_t\tilde{e}^k + \delta_t\tilde{e}^{k-1}).$$

Using the relation $\tilde{\psi}^k = \delta_t\mathsf{u}(t^{k+1}) + \tilde{e}^k$, we obtain

$$\begin{aligned} \frac{3}{2\delta t}I_2 &= 2(D_t\delta_t\tilde{e}^{k+1}, 3\delta_t\tilde{e}^{k+1} - 4\delta_t\tilde{e}^k + \delta_t\tilde{e}^{k-1}) \\ &\quad - 2(D_t\delta_t\mathsf{u}(t^{k+1}), 3\delta_t\tilde{e}^{k+1} - 4\delta_t\tilde{e}^k + \delta_t\tilde{e}^{k-1}). \end{aligned}$$

By denoting as I_{21} and I_{22} the two terms in the right-hand side, and by using the identities (3.14) and (3.18), we infer

$$\begin{aligned}
I_{21} &= 2(\nabla\cdot\delta_t\tilde{e}^{k+1}, \nabla\cdot(3\delta_t\tilde{e}^{k+1} - 4\delta_t\tilde{e}^{k} + \delta_t\tilde{e}^{k-1}))\\
&\quad + 2(\nabla\times\delta_{tt}\tilde{e}^{k+1}, 3\nabla\times\delta_{tt}\tilde{e}^{k+1} - \nabla\times\delta_{tt}\tilde{e}^{k})\\
&= \|\nabla\cdot\delta_t\tilde{e}^{k+1}\|^2_{0,\Omega} + \|\nabla\cdot(2\delta_t\tilde{e}^{k+1} - \delta_t\tilde{e}^{k})\|^2_{0,\Omega} + \|\nabla\cdot(\delta_{ttt}\tilde{e}^{k+1})\|^2_{0,\Omega}\\
&\quad - \|\nabla\cdot\delta_t\tilde{e}^{k}\|^2_{0,\Omega} - \|\nabla\cdot(2\delta_t\tilde{e}^{k} - \delta_t\tilde{e}^{k-1})\|^2_{0,\Omega} + 3\|\nabla\times\delta_{tt}\tilde{e}^{k+1}\|^2_{0,\Omega}\\
&\quad + \frac{1}{3}\|\nabla\times(3\delta_t\tilde{e}^{k+1} - 4\delta_t\tilde{e}^{k} + \delta_t\tilde{e}^{k-1})\|^2_{0,\Omega} - \frac{1}{3}\|\nabla\times\delta_{tt}\tilde{e}^{k}\|^2_{0,\Omega},\\
I_{22} &= -\,2(\nabla\times\delta_{tt}\mathsf{u}(t^{k+1}), \nabla\times(3\delta_t\tilde{e}^{k+1} - 4\delta_t\tilde{e}^{k} + \delta_t\tilde{e}^{k-1}))\\
&\geq -\,c\delta t^4 - \frac{1}{6}\|\nabla\times(3\delta_t\tilde{e}^{k+1} - 4\delta_t\tilde{e}^{k} + \delta_t\tilde{e}^{k-1})\|^2_{0,\Omega}.
\end{aligned}$$

By combining all the results above, we deduce the following bound:

$$\begin{aligned}
3\|\delta_t e^{k+1}\|^2_{0,\Omega} &- 3\|\delta_t\tilde{e}^{k+1}\|^2_{0,\Omega} + 3(1-\delta t)\|\delta_t e^{k+1} - \delta_t\tilde{e}^{k+1}\|^2_{0,\Omega}\\
&+ \|\delta_t\tilde{e}^{k+1}\|^2_{0,\Omega} + \|2\delta_t\tilde{e}^{k+1} - \delta_t\tilde{e}^{k}\|^2_{0,\Omega} + \|\delta_{ttt}\tilde{e}^{k+1}\|^2_{0,\Omega}\\
&+ \frac{2\delta t}{3}\Big(\|\nabla\cdot\delta_t\tilde{e}^{k+1}\|^2_{0,\Omega} + \|\nabla\cdot(2\delta_t\tilde{e}^{k+1} - \delta_t\tilde{e}^{k})\|^2_{0,\Omega} + \|\nabla\cdot(\delta_{ttt}\tilde{e}^{k+1})\|^2_{0,\Omega}\\
&\qquad + 3\|\nabla\times\delta_{tt}\tilde{e}^{k+1}\|^2_{0,\Omega} + \frac{1}{6}\|\nabla\times(3\delta_t\tilde{e}^{k+1} - 4\delta_t\tilde{e}^{k} + \delta_t\tilde{e}^{k-1})\|^2_{0,\Omega}\Big)\\
&+ 4\delta t(\delta_t e^{k+1}, D_t\tilde{\psi}^k)\\
&\leq \ \delta t\|\delta_t\tilde{e}^{k+1}\|^2_{1,\Omega} + \|\delta_t\tilde{e}^{k}\|^2_{0,\Omega} + \|2\delta_t\tilde{e}^{k} - \delta_t\tilde{e}^{k-1}\|^2_{0,\Omega}\\
&+ \frac{2\delta t}{3}\left(\|\nabla\cdot\delta_t\tilde{e}^{k}\|^2_{0,\Omega} + \|\nabla\cdot(2\delta_t\tilde{e}^{k} - \delta_t\tilde{e}^{k-1})\|^2_{0,\Omega} + \frac{1}{3}\|\nabla\times\delta_{tt}\tilde{e}^{k}\|^2_{0,\Omega}\right)\\
&+ c\delta t^5.
\end{aligned}$$

Step 3: By taking the square of (3.17), multiplying the result by $\frac{4}{3}\delta t^2$, and integrating over the domain, we have

$$\begin{aligned}
3\|\delta_t\tilde{e}^{k+1}\|^2_{0,\Omega} &+ 4\delta t(\delta_t\tilde{e}^{k+1}, D_t\tilde{e}^{k+1}) + \frac{4\delta t^2}{3}\|D_t\tilde{e}^{k+1}\|^2_{0,\Omega}\\
&= 3\|\delta_t e^{k+1}\|^2_{0,\Omega} + 4\delta t(\delta_t e^{k+1}, D_t\tilde{\psi}^k) + \frac{4\delta t^2}{3}\|D_t\tilde{\psi}^k\|^2_{0,\Omega}.
\end{aligned}$$

Owing to (3.14), we deduce

$$\begin{aligned}
3\|\delta_t\tilde{e}^{k+1}\|^2_{0,\Omega} &- 3\|\delta_t e^{k+1}\|^2_{0,\Omega} + 2\delta t\|\nabla\cdot\tilde{e}^{k+1}\|^2_{0,\Omega}\\
&+ 2\delta t\|\nabla\cdot\delta_t\tilde{e}^{k+1}\|^2_{0,\Omega} + 4\delta t\|\nabla\times\delta_t\tilde{e}^{k+1}\|^2_{0,\Omega} + \frac{4\delta t^2}{3}\|D_t\tilde{e}^{k+1}\|^2_{0,\Omega}\\
&= 2\delta t\|\nabla\cdot\tilde{e}^{k}\|^2_{0,\Omega} + 4\delta t(\delta_t e^{k+1}, D_t\tilde{\psi}^k) + \frac{4\delta t^2}{3}\|D_t\tilde{\psi}^k\|^2_{0,\Omega}.
\end{aligned}$$

A control on $\|D_t\tilde{\psi}^k\|^2_{0,\Omega}$ is obtained as follows:

$$\begin{aligned}
\|D_t\tilde{\psi}^k\|^2_{0,\Omega} &\leq \big(\|D_t\delta_t\mathsf{u}(t^{k+1})\|_{0,\Omega} + \|D_t\tilde{e}^k\|_{0,\Omega}\big)^2\\
&\leq \big(c\delta t^2 + \|D_t\tilde{e}^k\|_{0,\Omega}\big)^2 = c^2\delta t^4 + 2\delta t c\delta t\|D_t\tilde{e}^k\|_{0,\Omega} + \|D_t\tilde{e}^k\|^2_{0,\Omega}\\
&\leq c^2\delta t^4 + \delta t\big(c^2\delta t^2 + \|D_t\tilde{e}^k\|^2_{0,\Omega}\big) + \|D_t\tilde{e}^k\|^2_{0,\Omega}\\
&\leq c\delta t^3 + (1+\delta t)\|D_t\tilde{e}^k\|^2_{0,\Omega}.
\end{aligned}$$

Note that it is at this very point that the splitting error spoils the optimality. Finally, we have

$$\begin{aligned} 3\|\delta_t\tilde{e}^{k+1}\|^2_{0,\Omega} - 3\|\delta_t e^{k+1}\|^2_{0,\Omega} + 2\delta t\|\nabla\cdot\tilde{e}^{k+1}\|^2_{0,\Omega} + 2\delta t\|\delta_t\tilde{e}^{k+1}\|^2_{1,\Omega} + \frac{4\delta t^2}{3}\|D_t\tilde{e}^{k+1}\|^2_{0,\Omega} \\ \leq 2\delta t\|\nabla\cdot\tilde{e}^{k}\|^2_{0,\Omega} + 4\delta t(\delta_t e^{k+1}, D_t\tilde{\psi}^k) + \frac{4\delta t^2}{3}(1+\delta t)\|D_t\tilde{e}^{k}\|^2_{0,\Omega} + c\delta t^5. \end{aligned}$$

Step 4: By combining the bounds obtained at Steps 2 and 3, and by dropping some nonessential positive terms on the left-hand side, we finally deduce

$$\begin{aligned} \|\delta_t\tilde{e}^{k+1}\|^2_{0,\Omega} &+ \|2\delta_t\tilde{e}^{k+1} - \delta_t\tilde{e}^{k}\|^2_{0,\Omega} + 2\delta t\|\nabla\cdot\tilde{e}^{k+1}\|^2_{0,\Omega} + \frac{4\delta t^2}{3}\|D_t\tilde{e}^{k+1}\|^2_{0,\Omega} \\ &+ \frac{2\delta t}{3}\left(\|\nabla\cdot\delta_t\tilde{e}^{k+1}\|^2_{0,\Omega} + \|\nabla\cdot(2\delta_t\tilde{e}^{k+1} - \delta_t\tilde{e}^{k})\|^2_{0,\Omega} + \frac{1}{3}\|\nabla\times\delta_{tt}\tilde{e}^{k+1}\|^2_{0,\Omega}\right) \\ &+ 3(1-\delta t)\|\delta_t e^{k+1} - \delta_t\tilde{e}^{k+1}\|^2_{0,\Omega} + \delta t\|\delta_t\tilde{e}^{k+1}\|^2_{1,\Omega} \\ \leq{}& \|\delta_t\tilde{e}^{k}\|^2_{0,\Omega} + \|2\delta_t\tilde{e}^{k} - \delta_t\tilde{e}^{k-1}\|^2_{0,\Omega} + 2\delta t\|\nabla\cdot\tilde{e}^{k}\|^2_{0,\Omega} + (1+\delta t)\frac{4\delta t^2}{3}\|D_t\tilde{e}^{k}\|^2_{0,\Omega} \\ &+ \frac{2\delta t}{3}\left(\nabla\cdot\delta_t\tilde{e}^{k^2} + \|\nabla\cdot(2\delta_t\tilde{e}^{k} - \delta_t\tilde{e}^{k-1})\|^2_{0,\Omega} + \frac{1}{3}\|\nabla\times\delta_{tt}\tilde{e}^{k}\|^2_{0,\Omega}\right) \\ &+ c\delta t^5. \end{aligned}$$

By applying the discrete Gronwall lemma and using the initialization hypothesis (H), we infer

$$\begin{aligned} \delta t\|\nabla\cdot\tilde{e}^{k+1}\|^2_{0,\Omega} + \delta t^2\|D_t\tilde{e}^{k+1}\|^2_{0,\Omega} + \sum_{l=0}^{k}\|\delta_t e^l - \delta_t\tilde{e}^l\|^2_{0,\Omega} \\ \leq c(\|\tilde{e}^2\|^2_{0,\Omega} + \delta t\|\tilde{e}^2\|^2_{1,\Omega} + \delta t^2\|\tilde{e}^2\|^2_{2,\Omega} + \delta t^4). \end{aligned}$$

Thanks to (H), it is an easy matter to show directly that

$$\|\tilde{e}^2\|^2_{0,\Omega} + \delta t\|\tilde{e}^2\|^2_{1,\Omega} + \delta t^2\|\tilde{e}^2\|^2_{2,\Omega} + \delta t^4 \leq c\delta t^4.$$

Finally, noticing that

$$\begin{aligned} \frac{3}{2}\|\tilde{e}^{k+1} - e^{k+1}\|_{0,\Omega} &= \delta t\|D_t\tilde{e}^{k+1} + \nabla\times\nabla\times\delta_t \mathsf{u}(t^{k+1})\|_{0,\Omega} \\ &\leq \delta t\|D_t\tilde{e}^{k+1}\|_{0,\Omega} + c\delta t^2, \end{aligned}$$

the desired result follows from the last three inequalities. □

Remark 3.1. The first result in the above lemma, namely, $\|\nabla\cdot\tilde{u}\|_{l^\infty(L^2)} \leq c\delta t^{\frac{3}{2}}$, is the key for obtaining error estimates that improve on those from the standard velocity-correction scheme. A remarkable property of the rotational velocity-correction scheme is that even if the time stepping in (3.1)–(3.2) is replaced by the first-order backward Euler stepping, the estimate on $\nabla\cdot\tilde{u}$ still holds.

3.4.2. The inverse of the Stokes operator. In this section we recall properties of the inverse of the Stokes operator that will be useful for proving estimates in the L^2-norm. This operator, which we shall denote by $S : H^{-1}(\Omega)^d \longrightarrow V$, is defined as follows. For all v in $H^{-1}(\Omega)^d$, $S(v) \in V$ is the solution to the following problem:

$$\begin{cases} (\nabla S(v), \nabla w) - (r, \nabla\cdot w) = \langle v, w\rangle \qquad \forall w \in H^1_0(\Omega)^d, \\ (q, \nabla\cdot S(v)) = 0 \qquad \forall q \in L^2_0(\Omega), \end{cases}$$

where $\langle\cdot,\cdot\rangle$ denotes the duality pairing between $H^{-1}(\Omega)^d$ and $H_0^1(\Omega)^d$. Obviously, we have

$$\forall v \in H^{-1}(\Omega)^d, \qquad \|S(v)\|_{1,\Omega} + \|r\|_{0,\Omega} \leq c\|v\|_{-1,\Omega}. \tag{3.19}$$

We shall assume hereafter that the domain Ω is such that the following regularity property holds:

$$\forall v \in L^2(\Omega)^d, \qquad \|S(v)\|_{2,\Omega} + \|r\|_{1,\Omega} \leq c\|v\|_{0,\Omega}. \tag{3.20}$$

The operator S has interesting properties, as listed below.

LEMMA 3.2. *For all v in $H_0^1(\Omega)^d$ and all $0 < \gamma < 1$ we have*

$$\forall v^\star \in H, \qquad (\nabla S(v), \nabla v) \geq (1-\gamma)\|v\|_{0,\Omega}^2 - c(\gamma)\|v - v^\star\|_{0,\Omega}^2.$$

In particular,

$$\forall v \in V, \qquad (\nabla S(v), \nabla v) = \|v\|_{0,\Omega}^2.$$

Proof. Owing to the definition of $S(v)$, we have

$$\begin{aligned}
(\nabla S(v), \nabla v) &= (r, \nabla\cdot v) + \|v\|_{0,\Omega}^2 \\
&= (r, \nabla\cdot(v - v^\star)) + \|v\|_{0,\Omega}^2 && \forall v^\star \in H \\
&= -(\nabla r, v - v^\star) + \|v\|_{0,\Omega}^2 \\
&\geq -\|r\|_{1,\Omega}\|v - v^\star\|_{0,\Omega} + \|v\|_{0,\Omega}^2 \\
&\geq -c(\gamma)\|v - v^\star\|_{0,\Omega}^2 + (1-\gamma)\|v\|_{0,\Omega}^2, && \text{owing to (3.20).}
\end{aligned}$$

This completes the proof. □

LEMMA 3.3. *The bilinear form*

$$H^{-1}(\Omega)^d \times H^{-1}(\Omega)^d \ni (v,w) \longmapsto \langle S(v), w\rangle := (\nabla S(v), \nabla S(w)) \in \mathbb{R}$$

induces a seminorm on $H^{-1}(\Omega)^d$ that we denote $|\cdot|_\star$, and

$$\forall v \in H^{-1}(\Omega)^d, \qquad |v|_\star = \|\nabla S(v)\|_{0,\Omega} \leq c\|v\|_{-1,\Omega}.$$

Proof. It is clear that it is symmetric $\langle S(v), w\rangle = (\nabla S(v), \nabla S(w)) = \langle S(w), v\rangle$ and positive $\langle S(v), v\rangle = \|\nabla S(v)\|_{0,\Omega}^2$; hence, $\langle S(v), w\rangle$ induces a seminorm on $H^{-1}(\Omega)^d$. Furthermore,

$$|v|_\star^2 = \langle S(v), v\rangle = (\nabla S(v), \nabla S(v)) = \|\nabla S(v)\|_{0,\Omega}^2 \leq c\|v\|_{-1,\Omega}^2.$$

The proof is complete. □

3.4.3. Proof of the L^2-estimate on the velocity. In this subsection we prove

$$\|\mathsf{u} - u\|_{l^2(L^2(\Omega)^d)} \leq c\delta t^2.$$

Proof. We begin by reconstructing the momentum equation at time t^{k+1} by adding the projection step to the viscous step. In terms of the errors, we obtain

$$\frac{3\tilde{e}^{k+1} - 4\tilde{e}^k + \tilde{e}^{k-1}}{2\delta t} - \nabla^2\tilde{e}^{k+1} + \nabla\epsilon^{k+1} = R^{k+1}. \tag{3.21}$$

By taking the L^2 scalar product with $4\delta t S(\tilde{e}^{k+1})$, we obtain

$$|\tilde{e}^{k+1}|_\star^2 + |2\tilde{e}^{k+1} - \tilde{e}^k|_\star^2 + |\delta_{tt}\tilde{e}^{k+1}|_\star^2 + 4\delta t(\nabla S(\tilde{e}^{k+1}), \nabla \tilde{e}^{k+1}) \\ = 4\delta t(R^{k+1}, S(\tilde{e}^{k+1})) + |\tilde{e}^k|_\star^2 + |2\tilde{e}^k - \tilde{e}^{k-1}|_\star^2.$$

Owing to Lemma 3.2 and the fact that e^{k+1} is in H, we infer

$$4\delta t(\nabla S(\tilde{e}^{k+1}), \nabla \tilde{e}^{k+1}) \geq 2\delta t\|\tilde{e}^{k+1}\|_{0,\Omega}^2 - c\delta t\|\tilde{e}^{k+1} - e^{k+1}\|_{0,\Omega}^2.$$

Thanks to (3.20), we have

$$4\delta t(R^{k+1}, S(\tilde{e}^{k+1})) \leq c\delta t\|R^{k+1}\|_{0,\Omega}^2 + \delta t\|\tilde{e}^{k+1}\|_{0,\Omega}.$$

As a result, we obtain

$$|\tilde{e}^{k+1}|_\star^2 + |2\tilde{e}^{k+1} - \tilde{e}^k|_\star^2 + |\delta_{tt}\tilde{e}^{k+1}|_\star^2 + \delta t\|\tilde{e}^{k+1}\|_{0,\Omega}^2 \leq c\delta t^5 + c'\delta t\|\tilde{e}^{k+1} - e^{k+1}\|_{0,\Omega}^2 \\ + |\tilde{e}^k|_\star^2 + |2\tilde{e}^k - \tilde{e}^{k-1}|_\star^2.$$

By applying the discrete Gronwall lemma and using the initialization hypothesis, we infer

$$\|e\|_{l^2(L^2(\Omega)^d)}^2 \leq c\|\tilde{e} - e\|_{l^2(L^2(\Omega)^d)}^2 + \delta t^4.$$

The desired result is now an easy consequence of Lemma 3.1. □

3.4.4. Proof of the H^1-estimate on the velocity. First we need to prove an estimate on the approximate time derivative. For any sequence of functions $\phi^0, \phi^1, \ldots,$ we set

$$\mathcal{D}_t\phi^{k+1} = \frac{1}{2}(3\phi^{k+1} - 4\phi^k + \phi^{k-1}).$$

LEMMA 3.4. *Under the hypotheses of Theorem* 3.1 *we have the following error estimates:*

$$\|\mathcal{D}_t\tilde{e}\|_{l^2(L^2(\Omega)^d)} \leq c\delta t^{5/2}.$$

Proof. We use the same argument as for the proof of the L^2-estimate, but we use it on the time increment $\delta_t\tilde{e}^{k+1}$. For $k \geq 2$ we have

$$\frac{1}{2\delta t}(3\delta_t\tilde{e}^{k+1} - 4\delta_t\tilde{e}^k + \delta_t\tilde{e}^{k-1}) - \nabla^2\delta_t\tilde{e}^{k+1} + \nabla\delta_t\epsilon^{k+1} = \delta_t R^{k+1}.$$

By taking the L^2 scalar product with $4\delta t S(\delta_t\tilde{e}^{k+1})$ and repeating the same arguments as above, we obtain

$$|\delta_t\tilde{e}^{k+1}|_\star^2 + |2\delta_t\tilde{e}^{k+1} - \delta_t\tilde{e}^k|_\star^2 + |\delta_{ttt}\tilde{e}^{k+1}|_\star^2 + \delta t\|\delta_t\tilde{e}^{k+1}\|_{0,\Omega}^2 \\ \leq c\delta t^7 + c'\delta t\|\delta_t\tilde{e}^{k+1} - \delta_t e^{k+1}\|_{0,\Omega}^2 + |\delta_t\tilde{e}^k|_\star^2 + |2\delta_t\tilde{e}^k - \delta_t\tilde{e}^{k-1}|_\star^2.$$

Owing to this inequality, the discrete Gronwall lemma, and the initialization hypotheses, we infer

$$\|\delta_t\tilde{e}\|_{l^2(L^2(\Omega)^d)}^2 \leq c\|\delta_t\tilde{e} - \delta_t e\|_{l^2(L^2(\Omega)^d)}^2 + c\delta t^7.$$

The conclusion is an easy consequence of Lemma 3.1 together with the bound

$$\|\mathcal{D}_t\tilde{e}\|_{l^2(L^2(\Omega)^d)} \leq 2\|\delta_t\tilde{e}\|^2_{l^2(L^2(\Omega)^d)}. \qquad \square$$

Now we are in position to prove the H^1-estimate for the velocity approximation and the L^2-estimate for the pressure approximation.

Consider the error equation corresponding to (3.8):

$$\begin{cases} \frac{1}{2\delta t}(3\tilde{e}^{k+1} - 4\tilde{e}^k + \tilde{e}^{k-1}) - \nabla^2\tilde{e}^{k+1} = R^{k+1} - \nabla\epsilon^{k+1}, \\ \tilde{e}^{k+1}|_\Gamma = 0. \end{cases} \tag{3.22}$$

We rewrite the above equation and (3.2) as a nonhomogeneous Stokes system for $(\tilde{e}^{k+1}, \epsilon^{k+1})$:

$$\begin{cases} -\nabla^2\tilde{e}^{k+1} + \nabla\epsilon^{k+1} = h^{k+1}, \quad \tilde{e}^{k+1}|_\Gamma = 0, \\ \nabla\cdot\tilde{e}^{k+1} = g^{k+1}, \end{cases} \tag{3.23}$$

where we have defined

$$\begin{aligned} h^{k+1} &= R^{k+1} - \frac{3e^{k+1} - 4e^k + e^{k-1}}{2\delta t}, \\ g^{k+1} &= -\nabla\cdot\tilde{u}^{k+1}. \end{aligned} \tag{3.24}$$

Owing to Lemma 3.1, we have

$$\|g^{k+1}\|_{0,\Omega} = \|\nabla\cdot\tilde{u}^{k+1}\|_{0,\Omega} \leq c\delta t^{\frac{3}{2}} \qquad \forall k. \tag{3.25}$$

We also have

$$\begin{aligned} \|h^{k+1}\|_{-1,\Omega} &\leq \|R^{k+1}\|_{-1,\Omega} + \left\|\frac{3\tilde{e}^{k+1} - 4\tilde{e}^k + \tilde{e}^{k-1}}{2\delta t}\right\|_{-1,\Omega} \\ &= \|R^{k+1}\|_{-1,\Omega} + \frac{1}{\delta t}\|\mathcal{D}_t\tilde{e}^{k+1}\|_{-1,\Omega}. \end{aligned} \tag{3.26}$$

Now, the standard result for the nonhomogeneous Stokes system (3.23) leads to

$$\|\tilde{e}^{k+1}\|_{1,\Omega} + \|\epsilon^{k+1}\|_{0,\Omega} \leq c\|h^{k+1}\|_{-1,\Omega} + \|g^{k+1}\|_{0,\Omega}. \tag{3.27}$$

Thanks to (3.25), (3.26), and Lemma 3.4, we derive

$$\|\tilde{e}\|_{l^2(H^1(\Omega)^d)} + \|\epsilon\|_{l^2(L^2(\Omega))} \leq c\delta t^{\frac{3}{2}}.$$

Thus, all the results in Theorem 3.1 have been proved.

4. Numerical results.

To test the two versions of the velocity-correction methods described above, we make convergence tests with respect to δt with finite elements and a Legendre spectral approximation.

4.1. Convergence tests with finite elements.

We test the finite element approximation on the Stokes problem (1.1) in $\Omega =]0,1[^2$. We set the source term so that the exact solution is

$$\begin{aligned} \mathsf{p}(x,y,t) &= \cos(\pi x)\sin(\pi y)\sin t, \\ \mathsf{u}(x,y,t) &= \pi\sin(2\pi y)\sin^2(\pi x)\sin t, \\ \mathsf{u}(x,y,t) &= -\pi\sin(2\pi x)\sin^2(\pi y)\sin t. \end{aligned}$$

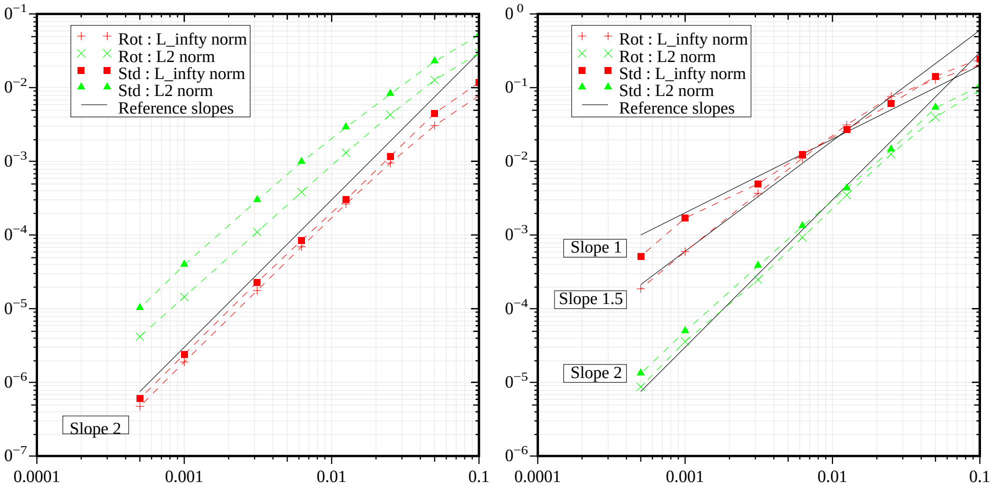


FIG. 1. *Convergence tests for the velocity-correction methods with BDF2 and finite elements. Left: velocity; right: pressure.*

We use mixed $\mathbb{P}_2/\mathbb{P}_1$ finite elements. The mesh used in the tests is composed of 3694 triangles so that the mesh size is $h \approx 1/40$. There are 1928 $\mathbb{P}_1$-nodes and 7549 $\mathbb{P}_2$-nodes. We make the tests on the range $5.10^{-4} \leq \delta t \leq 10^{-1}$ so that the approximation error in space is far smaller than the time splitting error.

We have tested the algorithms (2.3)–(2.4) and (3.1)–(3.2); the results are reported in the Figure 1. In the left panel we show the errors on the velocity in the L^∞- and L^2-norms as functions of δt. The $+$ and $\times$ symbols are for the results from the velocity-correction method in rotational form, whereas the black symbols are for the results from the standard form of the method. It is clear that for the velocity, the improvement brought by the rotational form is marginal and both schemes are second-order accurate in the L^2-norm. Note, however, that for any given δt the results from the rotational form of the algorithm are systematically more accurate than their standard counterparts. The situation is somewhat different for the pressure. The convergence results for this quantity in the L^∞- and L^2-norms are reported in the right panel of Figure 1, the $+$ and $\times$ symbols being for the rotational form of the method and the black symbols for the standard form. The behavior of the errors in the L^2-norm seems to be identical for both variants of the method with a slope slightly less than 2; however, the rotational form results are systematically better than the standard ones. For the L^∞-norm the picture is different. The results from the rotational form seem to behave like $\delta t^{3/2}$, whereas those from the standard form of the algorithm behave more or less like δt.

The difference between the standard form and the rotational form of the velocity-correction algorithm is more spectacular when looking directly at the error fields. We show in Figure 2 the error on the pressure obtained by both algorithms at time $T = 1$ with $\delta t = 0.01$, using the same scale on both graphs to emphasize the differences. It is clear from this picture that the pressure field from the standard method is polluted by a numerical boundary layer, whereas that from the rotational form is smooth.

4.2. Legendre spectral approximation. We have also implemented the second-order standard and rotational velocity-correction schemes with a Legendre–Galerkin approximation [13] using 32×32 modes. We tested the same analytical

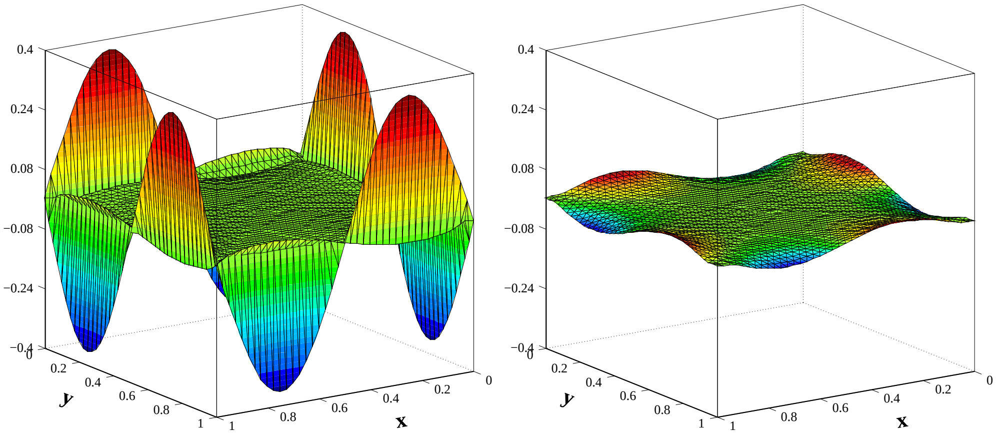


FIG. 2. *Pressure error fields at $T = 1$ with finite elements. Left: standard form; right: rotational form.*

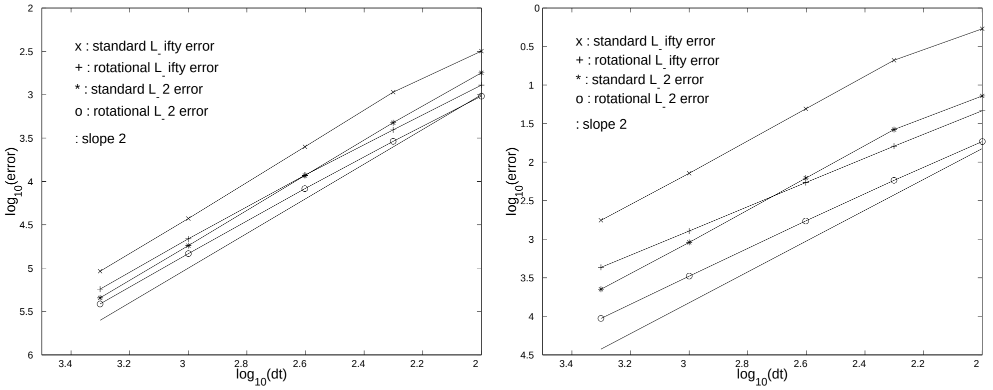


FIG. 3. *Convergence tests for the velocity-correction methods with BDF2 and the Legendre–Galerkin method. Left: velocity; right: pressure.*

solution as above but with $\Omega = \,]-1,+1[^2$. The convergence rates and the pressure error fields are presented in Figures 3 and 4. We observe that the results are similar to those obtained with the finite element approximation and are consistent with our theoretical analysis.

5. Connection with the schemes in [10, 9]. In this section we show how the schemes proposed by Orszag, Israeli, and Deville [10] and Karniadakis, Israeli, and Orszag [9] can be interpreted as the rotational form of our velocity-correction methods.

Let us denote by $\frac{1}{\delta t}(\beta_q u^{k+1} - \sum_{j=0}^{q-1} \beta_j u^{k-j})$ the qth-order BDF approximation for $\partial_t \mathsf{u}(t^{k+1})$. Then, the scheme originally proposed in [10] and [9] (with an Adams–Moulton-type scheme replacing our BDF scheme—note that this replacement is made for the convenience of our presentation only; it does not change the essential error

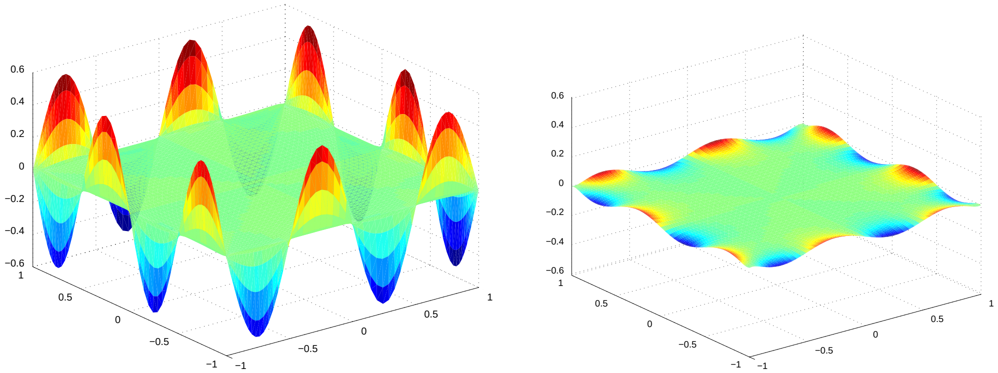


FIG. 4. *Pressure error fields at $T = 1$ with a Legendre–Galerkin approximation. Left: standard form; right: rotational form.*

behaviors) can be written as follows:

$$
(5.1) \qquad \begin{cases} \frac{1}{\delta t}(\beta_q \hat{u}^{k+1} - \sum_{j=0}^{q-1} \beta_j \tilde{u}^{k-j}) + \nabla p^{k+1} = f(t^{k+1}), \\ \nabla\cdot \hat{u}^{k+1} = 0, \\ \hat{u}^{k+1} \cdot n|_\Gamma = -\delta t (\nabla^2 u)^{\star,k} \cdot n|_\Gamma. \end{cases}
$$

We then correct the velocity $\hat{u}^{k+1}$ by computing $\tilde{u}^{k+1}$ as follows:

$$
(5.2) \qquad \begin{cases} \frac{\beta_q}{\delta t}(\tilde{u}^{k+1} - \hat{u}^{k+1}) - \nabla^2 \tilde{u}^{k+1} = 0, \\ \tilde{u}^{k+1}|_\Gamma = 0, \end{cases}
$$

where $(\nabla^2 u)^{\star,k}$ is some approximate value of $\nabla^2 \mathsf{u}(t^{k+1})$. The authors in [10, 9] proposed the followings choices:

$$
(5.3) \qquad (\nabla^2 u)^{\star,k} = \begin{cases} 0 & \text{for } \mathcal{O}(\delta t) \text{ accuracy}, \\ -\nabla\times\nabla\times \tilde{u}^k & \text{for } \mathcal{O}(\delta t^2) \text{ accuracy}, \\ -\nabla\times\nabla\times(2\tilde{u}^k - \tilde{u}^{k-1}) & \text{for } \mathcal{O}(\delta t^3) \text{ accuracy}. \end{cases}
$$

In practice, problem (5.1) is solved as a Poisson equation with the Neumann boundary condition

$$
\partial_n p^{k+1}|_\Gamma = (f(t^{k+1}) + (\nabla^2 u)^{\star,k}) \cdot n.
$$

These methods differ from the standard pressure-correction projection methods in the sense that a consistent pressure boundary condition is enforced. Hence, in principle, these methods should achieve better convergence properties. To the best of our knowledge, no proof of stability or convergence is available in the literature for this class of methods. Furthermore, since second derivatives of the velocity are used in the Neumann boundary condition for the pressure, this class of methods cannot be applied directly with a finite element method where these derivatives are usually not available. This is the main reason why successful implementations of these methods are reported only with spectral or spectral-element approximations where the trace of the second-order derivatives of the velocity are available. On the other hand, the explicit treatment of second derivatives of the velocity leads one to suspect that this type of algorithm can be only conditionally stable, with a stability condition of type $\delta t \leq ch^2$ for finite element approximations and $\delta t \leq cN^{-4}$ for spectral or spectral element approximations.

We shall see in what follows that the boundary condition ambiguity can be removed by rewriting the algorithm in the L^2 weak framework, and that the resulting algorithm is indeed *unconditionally stable*, for it is a velocity-correction algorithm.

5.1. The weak setting. Let us now rewrite (5.1), (5.2) in L^2. Let us assume for the time being that $\nabla\cdot(\nabla^2\tilde{u})^{\star,k} = 0$. By setting $u^{k+1} = \hat{u}^{k+1} + \delta t(\nabla^2\tilde{u})^{\star,k}$ and observing that $\nabla\cdot u^{k+1} = 0$ and $u^{k+1}\cdot n|_\Gamma = 0$, the system (5.1) can be rewritten

$$\tag{5.4} \begin{cases} \frac{1}{\delta t}(\beta_q u^{k+1} - \sum_{j=0}^{q-1}\beta_j\tilde{u}^{k-j}) - (\nabla^2\tilde{u})^{\star,k} + \nabla p^{k+1} = f(t^{k+1}), \\ \nabla\cdot u^{k+1} = 0, \\ u^{k+1}\cdot n|_\Gamma = 0. \end{cases}$$

Now, inserting the definition of u^{k+1} back into (5.2), we obtain

$$\tag{5.5} \begin{cases} \dfrac{\beta_q(\tilde{u}^{k+1} - u^{k+1})}{\delta t} - \nabla^2\tilde{u}^{k+1} + (\nabla^2\tilde{u})^{\star,k} = 0, \\ \tilde{u}^{k+1}|_\Gamma = 0. \end{cases}$$

Note that for $q = 2$ and $(\nabla^2\tilde{u})^{\star,k} = -\nabla\times\nabla\times\tilde{u}^k$, the scheme (5.4)–(5.5) is exactly the velocity-correction algorithm in rotational form (3.1)–(3.2), while the case $q = 2$ and $(\nabla^2\tilde{u})^{\star,k} = \nabla^2\tilde{u}^k$ corresponds to the velocity-correction algorithm in standard form (2.3)–(2.4).

5.2. First-order schemes. It is interesting to consider the case $q = 1$ and $(\nabla^2\tilde{u})^{\star,k} = 0$, the resulting scheme being

$$\tag{5.6} \begin{cases} \dfrac{u^{k+1} - \tilde{u}^k}{\delta t} + \nabla p^{k+1} = f(t^{k+1}), \\ \nabla\cdot u^{k+1} = 0, \\ u^{k+1}\cdot n|_\Gamma = 0, \end{cases}$$

$$\tag{5.7} \begin{cases} \dfrac{\tilde{u}^{k+1} - u^{k+1}}{\delta t} - \nabla^2\tilde{u}^{k+1} = 0, \\ \tilde{u}^{k+1}|_\Gamma = 0. \end{cases}$$

In this case, the standard version and the rotational version coincide, and this method can be viewed as the dual of the original Chorin–Temam method. Of course, it suffers from the dual ailments of the Chorin–Temam algorithm, i.e., it enforces $\partial_n p^{k+1}|_\Gamma = f(t^{k+1})\cdot n$ and $\nabla^2\tilde{u}^{k+1}|_\Gamma = 0$, whereas the Chorin–Temam scheme enforces $\nabla^2\tilde{u}^{k+1}|_\Gamma = f(t^{k+1})$ and $\partial_n p^{k+1}|_\Gamma = 0$.

From the point of view of accuracy, the two algorithms are equivalent.

THEOREM 5.1. *If* (u,p)*, the solution to* (1.1)*, is smooth enough in space and time, the solution to* (5.6)–(5.7) *satisfies the following error estimates:*

$$\|\mathsf{u} - u\|_{l^\infty(L^2(\Omega)^d)} + \|\mathsf{u} - \tilde{u}\|_{l^\infty(L^2(\Omega)^d)} \leq c(\mathsf{u},\mathsf{p},T)\,\delta t,$$
$$\|\mathsf{p} - p\|_{l^\infty(L^2(\Omega))} + \|\mathsf{u} - \tilde{u}^k\|_{l^\infty(H^1(\Omega)^d)} \leq c(\mathsf{u},\mathsf{p},T)\,\delta t^{1/2}.$$

Proof. Since the proof is very similar to that of the Chorin–Temam algorithm, we refer the reader to Shen [12], Rannacher [11], Guermond [6], or to the proof of second-order accuracy in section 3.4. ◻

6. Treatment of nonlinear terms.

6.1. Semi-implicit treatment. We now describe briefly how the nonlinear terms can be properly treated. Taking the second-order rotational velocity-correction scheme as an example, one way to treat the nonlinear term semi-implicitly is as follows:

$$\tag{6.1} \begin{cases} \dfrac{3u^{k+1}-4\tilde{u}^k+\tilde{u}^{k-1}}{2\delta t} + \nu\nabla\times\nabla\times\tilde{u}^k \\ \qquad + d(2\tilde{u}^{k-1}-\tilde{u}^{k-2},\tilde{u}^k) + \nabla p^{k+1} = f(t^{k+1}), \\ \nabla\cdot u^{k+1} = 0, \\ u^{k+1}\cdot n_{|\Gamma} = 0 \end{cases}$$

and

$$\tag{6.2} \begin{cases} \dfrac{3\tilde{u}^{k+1}-3u^{k+1}}{2\delta t} - \nu\nabla^2\tilde{u}^{k+1} + d(2\tilde{u}^k-\tilde{u}^{k-1},\tilde{u}^{k+1}) \\ \qquad\qquad - \nu\nabla\times\nabla\times\tilde{u}^k - d(2\tilde{u}^{k-1}-\tilde{u}^{k-2},\tilde{u}^k) = 0, \\ \tilde{u}^{k+1}_{|\Gamma} = 0, \end{cases}$$

where the bilinear form d accounts for the advection and can take various forms to ensure unconditional stability. For instance, we can use

$$\tag{6.3} d(v,w) = \begin{cases} v\cdot\nabla w + \frac{1}{2}(\nabla\cdot v)w & \text{or} \\ (\nabla\times v)\times w, \end{cases}$$

where in the second case, p^{k+1} is the total pressure, i.e., the kinetic energy has to be subtracted from p^{k+1} to get the real pressure. One can show, just as in the linear case, that the scheme (6.1)–(6.2) is unconditionally stable and that Theorem 3.1 holds.

Note that with the presence of the nonlinear term, the projection step is once again given by (3.4) in strong form or (3.9) in weak form. By adding (6.2) to (6.1), one obtains

$$\tag{6.4} \frac{3\tilde{u}^{k+1}-4\tilde{u}^k+\tilde{u}^{k-1}}{2\delta t} - \nu\nabla^2\tilde{u}^{k+1} + d(2\tilde{u}^k-\tilde{u}^{k-1},\tilde{u}^{k+1}) + \nabla p^{k+1} = f(t^{k+1}),$$

with $\tilde{u}^{k+1}_{|\Gamma} = 0$, which is a linear elliptic equation for $\tilde{u}^{k+1}$ that can be solved by standard procedures. As a result, a simple way to code the semi-implicit velocity-correction algorithm in rotational form with the projected velocity eliminated is (3.9), (3.7), (6.4).

6.2. Explicit treatment. One can also treat the nonlinear term totally explicitly as is done usually with spectral approximations [2]:

$$\tag{6.5} \begin{cases} \dfrac{3u^{k+1}-4\tilde{u}^k+\tilde{u}^{k-1}}{2\delta t} + \nu\nabla\times\nabla\times\tilde{u}^k \\ \qquad + (2d(\tilde{u}^k,\tilde{u}^k) - d(\tilde{u}^{k-1},\tilde{u}^{k-1})) + \nabla p^{k+1} = f(t^{k+1}), \\ \nabla\cdot u^{k+1} = 0, \\ u^{k+1}\cdot n_{|\Gamma} = 0 \end{cases}$$

and

$$\tag{6.6} \begin{cases} \dfrac{3\tilde{u}^{k+1}-3u^{k+1}}{2\delta t} - \nu\nabla^2\tilde{u}^{k+1} - \nu\nabla\times\nabla\times\tilde{u}^k = 0, \\ \tilde{u}^{k+1}_{|\Gamma} = 0. \end{cases}$$

In this case, the scheme is only conditionally stable with a usual CFL condition.

In practice the projected velocity can be completely eliminated from the algorithm as follows. Upon substituting $f(t^{k+1})$ into (3.9) by $f(t^{k+1}) - (2d(\tilde{u}^k, \tilde{u}^k) - d(\tilde{u}^{k-1}, \tilde{u}^{k-1}))$, the projection step is still (3.9). After updating the pressure according to (3.7), the new velocity $\tilde{u}^{k+1}$ is obtained by solving

$$\begin{aligned} &\frac{3\tilde{u}^{k+1} - 4\tilde{u}^k + \tilde{u}^{k-1}}{2\delta t} - \nu\nabla^2\tilde{u}^{k+1} + (2d(\tilde{u}^k, \tilde{u}^k) - d(\tilde{u}^{k-1}, \tilde{u}^{k-1})) + \nabla p^{k+1} \\ &\qquad = f(t^{k+1}) \end{aligned} \tag{6.7}$$

with $\tilde{u}^{k+1}_{|\Gamma} = 0$.

7. Concluding remarks. We have introduced a class of velocity-correction schemes in standard and rotational form. We proved stability and $\mathcal{O}(\delta t^2)$ convergence in the L^2-norm of the velocity for both versions. We also proved improved error estimates for the rotational form, i.e., $\mathcal{O}(\delta t^{3/2})$ convergence in the H^1-norm of the velocity and the L^2-norm of the pressure. Our numerical results indicate that these estimates appear to be the best possible under the general assumptions on Ω considered in this paper.

We have also shown that the schemes introduced in [10] and [9] are formally equivalent, in the spatial continuous case, to the velocity-correction projection methods in rotational form. Thus, our results provide the first rigorous proof of stability and convergence for these schemes. In addition, contrary to the original form of these methods which involve the normal trace of second-order derivatives of the velocity at the boundary, the new velocity-correction projection methods, being set in the standard L^2 weak setting, can be easily implemented by using any variational approximation techniques, including finite element methods.

REFERENCES

[1] A. Batoul, H. Khallouf, and G. Labrosse, *Une méthode de résolution directe (pseudo-spectrale) du problème de Stokes 2D/3D instationnaire. Application à la cavité entrainée carrée*, C. R. Acad. Sci. Paris, Sér. I, 319 (1994), pp. 1455–1461.

[2] C. Canuto, M. Hussaini, A. Quarteroni, and T. Zang, *Spectral methods in fluid dynamics*, Comput. Phys., Springer-Verlag, New York, 1987.

[3] A. J. Chorin, *Numerical solution of the Navier-Stokes equations*, Math. Comp., 22 (1968), pp. 745–762.

[4] W. E and J.-G. Liu, *Projection method* I: *Convergence and numerical boundary layers*, SIAM J. Numer. Anal., 32 (1995), pp. 1017–1057.

[5] K. Goda, *A multistep technique with implicit difference schemes for calculating two- or three-dimensional cavity flows*, J. Comput. Phys., 30 (1979), pp. 76–95.

[6] J.-L. Guermond, *Some practical implementations of projection methods for Navier–Stokes equations*, M2AN Modél. Math. Anal. Numér., 30 (1996), pp. 637–667.

[7] J.-L. Guermond, *Un résultat de convergence à l'ordre deux en temps pour l'approximation des équations de Navier–Stokes par une technique de projection*, M2AN Modél. Math. Anal. Numér., 33 (1999), pp. 169–189.

[8] J. van Kan, *A second-order accurate pressure-correction scheme for viscous incompressible flow*, SIAM J. Sci. Stat. Comput., 7 (1986), pp. 870–891.

[9] K. E. Karniadakis, M. Israeli, and S. A. Orszag, *High-order splitting methods for the incompressible Navier–Stokes equations*, J. Comput. Phys., 97 (1991), pp. 414–443.

[10] S. A. Orsag, M. Israeli, and M. Deville, *Boundary conditions for incompressible flows*, J. Sci. Comput., 1 (1986), pp. 75–111.

[11] R. Rannacher, *On Chorin's projection method for the incompressible Navier–Stokes equations*, in The Navier–Stokes Equations II. Theory and Numerical Methods, Lectures Notes in Math. 1530, Springer–Verlag, Berlin, 1992, pp. 167–183.

[12] J. Shen, *On error estimates of projection methods for Navier–Stokes equations: First-order schemes*, SIAM J. Numer. Anal, 29 (1992), pp. 57–77.

[13] J. Shen, *Efficient spectral-Galerkin method* I. *Direct solvers of second- and fourth-order equations using Legendre polynomials*, SIAM J. Sci. Comput., 15 (1994), pp. 1489–1505.

[14] J. Shen, *On error estimates of the projection methods for Navier–Stokes equations: Second-order schemes*, Math. Comp., 65 (1996), pp. 1039–1065.

[15] R. Temam, *Sur l'approximation de la solution des équations de Navier-Stokes par la méthode des pas fractionnaires* II, Arch. Ration. Mech. Anal., 33 (1969), pp. 377–385.

[16] R. Temam, *Navier–Stokes Equations*, Stud. Math. Appl. 2, North-Holland, Amsterdam, 1977.

SIAM J. NUMER. ANAL.
Vol. 41, No. 1, pp. 135–158

NUMERICAL APPROXIMATIONS OF PRESSURELESS AND ISOTHERMAL GAS DYNAMICS[*]

FRANÇOIS BOUCHUT[†], SHI JIN[‡], AND XIANTAO LI[‡]

Abstract. We study several schemes of first- or second-order accuracy based on kinetic approximations to solve pressureless and isothermal gas dynamics equations. The pressureless gas system is weakly hyperbolic, giving rise to the formation of density concentrations known as delta-shocks. For the isothermal gas system, the infinite speed of expansion into vacuum leads to zero timestep in the Godunov method based on exact Riemann solver. The schemes we consider are able to bypass these difficulties. They are proved to satisfy positiveness of density and discrete entropy inequalities, to capture the delta-shocks, and to treat data with vacuum.

Key words. pressureless gas, isothermal gas, kinetic schemes, positivity preserving, delta-shock, vacuum, entropy inequalities

AMS subject classifications. 65M06, 35L65

PII. S0036142901398040

1. Introduction.

The purpose of this work is to study some numerical approximations of the pressureless gas and isothermal gas dynamics equations. These equations take the form

$$\begin{cases} \rho_t + \operatorname{div}(\rho u) = 0, \\ (\rho u)_t + \operatorname{div}(\rho u \otimes u + \nu^2 \rho \ \mathrm{I}) = 0, \end{cases} \tag{1}$$

where $t > 0$, $x \in \mathbb{R}^N$, $\rho(x,t) \geq 0$, $u(x,t) \in \mathbb{R}^N$, with initial data

$$\rho(x,0) = \rho^0(x), \qquad \rho(x,0)u(x,0) = \rho^0(x)u^0(x). \tag{2}$$

The momentum will also be denoted by $q \equiv \rho u$. When $\nu = 0$, (1) is referred to as the pressureless gas equations, which arise in the modeling of sticky particles to explain the formation of large scale structures in the universe [26], [24]. The pressureless system has been studied at the theoretical level by several authors; see [2], [11], [14], [7], [21], [25], [23], [22], and for related problems see [8], [6], [1]. In the pressureless gas system, the Jacobian matrix of the flux is a Jordan block, thus the system is *weakly* hyperbolic. A main feature of this weakly hyperbolic system is the development of delta-shocks and the emergence of the vacuum state. For the isothermal gas equations, $\nu > 0$, the density profile forms a concentration that turns to a delta-wave as $\nu \to 0$ [12], [13]. Vacuum state cannot form for this system, but, if we start from vacuum, an expansion occurs at infinite velocity. Thus the Godunov method based on the exact Riemann solver requires the timestep to be zero for numerical stability, resulting in the failure of the method.

Some numerical schemes were introduced in [2], [4], [5] for the pressureless gas and related equations, and this paper can be viewed as a continuation of these earlier

[*] Received by the editors November 12, 2001; accepted for publication (in revised form) June 4, 2002; published electronically March 13, 2003. This research was supported in part by NSF grant DMS-0196106.

http://www.siam.org/journals/sinum/41-1/39804.html

[†] Département de Mathématiques et Applications, Ecole Normale Supérieure et CNRS, UMR 8553, 45 Rue d'Ulm, 75230 Paris Cedex 05, France (Francois.Bouchut@ens.fr).

[‡] Department of Mathematics, University of Wisconsin-Madison, 480 Lincoln Dr., Madison, WI 53706-1388 (jin@math.wisc.edu, xli@math.wisc.edu).

investigations. We consider here several first-order and second-order schemes based on kinetic approximations that are able to treat delta-waves and vacuum, and we compare their numerical performances. The maximum principle on the velocity u is also established.

In the case of isothermal gas dynamics, we introduce a kinetic scheme that reduces to the kinetic scheme for pressureless gas as $\nu \to 0$. This kinetic scheme is able to treat vacuum state since it takes the form of a flux vector splitting, giving some approximate Riemann solver that can deal with the vacuum state with nonzero timestep for numerical stability. Indeed, we establish an entropy inequality under some natural CFL condition.

Weakly hyperbolic systems also arise in mathematical modelings of multiphase geometrical optics [15, 18]. Simple and efficient numerical schemes for such problems include the Lax–Friedrichs scheme and its second-order extension [15], [17]. Compared with upwind-type schemes, including the Godunov and kinetic schemes, the central schemes offer greater simplicity and efficiency, yet they are slightly more diffusive. See more numerical experiments carried out in [16], [18].

The paper is organized as follows. In section 2, we study several schemes for the one-dimensional pressureless gas equations. In section 3, we generalize some kinetic schemes to two dimensions. In section 4, we study a kinetic scheme for an isothermal gas, and in section 5 we perform numerical tests.

2. Kinetic schemes for one-dimensional pressureless gas. As is now classical, gas dynamics equations can be solved by kinetic schemes; see [9], [19]. We refer, for example, to [3] for general properties of such schemes. For the pressureless system (1) with $\nu = 0$, a simple δ function can be taken as a Maxwellian [2], and the scheme can be written as follows in one space dimension. Starting from functions $\rho^n(x)$ and $u^n(x)$ at time t_n, one solves the transport equation

$$\left\{ \begin{array}{l} \partial_t f + \xi\, \partial_x f = 0 \qquad \text{in} \quad \mathbb{R}_x \times \mathbb{R}_\xi \times]t_n, t_{n+1}[, \\ f(x,\xi,t_n) = f^n(x,\xi) = M(\rho^n(x), u^n(x), \xi), \end{array} \right. \tag{3}$$

where the Maxwellian is defined for any $\rho \geq 0$, $u \in \mathbb{R}$, and $\xi \in \mathbb{R}$ by

$$M(\rho, u, \xi) = \rho\, \delta(\xi - u). \tag{4}$$

The transport equation has the exact solution $f(x,\xi,t) = f^n(x - \xi(t-t_n), \xi)$ for $t_n \leq t < t_{n+1}$. In the projection step, let

$$\begin{pmatrix} \rho^{n+1-}(x) \\ q^{n+1-}(x) \end{pmatrix} = \int_{\mathbb{R}} \begin{pmatrix} 1 \\ \xi \end{pmatrix} f(x,\xi,t_{n+1}-)\, d\xi. \tag{5}$$

In order to obtain discrete values over a mesh of constant size Δx, one defines the new averages ρ_j^{n+1-}, q_j^{n+1-}, with the usual definition

$$w_j^n = \frac{1}{\Delta x} \int_{x_{j-1/2}}^{x_{j+1/2}} w(x, t_n)\, dx. \tag{6}$$

It remains only to define functions $\rho^n(x)$ and $u^n(x)$ from the average values ρ_j^{n-}, q_j^{n-}. This is the reconstruction step, which needs to conserve $\rho_j^n = \rho_j^{n-}$, $q_j^n = q_j^{n-}$. For a first-order scheme, one just takes $\rho^n(x)$ and $u^n(x)$ (or, equivalently, $q^n(x)$) piecewise constant.

Notice that this algorithm keeps the nonnegativeness of the density ρ since f itself is nonnegative, and satisfies the maximum principle on the velocity u, in the following form:

$$\inf_x u^n(x) \le \frac{q^{n+1-}(x)}{\rho^{n+1-}(x)} \le \sup_x u^n(x), \tag{7}$$

which is easily seen from (5) and by the fact that from (3) and (4) the support in ξ of f lies in the range of u^n. The same inequalities (7) obviously hold also for q_j^{n+1-}/ρ_j^{n+1-}.

In order to obtain an explicit formula, one integrates the transport equation (3) over $(x_{j-1/2}, x_{j+1/2}) \times \mathbb{R}_\xi \times (t_n, t_{n+1})$, and a conservative numerical scheme is obtained as

$$\begin{aligned} \rho_j^{n+1} &= \rho_j^n - \frac{\Delta t}{\Delta x}\left(F^{(1)}_{j+1/2} - F^{(1)}_{j-1/2}\right), \\ q_j^{n+1} &= q_j^n - \frac{\Delta t}{\Delta x}\left(F^{(2)}_{j+1/2} - F^{(2)}_{j-1/2}\right), \end{aligned} \tag{8}$$

where

$$F_{j+1/2} = \begin{pmatrix} F^{(1)}_{j+1/2} \\ F^{(2)}_{j+1/2} \end{pmatrix} = \frac{1}{\Delta t}\int_{t_n}^{t_{n+1}} \int_{\mathbb{R}} \xi \begin{pmatrix} 1 \\ \xi \end{pmatrix} f(x_{j+1/2}, \xi, t)\, d\xi dt. \tag{9}$$

In addition, one can compute the integrals over $\mathbb{R}^+$ and $\mathbb{R}^-$ separately, and then the numerical flux can be written in the flux vector splitting form,

$$F_{j+1/2} = F^+_{j+1/2} + F^-_{j+1/2}, \tag{10}$$

with

$$F^\pm_{j+1/2} = \frac{1}{\Delta t}\int_{t_n}^{t_{n+1}} \int_{\pm\xi>0} \xi \begin{pmatrix} 1 \\ \xi \end{pmatrix} f^n\big(x_{j+1/2} - \xi(t-t_n), \xi, t\big)\, d\xi dt. \tag{11}$$

2.1. First-order kinetic scheme. For a first-order scheme, one uses piecewise constant data:

$$\rho(x) = \rho_j, \quad u(x) = u_j \qquad \text{for} \quad x_{j-1/2} < x < x_{j+1/2}. \tag{12}$$

Note that under the CFL condition

$$|u_j|\Delta t \le \Delta x \tag{13}$$

the integrals involved in (11) have support on at most one point. For instance, the integral for $F^+_{j+1/2}$ has support $\{\xi = u_j\}$ only if u_j is nonnegative. Therefore,

$$F^{+(1)}_{j+1/2} = \frac{1}{\Delta t}\int_{t_n}^{t_{n+1}} \int_0^\infty \xi\, \rho_j\, \delta(\xi - u_j)\, d\xi dt = \rho_j (u_j)_+. \tag{14}$$

Similarly,

$$F^{-(1)}_{j+1/2} = \rho_{j+1}(u_{j+1})_-, \;\; F^{+(2)}_{j+1/2} = \rho_j u_j (u_j)_+, \;\; F^{-(2)}_{j+1/2} = \rho_{j+1} u_{j+1} (u_{j+1})_-, \tag{15}$$

with the convention $u_+ + u_- = u$, $u_+ - u_- = |u|$. Thus the final scheme takes the form (8), (10), (14), (15), with $q_j^n = \rho_j^n u_j^n$. As is easily seen, under the CFL condition (13), this first-order kinetic scheme keeps the density ρ nonnegative and satisfies the maximum principle on the velocity u. Other properties, like entropy inequalities or the TVD property on u, can also be proved; see [2].

2.2. Godunov scheme. To compare with the kinetic scheme, we also derive the first-order Godunov scheme in this section. As was derived in [2], [23], the Riemann problem (1) with $\nu = 0$ and initial data

$$(\rho^0(x), u^0(x)) = \begin{cases} (\rho_j, u_j) & x < x_{j+1/2}, \\ (\rho_{j+1}, u_{j+1}) & x > x_{j+1/2}, \end{cases} \tag{16}$$

where $u_j \geq u_{j+1}$, has the delta-shock solution

$$(\rho(x,t), u(x,t)) = \begin{cases} (\rho_j, u_j) & x < x_{j+1/2} + u_\delta t, \\ (w(t)\,\delta(x - x_{j+1/2} - u_\delta t), u_\delta) & x = x_{j+1/2} + u_\delta t, \\ (\rho_{j+1}, u_{j+1}) & x > x_{j+1/2} + u_\delta t, \end{cases} \tag{17}$$

where

$$w(t) = \sqrt{\rho_j \rho_{j+1}}(u_j - u_{j+1})t, \quad u_\delta = \frac{\sqrt{\rho_j}u_j + \sqrt{\rho_{j+1}}u_{j+1}}{\sqrt{\rho_j} + \sqrt{\rho_{j+1}}}. \tag{18}$$

Therefore, still in the case when $u_j \geq u_{j+1}$, we obtain the Godunov flux

$$F_{j+1/2} = \begin{cases} (\rho_j u_j, \rho_j u_j^2) & \text{if } u_\delta > 0, \\ (\rho_{j+1} u_{j+1}, \rho_{j+1} u_{j+1}^2) & \text{if } u_\delta < 0, \\ ((\rho_j u_j + \rho_{j+1} u_{j+1})/2, \rho_j u_j^2 = \rho_{j+1} u_{j+1}^2) & \text{if } u_\delta = 0. \end{cases} \tag{19}$$

In the case when $u_\delta = 0$, there is a stationary delta-shock at the interface, thus the formula has to be explained a bit. We choose arbitrarily to put half of the mass to the left and the other half to the right of the line $x = x_{j+1/2}$. This means that

$$\rho_j^{n+1} = \frac{1}{\Delta x}\left(\int_{x_{j-1/2}}^{x_{j+1/2}-} \rho^{n+1-}(x)\,dx + \frac{1}{2}w(\Delta t)\right).$$

In order to see which numerical flux we get, we write down the finite difference formula

$$\rho_j^n + \frac{1}{2}\frac{\Delta t}{\Delta x}\sqrt{\rho_j \rho_{j+1}}(u_j - u_{j+1}) - \rho_j^n + \frac{\Delta t}{\Delta x}(F_{j+1/2}^{(1)} - F_{j-1/2}^{(1)}) = 0,$$

and assuming that $F_{j-1/2}^{(1)} = \rho_j u_j$ this yields by using the fact that $u_\delta = 0$:

$$F_{j+1/2}^{(1)} = \rho_j u_j - \frac{1}{2}\sqrt{\rho_j \rho_{j+1}}(u_j - u_{j+1}) = \frac{1}{2}(\rho_j u_j + \rho_{j+1} u_{j+1}).$$

On the contrary, there is no momentum on the delta-shock since $u_\delta = 0$, and a similar computation as above gives the flux $F_{j+1/2}^{(2)} = \rho_j u_j^2 = \rho_{j+1} u_{j+1}^2$.

In the case where $u_j < u_{j+1}$, the exact solution of the Riemann problem contains vacuum and is given by

$$(\rho(x,t), u(x,t)) = \begin{cases} (\rho_j, u_j) & x < x_{j+1/2} + u_j t, \\ (0, \text{undefined}) & x_{j+1/2} + u_j t < x < x_{j+1/2} + u_{j+1} t, \\ (\rho_{j+1}, u_{j+1}) & x > x_{j+1/2} + u_{j+1} t, \end{cases} \tag{20}$$

and the numerical flux takes the form

$$F_{j+1/2} = \begin{cases} (\rho_j u_j, \rho_j u_j^2) & \text{if } u_j > 0, \\ (\rho_{j+1} u_{j+1}, \rho_{j+1} u_{j+1}^2) & \text{if } u_{j+1} < 0, \\ (0, 0) & \text{otherwise.} \end{cases} \tag{21}$$

Again under the CFL condition (13), it is easy to see that the Godunov scheme (8), (19), (21), with $q_j^n = \rho_j^n u_j^n$, keeps the density ρ nonnegative and satisfies the maximum principle on the velocity u.

2.3. A second-order kinetic scheme. For a second-order scheme, one can use a piecewise linear reconstruction,

$$\left\{ \begin{array}{l} \rho(x) = \rho_j + D\rho_j(x - x_j) \\ u(x) = \overline{u}_j + Du_j(x - x_j) \end{array} \right. \qquad \text{for} \quad x_{j-1/2} < x < x_{j+1/2}, \tag{22}$$

where $\overline{u}_j$ is chosen as

$$\overline{u}_j = u_j - \frac{D\rho_j Du_j}{12\rho_j}\Delta x^2, \tag{23}$$

$$u_j = q_j/\rho_j, \tag{24}$$

in order to have the conservation property; namely, the cell average of $q(x) = \rho(x)u(x)$ must be q_j. There are several limitations on $D\rho_j$ and Du_j, which need to be computed from ρ_j and q_j in the reconstruction step. We shall collect and summarize all the restrictions at the end of this section.

Let us first explain the evolution step (3)–(6), starting from the initial data (22). We need to compute $F^+_{j+1/2}$ in (11). We use the CFL condition

$$\Delta t \sup_x |u^n(x)| \le \Delta x. \tag{25}$$

To get an explicit expression, we assume the piecewise nonovertaking condition

$$\Delta t Du^n_j > -1. \tag{26}$$

Then we can proceed with the calculation,

$$\begin{aligned} F^+_{j+1/2} &= \frac{1}{\Delta t}\int_0^{\Delta t}\int_0^\infty \xi \begin{pmatrix} 1 \\ \xi \end{pmatrix} \rho^n(x_{j+1/2} - \xi t)\,\delta(\xi - u^n(x_{j+1/2} - \xi t))\,d\xi dt \\ &= \frac{1}{\Delta t}\int_0^\infty \int_{x_{j+1/2}-\xi\Delta t}^{x_{j+1/2}} \begin{pmatrix} 1 \\ \xi \end{pmatrix} \rho^n(x)\,\delta(\xi - u^n(x))\,dx d\xi \\ &= \frac{1}{\Delta t}\int_{x_{j-1/2}}^{x_{j+1/2}} \rho^n(x) \begin{pmatrix} 1 \\ u^n(x) \end{pmatrix} \mathbb{I}_{x+\Delta t u^n(x) > x_{j+1/2}}\,dx \\ &= \frac{1}{\Delta t}\int_{x^L_{j+1/2}}^{x_{j+1/2}} \begin{pmatrix} \rho^n(x) \\ \rho^n(x)u^n(x) \end{pmatrix} dx, \end{aligned} \tag{27}$$

where

$$x^L_{j+1/2} = x_{j+1/2} - \Delta t \frac{(\overline{u}_j + \frac{\Delta x}{2}Du_j)_+}{1 + \Delta t Du_j}. \tag{28}$$

Thus we obtain

$$\begin{aligned} F^{+(1)}_{j+1/2} &= \frac{\rho^L_{j+1/2}(u^L_{j+1/2})_+}{1 + \Delta t Du_j} - \frac{\Delta t}{2} D\rho_j \frac{u^L_{j+1/2}(u^L_{j+1/2})_+}{(1 + \Delta t Du_j)^2}, \\ F^{+(2)}_{j+1/2} &= \rho^L_{j+1/2} u^L_{j+1/2}(u^L_{j+1/2})_+ \frac{1 + \frac{\Delta t}{2}Du_j}{(1 + \Delta t Du_j)^2} \\ &\qquad - \frac{\Delta t}{6} D\rho_j \frac{(u^L_{j+1/2})^2 (u^L_{j+1/2})_+}{(1 + \Delta t Du_j)^3}(3 + \Delta t Du_j), \end{aligned} \tag{29}$$

with

$$\begin{aligned}
\rho^L_{j+1/2} &= \rho_j + \tfrac{\Delta x}{2} D\rho_j,\\
\rho^R_{j+1/2} &= \rho_{j+1} - \tfrac{\Delta x}{2} D\rho_{j+1},\\
u^L_{j+1/2} &= \overline{u}_j + \tfrac{\Delta x}{2} Du_j,\\
u^R_{j+1/2} &= \overline{u}_{j+1} - \tfrac{\Delta x}{2} Du_{j+1}.
\end{aligned}\tag{30}$$

Similarly,

$$F^-_{j+1/2} = -\frac{1}{\Delta t}\int_{x_{j+1/2}}^{x^R_{j+1/2}} \begin{pmatrix} \rho^n(x) \\ \rho^n(x)u^n(x) \end{pmatrix} dx,\tag{31}$$

$$x^R_{j+1/2} = x_{j+1/2} - \Delta t \frac{(u^R_{j+1/2})_-}{1+\Delta t Du_{j+1}},\tag{32}$$

and

$$\begin{aligned}
F^{-(1)}_{j+1/2} &= \frac{\rho^R_{j+1/2}(u^R_{j+1/2})_-}{1+\Delta t Du_{j+1}} - \frac{\Delta t}{2} D\rho_{j+1} \frac{u^R_{j+1/2}(u^R_{j+1/2})_-}{(1+\Delta t Du_{j+1})^2},\\
F^{-(2)}_{j+1/2} &= \rho^R_{j+1/2} u^R_{j+1/2} (u^R_{j+1/2})_- \frac{1+\frac{\Delta t}{2} Du_{j+1}}{(1+\Delta t Du_{j+1})^2}\\
&\quad - \frac{\Delta t}{6} D\rho_{j+1} \frac{(u^R_{j+1/2})^2 (u^R_{j+1/2})_-}{(1+\Delta t Du_{j+1})^3}(3+\Delta t Du_{j+1}).
\end{aligned}\tag{33}$$

An interpretation of (27) is that we put in the flux $F^+_{j+1/2}$ the total mass and momentum of all particles located at $x \in (x_{j-1/2}, x_{j+1/2})$ at time t_n that pass through the node $x_{j+1/2}$ (in the sense of characteristics) between times t_n and t_{n+1}. Indeed, $x + (t - t_n)u^n(x) = x_{j+1/2}$ for some $t_n < t < t_{n+1}$ is equivalent to $x + \Delta t\, u^n(x) > x_{j+1/2}$. Under condition (26), these characteristics do not cross; however, this is done regardless of the trajectories of particles coming from the right.

Finally, we have to specify the reconstruction step. In order to ensure the nonnegativity of ρ, one needs that

$$|D\rho_j \Delta x/2| \le \rho_j.\tag{34}$$

To guarantee the maximum principle property on the velocity

$$m_j \le u(x) \le M_j, \qquad x \in (x_{j-1/2}, x_{j+1/2}),\tag{35}$$

where

$$m_j = \min\{u_{j-1}, u_j, u_{j+1}\}, \qquad M_j = \max\{u_{j-1}, u_j, u_{j+1}\},\tag{36}$$

we need that

$$\begin{aligned}
m_j &\le u_j + \frac{\Delta x}{2} Du_j (1 - D\rho_j \Delta x / 6\rho_j) \le M_j,\\
m_j &\le u_j - \frac{\Delta x}{2} Du_j (1 + D\rho_j \Delta x / 6\rho_j) \le M_j.
\end{aligned}\tag{37}$$

Therefore we have the restrictions on the slopes of u,

$$Du_j \le \min\left\{ \frac{M_j - u_j}{(1 - D\rho_j \Delta x/6\rho_j)\frac{\Delta x}{2}}, \frac{u_j - m_j}{(1 + D\rho_j \Delta x/6\rho_j)\frac{\Delta x}{2}} \right\} \tag{38}$$

and

$$Du_j \ge \max\left\{ \frac{m_j - u_j}{(1 - D\rho_j \Delta x/6\rho_j)\frac{\Delta x}{2}}, \frac{u_j - M_j}{(1 + D\rho_j \Delta x/6\rho_j)\frac{\Delta x}{2}} \right\}. \tag{39}$$

In practice, we can choose the following limiters to satisfy all the constraints:

$$\begin{aligned} D\rho_j &= \frac{1}{2}\Big(\operatorname{sgn}(\rho_{j+1} - \rho_j) + \operatorname{sgn}(\rho_j - \rho_{j-1})\Big) \\ &\qquad \times \min\left\{ \frac{|\rho_{j+1} - \rho_j|}{\Delta x}, \frac{|\rho_j - \rho_{j-1}|}{\Delta x}, \frac{2\rho_j}{\Delta x} \right\}, \\ Du_j &= \frac{1}{2}\Big(\operatorname{sgn}(u_{j+1} - u_j) + \operatorname{sgn}(u_j - u_{j-1})\Big) \\ &\qquad \times \min\left\{ \frac{|u_{j+1} - u_j|}{(1 - \Delta x D\rho_j/6\rho_j)\Delta x}, \frac{|u_j - u_{j-1}|}{(1 + \Delta x D\rho_j/6\rho_j)\Delta x}, \frac{1}{\Delta t} \right\}. \end{aligned} \tag{40}$$

The scheme we obtain is then second-order in space and time. The second-order accuracy in time here comes from the very special property of the system of pressureless gas, which is that (3) is indeed a kinetic formulation of (1) with $\nu = 0$, in the sense that as soon as we have smooth solutions to (1) the solution to (3) remains Maxwellian. This property was proved in [2] (see also [10]).

2.4. A simplified second-order kinetic scheme. In order to obtain an easy extension of our second-order method to two-dimensional problems, we can rather use the standard extension to second-order of the scheme we have obtained in section 2.1. It can be interpreted as using the transport-projection method (3)–(6) with initial data that are piecewise constant over half-cells,

$$\begin{aligned} \rho(x) &= \rho^R_{j-1/2}, \quad u(x) = u^R_{j-1/2} \qquad \text{for} \quad x_{j-1/2} < x < x_j, \\ \rho(x) &= \rho^L_{j+1/2}, \quad u(x) = u^L_{j+1/2} \qquad \text{for} \quad x_j < x < x_{j+1/2}, \end{aligned} \tag{41}$$

with as before

$$\begin{aligned} \rho^L_{j+1/2} &= \rho_j + D\rho_j \Delta x/2, \\ \rho^R_{j+1/2} &= \rho_{j+1} - D\rho_{j+1} \Delta x/2, \\ u^L_{j+1/2} &= \overline{u}_j + Du_j \Delta x/2, \\ u^R_{j+1/2} &= \overline{u}_{j+1} - Du_{j+1} \Delta x/2. \end{aligned} \tag{42}$$

Now, to have conservation of momentum, $\overline{u}_j$ is chosen as

$$\overline{u}_j = u_j - \frac{D\rho_j Du_j}{4\rho_j}\Delta x^2, \qquad u_j = q_j/\rho_j. \tag{43}$$

By restricting the CFL number to 1/2,

$$\Delta t \sup_x |u^n(x)| \le \Delta x/2, \tag{44}$$

we obtain the same numerical flux as in the first-order scheme,

$$
\begin{aligned}
F_{j+1/2} = \Big(& \rho^L_{j+1/2}(u^L_{j+1/2})_+ + \rho^R_{j+1/2}(u^R_{j+1/2})_-, \\
& (\rho u)^L_{j+\frac{1}{2}}(u^L_{j+1/2})_+ + (\rho u)^R_{j+1/2}(u^R_{j+\frac{1}{2}})_- \Big).
\end{aligned} \tag{45}
$$

From this construction, we obviously achieve second-order accuracy in space. Concerning time, a simple second-order Runge–Kutta method can be used,

$$
\begin{cases}
U^*_j = U^n_j - \frac{\Delta t}{\Delta x}(F^n_{j+1/2} - F^n_{j-1/2}), \\
U^{**}_j = U^*_j - \frac{\Delta t}{\Delta x}(F^*_{j+1/2} - F^*_{j-1/2}), \\
U^{n+1}_j = (U^n_j + U^{**}_j)/2.
\end{cases} \tag{46}
$$

In the reconstruction step, the restrictions on slopes are similar to the previous section,

$$
\begin{aligned}
D\rho_j = {} & \frac{1}{2}\Big(\operatorname{sgn}(\rho_{j+1} - \rho_j) + \operatorname{sgn}(\rho_j - \rho_{j-1})\Big) \\
& \times \min\left\{ \frac{|\rho_{j+1} - \rho_j|}{\Delta x}, \frac{|\rho_j - \rho_{j-1}|}{\Delta x}, \frac{2\rho_j}{\Delta x} \right\}, \\
Du_j = {} & \frac{1}{2}\Big(\operatorname{sgn}(u_{j+1} - u_j) + \operatorname{sgn}(u_j - u_{j-1})\Big) \\
& \times \min\left\{ \frac{|u_{j+1} - u_j|}{(1 - D\rho_j\Delta x/2\rho_j)\Delta x}, \frac{|u_j - u_{j-1}|}{(1 + D\rho_j\Delta x/2\rho_j)\Delta x} \right\}.
\end{aligned} \tag{47}
$$

Again, this reconstruction preserves the nonnegativity of the density and satisfies the maximum principle on the velocity. Similar properties were established in [20] for such a scheme.

2.5. An improved second-order kinetic scheme. A scheme that is even more precise than that of subsection 2.3 can be derived in the following way. We use again piecewise linear functions (22) with (23), (24) but perform a different reconstruction. By considering the mesh with half the cells of the original one, one can compute, under a half CFL condition, the new averages on half-cells $\rho^{n+1-}_{j\pm 1/4}$, $q^{n+1-}_{j\pm 1/4}$, with formulas similar to those of subsection 2.3. Then the averages are given by

$$
\rho^{n+1}_j = \frac{1}{2}\Big(\rho^{n+1-}_{j+1/4} + \rho^{n+1-}_{j-1/4}\Big), \qquad q^{n+1}_j = \frac{1}{2}\Big(q^{n+1-}_{j+1/4} + q^{n+1-}_{j-1/4}\Big). \tag{48}
$$

Formulas (23), (24) still hold for conservativity, but now the slopes are computed in a different way. Indeed, one evolves both the averages and the slopes, and therefore one has to give the value of the new slopes $D\rho^{n+1}_j$, Du^{n+1}_j from the evolution step. We use

$$
D\rho^{n+1}_j = \operatorname{sgn}\Big(\rho^{n+1-}_{j+1/4} - \rho^{n+1-}_{j-1/4}\Big)\frac{2}{\Delta x}\min\Big(\Big|\rho^{n+1-}_{j+1/4} - \rho^{n+1-}_{j-1/4}\Big|\ , \rho^{n+1}_j\Big), \tag{49}
$$

$$
\begin{aligned}
Du^{n+1}_j = {} & \operatorname{sgn}\left(\frac{q^{n+1-}_{j+1/4}}{\rho^{n+1-}_{j+1/4}} - \frac{q^{n+1-}_{j-1/4}}{\rho^{n+1-}_{j-1/4}}\right) \\
& \times \frac{2}{\Delta x}\min\left(\left|\frac{q^{n+1-}_{j+1/4}}{\rho^{n+1-}_{j+1/4}} - \frac{q^{n+1-}_{j-1/4}}{\rho^{n+1-}_{j-1/4}}\right|\ , \frac{\min(M_j - u^{n+1}_j, u^{n+1}_j - m_j)}{1 + |D\rho^{n+1}_j|\Delta x/6\rho^{n+1}_j}\right),
\end{aligned} \tag{50}
$$

where m_j and M_j are a bit less restrictive than in (36),

$$
\begin{aligned}
m_j &= \inf_{x_{j-3/2}<x<x_{j+3/2}} u^n(x) \\
&= \min\Big(\overline{u}^n_{j-1} - |Du^n_{j-1}|\frac{\Delta x}{2}, \overline{u}^n_j - |Du^n_j|\frac{\Delta x}{2}, \overline{u}^n_{j+1} - |Du^n_{j+1}|\frac{\Delta x}{2}\Big),
\end{aligned} \tag{51}
$$

$$
\begin{aligned}
M_j &= \sup_{x_{j-3/2}<x<x_{j+3/2}} u^n(x) \\
&= \max\Big(\overline{u}^n_{j-1} + |Du^n_{j-1}|\frac{\Delta x}{2}, \overline{u}^n_j + |Du^n_j|\frac{\Delta x}{2}, \overline{u}^n_{j+1} + |Du^n_{j+1}|\frac{\Delta x}{2}\Big).
\end{aligned} \tag{52}
$$

3. Two-dimensional pressureless gas. For convenience, we write the two-dimensional pressureless gas equations as

$$
\begin{cases}
\rho_t + p_x + q_y = 0, \\
p_t + (pu)_x + (pv)_y = 0, \\
q_t + (qu)_x + (qv)_y = 0,
\end{cases} \tag{53}
$$

with $p = \rho u$, $q = \rho v$. We shall use a conservative scheme

$$
U^{n+1}_{j,k} - U^n_{j,k} + \frac{\Delta t}{\Delta x}(F_{j+1/2,k} - F_{j-1/2,k}) + \frac{\Delta t}{\Delta y}(G_{j,k+1/2} - G_{j,k-1/2}) = 0, \tag{54}
$$

with $U = (\rho, p, q) = (\rho, \rho u, \rho v)$. The numerical fluxes have positive and negative contributions

$$
F_{j+1/2,k} = F^+_{j+1/2,k} + F^-_{j+1/2,k}, \qquad G_{j,k+1/2} = G^+_{j,k+1/2} + G^-_{j,k+1/2}. \tag{55}
$$

3.1. The first-order scheme. As usual, we use a piecewise constant data for $x_{j-1/2} < x < x_{j+1/2},\ y_{k-1/2} < y < y_{k+1/2}$,

$$
\begin{cases}
\rho(x,y) = \rho_{j,k}, \\
u(x,y) = u_{j,k}, \\
v(x,y) = v_{j,k}.
\end{cases} \tag{56}
$$

We use the same kinetic model (3)–(4), except that now $\xi \in \mathbb{R}^2$. The following computational lemma in one dimension will be useful.

LEMMA 3.1. *Suppose that $u(x)$ and $r(x)$ are piecewise constant functions,*

$$
u(x) = u_j,\ r(x) = r_j \quad \text{as } x \in (x_{j-1/2}, x_{j+1/2}), \tag{57}
$$

and the CFL condition is

$$
\frac{\Delta t}{\Delta x}|u| \le 1. \tag{58}
$$

Then

$$
\begin{aligned}
&\frac{1}{\Delta t}\int_0^{\Delta t} \frac{1}{\Delta x}\int_{x_{j-1/2}}^{x_{j+1/2}} \int_{-\infty}^{\infty} r(x-\xi t)\,\delta(\xi - u(x-\xi t))\,dt dx d\xi \\
&= \frac{\Delta t}{2\Delta x}\, r_{j-1}(u_{j-1})_+ + \left(1 - |u_j|\frac{\Delta t}{2\Delta x}\right) r_j - \frac{\Delta t}{2\Delta x}\, r_{j+1}(u_{j+1})_-.
\end{aligned} \tag{59}
$$

In order to get the numerical fluxes, we integrate the kinetic equation (3) over $(x_{j-1/2}, x_{j+1/2}) \times (y_{k-1/2}, y_{k+1/2}) \times \mathbb{R}^2 \times (0, \Delta t)$. Under the CFL conditions

$$\frac{\Delta t}{\Delta x}|u| \le 1, \qquad \frac{\Delta t}{\Delta y}|v| \le 1, \tag{60}$$

we find

$$\begin{aligned} F^{(1)+}_{j+1/2,k} &= \frac{1}{\Delta t \Delta y}\int_0^{\Delta t}\int_{y_{k-1/2}}^{y_{k+1/2}}\int_{\mathbb{R}} p(x_j, y-\xi_2 t)_+ \delta(\xi_2 - v(x_j, y-\xi_2 t))\, dt dy d\xi_2 \\ &= (p_{j,k})_+ + \frac{\Delta t}{2\Delta y}\Big((p_{j,k-1})_+(v_{j,k-1})_+ \\ &\qquad -(p_{j,k+1})_+(v_{j,k+1})_- - (p_{j,k})_+|v_{j,k}|\Big), \end{aligned}$$

$$\begin{aligned} F^{(1)-}_{j+1/2,k} &= \frac{1}{\Delta t \Delta y}\int_0^{\Delta t}\int_{y_{k-1/2}}^{y_{k+1/2}}\int_{\mathbb{R}} p(x_{j+1}, y-\xi_2 t)_- \delta(\xi_2 - v(x_{j+1}, y-\xi_2 t))\, dt dy d\xi_2 \\ &= (p_{j+1,k})_- + \frac{\Delta t}{2\Delta y}\Big((p_{j+1,k-1})_-(v_{j+1,k-1})_+ \\ &\qquad -(p_{j+1,k+1})_-(v_{j+1,k+1})_- - (p_{j+1,k})_-|v_{j+1,k}|\Big). \end{aligned}$$

Therefore, in compact form, the first-order kinetic scheme reads

$$\begin{aligned} F^{+}_{j+1/2,k} &= U_{j,k}(u_{j,k})_+ \\ &\quad + \frac{\Delta t}{2\Delta y}\Big(U_{j,k-1}(u_{j,k-1})_+(v_{j,k-1})_+ \\ &\qquad -U_{j,k+1}(u_{j,k+1})_+(v_{j,k+1})_- - U_{j,k}(u_{j,k})_+|v_{j,k}|\Big), \\ F^{-}_{j+1/2,k} &= U_{j+1,k}(u_{j+1,k})_- \\ &\quad + \frac{\Delta t}{2\Delta y}\Big(U_{j+1,k-1}(u_{j+1,k-1})_-(v_{j+1,k-1})_+ \\ &\qquad -U_{j+1,k+1}(u_{j+1,k+1})_-(v_{j+1,k+1})_- - U_{j+1,k}(u_{j+1,k})_-|v_{j+1,k}|\Big), \\ G^{+}_{j,k+1/2} &= U_{j,k}(v_{j,k})_+ \\ &\quad + \frac{\Delta t}{2\Delta x}\Big(U_{j-1,k}(v_{j-1,k})_+(u_{j-1,k})_+ \\ &\qquad -U_{j+1,k}(v_{j+1,k})_+(u_{j+1,k})_- - U_{j,k}(v_{j,k})_+|u_{j,k}|\Big), \\ G^{-}_{j,k+1/2} &= U_{j,k+1}(v_{j,k+1})_- \\ &\quad + \frac{\Delta t}{2\Delta x}\Big(U_{j-1,k+1}(v_{j-1,k+1})_-(u_{j-1,k+1})_+ \\ &\qquad -U_{j+1,k+1}(v_{j+1,k+1})_-(u_{j+1,k+1})_- - U_{j,k+1}(v_{j,k+1})_-|u_{j,k+1}|\Big). \end{aligned}$$

3.2. A second-order scheme. As in subsection 3.1, we use a simplified second-order scheme. Define

$$u^I_{j,k} = \overline{u}_{j,k} + \frac{\Delta x}{2} D_x u_{j,k} + \frac{\Delta y}{2} D_y u_{j,k}, \tag{61}$$

$$u^{II}_{j,k} = \overline{u}_{j,k} - \frac{\Delta x}{2} D_x u_{j,k} + \frac{\Delta y}{2} D_y u_{j,k}, \tag{62}$$

$$u^{III}_{j,k} = \overline{u}_{j,k} - \frac{\Delta x}{2} D_x u_{j,k} - \frac{\Delta y}{2} D_y u_{j,k}, \tag{63}$$

$$u^{IV}_{j,k} = \overline{u}_{j,k} + \frac{\Delta x}{2} D_x u_{j,k} - \frac{\Delta y}{2} D_y u_{j,k}, \tag{64}$$

where

$$\overline{u}_{j,k} = u_{j,k} - \frac{\Delta x^2}{4\rho_{j,k}} D_x \rho_{j,k} D_x u_{j,k} - \frac{\Delta y^2}{4\rho_{j,k}} D_y \rho_{j,k} D_y u_{j,k}. \tag{65}$$

We can define ρ and v in a similar way. Then it is easy to derive the following formulas, valid under half CFL conditions:

$$\begin{aligned}
F^{+}_{j+1/2,k} = {} & \frac{1}{2}\Big((Uu_+)^{I}_{j,k} + (Uu_+)^{IV}_{j,k}\Big) \\
& + \frac{\Delta t}{2\Delta y}\Big(-(Uu_+v_+)^{I}_{j,k} + (Uu_+v_-)^{IV}_{j,k} \\
& \qquad + (Uu_+v_+)^{I}_{j,k-1} - (Uu_+v_-)^{IV}_{j,k+1}\Big), \\
F^{-}_{j+1/2,k} = {} & \frac{1}{2}\Big((Uu_-)^{II}_{j+1,k} + (Uu_-)^{III}_{j+1,k}\Big) \\
& + \frac{\Delta t}{2\Delta y}\Big(-(Uu_-v_+)^{II}_{j+1,k} + (Uu_-v_-)^{III}_{j+1,k} \\
& \qquad + (Uu_-v_+)^{II}_{j+1,k-1} - (Uu_-v_-)^{III}_{j+1,k+1}\Big),
\end{aligned}$$

$$\begin{aligned}
G^{+}_{j,k+1/2} = {} & \frac{1}{2}\Big((Uv_+)^{I}_{j,k} + (Uv_+)^{II}_{j,k}\Big) \\
& + \frac{\Delta t}{2\Delta x}\Big(-(Uv_+u_+)^{I}_{j,k} + (Uv_+u_-)^{II}_{j,k} \\
& \qquad + (Uv_+u_+)^{I}_{j-1,k} - (Uv_+u_-)^{II}_{j+1,k}\Big), \\
G^{-}_{j,k+1/2} = {} & \frac{1}{2}\Big((Uv_-)^{IV}_{j,k+1} + (Uv_-)^{III}_{j,k+1}\Big) \\
& + \frac{\Delta t}{2\Delta x}\Big(-(Uv_-u_+)^{IV}_{j,k+1} + (Uv_-u_-)^{III}_{j,k+1} \\
& \qquad + (Uv_-u_+)^{IV}_{j-1,k+1} - (Uv_-u_-)^{III}_{j+1,k+1},\Big).
\end{aligned}$$

4. Isothermal gas dynamics. We now wish to generalize the previous schemes to the system of isothermal gas dynamics

$$\begin{cases} \partial_t \rho + \partial_x(\rho u) = 0, \\ \partial_t(\rho u) + \partial_x(\rho u^2 + \nu^2 \rho) = 0, \end{cases} \tag{66}$$

with $\rho(x,t) \geq 0$, $u(x,t) \in \mathbb{R}$, and $\nu > 0$. Features that are desirable for a numerical scheme to solve (66) are positivity preserving for the density ρ, the ability to treat data with vacuum, and a discrete entropy inequality. We shall achieve these goals in such a way that (at least for first-order), as $\nu \to 0$, the scheme reduces to the one presented in section 2.1. We recall that the physical energy

$$\eta(\rho, u) = \rho u^2/2 + \nu^2 \rho \ln \frac{\rho}{\sqrt{2\pi\nu}} \tag{67}$$

is a convex entropy for (66), with entropy flux

$$\vartheta(\rho,u)=\big(\eta(\rho,u)+\nu^2\rho\big)\,u, \tag{68}$$

i.e., $\vartheta'=\eta' F'$, where $F=(\rho u,\rho u^2+\nu^2\rho)$, and prime stands for differentiation with respect to $(\rho, q=\rho u)$.

Since second-order accuracy can be obtained as usual with a half CFL condition (as in section 2.4), we give only the first-order scheme. We shall use again a kinetic method, as described in [3], with the physical Maxwellian equilibrium

$$M(\rho,u,\xi)=\begin{pmatrix}1\\ \xi\end{pmatrix}\frac{\rho}{\sqrt{2\pi}\nu}e^{-(\xi-u)^2/2\nu^2},\qquad \rho\ge 0,\ u,\xi\in\mathbb{R}. \tag{69}$$

We can decompose the flux F of (66) as

$$F(\rho,u)\equiv(\rho u,\rho u^2+\nu^2\rho)=F^+(\rho,u)+F^-(\rho,u), \tag{70}$$

with

$$F^+(\rho,u)=\int_0^\infty \xi M(\rho,u,\xi)\,d\xi,\quad F^-(\rho,u)=\int_{-\infty}^0 \xi M(\rho,u,\xi)\,d\xi. \tag{71}$$

These half-fluxes are homogeneous of degree 1 with respect to $(\rho,\rho u)$, and this enables the scheme to treat vacuum data. The numerical scheme then takes the form

$$U_j^{n+1}-U_j^n+\frac{\Delta t}{\Delta x}(F_{j+1/2}-F_{j-1/2})=0, \tag{72}$$

with $U_j^n=(\rho_j^n,\rho_j^n u_j^n)$ and

$$F_{j+1/2}=F^+(\rho_j^n,u_j^n)+F^-(\rho_{j+1}^n,u_{j+1}^n). \tag{73}$$

Since the Maxwellian M in (69) does not have a compact support in ξ, we cannot apply the classical CFL condition $\Delta t\sup_{\xi\in\mathrm{supp}M}|\xi|\le\Delta x$ to derive positiveness of density and entropy inequalities. Therefore, we are going to use the analysis of [3] to derive explicitly a sufficient CFL condition. The scheme will satisfy the discrete entropy inequality

$$\eta(\rho_j^{n+1},u_j^{n+1})-\eta(\rho_j^n,u_j^n)+\frac{\Delta t}{\Delta x}(\vartheta_{j+1/2}-\vartheta_{j-1/2})\le 0, \tag{74}$$

with

$$\vartheta_{j+1/2}=\vartheta^+(\rho_j^n,u_j^n)+\vartheta^-(\rho_{j+1}^n,u_{j+1}^n), \tag{75}$$

and ϑ^+, ϑ^- verify $(\vartheta^+)'=\eta'(F^+)'$, $(\vartheta^-)'=\eta'(F^-)'$ and are given by

$$\vartheta^+(\rho,u)=\int_0^\infty \xi H(M(\rho,u,\xi),\xi)\,d\xi,\ \vartheta^-(\rho,u)=\int_{-\infty}^0 \xi H(M(\rho,u,\xi),\xi)\,d\xi, \tag{76}$$

where, for any $f=(f_0,f_1)$ with $f_0\ge 0$, $f_1=\xi f_0$,

$$H(f,\xi)=f_0\xi^2/2+\nu^2 f_0\ln f_0. \tag{77}$$

Let us recall the results of [3], which apply to our case.

DEFINITION 4.1. *A vector function $W(U)$ satisfying*

$$W'(U)^t\eta''(U) \quad \textit{is symmetric} \tag{78}$$

is said to be η-dissipative in a set $\mathcal{U}_{stab}$ if

$$D_\eta[W](U,V) \le 0 \quad \textit{for all } U,V \in \mathcal{U}_{stab}, \tag{79}$$

where the elementary entropy dissipation is defined by

$$D_\eta[W](U,V) = G_\eta[W](U) - G_\eta[W](V) + \eta'(U)(W(V) - W(U)), \tag{80}$$

and

$$G_\eta[W]' = \eta' W'. \tag{81}$$

PROPOSITION 4.2 (see [3]). *Let F^+, F^- be given by* (70), *and assume that*

$$F^+, \quad -F^- \quad \textit{are } \eta\textit{-dissipative in } \mathcal{U}_{stab}, \tag{82}$$

and the CFL conditions are

$$c\Delta t \le \Delta x, \tag{83}$$

$$\mathrm{Id} - (F^+ - F^-)/c \quad \textit{is } \eta\textit{-dissipative in } \mathcal{U}_{stab} \tag{84}$$

for some $c > 0$. Then the scheme (72)–(73) *satisfies the discrete entropy inequality* (74)–(75) *as soon as $U_j^n \in \mathcal{U}_{stab}$, $j \in \mathbb{Z}$, and $U_j^{n+1} \in \mathcal{U}_{stab}$, $j \in \mathbb{Z}$.*

The main result of this section is the following. We state it for positive densities, but by continuity it also holds for data with vacuum.

THEOREM 4.3. *Assume that at time t_n, $U_j^n = (\rho_j^n, \rho_j^n u_j^n)$ satisfy for some $c > 0$*

$$\rho_j^n > 0, \quad |u_j^n| + \left(\frac{4}{\sqrt{2\pi}} + \frac{\sqrt{2\pi}}{2}\right)\nu \le c \quad \textit{for } j \in \mathbb{Z}. \tag{85}$$

Define $\mathcal{U}_{stab}$ as

$$\mathcal{U}_{stab} = \left\{(\rho, \rho u)\ ;\ \rho > 0,\ |u| + \frac{4}{\sqrt{2\pi}}\nu \le c\right\}. \tag{86}$$

Then the stability requirements (82), (84) *hold, and under the CFL condition* (83) *the $(U_j^{n+1})_{j\in\mathbb{Z}}$ defined by* (72) *belong to $\mathcal{U}_{stab}$. Therefore, by Proposition* 4.2, *the entropy inequalities* (74) *hold.*

Before going into the proof of Theorem 4.3, let us first give explicit formulas for the fluxes and entropy fluxes. Let us define

$$\mathcal{M}(\xi) = \frac{e^{-\xi^2/2}}{\sqrt{2\pi}}, \tag{87}$$

$$\mathrm{erf}(y) = \int_y^\infty \mathcal{M}(\xi)\, d\xi. \tag{88}$$

Then

$$\begin{gathered}\int_y^\infty \xi\mathcal{M}(\xi)\,d\xi = \mathcal{M}(y), \quad \int_y^\infty \xi^2\mathcal{M}(\xi)\,d\xi = y\mathcal{M}(y) + \operatorname{erf}(y),\\ \int_y^\infty \xi^3\mathcal{M}(\xi)\,d\xi = (y^2+2)\mathcal{M}(y).\end{gathered}\tag{89}$$

With these formulas, one can easily check that

$$\begin{aligned}F^+(\rho,u) = \Big(&\rho u\operatorname{erf}(-u/\nu) + \rho\nu\mathcal{M}(u/\nu),\\ &(\rho u^2+\rho\nu^2)\operatorname{erf}(-u/\nu) + \nu\rho u\mathcal{M}(u/\nu)\Big).\end{aligned}\tag{90}$$

Noticing that, with $M = (M_0, M_1)$,

$$M(\rho,u,-\xi) = \Big(M_0(\rho,-u,\xi), -M_1(\rho,-u,\xi)\Big),\tag{91}$$

we get

$$F^-(\rho,u) = \Big(-F_0^+(\rho,-u), F_1^+(\rho,-u)\Big),\tag{92}$$

where $F^+ = (F_0^+, F_1^+)$, and therefore

$$\begin{aligned}F^+ - F^- = \Big(&\rho u\,(\operatorname{erf}(-u/\nu) - \operatorname{erf}(u/\nu)) + 2\nu\rho\mathcal{M}(u/\nu),\\ &(\rho u^2+\nu^2\rho)\,(\operatorname{erf}(-u/\nu) - \operatorname{erf}(u/\nu)) + 2\nu\rho u\mathcal{M}(u/\nu)\Big).\end{aligned}\tag{93}$$

For the computation of ϑ^+, we have

$$\begin{aligned}\vartheta^+(\rho,u) &= \int_0^\infty \xi H(M(\rho,u,\xi),\xi)\,d\xi\\ &= \int_0^\infty \xi\frac{\rho}{\nu}\mathcal{M}\left(\frac{\xi-u}{\nu}\right)\left(\nu^2\ln\frac{\rho}{\sqrt{2\pi}\nu} - u^2/2 + u\xi\right)d\xi\\ &= \left(\nu^2\ln\frac{\rho}{\sqrt{2\pi}\nu} - u^2/2\right)F_0^+(\rho,u) + uF_1^+(\rho,u)\end{aligned}\tag{94}$$

and

$$\vartheta^-(\rho,u) = -\vartheta^+(\rho,-u).\tag{95}$$

Proof of Theorem 4.3. Let us first prove that (82) and (84) hold. We notice that η is strictly convex, with

$$\eta' = \left(\nu^2\left(1+\ln\frac{\rho}{\sqrt{2\pi}\nu}\right) - u^2/2, u\right) \equiv (v_0, v_1),\tag{96}$$

and that $\eta'(\mathcal{U}_{stab})$ is convex. Therefore, as pointed out in [3], a function W is η-dissipative in $\mathcal{U}_{stab}$ if and only if it satisfies

$$W'(U)^t\eta''(U) \text{ is symmetric nonnegative for any } U\in\mathcal{U}_{stab}.\tag{97}$$

This is also equivalent to assert that

$$W = \nabla_{v_0,v_1}\psi_W, \qquad \psi_W \text{ convex in } \eta'(\mathcal{U}_{stab}).\tag{98}$$

We observe that

$$\nu^2 M_0(\rho,u,\xi) = \nu \frac{\rho}{\sqrt{2\pi}} e^{-(\xi-u)^2/2\nu^2} = \nu^2 e^{(v_0+v_1\xi-\xi^2/2)/\nu^2-1}, \tag{99}$$

thus

$$M(\rho,u,\xi) = \nabla_{v_0,v_1}\psi(\rho,u,\xi), \qquad \psi(\rho,u,\xi) = \nu^2 M_0(\rho,u,\xi), \tag{100}$$

and therefore $F^+ = \nabla_{v_0,v_1}\psi^+$, $F^- = \nabla_{v_0,v_1}\psi^-$, with

$$\psi^+ = \int_0^\infty \xi\psi(\rho,u,\xi)\,d\xi, \qquad \psi^- = \int_{-\infty}^0 \xi\psi(\rho,u,\xi)\,d\xi. \tag{101}$$

Since $(\rho,\rho u) = \nabla_{v_0,v_1}\eta^*(v_0,v_1)$ with $\eta^*(v_0,v_1) = \eta'\cdot U - \eta$, the dual convex function of η (here, $\eta^* = \nu^2\rho$), we have to check that ψ^+, $-\psi^-$ and $\eta^* - (\psi^+ - \psi^-)/c$ are convex with respect to $(v_0,v_1) \in \eta'(\mathcal{U}_{stab})$. In other words, we have to check the nonnegativity of $D^2_{v_0,v_1}\psi^+$, $-D^2_{v_0,v_1}\psi^-$, $(\eta'')^{-1} - D^2_{v_0,v_1}(\psi^+ - \psi^-)/c$ in $\eta'(\mathcal{U}_{stab})$.

From (101) and (99), we get

$$D^2_{v_0,v_1}\psi^+ = \int_0^\infty \xi\,\psi(\rho,u,\xi)\begin{pmatrix} 1 & \xi \\ \xi & \xi^2 \end{pmatrix}\frac{d\xi}{\nu^4}, \tag{102}$$

$$D^2_{v_0,v_1}\psi^- = \int_{-\infty}^0 \xi\,\psi(\rho,u,\xi)\begin{pmatrix} 1 & \xi \\ \xi & \xi^2 \end{pmatrix}\frac{d\xi}{\nu^4}, \tag{103}$$

and the two first properties follow. Next, since

$$\begin{aligned}(\eta'')^{-1} &= \nabla_{v_0,v_1}\begin{pmatrix} \rho \\ \rho u \end{pmatrix} = \nabla_{v_0,v_1}\int_{\mathbb{R}} M(\rho,u,\xi)\,d\xi \\ &= D^2_{v_0,v_1}\int_{\mathbb{R}}\psi(\rho,u,\xi)\,d\xi \\ &= \int_{\mathbb{R}}\psi(\rho,u,\xi)\begin{pmatrix} 1 & \xi \\ \xi & \xi^2 \end{pmatrix}\frac{d\xi}{\nu^4},\end{aligned} \tag{104}$$

we obtain for $(\rho,\rho u) \in \mathcal{U}_{stab}$

$$\begin{aligned}&c(\eta'')^{-1} - D^2_{v_0,v_1}(\psi^+ - \psi^-) \\ &= \frac{1}{\nu^4}\int_{\mathbb{R}}(c - |\xi|)\psi(\rho,u,\xi)\begin{pmatrix} 1 & \xi \\ \xi & \xi^2 \end{pmatrix}d\xi \\ &\geq \frac{1}{\nu^4}\int_{\mathbb{R}}\left(|u| + \frac{4}{\sqrt{2\pi}}\nu - |\xi|\right)\psi(\rho,u,\xi)\begin{pmatrix} 1 & \xi \\ \xi & \xi^2 \end{pmatrix}d\xi \\ &\geq \frac{1}{\nu^4}\int_{\mathbb{R}}\left(\frac{4}{\sqrt{2\pi}}\nu - |\xi - u|\right)\psi(\rho,u,\xi)\begin{pmatrix} 1 & \xi \\ \xi & \xi^2 \end{pmatrix}d\xi.\end{aligned} \tag{105}$$

Noticing that

$$\begin{pmatrix} 1 & \xi \\ \xi & \xi^2 \end{pmatrix}\cdot\begin{pmatrix} y_0 \\ y_1 \end{pmatrix}\cdot\begin{pmatrix} y_0 \\ y_1 \end{pmatrix} = \begin{pmatrix} 1 & (\xi-u)/\nu \\ (\xi-u)/\nu & (\xi-u)^2/\nu^2 \end{pmatrix}\cdot\begin{pmatrix} y_0' \\ y_1' \end{pmatrix}\cdot\begin{pmatrix} y_0' \\ y_1' \end{pmatrix}$$

with $y_0' = y_0 + uy_1$ and $y_1' = \nu y_1$, we observe that the positiveness of (105) is equivalent to

$$\int_{\mathbb{R}} \left(\frac{4}{\sqrt{2\pi}}\nu - |\xi - u| \right) \psi(\rho, u, \xi) \begin{pmatrix} 1 & (\xi - u)/\nu \\ (\xi - u)/\nu & (\xi - u)^2/\nu^2 \end{pmatrix} d\xi \geq 0. \tag{106}$$

This is checked easily with (89); the computation of this integral is left to the reader; and this concludes the first part of the theorem.

It now remains to prove that $U_j^{n+1} \in \mathcal{U}_{stab}$.

Lemma 4.4. *For any $\rho \geq 0$, $u \in \mathbb{R}$, we have $F_0^+(\rho, u) \geq 0$, $F_0^-(\rho, u) \leq 0$, and*

$$0 \leq F_1^+(\rho, u) \leq \left(u_+ + \nu\sqrt{2\pi}/2\right) F_0^+(\rho, u), \tag{107}$$

$$0 \leq F_1^-(\rho, u) \leq \left((-u)_+ + \nu\sqrt{2\pi}/2\right)(-F_0^-(\rho, u)). \tag{108}$$

Moreover, whenever $(\rho, \rho u) \in \mathcal{U}_{stab}$,

$$\rho - \frac{F_0^+(\rho, u) - F_0^-(\rho, u)}{c} > 0, \tag{109}$$

$$\left| \rho u - \frac{F_1^+(\rho, u) - F_1^-(\rho, u)}{c} \right| \leq \left(\rho - \frac{F_0^+(\rho, u) - F_0^-(\rho, u)}{c} \right) |u|. \tag{110}$$

Let us postpone the proof of the lemma and conclude with Theorem 4.3. Since $U \equiv U_j^n$ satisfies (85), we have $U \in \mathcal{U}_{stab}$, and by (109)–(110)

$$\left| \frac{\rho u - (F_1^+ - F_1^-)/c}{\rho - (F_0^+ - F_0^-)/c} \right| + \frac{4}{\sqrt{2\pi}}\nu \leq c. \tag{111}$$

Then, from (107), (108), and (85)

$$\frac{F_1^+}{F_0^+} + \frac{4}{\sqrt{2\pi}}\nu \leq u_+ + \frac{\sqrt{2\pi}}{2}\nu + \frac{4}{\sqrt{2\pi}}\nu \leq c, \tag{112}$$

$$\frac{F_1^-}{-F_0^-} + \frac{4}{\sqrt{2\pi}}\nu \leq (-u)_+ + \frac{\sqrt{2\pi}}{2}\nu + \frac{4}{\sqrt{2\pi}}\nu \leq c. \tag{113}$$

We deduce that

$$F^+(\rho, u)/c, \quad -F^-(\rho, u)/c, \quad U - (F^+(\rho, u) - F^-(\rho, u))/c \ \in \mathcal{U}_{stab}. \tag{114}$$

Finally, we write

$$\begin{aligned} U_j^{n+1} &= U_j^n - \frac{\Delta t}{\Delta x}\left(F^+(U_j^n) + F^-(U_{j+1}^n) - F^+(U_{j-1}^n) - F^-(U_j^n)\right) \\ &= \left(1 - \frac{c\Delta t}{\Delta x}\right) U_j^n + \frac{c\Delta t}{\Delta x}\left(U_j^n - F^+(U_j^n)/c + F^-(U_j^n)/c\right) \\ &\quad + \frac{\Delta t}{\Delta x} F^+(U_{j-1}^n) - \frac{\Delta t}{\Delta x} F^-(U_{j+1}^n), \end{aligned} \tag{115}$$

and since $\mathcal{U}_{stab}$ is a convex cone we conclude with the CFL condition (83) that $U_j^{n+1} \in \mathcal{U}_{stab}$. □

Proof of Lemma 4.4. From (69) and (71), we obviously have $F_0^+ \geq 0$, $F_1^+ \geq 0$. However, from (90),

$$F_1^+ = uF_0^+ + \nu^2 \rho \operatorname{erf}(-u/\nu). \tag{116}$$

One can check that

$$\sup_{y \geq 0} \frac{\operatorname{erf}(-y)}{y \operatorname{erf}(-y) + \mathcal{M}(y)} = \frac{\sqrt{2\pi}}{2}, \tag{117}$$

the maximum being attained at $y = 0$. Therefore,

$$\text{for all } u \geq 0, \qquad F_1^+(\rho, u) \leq \left(u + \frac{\sqrt{2\pi}}{2} \nu \right) F_0^+(\rho, u). \tag{118}$$

Then we observe that if $\rho > 0$,

$$F_0^+ = \int_0^\infty \xi e^{(v_0 + v_1 \xi - \xi^2/2)/\nu^2 - 1} \, d\xi, \tag{119}$$

with (v_0, v_1) defined in (96), thus this function is log-convex in v_1 as an integral of a log-convex function. Next,

$$F_1^+ = \int_0^\infty \xi^2 e^{(v_0 + v_1 \xi - \xi^2/2)/\nu^2 - 1} \, d\xi = \nu^2 \partial_{v_1} F_0^+. \tag{120}$$

Therefore, $F_1^+/F_0^+ = \nu^2 \partial_{v_1} (\ln F_0^+)$ is nondecreasing in v_1, and we deduce that, for $v_1 \leq 0$, F_1^+/F_0^+ is upper bounded by the value at $v_1 = 0$, that is, $\nu\sqrt{2\pi}/2$. Together with (118), we conclude that (107) holds. Obviously, (108) follows from (92). In order to prove (109)–(110), let us first establish that, for any $u \in \mathbb{R}$,

$$(u^2 + \nu^2) \frac{2}{u} \int_0^{u/\nu} \frac{e^{-\xi^2/2}}{\sqrt{2\pi}} d\xi \leq |u| + \frac{2}{\sqrt{2\pi}} \nu. \tag{121}$$

Indeed, the left-hand side is an even nonnegative function of u. On one hand, if $u \geq \nu\sqrt{2\pi}/2$, by bounding the integral by $1/2$, we can estimate the left-hand side by $(u^2 + \nu^2)/u \leq u + 2\nu/\sqrt{2\pi}$. On the other hand, if $0 < u \leq \nu\sqrt{2\pi}/2$, we can estimate the integral by $u/\sqrt{2\pi}\nu$, thus the left-hand side is less than $(u^2 + \nu^2)2/\sqrt{2\pi}\nu \leq u + 2\nu/\sqrt{2\pi}$. This proves (121).

Now we use (93) and get that

$$\rho - \frac{F_0^+ - F_0^-}{c} = \rho \left(1 - \left(2u \int_0^{u/\nu} \mathcal{M}(\xi) \, d\xi + 2\nu \mathcal{M}(u/\nu) \right) /c \right), \tag{122}$$

and using (121) and $\mathcal{M}(u/\nu) \leq 1/\sqrt{2\pi}$ we get with (86) that (122) is positive. Finally, from (93),

$$\rho u - \frac{F_1^+ - F_1^-}{c} = u \left(\rho - \frac{F_0^+ - F_0^-}{c} - \frac{\nu^2 \rho}{c} \frac{2}{u} \int_0^{u/\nu} \mathcal{M}(\xi) \, d\xi \right). \tag{123}$$

By (121) and (86) again,

$$0 \le \rho - \frac{F_0^+ - F_0^-}{c} - \frac{\nu^2\rho}{c}\frac{2}{u}\int_0^{u/\nu} \mathcal{M}(\xi)\,d\xi \le \rho - \frac{F_0^+ - F_0^-}{c}, \tag{124}$$

which yields (110). □

Remark. When using the Godunov method to solve the isothermal problem with vacuum initial data, the speed defined by the exact Riemann solution is infinity; see the third test in section 5. The kinetic scheme, which takes the form of a flux vector splitting, can be interpreted as an approximate Riemann solver with a finite speed [3], thus the timestep can be nonzero. The speed of this approximate Riemann solver is the c in Theorem 4.3, which needs to bound the eigenvalues, but *only on initial data* rather than the exact solution (which would be infinity). The counterpart is that the sound speed is increased by a factor greater than one, $4/\sqrt{2\pi} + \sqrt{2\pi}/2$ in (85). Then, at the end of the first timestep, the new velocity will be greater than its initial value, thus c needs to be recomputed at each timestep. A rough estimate of this increase is contained in the statement in Theorem 4.3 that says that U_j^{n+1} belongs to $\mathcal{U}_{stab}$. Thus the maximal velocity increases by at most a constant, $\nu\sqrt{2\pi}/2$, at each timestep (that is chosen to satisfy the CFL condition). This gives the following rough estimates:

$$\|u^n\| \approx \|u^0\| + n\nu, \qquad \Delta t_n \approx \Delta x/(\|u^0\| + n\nu)\,.$$

The sum over all n of these timesteps is infinity. This justifies that we can attain any given time T in a finite number of timesteps, roughly $\Delta x/\nu \ln(1+n\nu/\|u^0\|) = T$, thus $n = \|u^0\|(e^{T\nu/\Delta x} - 1)/\nu$. The size of the support is of the order of $n\Delta x$.

5. Numerical tests.

We now conduct several numerical tests using the numerical schemes we have derived. In the first two tests, we solve the pressureless gas system and compare the first-order kinetic scheme with the first-order Godunov scheme and the second-order kinetic schemes. In the third test, we solve the isothermal system with vacuum data by both the first- and the second-order kinetic schemes.

In the first numerical test, we take the initial data to be

$$\rho^0 = 0.5\,, \qquad u^0(x) = \begin{cases} -0.5, & x < -0.5, \\ 0.4, & -0.5 < x < 0, \\ 0.4 - x, & 0 < x < 0.8, \\ -0.4, & x > 0.8. \end{cases}$$

The exact solution at time $t = 0.5$ is

$$u(x, 0.5) = \begin{cases} -0.5, & x < -0.75, \\ \text{undefined} & -0.75 < x < -0.3, \\ 0.4, & -0.3 < x < 0.2, \\ 0.8 - 2x, & 0.2 < x < 0.6, \\ -0.4, & x > 0.6, \end{cases}$$

and

$$\rho(x, 0.5) = \begin{cases} 0, & -0.75 < x < -0.3, \\ 1, & 0.2 < x < 0.6, \\ 0.5 & \text{otherwise.} \end{cases}$$

The initial velocity jumps to a higher value at $x = -0.5$, which leads to a vacuum state, followed by a linearly decreasing part, where the mass accumulates and causes

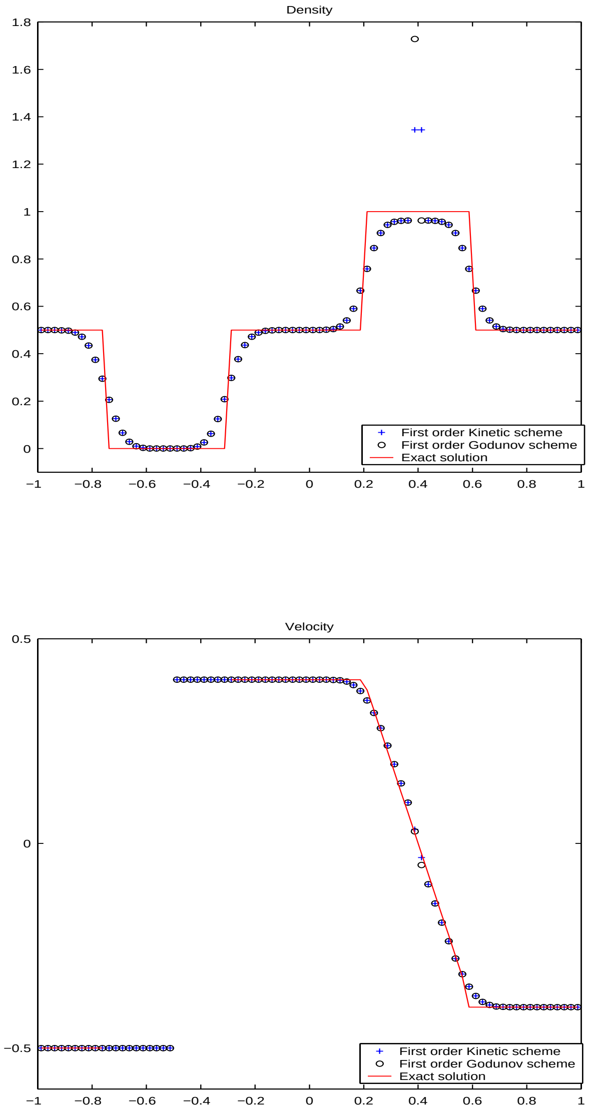


FIG. 1. *Numerical test* I. $\Delta t/\Delta x = 5/3$, $\Delta x = 0.025$.

the density to increase. In the density profile, first-order kinetic and Godunov schemes form a spike, due to some inconsistency at sonic points that was noticed in [2]. See Figure 1. This problem goes away with all second-order kinetic schemes, among which the improved scheme slightly outperforms the other two and the simplified scheme is slightly inferior to the other two, as shown by Figure 2.

A Riemann problem is solved in the second test. We take

$$(\rho^0(x), u^0(x)) = \begin{cases} (1, 0.5) & x < 0, \\ (0.25, -0.4) & x > 0. \end{cases}$$

A delta-shock immediately develops and by (18) the shock speed is 0.2. As shown in Figures 3 and 4, all the numerical schemes are able to capture the delta-shock with

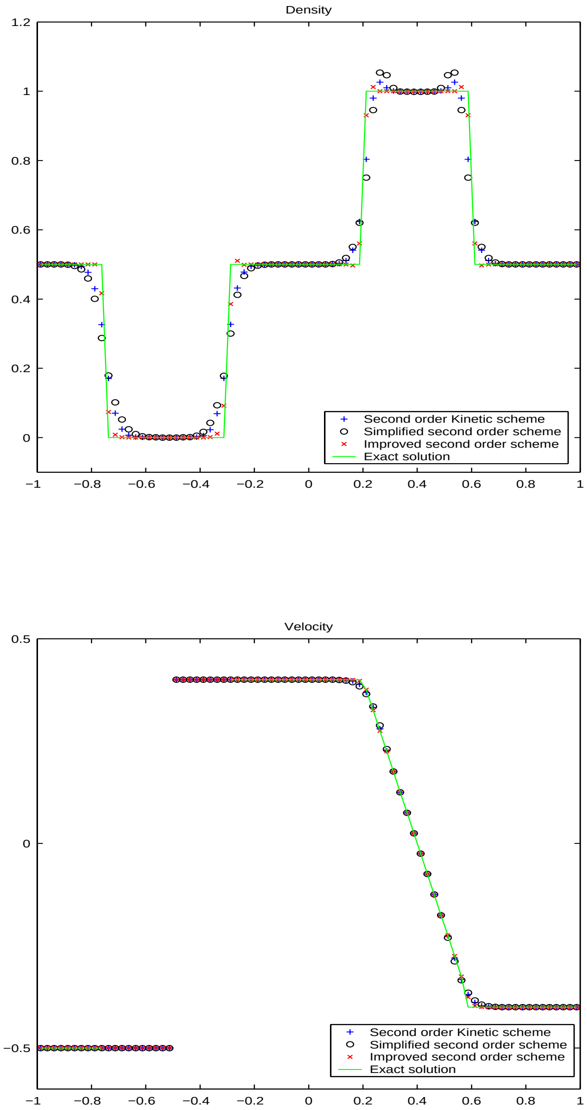


FIG. 2. *Numerical test* I. $\Delta t/\Delta x = 5/3$, $\Delta x = 0.025$.

the correct propagation speed. The numerical approximations provided by the second-order schemes give a slightly sharper profile across the velocity discontinuity.

In the third test, we solve the isothermal gas equations (66) with $\nu = 0.2$ and the vacuum initial data

$$\rho^0(x) = \begin{cases} 0 & x < 0, \\ 1 & x > 0, \end{cases} \qquad u^0(x) = 0.$$

The exact solution at time t is given by

$$(\rho(x,t), u(x,t)) = \begin{cases} (\mathrm{e}^{x/\nu t-1}, x/t - \nu) & x < \nu t, \\ (1,0) & x > \nu t. \end{cases}$$

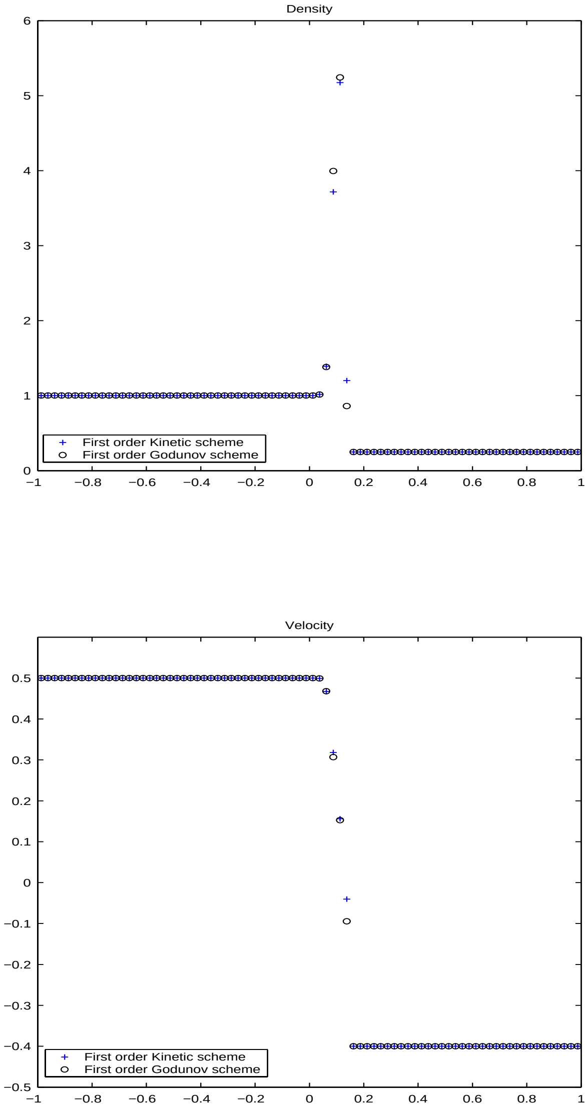


FIG. 3. *Numerical test* II. $\Delta t/\Delta x = 5/3$, $\Delta x = 0.025$.

The density decays exponentially to zero while the velocity goes to minus infinity linearly in the vacuum state. The numerical results for $t = 0.5$ are shown in Figure 5. Both first- and second-order kinetic schemes give accurate density profile, while for the velocity the second-order scheme yields more accurate velocity slope than the first-order one.

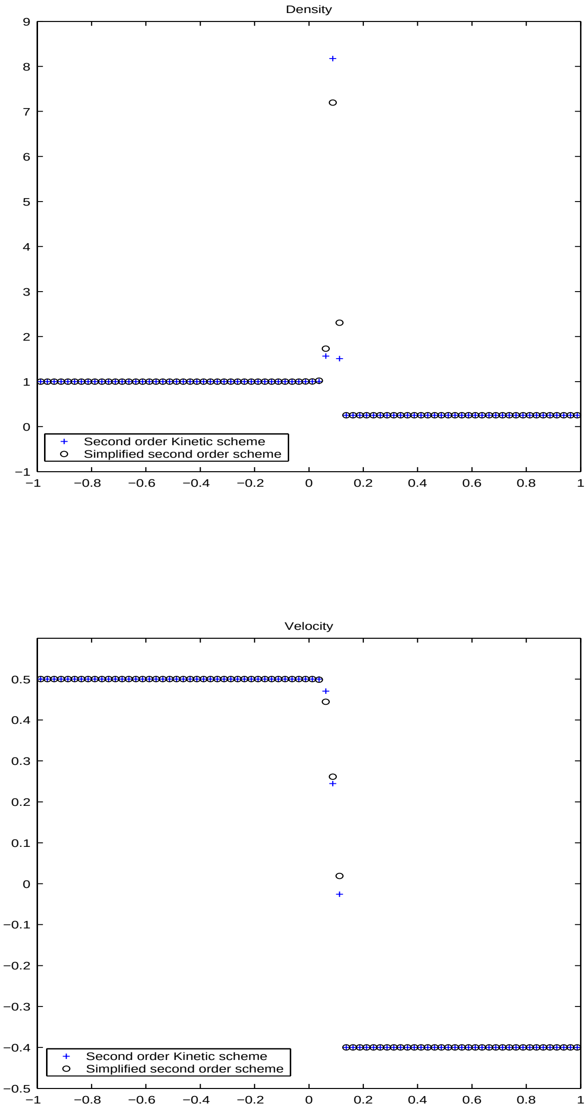


FIG. 4. *Numerical test* II. $\Delta t/\Delta x = 5/3$, $\Delta x = 0.025$.

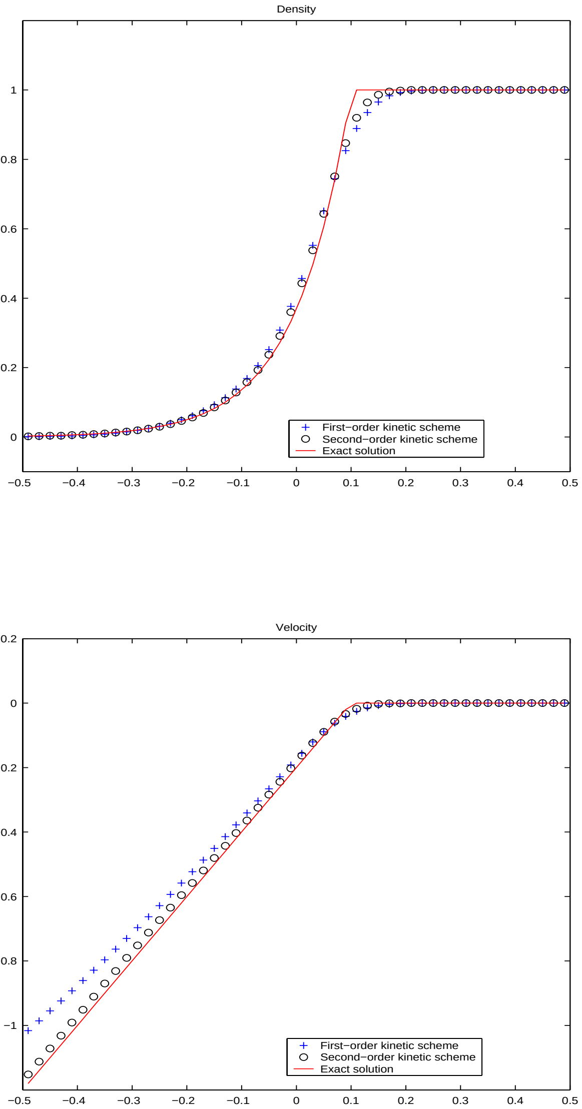


FIG. 5. *Numerical test* III. $t = 0.5$, $\Delta x = 0.02$.

REFERENCES

[1] F. BERTHELIN, *Existence and weak stability for a pressureless model with unilateral constraint*, Math. Models Methods Appl. Sci., 12 (2002), pp. 249–272.

[2] F. BOUCHUT, *On zero pressure gas dynamics*, in Advances in Kinetic Theory and Computing, Ser. Adv. Math. Appl. Sci. 22, World Scientific, River Edge, NJ, 1994, pp. 171–190.

[3] F. BOUCHUT, *Entropy satisfying flux vector splittings and kinetic BGK models*, Numer. Math., to appear.

[4] F. BOUCHUT AND G. BONNAUD, *Numerical simulation of relativistic plasmas in hydrodynamic regime*, Z. Angew. Math. Mech., 76 (1996), pp. 287–290.

[5] F. BOUCHUT, G. BONNAUD, S. DUSSY, AND E. LEFEBVRE, *Comportement électromagnétique d'un plasma en régimes hydrodynamique et relativiste: code de simulation RHEA*, Technical report CEA-R-5807, CEA, France, 1998.

[6] F. BOUCHUT, Y. BRENIER, J. CORTES, AND J.-F. RIPOLL, *A hierarchy of models for two-phase*

flows, J. Nonlinear Sci., 10 (2000), pp. 639–660.

[7] F. Bouchut and F. James, *Duality solutions for pressureless gases, monotone scalar conservation laws, and uniqueness*, Comm. Partial Differential Equations, 24 (1999), pp. 2173–2189.

[8] L. Boudin, *A solution with bounded expansion rate to the model of viscous pressureless gases*, SIAM J. Math. Anal., 32 (2000), pp. 172–193.

[9] Y. Brenier, *Résolution d'équations d'évolution quasilinéaires en dimension N d'espace à l'aide d'équations linéaires en dimension $N+1$*, J. Differential Equations, 50 (1983), pp. 375–390.

[10] Y. Brenier, *Equations de moment et conditions d'entropie pour des modèles cinétiques*, in Séminaire sur les Equations aux Dérivées Partielles, 1994–1995, Exp. No. XXII, Ecole Polytech., Palaiseau, France, 1995.

[11] Y. Brenier and E. Grenier, *Sticky particles and scalar conservation laws*, SIAM J. Numer. Anal., 35 (1998), pp. 2317–2328.

[12] G.Q. Chen and H.L. Liu, *Concentration and Cavitation in Vanishing Pressure Limit for Compressible Flow*, preprint.

[13] G.Q. Chen and H.L. Liu, *Formation of δ-Shocks and vacuum states in the vanishing pressure limit of solutions to the Euler equations for isentropic fluids*, SIAM J. Math. Anal., to appear.

[14] W. E, Y.G. Rykov, and Y.G. Sinai, *Generalized variational principles, global weak solutions and behavior with random initial data for systems of conservation laws arising in adhesion particle dynamics*, Comm. Math. Phys., 177 (1996), pp. 349–380.

[15] B. Engquist and O. Runborg, *Multi-phase computations in geometric optics*, J. Comput. Appl. Math., 74 (1996), pp. 175–192.

[16] L. Gosse, *Using K-branch entropy solutions for multivalued geometric optics computations*, J. Comput. Phys., 180 (2002), pp. 155–182.

[17] G.-S. Jiang and E. Tadmor, *Nonoscillatory central schemes for multidimensional hyperbolic conservation laws*, SIAM J. Sci. Comput., 19 (1998), pp. 1892–1917.

[18] S. Jin and X.T. Li, *Multi-phase computations of the semiclassical limit of the Schrödinger equation and related problems: Whitham vs. Wigner*, Phys. D, submitted.

[19] B. Perthame, *Boltzmann type schemes for gas dynamics and the entropy property*, SIAM J. Numer. Anal., 27 (1990), pp. 1405–1421.

[20] B. Perthame and C.-W. Shu, *On positivity preserving finite volume schemes for Euler equations*, Numer. Math., 73 (1996), pp. 119–130.

[21] Y.G. Rykov, *Propagation of singularities of shock wave type in a system of equations of two-dimensional pressureless gas dynamics*, Mat. Zametki, 66 (1999), pp. 760–769 (in Russian); Math. Notes, 66 (1999), pp. 628–635 (2000) (in English).

[22] M. Sever, *An existence theorem in the large for zero-pressure gas dynamics*, Differential Integral Equations, 14 (2001), pp. 1077–1092.

[23] W. Sheng and T. Zhang, *The Riemann problem for the transportation equations in gas dynamics*, Mem. Amer. Math. Soc., 137 (1999).

[24] M. Vergassola, B. Dubrulle, U. Frisch, and A. Noullez, *Burgers' equation, devil's staircases and the mass distribution function for large-scale structures*, Astron. and Astrophys., 289 (1994), pp. 325–356.

[25] Z. Wang and X. Ding, *Uniqueness of generalized solution for the Cauchy problem of transportation equations*, Acta Math. Sci. (English Ed.), 17 (1997), pp. 341–352.

[26] Ya. B. Zeldovich, *Gravitational instability: An approximate theory for large density perturbations*, Astron. and Astrophys., 5 (1970), pp. 84–89.

SIAM J. NUMER. ANAL.
Vol. 41, No. 1, pp. 159–180

WAVELET DISCRETIZATIONS OF PARABOLIC INTEGRODIFFERENTIAL EQUATIONS*

T. VON PETERSDORFF[†] AND C. SCHWAB[‡]

Abstract. We consider parabolic problems $\dot{u} + Au = f$ in $(0,T) \times \Omega$, $T < \infty$, where $\Omega \subset \mathbb{R}^d$ is a bounded domain and A is a strongly elliptic classical pseudodifferential operator of order $\rho \in [0,2]$ in $\tilde{H}^{\rho/2}(\Omega)$. We use a θ-scheme for time discretization and a Galerkin method with N degrees of freedom for space discretization. The full Galerkin matrix for A can be replaced with a sparse matrix using a wavelet basis, and the linear systems for each time step are solved approximatively with GMRES. We prove that the total cost of the algorithm for M time steps is bounded by $O(MN(\log N)^\beta)$ operations and $O(N(\log N)^\beta)$ memory. We show that the algorithm gives optimal convergence rates (up to logarithmic terms) for the computed solution with respect to L^2 in time and the energy norm in space.

Key words. parabolic integrodifferential equation, Lévy-process, wavelet compression, discontinuous Galerkin method

AMS subject classifications. 35K15, 45K05, 65M12, 65M60, 65T60

PII. S0036142901394844

1. Introduction.

Fast algorithms such as wavelets and multipole or clustering methods for the numerical solution of elliptic integrodifferential equations

$$A[u](x) = \int_\Omega k(x, x-y)u(y)dy = f, \qquad x \in \Omega, \tag{1.1}$$

with kernel function $k(x,z)$, have been introduced and analyzed in recent years (see, e.g., [3, 4, 7]). In the present paper we investigate the numerical solution of a class of parabolic integrodifferential equations

$$u_t = A[u](x) + f \quad \text{ in } (0,T) \times \Omega, \tag{1.2}$$

with suitable initial and boundary conditions. Such equations arise as Kolmogorov forward equations for Lévy processes X_t with infinitesimal generators $A[u]$. Brownian motion B_t with diffusion $\sigma(x)$ and drift $r(x)$ is a particular Lévy process. The infinitesimal generator of B_t in dimension $d = 1$ is the second order elliptic differential operator

$$A_B[u](x) = -\frac{d}{dx}\left(\sigma(x)\frac{du}{dx}(x)\right) + r(x)\frac{du}{dx}, \tag{1.3}$$

and the Kolmogorov forward equation is the diffusion equation with drift. The decomposition theorem of Lévy states that the infinitesimal generator A of any Lévy-process X_t is the sum of a differential operator A_B as in (1.3), which accounts for the diffusion

[*] Received by the editors August 31, 2001; accepted for publication (in revised form) June 24, 2002; published electronically March 13, 2003. This research was supported by the Swiss government as part of the TMR-network “Wavelets and Multiscale Methods in Numerical Analysis” under BBW 97.404.

http://www.siam.org/journals/sinum/41-1/39484.html

[†] Department of Mathematics, University of Maryland, College Park, MD 20742 (tvp@math.umd.edu).

[‡] Seminar für Angewandte Mathematik, ETH Zürich, HG G 58.1, CH 8092 Zürich, Switzerland (christoph.schwab@sam.math.ethz.ch).

part of X_t and could possibly vanish, and a nonlocal operator A_L of the form (1.1), which corresponds to the pure jump part of the process (see, e.g., [2, 12]). The order ρ of the infinitesimal generator A of a Lévy-process always satisfies

$$0 \leq \rho \leq 2. \tag{1.4}$$

We emphasize that due to (1.4) the kernels $k(x, z)$ are not integrable near $z = 0$ and that the integral in (1.1) has to be understood as a finite part or principal value, i.e., in the sense of distributions [16]. Interpretations of the integral operators A in the distribution sense can naturally be accounted for in Galerkin discretizations.

While the initial-boundary value problems (1.2) with (1.3) and constant σ, r can be solved analytically for certain initial conditions, numerical solutions are required for nonconstant coefficients, general Lévy-processes, and free boundary problems arising with optimal stopping of X_t. In a numerical solution, $u(x, t)$ is approximated by finite differences or finite elements in x with N degrees of freedom, reducing (1.2) to a system of N ordinary differential equations for the approximation u_N which must be integrated in t by a time-stepping scheme. We consider the θ-scheme for time discretization, which includes as special cases the forward Euler method ($\theta = 0$), the backward Euler method ($\theta = 1$), and the Crank–Nicolson method ($\theta = \frac{1}{2}$). In general this leads to implicit methods where a linear system has to be solved for each time step. For the differential operator (1.3) in dimension $d = 1$, the matrices to be inverted in each time step are banded and can be factored in $O(N)$ operations. If the operator A is nonlocal, however, standard Galerkin discretizations of u with N degrees of freedom entail dense stiffness matrices and hence at least $O(N^2)$ complexity per time step for the numerical solution of (1.2). We reduce this complexity by a wavelet-based matrix compression as in [10, 11, 14, 4]. The basic idea behind this compression is to represent the Galerkin approximation u_N of (1.2) in a wavelet basis. Wavelet matrix compression exploits the fact that the generators A are often classical pseudodifferential operators, which implies special properties of their Schwartz kernel function $k(x, z)$ such as

$$\text{sing supp}(k(x, z)) \subset \{z = 0\}$$

and even analyticity of $k(x, z)$ off the origin $z = 0$. Wavelet matrix compression requires only finite differentiability of $k(x, z)$ for $z \neq 0$ and allows the generation of an approximate stiffness matrix of the nonlocal operator A in (1.1) in $O(N(\log N)^a)$ memory and operations where $a \geq 0$ is a small integer (see, e.g., [3, 4, 10, 9, 11, 14] and the references there).

The analysis of the impact of this truncation error on stability and consistency of the θ time-stepping scheme for the nonlocal parabolic initial-boundary value problems (1.1), (1.2) is the purpose of the present paper. A large body of literature on time-stepping for parabolic problems with Galerkin discretization is available; see [18] and the references there. The impact of quadrature errors on spatial semidiscretizations was also investigated early on, for example, in [13]. However, the present fully discrete setting with integral operators and wavelet matrix compression causes consistency errors which do not fit readily into existing error analyses of fully discrete schemes: For consistency errors resulting from numerical integration, one has that the difference between the energy bilinear form $a(u_h, v_h)$ and the perturbed bilinear form $\tilde{a}_h(u_h, v_h)$ goes to zero with respect to the energy norms of u_h, v_h; i.e.,

$$|a(u_h, v_h) - \tilde{a}_h(u_h, v_h)| \leq \eta(h) \|u_h\|_V \|v_h\|_V , \quad \text{with } \eta(h) \to 0 \text{ as } h \to 0.$$

For a perturbed bilinear form $\tilde{a}_h(u_h, v_h)$ resulting from wavelet compression, however, this is no longer true.

Here, we develop a framework for perturbation error analysis of the θ-scheme which accommodates consistency errors due to wavelet compression of the spatial nonlocal operator A without any loss in order. We note in passing that our framework can accommodate other matrix compression techniques, such as, e.g., multipole-based methods. Unlike wavelet-based compression, these techniques exploit the analyticity of the kernel function and yield exponentially convergent matrix approximations. We also give a stability analysis for the perturbed θ-schemes that does not require symmetry or resort to eigenfunction expansions of the spatial operator. As is well known, the stability of explicit time-stepping schemes for Galerkin discretizations for the parabolic problems (1.2), (1.4) requires a CFL condition which, as we will show, depends on the order ρ of the operator A and which takes the form

$$\Delta t \leq C(\Delta x)^{\rho}, \quad \rho \in [0, 2]. \tag{1.5}$$

For $\rho = 2$, e.g., the heat equation, we recover the classical CFL condition which forces small time steps Δt in explicit schemes when the meshwidth $h = \Delta x$ of the space discretization is reduced. If, however, the order of A is $\rho \leq 1$, condition (1.5) is of the type usually encountered in time-stepping for first order hyperbolic equations, and explicit time-stepping schemes appear competitive.

Next, we present classes of spline wavelets and a matrix compression strategy which leads to sparse approximations for the stiffness matrix of A with $O(N \log N)$ (rather than $O(N^2)$ for standard Galerkin schemes) nonvanishing entries. We prove that this compression preserves the asymptotic convergence rates of the full Galerkin scheme.

In the θ-scheme, a linear system of equations at each implicit time step must be solved. Since the compressed matrices are not banded and possibly nonsymmetric (due to the presence of a drift term or if $k(x, z)$ is asymmetric for $z \to \pm\infty$), we propose inexact equation solution by GMRES iteration. Using wavelets, we precondition the compressed matrix in dependence on the discretization parameters and the order ρ of A. We relate the GMRES stopping criterion to the discretization error of the scheme and prove that the resulting method converges still with optimal order in space and time while its complexity is essentially $O(N)$ memory and operations per (explicit or implicit) time step. This is comparable to the complexity for the heat equation using the backward Euler method in time and banded matrices in space.

We emphasize that our analysis is applicable to general kernels $k(x, z)$, translation invariant or not, of any order $\rho \geq 0$. Therefore, the θ-scheme with wavelet compression allows the numerical solution of the Kolmogorov equations (1.2) for a large class of Lévy-processes with complexity comparable to standard finite differences for the heat equation in one dimension.

The outline of the paper is as follows. In section 2, we present the class of parabolic problems and the class of spatial integrodifferential operators A admissible in our analysis. In section 3, we discuss the fully discrete θ-scheme. We describe wavelet Galerkin discretization of A and give several examples of wavelets. Section 4 is devoted to the stability analysis of the θ-scheme with compression in the "explicit" case $0 \leq \theta < \frac{1}{2}$ as well as in the implicit case $\theta > \frac{1}{2}$. In section 5, we prove our convergence estimates with particular attention to the error due to wavelet compression of the stiffness matrix $\mathbf{A}$ of A. Section 6 is devoted to the complexity estimates, the matrix preconditioning in the implicit time-stepping schemes, and the error analysis in the presence of incomplete GMRES iterations. Throughout, C will denote a generic positive constant independent of the discretization parameters taking different values in different places. If the value of C is relevant, we write also C_i.

2. Problem formulation. In the time interval $J = (0, T)$ with $T > 0$, we consider parabolic evolution problems of the form

$$\dot{u}(t) + Au(t) = g(t), \qquad t \in J, \tag{2.1}$$

$$u(0) = u_0, \tag{2.2}$$

where A is a possibly nonlocal operator of order $\rho > 0$.

For a variational formulation of this problem we introduce Sobolev spaces. Let $\Omega \subset \mathbb{R}^d$ be a bounded domain with Lipschitz boundary $\Gamma = \partial\Omega$. We denote by $H = L^2(\Omega)$ the usual square integrable functions with inner product $(.,.)$, and by $H^s(\Omega)$, $s \geq 0$, the corresponding Sobolev spaces (see, e.g., [1]). Further, for $s \geq 0$, we define the space

$$\tilde{H}^s(\Omega) = \{\, u|_\Omega \mid u \in H^s(\mathbb{R}^d),\, u|_{\mathbb{R}^d \setminus \Omega} = 0\}. \tag{2.3}$$

If $s + 1/2 \notin \mathbb{N}$, then $\tilde{H}^s(\Omega)$ coincides with $H_0^s(\Omega)$, the closure of $C_0^\infty(\Omega)$ with respect to the norm in $H^s(\Omega)$. We identify $L^2(\Omega)$ with its dual and define

$$V = \widetilde{H}^{\rho/2}(\Omega). \tag{2.4}$$

Then $V \overset{d}{\hookrightarrow} L^2(\Omega)$ with dense injection, and V^*, the dual of V, satisfies

$$V \overset{d}{\hookrightarrow} L^2(\Omega) \overset{d}{\hookrightarrow} V^*. \tag{2.5}$$

We assume that $A \in \mathcal{L}(V; V^*)$. By $(\cdot,\cdot)_{V^* \times V}$ we denote the extension of $(.,.)$ as duality pairing in $V^* \times V$, and by $\|\cdot\|$, $\|\cdot\|_V$, $\|\cdot\|_{V^*}$ the norms in $L^2(\Omega), V, V^*$, respectively. We associate with A the bilinear form $a(\cdot,\cdot)\colon V \times V \to \mathbb{C}$ via

$$a(u, v) := (Au, v)_{V^* \times V}, \qquad u, v \in V. \tag{2.6}$$

Then the form $a(\cdot,\cdot)$ is continuous,

$$\forall u, v \in V: \quad |a(u, v)| \leq \alpha \|u\|_V \|v\|_V, \tag{2.7}$$

and we assume that it is coercive in the sense that

$$\forall u \in V: \quad a(u, u) \geq \beta \|u\|_V^2 \tag{2.8}$$

for some $0 < \beta \leq \alpha < \infty$. Then $A \in \mathcal{L}(V, V^*)$ is an isomorphism and $\|A\|_{\mathcal{L}(V,V^*)} \leq \alpha$, $\|A^{-1}\|_{\mathcal{L}(V^*,V)} \leq \frac{1}{\beta}$. The time derivative $\dot{u}(t)$ in (2.1) is understood in the weak sense; i.e., for $u \in L^2(J; V)$ we have $\dot{u} \in L^2(J; V^*)$ defined by

$$\int_J (\dot{u}(t), v)_{V^* \times V}\, \varphi(t) dt = -\int_J (u(t), v)\, \dot{\varphi}(t) dt \tag{2.9}$$

for every $v \in V$, $\varphi \in C_0^\infty(J)$. The weak form of (2.1), (2.2) reads: Given

$$u_0 \in H, \qquad g \in L^2(J; H), \tag{2.10}$$

find $u \in L^2(J;V) \cap H^1(J;V^*)$ such that $u(0) = u_0$ and, for every $v \in V$, $\varphi \in C_0^\infty(J)$,

$$-\int_J (u(t),v)\,\dot{\varphi}(t)dt + \int_J a(u,v)\,\varphi(t)dt = \int_J (g(t),v)_{V^*\times V}\,\varphi(t)dt. \tag{2.11}$$

Note that the initial condition is well defined since (see [8])

$$L^2(J;V) \cap H^1(J;V^*) \subset C^0([0,T];\ H). \tag{2.12}$$

Under the assumption (2.10), problem (2.11) has a unique weak solution $u(t)$, and there holds the a priori estimate (see, e.g., [8])

$$\|u\|_{C(\overline{J};H)} + \|u\|_{L^2(J;V)} + \|\dot{u}\|_{L^2(J;V^*)} \le C(\|g\|_{L^2(J;H)} + \|u_0\|_H). \tag{2.13}$$

Remark 2.1.

(i) We do not assume A to be self-adjoint. The form $a(\cdot,\cdot)$ need not be symmetric.

(ii) Properties (2.7) and (2.8) allow us to define on V an equivalent norm by

$$\|u\|_a := (a(u,u))^{1/2} \sim \|u\|_V, \tag{2.14}$$

to which we shall refer below as the “energy-norm.”

(iii) Testing (2.1) with $u(t)$ in the $(\cdot,\cdot)$ inner product, we find with (2.6) that for almost every $t \in (0,T)$

$$(u,\dot{u}) + a(u,u) = (u,g),$$

and integrating from $t=0$ to $t=T$, we find

$$\frac{1}{2}\|u(T)\|^2 - \frac{1}{2}\|u(0)\|^2 + \int_0^T a(u(t),\,u(t))dt = \int_0^T (u,g)dt$$
$$\le \int_0^T \|u(t)\|_a \sup_{v\in V} \frac{(v,g)}{\|v\|_a}\,dt \le \frac{1}{2}\int_0^T \|u(t)\|_a^2\,dt + \frac{1}{2}\int_0^T \|g(t)\|_{V^*}^2\,dt,$$

which implies the a priori estimate

$$\|u(T)\|^2 + \int_0^T \|u(t)\|_a^2\,dt \le \|u(0)\|^2 + \int_0^T \|g(t)\|_{V^*}^2\,dt, \tag{2.15}$$

where we have set, for any $g \in V^*$,

$$\|g\|_{V^*} = \sup_{v\in V} \frac{(g,v)}{\|v\|_a}\,.$$

Some examples follow.

Example 2.2 (diffusion problem). Here $\rho = 2$ and

$$A = -\nabla\cdot\mathbf{D}(x)\nabla,\ \ V = H_0^1(\Omega) \overset{d}{\hookrightarrow} L^2(\Omega) = H, \quad a(u,v) = \int_\Omega \nabla v\cdot\mathbf{D}(x)\nabla u\,dx,$$

where $\mathbf{D} \in L^\infty(\Omega)^{d\times d}$ satisfies for some $\gamma > 0$

$$\xi^T\,\mathbf{D}(x)\xi \ge \gamma\,|\xi|^2 \qquad \forall\underline{\xi} \in \mathbb{R}^n,\ \text{a.e. } x \in \Omega\,.$$

Then (2.1), (2.2) is the Dirichlet problem for the heat equation in $\Omega \times (0,T)$.

In this example, the operators A are differential operators and, in particular, local. The nonlocal operators A of interest to us are classical pseudodifferential operators.

Example 2.3. For $0 \leq \rho \in \mathbb{R}$, $\Omega \subset \mathbb{R}^d$ open, bounded, and Lipschitz, we consider classical pseudodifferential operators of order $\rho \in [0,2]$ in Ω, i.e., $A \in \Psi^\rho(\Omega)$, which acts from $V \to V^*$, where $V = \widetilde{H}^{\frac{\rho}{2}}(\Omega)$. By the Schwartz kernel theorem (see, e.g., [16, 15]), $A \in \Psi^\rho(\Omega)$ has a representation in terms of a distributional kernel

$$k(x, x-y) \in \mathcal{D}'(\Omega \times \Omega) \tag{2.16}$$

plus a C^∞ kernel $c(x,y)$. To the singular kernel, we associate a bilinear form

$$a(u,v) = (Au, v)_{V^* \times V} = \left\langle k(x,x-y),\ v(x) \otimes u(y) \right\rangle. \tag{2.17}$$

Moreover, the singular kernel $k(x,x-y) \in C^\infty(\Omega \times \Omega \backslash \{x=y\})$ satisfies the so-called *Calderón–Zygmund estimates*: for all $\alpha, \beta \in \mathbb{N}_0^n$, $(x,y) \in \Omega \times \Omega \backslash \{x=y\}$,

$$|\partial_x^\alpha \, \partial_y^\beta \, k(x,x-y)| \leq C(\alpha,\beta)|x-y|^{-(d+\rho+|\alpha|+|\beta|)}. \tag{2.18}$$

A particular example for a nonlocal operator of order $\rho = 1$ is given by $\Omega = (-1/2, 1/2) \subset \mathbb{R}$ and, for $u \in V = \widetilde{H}^{1/2}(\Omega)$,

$$(Wu)(x) = -\text{p.f.} \int_\Omega \frac{u(y)}{|x-y|^2} \, dy, \tag{2.19}$$

where the integral is to be understood in the finite-part sense (see, e.g., [16]). For the bilinear form $a(u,v)$ corresponding to the hypersingular operator W in (2.19), integration by parts yields the representation

$$\forall u, v \in \widetilde{H}^{1/2}(\Omega): \ a(u,v) = -\int_\Omega v'(x) \int_\Omega \log(x-y)\, u'(y)\, dy\, dx, \tag{2.20}$$

and one can show that there are $\beta, \gamma > 0$ with

$$\forall u \in \widetilde{H}^{1/2}(\Omega): \ a(u,u) \geq \beta \, \|u\|^2_{\widetilde{H}^{1/2}(\Omega)}. \tag{2.21}$$

Remark 2.4. In the setting (2.17), we often do not have the coercivity (2.8), but rather a (weaker) *Gårding inequality:* There is $\gamma \geq 0$ such that

$$\forall u \in V: \ a(u,u) + \gamma \, \|u\|^2 \geq \beta \, \|u\|_V^2 \tag{2.22}$$

(where $\|u\|^2$ is, e.g., due to the C^∞ part of the kernel of A).

This case can be reduced to (2.8) by the substitution $w = \exp(-\gamma t)u$, since then (2.1) implies that w solves the problem

$$\dot{w} + (A + \gamma I)\, w = \exp(-\gamma t)\, g \ \text{ in } \ (0,T),$$

and the operator $A + \gamma I$ is, by (2.22), once again coercive.

3. Discretization. We discretize (2.1) in time using the so-called θ-scheme, and in space by a finite element method. We describe wavelet finite element bases and the compression of the stiffness matrix.

3.1. Space discretization. To discretize the parabolic problem (2.11) in space, we use an elliptic projection onto a family $\{V_h\}_h \subset V$ of finite dimensional subspaces of V, based on piecewise polynomials of degree $p \geq 0$ on a quasi-uniform family of triangulations $\{\mathcal{T}_h\}_h$ of Ω.

The semidiscrete problem reads: Given $u_0 \in H$, $g \in L^2(J; H)$, first choose an approximation $u_{0,h} \in V_h$ for the initial data u_0. Then find $u_h \in H^1(J; V_h)$ such that

$$u_h(0) = u_{0,h} \tag{3.1}$$

and

$$\frac{d}{dt}(u_h, v_h) + a(u_h, v_h) = (g(t), v_h) \qquad \forall v_h \in V_h. \tag{3.2}$$

Let $P_h \colon L^2 \to V_h$ be a projector. Then the approximation of the initial data could be chosen as $u_{0,h} = P_h u_0$ or as an interpolant of u_0.

The semidiscrete problem (3.1), (3.2) is an initial value problem for $N = \dim V_h$ ordinary differential equations

$$\mathbf{K}\frac{d}{dt}\underline{u} + \mathbf{A}\underline{u} = \underline{g}(t), \qquad \underline{u}(0) = \underline{u}_0,$$

where $\underline{u}(t)$ denotes the coefficient vector of $u_h(t)$ with respect to some basis of V_h. Likewise $\underline{u}_0$ denotes the coefficient vector of $u_{0,h}$, and $\mathbf{K}, \mathbf{A}$ denote the mass- and stiffness matrix, respectively, with respect to the basis of V_h.

In the ensuing error analysis, we need to consider functions in V which have additional regularity and introduce for this purpose the spaces $\mathcal{H}^s(\Omega)$ which are defined as

$$\mathcal{H}^s(\Omega) = \begin{cases} V = \tilde{H}^{\rho/2}(\Omega) & \text{for } s = \rho/2, \\ V \cap H^s(\Omega) & \text{for } s > \rho/2. \end{cases}$$

We assume the *approximation property*: For all $u \in \mathcal{H}^t$ with $t \geq \rho/2$ there exists a $u_h \in V_h$ such that for $0 \leq s \leq \frac{\rho}{2}$ and $\rho/2 \leq t \leq p+1$

$$\|u - u_h\|_{\tilde{H}^s(\Omega)} \leq ch^{t-s} \|u\|_{\mathcal{H}^t(\Omega)}. \tag{3.3}$$

We assume that the projector $P_h \colon V \to V_h$ satisfies (3.3) with $u_h = P_h u$.

We shall also need the *inverse property*: There is $c > 0$ independent of h such that

$$\forall u_h \in V_h \quad \|u_h\|_{\tilde{H}^s(\Omega)} \leq ch^{-s} \|u_h\|_{L^2(\Omega)}, \quad 0 \leq s \leq \frac{\rho}{2}. \tag{3.4}$$

3.2. Time discretization using the $\boldsymbol{\theta}$-scheme. For $T < \infty$ and $M \in \mathbb{N}$, define the time step

$$k = \frac{T}{M}$$

and $t^m = mk$, $m = 0, \ldots, M$. The fully discrete θ-scheme reads: Given $u_0 \in H$, find $u_h^m \in V_h$ satisfying

$$u_h^0 = u_{0,h}, \tag{3.5}$$

and, for $m = 0, 1, \dots, M-1$, find $u_h^{m+1} \in V$ such that for all $v_h \in V_h$

$$\left(\frac{u_h^{m+1} - u_h^m}{k}, v_h \right) + a\big(u_h^{m+\theta}, v_h\big) = \big(g^{m+\theta}, v_h\big) \tag{3.6}$$

holds. Here $u_h^{m+\theta} := \theta u_h^{m+1} + (1-\theta) u_h^m$ and $g^{m+\theta} := \theta g(t^{m+1}) + (1-\theta) g(t^m)$. In matrix form, (3.6) reads

$$(k^{-1}\mathbf{K} + \theta \mathbf{A}) \underline{u}^{m+1} = k^{-1}\mathbf{K}\underline{u}^m - (1-\theta)\mathbf{A}\underline{u}^m + \underline{g}^{m+\theta}, \qquad m = 0, 1, \dots, M-1,$$

where $\underline{u}^m$ is the coefficient vector of u_h^m with respect to a basis of V_h.

Remark 3.1. Even for the forward Euler method (i.e., for $\theta = 0$), we have to solve at each time step a linear system with the mass matrix. For $0 \le \rho < 1$ the spaces $H^{\rho/2}(\Omega)$ allow for the use of discontinuous multiwavelets to obtain a diagonal mass matrix. In this case, each time step requires only one matrix-vector product with the matrix $\mathbf{A}$.

3.3. Perturbation. Previous analyses of the θ-scheme (3.5) assumed that the form $a(\cdot,\cdot)\colon V_h \times V_h \to \mathbb{R}$ can be evaluated exactly, i.e., that the corresponding stiffness matrix $\mathbf{A}$ is available. In practice, this is unrealistic. Even for the heat equation, numerical integration and isoparametric boundary approximations allow one to realize only an approximation of $\mathbf{A}$. The impact of the resulting consistency error in the context of semidiscrete schemes was investigated early on [13].

Here, we are interested in wavelet compression of $\mathbf{A}$ as in, e.g., [10, 11, 14], resulting in a compressed matrix $\tilde{\mathbf{A}}$. With the compressed matrix $\tilde{\mathbf{A}}$ we associate the perturbed bilinear form $\tilde{a}(\cdot,\cdot)$. (Other perturbations, e.g., due to numerical integration or domain approximation by isoparametric elements in the context of Example 2.2, can be treated in the same way.) Using $\tilde{a}(\cdot,\cdot)$ in place of $a(\cdot,\cdot)$ in (3.6) gives *perturbed θ-schemes*

$$\widetilde{u}_h^0 = u_{0,h}, \tag{3.7a}$$

$$\left(\frac{\widetilde{u}_h^{m+1} - \widetilde{u}_h^m}{k}, v_h \right) + \widetilde{a}\big(\widetilde{u}_h^{m+\theta}, v_h\big) = \big(g^{m+\theta}, v_h\big) \tag{3.7b}$$

for $m = 0, 1, 2, \dots, M-1$ and every $v_h \in V_h$, where $\widetilde{u}_h^{m+\theta} := \theta \widetilde{u}_h^{m+1} + (1-\theta)\,\widetilde{u}_h^m$. In matrix form, (3.7b) reads

$$(k^{-1}\mathbf{K} + \theta \tilde{\mathbf{A}}) \underline{\tilde{u}}^{m+1} = k^{-1}\mathbf{K}\underline{\tilde{u}}^m - (1-\theta)\tilde{\mathbf{A}}\underline{\tilde{u}}^m + \underline{g}^{m+\theta}, \qquad m = 0, 1, \dots, M-1,$$

where $\underline{\tilde{u}}^m$ is the coefficient vector of $\tilde{u}_h^m$ with respect to a basis of V_h.

We shall assume for $\widetilde{a}(\cdot,\cdot)$ the following *consistency conditions*: There is $\delta < 1$ independent of h such that

$$|a(u_h, v_h) - \widetilde{a}(u_h, v_h)| \le \delta \, \|u_h\|_a \, \|v_h\|_a \qquad \forall u_h, v_h \in V_h, \tag{3.8}$$

and there is $C > 0$ independent of h such that

$$\begin{aligned} &|a(P_h u, v_h) - \widetilde{a}(P_h u, v_h)| \\ &\quad \le C h^{p+1-\rho/2} |\log h|^\nu \|u\|_{\mathcal{H}^{p+1}(\Omega)} \, \|v_h\|_{\tilde{H}^{\rho/2}(\Omega)} \qquad \forall u \in \mathcal{H}^{p+1}(\Omega),\ v_h \in V_h, \end{aligned} \tag{3.9}$$

with some $\nu \ge 0$.

Condition (3.8) shows that on $V_h \times V_h$ the form $\widetilde{a}(\cdot,\cdot)$ is equivalent to $a(\cdot,\cdot)$ in the following sense.

PROPOSITION 3.2. *For* $\delta < 1$ *in* (3.8), *we have for some constants* $0 < \widetilde{\beta} \leq \widetilde{\alpha} < \infty$ *independent of* h

$$\forall u_h, v_h \in V_h : \ |\widetilde{a}(u_h, v_h)| \leq \widetilde{\alpha} \, \|u_h\|_a \, \|v_h\|_a \tag{3.10}$$

and

$$\forall u_h \in V_h : \ |\widetilde{a}(u_h, u_h)| \geq \widetilde{\beta} \, \|u_h\|_a^2 \, . \tag{3.11}$$

Proof. Consider (3.11). We have for $u_h \in V_h$

$$|\tilde{a}(u_h, u_h)| \geq |a(u_h, u_h)| - |a(u_h, u_h) - \tilde{a}(u_h, u_h)| = \|u_h\|_a^2 - |a(u_h, u_h) - \tilde{a}(u_h, u_h)| \, ,$$

and, using the definition of $\|\cdot\|_a$ and the consistency condition (3.8), we get (3.11) with $\tilde{\beta} = 1 - \delta$. The continuity (3.10) is proved in the same way. □

3.4. Wavelet compression. In the context of Example 2.3, perturbed bilinear forms $\widetilde{a}$ are obtained by various matrix compression techniques which reduce the dense matrices $\mathbf{A}$ to sparse ones that can be manipulated in linear complexity. We illustrate this by the wavelet compression of operators of order $0 \leq \rho \leq 2$ in dimensions $d = 1, 2$; we present here only the main principles—for details and proofs, see [9, 14, 4]. All results carry over to dimensions $d > 2$ if a suitable wavelet basis is used.

3.4.1. Subspaces V_h. For $d = 1$ the domain Ω is an interval. For $d = 2$ we assume that Ω is a polygon. Let T_0 be a fixed coarse triangulation of the domain. We then define the triangulation T_l for $l > 0$ by bisection of each interval in T_{l-1} for $d = 1$, or by subdivision of a triangle in T_{l-1} into four congruent subtriangles for $d = 2$. We assume that the triangulation $\{\mathcal{T}_h\}$ is obtained in this way as T_L, for some $L > 0$ so that $h = C2^{-L}$.

For $0 \leq \rho < 1$ we define V_h as the space of piecewise polynomials of total degree $p \geq 0$ (without any continuity restriction) on the triangulation T_L.

For $1 \leq \rho \leq 2$ the space V_h is defined as the space of continuous piecewise polynomials of degree $p \geq 1$ on the triangulation with zero values on the boundary $\partial\Omega$.

In the same way we define the spaces V^l corresponding to the triangulation T_l, so that we have

$$V^0 \subset V^1 \subset \cdots \subset V^L = V_h.$$

Let $N^l = \dim V^l$ and $M^l := N^l - N^{l-1}$ so that $N = \dim V_h = N^L = C2^L$.

3.4.2. Wavelet basis. By choosing a suitable basis for V_h we will be able to represent the bilinear form $a(\cdot,\cdot)$ as a matrix where most elements are small and can be neglected, yielding the approximate bilinear form $\tilde{a}(\cdot,\cdot)$. The basis will also allow optimal preconditioning. We will use so-called biorthogonal wavelets. (Note that the dual wavelets described below will not be used in the computation.)

We will use a hierarchical basis of functions ψ_j^l with $j = 1, \ldots, M^l$ and $l = 0, 1, \ldots$ with the following properties:

1. $V^l = \operatorname{span}\{\psi_j^l \mid 0 \leq l \leq L, 1 \leq j \leq M^l\}$.
2. The function ψ_j^l has support $S_j^l := \operatorname{supp} \psi_j^l$ of diameter bounded by $C \, 2^{-l}$.

3. Wavelets ψ_j^l with $\bar{S}_j^l \cap \partial\Omega = \emptyset$ have vanishing moments up to order p, i.e., $(\psi_j^l, q) = 0$ for all polynomials q of total degree p or less.
4. The functions ψ_j^l for $l \geq l_0$ are obtained by scaling and translation of the functions $\psi_j^{l_0}$.
5. A function $v \in V_h$ has the representation

$$v = \sum_{l=0}^{L} \sum_{j=1}^{M^l} v_j^l \psi_j^l$$

with $v_j^l = (v, \tilde{\psi}_j^l)$, where $\tilde{\psi}_j^l$ are the so-called dual wavelets. For $v \in V$ one obtains an infinite series

$$v = \sum_{l=0}^{\infty} \sum_{j=1}^{M^l} v_j^l \psi_j^l$$

with $v_j^l = (v, \tilde{\psi}_j^l)$, which converges in $\tilde{H}^s$ for $0 \leq s \leq \rho/2$.
6. There holds the norm equivalence

$$c_1 \|v\|_{\tilde{H}^s(\Omega)}^2 \leq \sum_{l=0}^{\infty} \sum_{j=1}^{M^l} \left|v_j^l\right|^2 2^{2ls} \leq c_2 \|v\|_{\tilde{H}^s(\Omega)}^2 \tag{3.12}$$

for $0 \leq s \leq \rho/2$, and for $\rho/2 < s \leq p+1$ we have the one-sided bounds

$$\sum_{l=0}^{L} \sum_{j=1}^{M^l} \left|v_j^l\right|^2 2^{2ls} \leq c_3 L^\nu \|v\|_{\mathcal{H}^s(\Omega)}^2 ,$$

where $c_i > 0$ are independent of L, $\nu = 0$ if $s < p+1$, and $\nu = 1$ if $s = p+1$.

We now define the projection $P_h : V \to V_h$ by truncating the wavelet expansion: For $v \in V$

$$P_h v := \sum_{l=0}^{L} \sum_{j=1}^{M^l} v_j^l \psi_j^l. \tag{3.13}$$

This projection satisfies the approximation property (3.3).

3.4.3. Examples for wavelets. In the case $0 \leq \rho < 1$ a multiwavelet basis as in [11] can be used: Let $\{p_k\}$ be a basis for polynomials of total degree p or less. Then the functions ψ_j^0 are the functions which are on one element equal to a function p_k, and zero elsewhere. For $l > 1$ we choose for ψ_j^l functions in V^l that are nonzero on one element of T_{l-1} and that are orthogonal on all polynomials of total degree p or less.

In all cases $0 \leq \rho \leq 2$, so-called prewavelets can be used; these are functions in V^l with small support which are orthogonal on V^{l-1}.

Another possibility are so-called biorthogonal wavelets, which need not be orthogonal on V^{l-1}. For piecewise linears the functions ψ_j^l in the interior of the interval have values $0, \dots, 0, -1, 2, -1, 0, \dots, 0$. In the case of Neumann boundary conditions the wavelet at the left boundary has values $-2, 2, 1, 0, \dots, 0$; in the case of Dirichlet conditions the values are $0, 2, -1, 0, \dots, 0$ (and similarly at the right boundary). Note that the boundary wavelets have fewer vanishing moments in general.

In dimension $d = 2$ the construction of piecewise linear prewavelets on arbitrary polygons is described, e.g., in [5, 17].

3.4.4. Matrix compression. The bilinear form a on $V_h \times V_h$ corresponds to a matrix $\mathbf{A}$ with elements $A_{(l,j),(l',j')} = a(\psi^l_j, \psi^{l'}_{j'})$.

We assumed that the kernel of the operator satisfies the estimates (2.18). This implies a decay of the matrix elements with increasing distance of their supports.

We define the compressed matrix $\tilde{\mathbf{A}}$ and the corresponding bilinear form $\tilde{a}$ by replacing certain small matrix elements in $\mathbf{A}$ with zero:

$$\tilde{A}_{(j,l),(j',l')} := \begin{cases} A_{(j,l),(j',l')} & \text{if } \operatorname{dist}(S^l_j, S^{l'}_{j'}) \leq \delta_{l,l'} \text{ or } S^l_j \cap \partial\Omega \neq \emptyset, \\ 0 & \text{otherwise.} \end{cases} \tag{3.14}$$

Here the truncation parameters $\delta_{l,l'}$ are given by

$$\delta_{l,l'} := c \max\{2^{-L+\hat{\alpha}(2L-l-l')}, 2^{-l}, 2^{-l'}\} \tag{3.15}$$

with some parameters $c > 0$ and $\hat{\alpha} > 0$. The consistency conditions (3.8), (3.9) can be satisfied as follows (see, e.g., [9, 10, 11, 14]).

PROPOSITION 3.3. *If c in* (3.15) *is chosen sufficiently large, then for all $L > 0$ condition* (3.8) *holds. If additionally*

$$\hat{\alpha} \geq \frac{2p+2}{2p+2+\rho} \tag{3.16}$$

holds, then condition (3.9) *holds with $\nu = \frac{3}{2}$ if equality holds in* (3.16)*, and $\nu = \frac{1}{2}$ otherwise.*

The matrix compression (3.14) reduces the number of nonzero elements from N^2 to N times a logarithmic term as follows (see [9, 10, 11, 14]).

PROPOSITION 3.4. *The compressed matrix $\tilde{\mathbf{A}}$ has $O(N \log N)$ nonzero elements if $\hat{\alpha} < 1$, and $O(N(\log N)^2)$ nonzero elements if $\hat{\alpha} = 1$.*

In particular, for operators of order $\rho > 0$ we can choose $\hat{\alpha}$ such that $\nu = \frac{1}{2}$ in (3.9) *and the number of nonzero elements in $\tilde{\mathbf{A}}$ is $O(N \log N)$. In the case of order $\rho = 0$ we have to choose $\hat{\alpha} = 1$, implying $\nu = \frac{3}{2}$ in* (3.9)*, and the number of nonzero elements in $\tilde{\mathbf{A}}$ is $O(N(\log N)^2)$.*

4. Stability. The stability of the θ-scheme is well known in the context of Example 2.2, i.e., if the spatial operator is elliptic and of second order. We investigate here general operators A of order $\rho \geq 0$ that are elliptic in the sense that (2.7), (2.8) hold in $V = \widetilde{H}^{\rho/2}(\Omega)$. We prove an $L^2(J;V)$ stability estimate for the approximate solutions obtained from the θ-scheme with wavelet-compressed space operator.

In the analysis, we will use for $f \in V_h^*$ the following notation:

$$\|f\|_* := \sup_{v_h \in V_h} \frac{(f, v_h)}{\|v_h\|_a} \,. \tag{4.1}$$

We will also need λ_A defined by

$$\lambda_A := \sup_{v_h \in V_h} \frac{\|v_h\|^2}{\|v_h\|_*^2}.$$

We first address the stability of the θ-scheme with exact bilinear form $a(\cdot,\cdot)$. In the case $\frac{1}{2} \leq \theta \leq 1$, the θ-scheme is stable for any time step $k > 0$, whereas in the case $0 \leq \theta < \frac{1}{2}$, the time step k must be sufficiently small.

Proposition 4.1. *In the case of* $\frac{1}{2} \le \theta \le 1$ *assume that*

$$0 < C_1 < 2, \qquad C_2 \ge \frac{1}{2 - C_1}, \tag{4.2}$$

and in the case of $0 \le \theta < \frac{1}{2}$ *assume that*

$$\sigma := k(1 - 2\theta)\lambda_A < 2, \tag{4.3}$$

$$0 < C_1 < 2 - \sigma, \qquad C_2 \ge \frac{1 + (4 - C_1)\sigma}{2 - \sigma - C_1}. \tag{4.4}$$

Then the sequence $\{u_h^m\}_{m=0}^M$ *of solutions of the* θ*-scheme* (3.5) *satisfies the stability estimate*

$$\|u_h^M\|^2 + C_1 k \sum_{m=0}^{M-1} \|u_h^{m+\theta}\|_a^2 \le \|u_h^0\|^2 + C_2 k \sum_{m=0}^{M-1} \|g^{m+\theta}\|_*^2. \tag{4.5}$$

Proof. Let

$$X^m := \|u_h^m\|^2 - \left\|u_h^{m+1}\right\|^2 + C_2 k \left\|g^{m+\theta}\right\|_*^2 - C_1 k \left\|u_h^{m+\theta}\right\|_a^2.$$

We want to show that $X^m \ge 0$. Then adding these inequalities for $m = 0, \dots, M-1$ will obviously give (4.5).

Let $w := u_h^{m+1} - u_h^m$; then $u_h^{m+\theta} = (u_h^m + u_h^{m+1})/2 + (\theta - \frac{1}{2})w$ and

$$\left\|u_h^{m+1}\right\|^2 - \|u_h^m\|^2 = (u_h^{m+1} - u_h^m, u_h^{m+1} + u_h^m) = (w, 2u_h^{m+\theta} - (2\theta - 1)w).$$

By the definition of the θ-scheme, we have

$$\begin{aligned}(w, u_h^{m+\theta}) &= k(-Au_h^{m+\theta} + g^{m+\theta}, u_h^{m+\theta}) = k\Big[-\left\|u_h^{m+\theta}\right\|_a^2 + (g^{m+\theta}, u_h^{m+\theta})\Big] \\ &\le k\Big[-\left\|u_h^{m+\theta}\right\|_a^2 + \left\|g^{m+\theta}\right\|_* \left\|u_h^{m+\theta}\right\|_a\Big].\end{aligned}$$

This gives

$$X^m \ge (2\theta - 1)\|w\|^2 + k\Big[(2 - C_1)\left\|u_h^{m+\theta}\right\|_a^2 - 2\left\|g^{m+\theta}\right\|_* \left\|u_h^{m+\theta}\right\|_a + C_2 \left\|g^{m+\theta}\right\|_*^2\Big].$$

In the case of $\frac{1}{2} \le \theta \le 1$ we now obtain $X^m \ge 0$ if the conditions (4.2) are satisfied.

In the case $0 \le \theta < \frac{1}{2}$ we have by the definition of the θ-scheme that $(w, v_h) = k(-Au_h^{m+\theta} + g^{m+\theta}, v_h)$, yielding

$$\|w\| \le \lambda_A^{1/2} \|w\|_* \le \lambda_A^{1/2} k\Big(\left\|Au_h^{m+\theta}\right\|_* + \left\|g^{m+\theta}\right\|_*\Big) = \lambda_A^{1/2} k\Big(\left\|u_h^{m+\theta}\right\|_a + \left\|g^{m+\theta}\right\|_*\Big),$$

since $(Au_h^{m+\theta}, v_h) \le \|u_h^{m+\theta}\|_a \|v_h\|_a$ gives $\|Au_h^{m+\theta}\|_* \le \|u_h^{m+\theta}\|_a$ and choosing $v_h := u_h^{m+\theta}$ gives $\|Au_h^{m+\theta}\|_* \ge \|u_h^{m+\theta}\|_a$. Hence

$$k^{-1}X^m \ge (2 - C_1 - \sigma)\left\|u_h^{m+\theta}\right\|_a^2 - 2(1 + \sigma)\left\|g^{m+\theta}\right\|_* \left\|u_h^{m+\theta}\right\|_a + (C_2 - \sigma)\left\|g^{m+\theta}\right\|_*^2.$$

Therefore we have $X^m \ge 0$ if conditions (4.3) hold. □

Remark 4.2. The a priori estimate (4.5) is, in a sense, the discrete analogue of the a priori estimate (2.15). Note, however, that $\|g^{m+\theta}\|_*$ is not identical to $\|g^{m+\theta}\|_{V^*}$—in fact, for $g \in V^*$ we have by $V_h \subset V$ that

$$\|g^{m+\theta}\|_* \leq \|g^{m+\theta}\|_{V^*}.$$

Consider now the sequence $\{\widetilde{u}_h^m\}_{m=0}^M$ of solutions to the perturbed θ-scheme (3.7a), (3.7b). We analogously define for $v_h \in V_h$ and $f \in V_h^*$

$$\|v_h\|_{\tilde{a}} := \tilde{a}(v_h, v_h), \qquad \|f\|_{\tilde{*}} := \sup_{v_h \in V_h} \frac{(f, v_h)}{\|v_h\|_{\tilde{a}}}, \qquad \lambda_{\tilde{A}} := \sup_{v_h \in V_h} \frac{\|v_h\|^2}{\|v_h\|_{\tilde{*}}^2}. \tag{4.6}$$

Due to the norm equivalence in Proposition 3.2, we obtain in the same way as in Proposition 4.1, with $\tilde{a}(\cdot,\cdot)$ in place of $a(\cdot,\cdot)$, the following result.

PROPOSITION 4.3. *Assume that* (3.8) *holds with* $\delta < 1$. *In the case of* $\frac{1}{2} \leq \theta \leq 1$ *assume that* (4.2) *holds. In the case of* $0 \leq \theta < \frac{1}{2}$ *assume that*

$$\sigma := k(1 - 2\theta)\lambda_{\tilde{A}} < 2 \tag{4.7}$$

and that (4.4) *holds.*

Then the sequence $\{\tilde{u}_h^m\}_{m=0}^M$ *of solutions of the perturbed* θ*-scheme* (3.7a), (3.7b) *satisfies the stability estimate*

$$\|\tilde{u}_h^M\|^2 + C_1 k \sum_{m=0}^{M-1} \|\tilde{u}_h^{m+\theta}\|_{\tilde{a}}^2 \leq \|\tilde{u}_h^0\|^2 + C_2 k \sum_{m=0}^{M-1} \|g^{m+\theta}\|_{\tilde{*}}^2. \tag{4.8}$$

Remark 4.4. By the inverse estimate (3.4) and the norm equivalence (2.14) we have for $w_h \in V_h$

$$\|w_h\|_a \leq C\,\|w_h\|_{\rho/2} \leq C' h^{-\rho/2}\,\|w_h\|,$$

and therefore for $v_h \in V_h$

$$\|v_h\|_* = \sup_{w_h \in V_h} \frac{(v_h, w_h)}{\|w_h\|_a} \geq C h^{\rho/2} \sup_{w_h \in V_h} \frac{(v_h, w_h)}{\|w_h\|} = C h^{\rho/2}\,\|v_h\|, \tag{4.9}$$

$$\lambda_A^{1/2} = \sup_{v_h \in V_h} \frac{\|v_h\|}{\|v_h\|_*} \leq C h^{-\rho/2}. \tag{4.10}$$

Hence there exists a positive constant C_* independent of h and θ such that the time-step restriction

$$k \leq C_* \frac{h^\rho}{1 - 2\theta} \tag{4.11}$$

is sufficient for stability (4.3). For $\rho = 2$ and $\theta < \frac{1}{2}$ (e.g., forward Euler and the heat equation) this reduces to the well-known time-step restriction $k \leq C_\theta h^2$ for explicit schemes. For smaller values of ρ the restriction is less severe, and in the limiting case $\rho = 0$ condition (4.11) gives $k \leq C_*/(1 - 2\theta)$ with a bound independent of h.

For the perturbed scheme (3.7) we can proceed in the same way and obtain, using Proposition 3.2, that (4.11) is a sufficient condition for (4.7) (with a different value of C_*).

Remark 4.5. As θ tends to $\frac{1}{2}$ from below the bound on k in the stability condition, (4.3) tends to infinity, and for $\theta \geq \frac{1}{2}$ the stability holds with $\sigma = 0$ and C_1, C_2 as in (4.4) for *all* values of k.

5. Convergence. Based on the stability results obtained in section 4 and the consistency (3.8), (3.9) of the compressed form $\widetilde{a}(\cdot,\cdot)$, we shall now obtain optimal convergence estimates of the compressed θ-scheme (sufficient regularity of the exact solution $u(x,t)$ in space and time provided). Throughout this section, we shall set

$$u^m = u(t^m) \in V. \tag{5.1}$$

We will estimate the error

$$\widetilde{e}_h^m := u^m - \widetilde{u}_h^m\,. \tag{5.2}$$

To this end, we split $\widetilde{e}_h^m$ as follows:

$$\widetilde{e}_h^m = \underbrace{(u^m - P_h u^m)}_{\eta^m} + \underbrace{(P_h u^m - \widetilde{u}_h^m)}_{\xi_h^m} = \eta^m + \xi_h^m, \tag{5.3}$$

where $P_h : V \to V_h$ is the quasi interpolant in (3.13) (realized as a truncated wavelet expansion; see section 3.4 or [9, 3] for details).

As η^m is a best approximation error, we focus now on $\xi_h^m \in V_h$.

LEMMA 5.1. *If $u \in C^1(\overline{J}; H)$, the $\{\xi_h^m\}_m$ are solutions of the θ-scheme*

$$\xi_h^0 = P_h u_0 - \widetilde{u}_h^0$$

for $m = 0, 1, \ldots, M-1$ and every $v_h \in V_h$;

$$k^{-1}(\xi_h^{m+1} - \xi_h^m, v_h) + \widetilde{a}\left(\theta\xi_h^{m+1} + (1-\theta)\,\xi_h^m, v_h\right) = (r^m, v_h), \tag{5.4}$$

where the weak residuals r^m: $V_h \to \mathbb{R}$ are given by

$$r^m = r_1^m + r_2^m + r_3^m + r_4^m \tag{5.5}$$

with

$$\begin{aligned}
(r_1^m, v_h) &:= \left(\frac{u^{m+1} - u^m}{k} - \dot{u}^{m+\theta}, v_h\right),\\
(r_2^m, v_h) &:= \left(\frac{P_h u^{m+1} - P_h u^m}{k} - \frac{u^{m+1} - u^m}{k}\,,\ v_h\right),\\
(r_3^m, v_h) &:= \widetilde{a}\left(P_h u^{m+\theta}, v_h\right) - a\left(P_h u^{m+\theta}, v_h\right),\\
(r_4^m, v_h) &:= a\left(P_h u^{m+\theta} - u^{m+\theta}, v_h\right).
\end{aligned}$$

Proof. We note that (2.11) and $u \in C^1(\overline{J}; H)$ imply

$$(\dot{u}^{m+\theta}, v) + a(u^{m+\theta}, v) = (g^{m+\theta}, v) \qquad \forall v \in V\,. \tag{5.6}$$

Since $V_h \subset V$, we get for every $v_h \in V_h$

$$\begin{aligned}
&k^{-1}(\xi_h^{m+1} - \xi_h^m, v_h) + \widetilde{a}(\theta\xi_h^{m+1} + (1-\theta)\,\xi_h^m, v_h)\\
&\quad= \left(\frac{(P_h u^{m+1} - \widetilde{u}_h^{m+1}) - (P_h u^m - \widetilde{u}_h^m)}{k}\,,\ v_h\right) + \widetilde{a}\left(P_h u^{m+\theta}, v_h\right) - \widetilde{a}\left(\widetilde{u}_h^{m+\theta}, v_h\right)\\
&\quad= \left(\frac{P_h u^{m+1} - P_h u^m}{k}\,,\ v_h\right) + \widetilde{a}(P_h u^{m+\theta}, v_h)
\end{aligned}$$

$$
\begin{aligned}
& - \left\{ \left(\frac{\widetilde{u}_h^{m+1} - \widetilde{u}_h^{m}}{k}\,,\; v_h \right) - \widetilde{a}\,(\widetilde{u}_h^{m+\theta}, v_h) \right\} \\
&\overset{(3.7)}{=} \left(\frac{P_h u^{m+1} - P_h u^{m}}{k}\,,\; v_h \right) + \widetilde{a}\,(P_h u^{m+\theta}, v_h) - (g^{m+\theta}, v_h) \\
&\overset{(5.6)}{=} \left(\frac{P_h u^{m+1} - P_h u^{m}}{k} - \dot{u}^{m+\theta}, v_h \right) + \widetilde{a}\,(P_h u^{m+\theta}, v_h) - a\,(u^{m+\theta}, v_h) =: (r, v_h)\,.
\end{aligned}
$$

The representation (5.5) of (r, v_h) is now evident. □

Lemma 5.1 implies, together with the stability result Proposition 4.3, the following estimate for the ξ_h^m.

COROLLARY 5.2. *Under the assumptions of Proposition* 4.3 *and if* $u \in C^1(\overline{J}; H)$, *we have*

$$
\|\xi_h^M\|^2 + C_1 k \sum_{m=0}^{M-1} \left\|\xi_h^{m+\theta}\right\|_{\widetilde{a}}^2 \le \|\xi_h^0\|^2 + C_2 k \sum_{m=0}^{M-1} \left\|r^m\right\|_{\widetilde{*}}^2. \tag{5.7}
$$

Based on (5.5), we must estimate the $\|r_j^m\|_{\widetilde{*}}$, $j = 1, \dots, 4$.

Estimate of r_1^m. This value is based on a Taylor expansion in t. Noting that for any $v_h \in V_h$

$$
|(r_1^m, v_h)| \le \|k^{-1}(u^{m+1} - u^m) - \dot{u}^{m+\theta}\|_{\tilde{*}}\; \|v_h\|_{\tilde{a}}
$$

and

$$
k^{-1}\left|(u^{m+1} - u^m) - \dot{u}^{m+\theta}\right| = k^{-1}\left|\int_{t_m}^{t_{m+1}} (s - (1-\theta)t_{m+1} - \theta\, t_m)\, \ddot{u}\, ds\right|,
$$

we get

$$
\begin{aligned}
\|k^{-1}(u^{m+1} - u^m) - \dot{u}^{m+\theta}\|_{\tilde{*}} &\le k^{-1} \int_{t_m}^{t_{m+1}} |s - (1-\theta)\, t_{m+1} - \theta\, t_m|\; \|\ddot{u}\|_{\tilde{*}}\, ds \\
&\le C_\theta \int_{t_m}^{t_{m+1}} \|\ddot{u}(s)\|_{\tilde{*}}\, ds \\
&\le C_\theta\, k^{\frac{1}{2}} \left(\int_{t_m}^{t_{m+1}} \|\ddot{u}(s)\|_{\tilde{*}}^2\, ds \right)^{\frac{1}{2}}.
\end{aligned} \tag{5.8}
$$

If $\theta = \frac{1}{2}$, an integration by parts gives

$$
\left|k^{-1}(u^{m+1} - u^m) - \dot{u}^{m+\theta}\right| = \frac{1}{2k}\left|\int_{t_m}^{t_{m+1}} (t_{m+1} - s)(t_m - s)\; \dddot{u}\; (s) ds\right|,
$$

and it follows that

$$
\|k^{-1}(u^{m+1} - u^m) - \dot{u}^{m+\theta}\|_{\tilde{*}} = C\, k^{\frac{3}{2}} \left(\int_{t_m}^{t_{m+1}} \|\; \dddot{u}\; (s)\|_{\tilde{*}}^2\, ds \right)^{\frac{1}{2}}.
$$

Estimate of r_2^m. Here

$$
\begin{aligned}
|(r_2^m, v_h)| &\le C\, \|k^{-1}\, [(u^{m+1} - u^m) - P_h(u^{m+1} - u^m)]\|_{\tilde{*}}\, \|v_h\|_{\widetilde{a}} \\
&= C\, k^{-1} \left\| (I - P_h) \int_{t_m}^{t_{m+1}} \dot{u}(s)\, ds \right\|_{\tilde{*}} \|v_h\|_{\widetilde{a}} \\
&\le C\, k^{-1} \int_{t_m}^{t_{m+1}} \|(I - P_h)\, \dot{u}\|_{\tilde{*}}\, ds\, \|v_h\|_{\widetilde{a}} \\
&\le C k^{-\frac{1}{2}} h^{p+1-\rho/2} \left(\int_{t_m}^{t_{m+1}} \|\dot{u}\|^2_{\mathcal{H}^{p+1-\rho/2}(\Omega)} ds \right)^{\frac{1}{2}},
\end{aligned}
\tag{5.9}
$$

where we used $\|w\|_{\tilde{*}} \le C\, \|w\|$ and the approximation property (3.3) of P_h pointwise in t.

Estimate of r_3^m. Here we use the consistency (3.9)

$$
|(r_3^m, v_h)| \le C\, h^{p+1-\rho/2} |\log h|^{\nu} \|u^{m+\theta}\|_{\mathcal{H}^{p+1}(\Omega)} \|v_h\|_{\tilde{H}^{\rho/2}(\Omega)}\,. \tag{5.10}
$$

By (2.14) and (3.11), we get $\|v_h\|_{\tilde{H}^{\rho/2}} \le C\, \|v_h\|_{\widetilde{a}}$ and hence a bound on $\|r_3^m\|_{\widetilde{*}}$.

Estimate on r_4^m. Using (3.11) gives

$$
|(r_4^m, v_h)| \le C\, \|u^{m+\theta} - P_h u^{m+\theta}\|_a\, \|v_h\|_{\widetilde{a}},
$$

and with the approximation property (3.3) we find

$$
|(r_4^m, v_h)| \le C\, h^{p+1-\rho/2} \|u^{m+\theta}\|_{\mathcal{H}^{p+1}(\Omega)}\, \|v_h\|_{\widetilde{a}}\,. \tag{5.11}
$$

Collecting the bounds (5.8)–(5.11) gives the following result.

LEMMA 5.3. *Assume that* (3.8), (3.9) *hold. If* $u(x,t)$ *is sufficiently smooth in* $\overline{J} \times \overline{\Omega}$, *we have for* r^m *given by* (5.5)

$$
\begin{aligned}
\|r^m\|_{\widetilde{*}} \le C &\begin{cases} k^{\frac{1}{2}} \left(\int_{t_m}^{t_{m+1}} \|\ddot{u}(s)\|_*^2 ds \right)^{\frac{1}{2}} & \forall\, \theta \in [0,1], \\ k^{\frac{3}{2}} \left(\int_{t_m}^{t_{m+1}} \|\dddot{u}(s)\|_*^2 ds \right)^{\frac{1}{2}} & \forall\, \theta = \dfrac{1}{2} \end{cases} \\
&+ C k^{-\frac{1}{2}} h^{p+1-\rho/2} \left(\int_{t_m}^{t_{m+1}} \|\dot{u}\|^2_{\mathcal{H}^{p+1-\rho/2}(\Omega)} ds \right)^{\frac{1}{2}} \\
&+ C h^{p+1-\rho/2} |\log h|^{\nu} \|u^{m+\theta}\|_{\mathcal{H}^{p+1}(\Omega)}\,.
\end{aligned}
\tag{5.12}
$$

THEOREM 5.4. *Assume that the consistency conditions* (3.8), (3.9) *hold. For* $\theta \in [0, \frac{1}{2})$ *assume* (4.7). *Assume further that the approximation* $u_{0,h} \in V_h$ *of the initial data* u^0 *is quasi-optimal in* $L^2(\Omega)$. *Then the following error estimate holds for* *the perturbed θ-scheme with $\theta \in [0,1]$:*

(5.13)

$$\left\|u^M - \widetilde{u}_h^M\right\|^2 + k \sum_{m=0}^{M-1} \|u^{m+\theta} - \widetilde{u}_h^{m+\theta}\|_a^2 \leq Ch^{2(p+1-\rho/2)} |\log h|^{2\nu} \max_{0 \leq t \leq T} \|u(t)\|^2_{\mathcal{H}^{p+1}(\Omega)}$$
$$+ C \begin{cases} k^2 \displaystyle\int_0^T \|\ddot{u}(s)\|_*^2 ds & \forall\, \theta \in [0,1], \\ k^4 \displaystyle\int_0^T \|\dddot{u}(s)\|_*^2 ds & \text{for } \theta = \frac{1}{2} \end{cases}$$
$$+ Ch^{2(p+1-\rho/2)} \int_0^T \|\dot{u}(s)\|^2_{\mathcal{H}^{p+1-\rho/2}(\Omega)} ds.$$

Proof. Based on (5.3), we have for every $M \geq 1$

$$\left\|\widetilde{e}_h^M\right\|^2 + k \sum_{m=0}^{M-1} \|\widetilde{e}_h^{m+\theta}\|_a^2$$
$$\leq 2 \left\{ \left\|\eta^M\right\|^2 + k \sum_{m=0}^{M-1} \|\eta^{m+\theta}\|_a^2 \right\} + 2 \left\{ \left\|\xi_h^M\right\|^2 + k \sum_{m=0}^{M-1} \|\xi_h^{m+\theta}\|_a^2 \right\}.$$

The first term can be estimated with the approximation property (3.3). The second term is treated using (3.11) and (4.8). We get

$$\begin{aligned} \left\|\xi_h^M\right\|^2 + k \sum_{m=0}^{M-1} \|\xi_h^{m+\theta}\|_a^2 &\leq \left\|\xi_h^M\right\|^2 + k\widetilde{\beta}^{-1} \sum_{m=0}^{M-1} \left\|\xi_h^{m+\theta}\right\|_{\widetilde{a}}^2 \\ &\leq \max\left\{1, \frac{1}{(\widetilde{\beta} C_1)}\right\} \left\{ \left\|\xi_h^M\right\|^2 + C_1 k \sum_{m=0}^{M-1} \left\|\xi_h^{m+\theta}\right\|_{\widetilde{a}}^2 \right\} \\ &\leq \max\left\{1, \frac{1}{(\widetilde{\beta} C_1)}\right\} \left\{ \left\|\xi_h^0\right\|^2 + C_2 k \sum_{m=0}^{M-1} \left\|r_h^{m+\theta}\right\|_{\widetilde{*}}^2 \right\}. \end{aligned}$$

Using now the bound (5.12) for $\left\|r_h^{m+\theta}\right\|_{\widetilde{*}}$, the quasi optimality of $u_{0,h}$ and the approximation property (3.3) with $s = 0$ to estimate $\left\|\xi_h^0\right\|$ give the assertion. □

6. Approximate solution of linear equations and complexity. In order to compute the approximate solution $\tilde{u}_h^m$ in (3.7) for $m = 1, \dots, M$, we proceed as follows.

First we compute the mass matrix $\mathbf{K}$ in the wavelet basis with elements $K_{(l,j),(l',j')}$, where $O(N \log N)$ elements are nonzero. Note that for discontinuous multiwavelets (which can be used for $0 \leq \rho < 1$) the mass matrix is diagonal.

Then we compute the compressed stiffness matrix $\tilde{\mathbf{A}}$, where $O(N(\log N)^r)$ elements are nonzero and $r = 1$ if $\rho \in (0,2]$, $r = 2$ if $\rho = 0$; see Proposition 3.4. If explicit antiderivatives of the kernel function are available (as is often the case), the total cost for computing the stiffness matrix $\tilde{\mathbf{A}}$ is $O(N(\log N)^r)$ operations. In other cases, quadrature as described in [10] can be used. This preserves the consistency conditions (3.8), (3.9), and the total cost of computing $\tilde{\mathbf{A}}$ is $O(N(\log N)^{r+d})$ for $d = 1, 2$.

For each time step we have to solve (3.7b): We have to find $\tilde{w}_h^m := \tilde{u}_h^{m+1} - \tilde{u}_h^m \in V_h$ satisfying

$$k^{-1}(\tilde{w}_h^m, v_h) + \theta\tilde{a}(\tilde{w}_h^m, v_h) = (g^{m+\theta}, v_h) - \tilde{a}(\tilde{u}_h^m, v_h) \quad \forall v_h \in V_h \tag{6.1}$$

and then update $\tilde{u}_h^{m+1} := \tilde{u}_h^m + \tilde{w}_h^m$. Let $\tilde{w}^m \in \mathbb{R}^N$ denote the coefficient vectors of $\tilde{w}_h^m$ with respect to the wavelet basis, and $\mathbf{K}, \tilde{\mathbf{A}} \in \mathbb{R}^{N\times N}$ the mass and stiffness matrices corresponding to $(\cdot,\cdot)$ and $\tilde{a}(\cdot,\cdot)$ in this basis. Then we obtain for $\tilde{w}^m$ a linear system $\mathbf{B}\tilde{w}^m = \tilde{b}^m$ with the matrix $\mathbf{B} = k^{-1}\mathbf{K} + \theta\tilde{\mathbf{A}}$ and a known right-hand side vector $\tilde{b}^m$.

For a standard finite element basis, the matrix $\mathbf{B}$ has a condition number of order $h^{-\rho}$ for small h and fixed k. For the matrix $\mathbf{B}$ in the wavelet basis we can achieve a uniformly bounded condition number if we scale the rows and columns of $\mathbf{B}$ as follows: Let $\mu_l := (k^{-1} + \theta 2^{l\rho})^{1/2}$ and let $\hat{B}_{(l,j),(l',j')} := \mu_l^{-1}\mu_{l'}^{-1}B_{(l,j),(l',j')}$. A similar scaling was proposed in [4]. In what follows let $\|\cdot\|$ denote the 2-norm of a vector or the 2-norm of a matrix. We will use the GMRES method with restart every $m_0 \geq 1$ iterations, denoted by GMRES(m_0).

LEMMA 6.1. *For the linear system $\hat{\mathbf{B}}x = b$ let x_j for $j \in \mathbb{N}$ denote the iterates obtained by the restarted GMRES(m_0) method with initial guess x_0. Then there holds*

$$\|x - x_j\| \leq Cq^j \|x - x_0\|, \tag{6.2}$$

where C and $q < 1$ are independent of L, k, θ.

Proof. Throughout the proof, C_i will denote generic positive constants independent of h, k and unrelated to C_i in the main body of the text above. Let $\mathbf{D}$ denote the diagonal matrix with entries $D_{(l,j),(l,j)} = 2^{l\rho/2}$. Because of the norm equivalence (3.12), we have for all $x, y \in \mathbb{R}^N$

$$C_1 \|x\|^2 \leq x^T\mathbf{K}x, \qquad x^T\mathbf{K}y \leq C_2 \|x\| \|y\|.$$

Using the consistency conditions (3.8) of the wavelet truncation and (2.7), (2.8), we obtain

$$C_3 \|\mathbf{D}x\|^2 \leq x^T\tilde{\mathbf{A}}x, \qquad x^T\tilde{\mathbf{A}}y \leq C_4 \|\mathbf{D}x\| \|\mathbf{D}y\|.$$

The constants $C_j > 0$ are independent of L. Therefore $\mathbf{B} = k^{-1}\mathbf{K} + \theta\tilde{\mathbf{A}}$ satisfies with $C_5 := \min\{C_1, C_3\}$ and $C_6 := \max\{C_2, C_4\}$

$$C_5\, x^T(k^{-1}\mathbf{I} + \theta\mathbf{D}^2)x \leq x^T\mathbf{B}x, \tag{6.3}$$

$$\begin{aligned} x^T\mathbf{B}y &\leq C_6[k^{-1}\|x\|\|y\| + \theta\|\mathbf{D}x\|\|\mathbf{D}y\|] \\ &\leq C_6\big[x^T(k^{-1}\mathbf{I} + \theta\mathbf{D}^2)x\big]^{1/2}\big[y^T(k^{-1} + \theta\mathbf{D}^2)y\big]^{1/2} \end{aligned} \tag{6.4}$$

using the Cauchy–Schwarz inequality for the last estimate. Hence scaling with the diagonal matrix $\mathbf{S} := (k^{-1}\mathbf{I}+\theta\mathbf{D}^2)^{1/2}$ yields with $\hat{\mathbf{B}} = \mathbf{S}^{-1}\mathbf{B}\mathbf{S}^{-1}$ and $\hat{x} := \mathbf{S}x$, $\hat{y} := \mathbf{S}y$ that

$$C_5 \|\hat{x}\|^2 \leq \hat{x}^T\hat{\mathbf{B}}\hat{x}, \qquad \hat{x}^T\hat{\mathbf{B}}\hat{y} \leq C_6 \|\hat{x}\| \|\hat{y}\| \tag{6.5}$$

for all $\hat{x}, \hat{y} \in \mathbb{R}^N$, and therefore

$$\lambda_{\min}\left(\frac{(\hat{\mathbf{B}}+\hat{\mathbf{B}}^T)}{2}\right) \geq C_5, \qquad \|\hat{\mathbf{B}}\| \leq C_6.$$

According to [6], the GMRES iterates $x_{j+\nu}$ and their residuals $r_{j+\nu} := b - \hat{\mathbf{B}}x_{j+\nu}$ after a restart satisfy for $\nu = 1, \dots, m_0$

$$\|r_{j+\nu}\| \leq \left(1 - \frac{C_5^2}{C_6^2}\right)^{\nu/2} \|r_j\| .$$

Because of $C_5 \|x_j - x\|^2 \leq (x_j - x)^T \hat{\mathbf{B}}(x_j - x) \leq C_6 \|x_j - x\| \|r_j\|$, a corresponding estimate holds for the errors $\|x_j - x\|$. □

Remark 6.2. If the operator A is symmetric, we can also use the conjugate gradient method for the symmetric matrix $\hat{\mathbf{B}}$. This will in general give the bound (6.2) with a smaller constant q than the GMRES method.

Note that, for a function $v_h \in V_h$ with coefficient vector v and scaled coefficient vector $\hat{v} = \mathbf{S}v$, we have from (6.5) that with $b(u,v) := k^{-1}(u,v) + \theta\tilde{a}(u,v)$ and $\|v\|_b^2 := b(v,v)$

$$\|\hat{v}\|^2 \sim \hat{v}^T \hat{\mathbf{B}} \hat{v} = \|v_h\|_b^2 .$$

A functional $f_h \in V_h^*$ corresponds to a coefficient vector f so that $(f_h, v_h) = f^T v$, and a scaled vector $\hat{f} = \mathbf{S}^{-1} f$ so that $(f_h, v_h) = \hat{f}^T \hat{v}$. Assume that we solve a linear system $\hat{B}\hat{v}_* = \hat{f}$ using n_G steps of GMRES(m_0), starting with initial guess 0, yielding an approximation $\hat{v}$. We then have

$$\|v_{h,*} - v_h\|_b \leq C q^{n_G} \|v_{h,*}\|_b ,$$

and for the residuals $\rho_h \in V_h^*$ defined by $(\rho_h, w_h) = (f, w_h) - b(v_h, w_h)$ it holds that

$$\|\rho_h\|_{b,*} \leq C q^{n_G} \|f_h\|_{b,*} ,$$

where for $g_h \in V_h^*$ with $\hat{\mathbf{B}}_s := (\hat{\mathbf{B}} + \hat{\mathbf{B}}^T)/2$

$$\|g_h\|_{b,*} := \sup_{w_h \in V_h} \frac{(g_h, w_h)}{\|w_h\|_b} = (\hat{g}^T \hat{\mathbf{B}}_s^{-1} \hat{g})^{1/2} \sim \|\hat{g}\| .$$

We have with the inverse inequality

$$(c_1 k^{-1} h^\rho + \theta)\tilde{a}(v_h, v_h) \leq b(v_h, v_h) \leq (c_2 k^{-1} + \theta)\tilde{a}(v_h, v_h)$$

implying

$$(c_2 k^{-1} + \theta)^{-1/2} \|f_h\|_{\tilde{*}} \leq \|f_h\|_{b,*} \leq (c_1 k^{-1} h^\rho + \theta)^{-1/2} \|f_h\|_{\tilde{*}}$$

and

$$\|v_{h,*} - v_h\|_{\tilde{a}} \leq C\gamma^{1/2} q^{n_G} \|v_{h,*}\|_{\tilde{a}} , \tag{6.6}$$

$$\|\rho_h\|_{\tilde{*}} \leq C\gamma^{1/2} q^{n_G} \|f_h\|_{\tilde{*}} , \tag{6.7}$$

where

$$\gamma := \frac{c_2 k^{-1} + \theta}{c_1 k^{-1} h^{\rho} + \theta}.$$

We now define the *perturbed θ-scheme with GMRES approximation* as follows. Pick a value $m_0 \geq 1$ for the restart number, e.g., $m_0 = 1$, and a value n_G for the number of GMRES iterations. Let $\breve{u}_h^0 := u_{0,h}$. At each time step we want to find an approximation of $w_{h,*}^m$ satisfying

$$b(w_{h,*}^m, v_h) = (g^{m+\theta}, v_h) - \tilde{a}(\breve{u}_h^m, v_h) \qquad \forall\ v_h \in V_h,$$

which corresponds to a scaled linear system $\hat{\mathbf{B}}\hat{w}_*^m = \hat{b}^m$. We solve this system approximately with n_G steps of GMRES(m_0), using zero as initial guess, yielding an approximation $\hat{w}^m$ of the exact solution $\hat{w}_*^m$. We then let $\breve{u}_h^{m+1} := \breve{u}_h^m + w_h^m$, where $w_h^m \in V_h$ is the function corresponding to the scaled vector $\hat{w}^m$. Then we have the following.

THEOREM 6.3. *Assume that the consistency conditions* (3.8), (3.9) *hold. For* $\theta \in [0, \frac{1}{2})$ *assume* (4.3). *Then the solution* $\breve{u}_h^m$ *of the perturbed* θ*-scheme with GMRES approximation satisfies the same error bound as* $\tilde{u}_h^m$ *in* (5.13) *if* $n_G \geq C\,|\log h|$.

Proof. Let $\tilde{u}_h^m$ denote the solution of (3.7) (with all linear systems solved exactly), and let $\breve{u}_h^m$ denote the corresponding solution where the linear system (6.1) for each time step is solved with n_G GMRES(m_0) steps, using zero as initial guess. Let $\rho_h^m \in V_h^*$ denote the residual of the approximate GMRES solution w_h^m: For all $v_h \in V_h$

$$\begin{aligned}(\rho_h^m, v_h) &= b(w_h^m, v_h) - (g^{m+\theta}, v_h) + \tilde{a}(\breve{u}_h^m, v_h)\\ &= k^{-1}(\breve{u}_h^{m+1} - \breve{u}_h^m, v_h) + \tilde{a}(\breve{u}_h^{m+\theta}, v_h) - (g^{m+\theta}, v_h).\end{aligned}$$

Then the difference $\zeta_h^m := \breve{u}_h^m - \tilde{u}_h^m$ satisfies $\zeta_h^0 = 0$ and a θ-scheme of the same form as (3.7b),

$$k^{-1}(\zeta_h^{m+1} - \zeta_h^m, v_h) + \tilde{a}(\zeta_h^{m+\theta}, v_h) = (\rho_h^m, v_h),$$

where $\zeta_h^{m+\theta} = (1-\theta)\zeta^m + \theta\zeta^{m+1}$.

We now apply Proposition 4.3 and obtain for $l = 0, \dots, M$

$$\begin{aligned}E_l := \left\|\zeta^l\right\|^2 + C_1 k \sum_{m=0}^{l-1} \left\|\zeta_h^{m+\theta}\right\|_{\tilde{a}}^2 &\leq C_2 k \sum_{m=0}^{l-1} \|\rho_h^m\|_{\tilde{*}}^2\\ &\leq C\gamma q^{2n_G} k \sum_{m=0}^{l-1} \left\|g^{m+\theta} - \tilde{a}(\breve{u}_h^m, \cdot\,)\right\|_{\tilde{*}}^2\\ &\leq C'\gamma q^{2n_G} k \sum_{m=0}^{l-1} \Big(\left\|g^{m+\theta}\right\|_{\tilde{*}}^2 + \|\zeta_h^m\|_{\tilde{a}}^2 + \|\tilde{u}_h^m\|_{\tilde{a}}^2\Big).\end{aligned}$$

We denote the right-hand side of (4.8) by Q.

Let us first assume that $\theta = 0$ or $\theta = 1$. In this case we choose n_G large enough so that $C'\gamma q^{2n_G} \leq C_1/2$ and obtain with $l = M$

$$\left\|\zeta^M\right\|^2 + \tfrac{1}{2} C_1 k \sum_{m=0}^{M-1} \left\|\zeta_h^{m+\theta}\right\|_{\tilde{a}}^2 \leq C\gamma q^{2n_G} Q, \tag{6.8}$$

since the terms $\|\zeta_h^m\|_{\tilde{a}}^2$ occur in E_M and the terms $\|\tilde{u}_h^m\|_{\tilde{a}}^2$ occur in the left-hand side of (4.8).

In the general case $\theta \in [0,1]$ we use

$$\|\tilde{u}_h^m\|_{\tilde{a}}^2 \le Ch^{-\rho} \|\tilde{u}_h^m\|^2 \le Ch^{-\rho} Q,$$
$$\|\zeta_h^m\|_{\tilde{a}}^2 \le Ch^{-\rho} \|\zeta_h^m\|^2 \le Ch^{-\rho} E_m,$$

yielding

$$E_l \le C\gamma q^{2n_G} \left((1+h^{-\rho})Q + k \sum_{m=0}^{l-1} h^{-\rho} E_m \right).$$

Therefore we have estimates of the form

$$E_0 = 0, \qquad E_l \le \mu + \nu \sum_{m=0}^{l-1} E_m,$$

from which we easily get by induction

$$E_l \le \mu(1+\nu)^{l-1}.$$

Here we have $\nu = C\gamma q^{2n_G} h^{-\rho} T/M$. We choose n_G large enough so that $C\gamma q^{2n_G} h^{-\rho} \le 1$ and get $(1+\nu)^M \le e^T$ and

$$E_M \le C\gamma q^{2n_G}(1+h^{-\rho})Qe^T. \tag{6.9}$$

Finally we have to choose n_G large enough so that the right-hand side in (6.8) or (6.9) is less than the bound in Theorem 5.4: If $k \le 1$, we have $\gamma \le Ch^{-\rho}$, and therefore we need n_G such that

$$q^{2n_G} h^{-2\rho} \le Ch^{2(p+1-\rho/2)},$$

which is satisfied for $n_G \ge C\,|\log h|$ with $C > 0$ sufficiently large, but independent of h, k. □

Theorem 6.3 allows us to estimate the complexity of the time-stepping scheme with incomplete GMRES solution of the linear systems.

Corollary 6.4. *Given the compressed stiffness matrix $\tilde{\mathbf{A}}$, the additional work for computing $\check{u}_h^1, \dots, \check{u}_h^M$ is bounded by $CMN(\log N)^{r+1}$, where $r = 1$ for $\rho \in (0,2]$, $r = 2$ for $\rho = 0$.*

The total work of the algorithm (for computing the compressed stiffness matrix and performing M time steps) is bounded by $CMN(\log N)^{r+1}$ operations if we use exact antiderivatives, and by $CN(\log N)^{r+d} + CMN(\log N)^{r+1}$ operations if we use quadrature for $d = 1, 2$.

We remark in closing that the powers of $\log N$ in these complexity estimates could be reduced by more elaborate compression techniques; see, e.g., [14].

REFERENCES

[1] R. A. Adams, *Sobolev Spaces*, Academic Press, New York, 1978.

[2] L. Breiman, *Probability*, Classics Appl. Math. 7, SIAM, Philadelphia, 1992.

[3] A. Cohen, *Wavelet methods in numerical analysis*, in Handb. Numer. Anal. 7, P. G. Ciarlet and J. L. Lions, eds., Elsevier/North–Holland, Amsterdam, 2000.

[4] W. Dahmen, *Wavelet solution of operator equations*, Acta Numer., 6 (1997), pp. 55–228.

[5] W. Dahmen and R. Stevenson, *Element-by-element construction of wavelets satisfying stability and moment conditions*, SIAM J. Numer. Anal., 37 (1999), pp. 319–352.

[6] S. C. Eisenstat, H. C. Elman, and M. H. Schultz, *Variational iterative methods for nonsymmetric systems of linear equations*, SIAM J. Numer. Anal., 20 (1983), pp. 345–357.

[7] L. Greengard and V. Rokhlin, *A new version of the fast multipole method for the Laplace equation in three dimensions*, Acta Numer., 6 (1997), pp. 229–269.

[8] J. L. Lions and E. Magenes, *Non-Homogeneous Boundary Value Problems and Applications*, Springer-Verlag, Berlin, 1972.

[9] T. von Petersdorff and C. Schwab, *Wavelet approximation for first kind boundary integral equations in polygons*, Numer. Math., 74 (1996), pp. 479–519.

[10] T. von Petersdorff and C. Schwab, *Fully discrete multiscale Galerkin BEM*, in Multiscale Wavelet Methods for Partial Differential Equations, Wavelet Anal. Appl. 6, W. Dahmen, P. Kurdila, and P. Oswald, eds., Academic Press, San Diego, 1997, pp. 287–346.

[11] T. von Petersdorff, C. Schwab, and R. Schneider, *Multiwavelets for second kind integral equations*, SIAM J. Numer. Anal., 34 (1997), pp. 2212–2227.

[12] P. Protter, *Stochastic Integration and Differential Equations*, Springer-Verlag, New York, Berlin, 1990.

[13] P. Raviart, *The use of numerical integration in finite element methods for solving parabolic equations*, in Topics in Numerical Analysis, J. Miller, ed., Academic Press, New York, 1973, pp. 353–382.

[14] R. Schneider, *Multiskalen- und Wavelet Matrixkompression*, Teubner Series in Applied Mathematics, Stuttgart, 1998.

[15] M. A. Shubin, *Pseudodifferential Operators and Spectral Theory*, Springer-Verlag, Berlin, Heidelberg, New York, 1987 (translated from Nauka, Moscow, 1978).

[16] L. Schwartz, *Theorie des distributions*, Hermann, Paris, 1957.

[17] R. Stevenson, *Stable three-point wavelet bases on general meshes*, Numer. Math., 80 (1998), pp. 131–158.

[18] V. Thomée, *Galerkin Finite Element Methods for Parabolic Problems*, Springer-Verlag, Berlin, Heidelberg, New York, 1997.

SIAM J. Numer. Anal.
Vol. 41, No. 1, pp. 181–194

DOMAIN DECOMPOSITION FOR A MIXED FINITE ELEMENT METHOD IN THREE DIMENSIONS*

Z. CAI[†], R. R. PARASHKEVOV[‡], T. F. RUSSELL[§], J. D. WILSON[¶], AND X. YE[‖]

Abstract. We consider the solution of the discrete linear system resulting from a mixed finite element discretization applied to a second-order elliptic boundary value problem in three dimensions. Based on a decomposition of the velocity space, these equations can be reduced to a discrete elliptic problem by eliminating the pressure through the use of substructures of the domain. The practicality of the reduction relies on a local basis, presented here, for the divergence-free subspace of the velocity space. We consider additive and multiplicative domain decomposition methods for solving the reduced elliptic problem, and their uniform convergence is established.

Key words. divergence-free basis, domain decomposition, second-order elliptic problems, mixed finite element method

AMS subject classifications. 65F10, 65F30, 65L60

PII. S0036142996296935

1. Introduction.

In [6], Ewing and Wang considered and analyzed a domain decomposition method for solving the discrete system of equations which result from mixed finite element approximation of second-order elliptic boundary value problems in two dimensions. The approach in [6] is first to seek a discrete velocity satisfying the discrete continuity equation through a variation of domain decomposition (static condensation), and then to apply a domain decomposition method to the reduced elliptic problem arising from elimination of the pressure in the saddle-point problem. For analogous work, see also [8], [10], and [4]. The crucial part of the approach in [6] is to characterize the divergence-free velocity subspaces. This is also the essential difference with those in [8], [10], and [4].

In this paper, we will use the domain decomposition approach in [6] for the solution of the algebraic system resulting from the mixed finite element method applied to second-order elliptic boundary value problems in three dimensions. As mentioned above, the basis of the divergence-free velocity subspace plays an essential role in the approach; hence we will construct a basis of this subspace for the lowest-order rectangular Raviart–Thomas–Nedelec velocity space [13], [12]. The construction in two dimensions (2-D) is more general and rather easier than in three dimensions (3-D) due to the fact that any divergence-free vector in 2-D can be expressed as the curl

*Received by the editors January 2, 1996; accepted for publication (in revised form) September 3, 2002; published electronically March 13, 2003.

http://www.siam.org/journals/sinum/41-1/29693.html

[†]Department of Mathematics, Purdue University, West Lafayette, IN 47907-1395 (zcai@math.purdue.edu).

[‡]Exxon Mobil Upstream Research Corporation, Houston, TX 77252-2189 (rossen.parashkevov@exxonmobil.com).

[§]Department of Mathematics, University of Colorado at Denver, P.O. Box 173364, Campus Box 170, Denver, CO 80217-3364 (trussell@carbon.cudenver.edu). The research of this author was supported in part by National Science Foundation grants DMS-9312752 and DMS-0084438, U.S. Army Research Office grants DAAH04-95-1-0398 and DAAG55-97-1-0171, and the U.S. Geological Survey, Water Resources Division, Denver, CO.

[¶]U.S. Geological Survey, M.S. 413, Denver, CO 80225-0046 (johndw@usgs.gov).

[‖]Department of Mathematics and Statistics, University of Arkansas at Little Rock, Little Rock, AR 72204 (xxye@ualr.edu).

of a scalar stream function. Extension of this work to triangular or irregular meshes and to multilevel domain decomposition will be discussed in a forthcoming paper.

This approach has several practical advantages. For an $n \times n \times n$ grid in 3-D, the number of discrete unknowns is approximately $4n^3$, essentially one pressure and three velocity components per cell. Using the divergence-free subspace, we decouple the system in such a manner that the velocity can be obtained by solving a symmetric positive definite system of order roughly $2n^3$. In contrast to some other proposed procedures, this does not require the introduction of Lagrange multipliers corresponding to pressures at cell interfaces, and it permits direct computation of the velocity, which is often the principal variable of interest, alone. If the pressure is also needed, it can be calculated inexpensively in an additional step. Furthermore, the approach deals readily with the case of full-tensor conductivity (cross-derivatives), where the mass matrix is fuller than tridiagonal and methods based on reduced integration (mass lumping) are difficult to apply. This case results, for example, from anisotropic permeabilities in flows in porous media, where highly discontinuous conductivity coefficients are also common. For such problems, mixed methods are known to produce more realistic velocities than standard techniques [11].

The outline of the remainder of this paper is as follows. In section 2, we review the mixed finite element method for elliptic problems with homogeneous Neumann boundary conditions. The domain decomposition method for the resulting algebraic system is discussed in section 3, and its uniform convergence is established in section 5. A computationally convenient, divergence-free basis with minimal support is constructed in section 4.

2. Mixed finite element method. In this section, we begin with a brief review of the mixed finite element method with lowest-order Raviart–Thomas–Nedelec [13], [12] (RTN) approximation space for second-order elliptic boundary value problems in three dimensions. For simplicity, we consider a homogeneous Neumann problem: find p such that

$$\left\{ \begin{array}{rcllc} -\nabla \cdot (k\nabla p) & = & f & \text{in} & \Omega = (0,\,1)^3, \\ (k\nabla p)\cdot \mathbf{n} & = & 0 & \text{on} & \partial\Omega, \end{array} \right. \tag{2.1}$$

where $f \in L^2(\Omega)$ satisfies the relation

$$\int_\Omega f \, dx\, dy\, dz = 0 \tag{2.2}$$

and $\mathbf{n}$ denotes the unit outward normal vector to $\partial\Omega$. The symbols $\nabla\cdot$ and ∇ stand for the divergence and gradient operators, respectively. Assume that $k = (k_{ij})_{3\times 3}$ is a given real-valued symmetric matrix function with bounded and measurable entries k_{ij} ($i,\, j = 1,\, 2,\, 3$) and satisfies the ellipticity condition; i.e., there exist positive constants α_1 and α_2 such that

$$\alpha_1 \xi^t\, \xi \le \xi^t k(x,\, y,\, z)\xi \le \alpha_2 \xi^t\, \xi \tag{2.3}$$

for all $\xi \in \mathbb{R}^3$ and almost all $(x,\, y,\, z) \in \bar{\Omega}$.

We shall use the following space to define the mixed variational problem. Let

$$H(div;\Omega) \equiv \{\mathbf{w} \in L^2(\Omega)^3 \,|\, \nabla\cdot\mathbf{w} \in L^2(\Omega)\},$$

which is a Hilbert space equipped with the norm

$$\|\mathbf{w}\|_{H(div;\Omega)} \equiv (\|\mathbf{w}\|^2_{L^2(\Omega)^3} + \|\nabla\cdot\mathbf{w}\|^2_{L^2(\Omega)})^{1/2}$$

and the associated inner product. By introducing the flux variable

$$\mathbf{v} = -k\nabla p,$$

which is of practical interest for many physical problems, we can rewrite the PDE of (2.1) as a first-order system

$$\left\{ \begin{array}{rcl} k^{-1}\mathbf{v} + \nabla p &=& 0, \\ \nabla \cdot \mathbf{v} &=& f \end{array} \right.$$

and obtain the mixed formulation of (2.1): find $(\mathbf{v},\, p) \in \mathbf{V} \times \Lambda$ such that

$$\left\{ \begin{array}{lcll} a(\mathbf{v},\, \mathbf{w}) - b(\mathbf{w},\, p) &=& 0 & \forall\, \mathbf{w} \in \mathbf{V}, \\ b(\mathbf{v},\, \lambda) &=& (f,\, \lambda) & \forall\, \lambda \in \Lambda. \end{array} \right. \tag{2.4}$$

Here $\mathbf{V} = H_0(div;\Omega) \equiv \{\mathbf{w} \in H(div;\Omega)\,|\,\mathbf{w}\cdot\mathbf{n} = 0 \text{ on } \partial\Omega\}$, Λ is the quotient space $L_0^2(\Omega) = L^2(\Omega)/\{\text{constants}\}$, the bilinear forms $a(\cdot\,,\cdot) : \mathbf{V}\times\mathbf{V} \to \mathbb{R}$ and $b(\cdot\,,\cdot) : \mathbf{V}\times\Lambda \to \mathbb{R}$ are defined by

$$a(\mathbf{w},\, \mathbf{u}) = \int_\Omega (k^{-1}\mathbf{w})\cdot\mathbf{u}\, dx\, dy\, dz \quad \text{and} \quad b(\mathbf{w},\, \lambda) = \int_\Omega (\nabla\cdot\mathbf{w})\lambda\, dx\, dy\, dz$$

for any $\mathbf{w},\, \mathbf{u} \in \mathbf{V}$ and $\lambda \in \Lambda$, respectively, and $(\cdot,\cdot)$ denotes the $L^2(\Omega)$ inner product.

To discretize the mixed formulation (2.4), we assume that we are given two finite element subspaces

$$\mathbf{V}^h \subset \mathbf{V} \quad \text{and} \quad \Lambda^h \subset \Lambda$$

defined on a uniform rectangular mesh with elements of size $O(h)$. The mixed approximation of $(\mathbf{v}\,, p)$ is defined to be the pair $(\mathbf{v}^h,\, p^h) \in \mathbf{V}^h \times \Lambda^h$ satisfying

$$\left\{ \begin{array}{lcll} a(\mathbf{v}^h,\, \mathbf{w}) - b(\mathbf{w}\,, p^h) &=& 0 & \forall\, \mathbf{w} \in \mathbf{V}^h, \\ b(\mathbf{v}^h,\, \lambda) &=& (f,\, \lambda) & \forall\, \lambda \in \Lambda^h. \end{array} \right. \tag{2.5}$$

We refer to [13] for the definition of a class of approximation subspaces $\mathbf{V}^h$ and Λ^h. In this paper, we shall consider only the lowest-order RTN space defined on a rectangular triangulation of Ω. Such a space for the velocity consists of vector functions whose ith component is continuous piecewise linear in the x_i variable and discontinuous piecewise constant in the x_j variable for $j \neq i$. The corresponding pressure space Λ^h consists of discontinuous piecewise constants with respect to the triangulation $\mathcal{T}^h$ with a fixed value on one element. Specifically, let $\mathcal{T}^h$ denote a uniform rectangular triangulation of Ω. Then the lowest-order RTN approximation space for the velocity on a rectangle $K \in \mathcal{T}^h$ is defined by

$$\mathbf{V}^h(K) = \mathcal{P}_{1,\,0,\,0} \times \mathcal{P}_{0,\,1,\,0} \times \mathcal{P}_{0,\,0,\,1}, \tag{2.6}$$

and the corresponding pressure space is

$$\Lambda^h(K) = \mathcal{P}_{0,\,0,\,0}, \tag{2.7}$$

where $\mathcal{P}_{i_1,\,i_2,\,i_3}(K)$ denotes the polynomials of degree i_j $(j = 1,\, 2,\, 3)$ with respect to x_j. It is well known that the above RTN space satisfies the Babŭska–Brezzi stability condition (cf. [13]): there exists a positive constant β independent of the mesh size h of $\mathcal{T}^h$ such that

$$\sup_{\mathbf{w}\in\mathbf{V}^h} \frac{b(\mathbf{w}\,,\lambda)}{\|\mathbf{w}\|_{H(div,\,\Omega)}} \geq \beta\|\lambda\|_{L^2(\Omega)} \quad \forall\,\lambda\in\Lambda^h. \tag{2.8}$$

Also, Raviart and Thomas in [13] demonstrated the existence of a projection operator $\mathbf{\Pi}_h : \mathbf{V} \longrightarrow \mathbf{V}^h$ such that, for any $\mathbf{v}\in\mathbf{V}$,

$$b(\mathbf{\Pi}_h\mathbf{v}\,,\lambda) = b(\mathbf{v}\,,\lambda) \quad \forall\,\lambda\in\Lambda^h, \tag{2.9}$$

$$\|\mathbf{\Pi}_h\mathbf{v}-\mathbf{v}\|_{L^2(\Omega)^3} \leq C\,h^s\|\mathbf{v}\|_{H^s(\Omega)^3}, \quad s=0,\,1. \tag{2.10}$$

3. Domain decomposition.

Problem (2.5) is clearly symmetric and indefinite. To reduce it to a symmetric positive definite problem, we need a discrete velocity $\mathbf{v}_I^h\in\mathbf{V}^h$ satisfying

$$b(\mathbf{v}_I^h,\,\lambda) = (f,\,\lambda) \quad \forall\,\lambda\in\Lambda^h. \tag{3.1}$$

Define the discretely (as opposed to pointwise) divergence-free subspace $\mathbf{D}^h$ of $\mathbf{V}^h$:

$$\mathbf{D}^h = \{\mathbf{w}\in\mathbf{V}^h \,|\, b(\mathbf{w},\lambda)=0 \quad \forall\,\lambda\in\Lambda^h\}, \tag{3.2}$$

and let

$$\mathbf{v}_D^h = \mathbf{v}^h - \mathbf{v}_I^h,$$

which is obviously in $\mathbf{D}^h$ by the second equation of (2.5) and which satisfies

$$a(\mathbf{v}_D^h,\,\mathbf{w}) = -a(\mathbf{v}_I^h,\,\mathbf{w}) \quad \forall\,\mathbf{w}\in\mathbf{D}^h, \tag{3.3}$$

by the first equation. This problem is symmetric and positive definite.

This suggests the following procedure for obtaining $\mathbf{v}^h$, the solution of (2.5): find $\mathbf{v}_I^h\in\mathbf{V}^h$ satisfying (3.1), compute the projection $\mathbf{v}_D^h\in\mathbf{D}^h$ satisfying (3.3), then set $\mathbf{v}^h=\mathbf{v}_I^h+\mathbf{v}_D^h$. This procedure will be the basis for Algorithms 3.1 and 3.2 below. Given $\mathbf{v}_I^h$, (3.3) leads to a unique $\mathbf{v}^h$, which is independent of the choice of $\mathbf{v}_I^h$. (A term added to a given $\mathbf{v}_I^h$ must be in $\mathbf{D}^h$, and it is canceled by the resulting change in $\mathbf{v}_D^h$.) For an $n\times n\times n$ grid, computing the projection $\mathbf{v}_D^h$ involves solving a system of order approximately $2n^3$. Solving for p^h is optional; if it is desired, it can be obtained from the first equation in (2.5) once $\mathbf{v}^h$ is known.

There are many discrete velocities in $\mathbf{V}^h$ satisfying (3.1), and several approaches have been discussed in the literature for seeking such a discrete velocity (e.g., [6], [8], and [10]). All of these approaches are based on a type of domain decomposition (static condensation) method applied to problem (2.5). In this paper, we will adopt the approach discussed in [6] by Ewing and Wang. This approach requires solving only a coarse-grid problem and some local problems of the form (2.5).

To compute $\mathbf{v}_I^h$ and define the domain decomposition method for problem (3.3), we start with a coarse initial rectangular triangulation $\mathcal{T}^H=\{K_j\}_{j=1}^J$ of the domain Ω (so that $\bar{\Omega}=\cup_{j=1}^J\bar{K}_j$), and a regular fine rectangular triangulation $\mathcal{T}^h$ obtained by further partitioning all of the elements in $\mathcal{T}^H$. Associated with the coarse triangulation $\mathcal{T}^H$, we construct a set of overlapping subdomains $\{\Omega_j\}_{j=1}^J$ by extending each element $K_j \in \mathcal{T}^H$ to a larger subdomain Ω_j, whose diameter is denoted by $H_j \leq C\,H$. Assume that the maximum number of subdomain overlaps is bounded, and further that the distance between the boundaries ∂K_j and $\partial\Omega_j$ is bounded below by $\zeta_1 H$ and above by $\zeta_2 H$; i.e., for all $j \in \{1, \ldots, J\}$ there exist constants $\zeta_1,\, \zeta_2 > 0$ such that

$$\zeta_1 H \leq \operatorname{dist}(\partial K_j,\, \partial\Omega_j) \leq \zeta_2 H.$$

Also assume that the boundaries of the Ω_j do not cut through any element in $\mathcal{T}^h$, i.e., they must coincide with boundaries of elements of $\mathcal{T}^h$. Thus, the restrictions of $\mathcal{T}^h$ on Ω_j and K_j provide two uniform triangulations $\mathcal{T}_j^h$ and $\tilde{\mathcal{T}}_j^h$ for Ω_j and K_j, respectively.

Let $\mathbf{V}_j \times \Lambda_j$ and $\tilde{\mathbf{V}}_j \times \tilde{\Lambda}_j$ be the lowest-order RTN approximation spaces corresponding to the triangulations $\mathcal{T}_j^h$ and $\tilde{\mathcal{T}}_j^h$, respectively. For convenience, let $\mathbf{V}^H = \mathbf{V}_0 = \tilde{\mathbf{V}}_0$ and $\Lambda^H = \Lambda_0 = \tilde{\Lambda}_0$. As in [6], let f^h and $f_0^h \equiv f^H$ be the L^2 projection of f in Λ^h and Λ^H, respectively, and $f_j^h \in \tilde{\Lambda}_j^h$ be the restriction of $f^h - f_0^h$ on K_j. Then the discrete velocity $\mathbf{v}_I^h$ satisfying (3.1) may be determined by the sum of $\mathbf{v}_j$'s which are the solutions of the following problems: find $(\mathbf{v}_j,\, p_j) \in \tilde{\mathbf{V}}_j \times \tilde{\Lambda}_j$ such that

$$\left\{ \begin{array}{lcll} (\tilde{k}\mathbf{v}_j,\, \mathbf{w}) - b(\mathbf{w},\, p_j) & = & 0 & \forall\, \mathbf{w} \in \tilde{\mathbf{V}}_j, \\ b(\mathbf{v}_j,\, \lambda) & = & (f_j^h,\, \lambda) & \forall\, \lambda \in \tilde{\Lambda}_j, \end{array} \right. \tag{3.4}$$

where $\tilde{k} \in \mathbb{R}^{3\times 3}$ is an arbitrary matrix-valued function which is symmetric positive definite and defined on Ω_j for all $j \in \{0,\, 1, \ldots, J\}$. Note that $\mathbf{v}_0$ is the solution of problem (2.5) corresponding to the coarse triangulation $\mathcal{T}^H$, and that $\mathbf{v}_j$ for $1 \leq j \leq J$ can be obtained by solving some local problems.

We shall use additive and multiplicative domain decomposition methods for approximate computation of the solution of problem (3.3). To this end, we define the family of discretely divergence-free velocity subspaces $\{\mathbf{D}_j\}_{j=0}^J$ by $\mathbf{D}_0 = \mathbf{D}^H$, and for $j \in \{1,\, 2, \ldots, J\}$,

$$\mathbf{D}_j = \{\mathbf{u} \in \mathbf{V}_j \,|\, b(\mathbf{u},\, \lambda) = 0 \quad \forall\, \lambda \in \Lambda_j\}.$$

For any $\mathbf{u} \in \mathbf{D}^h$, we define the projection operators $\mathbf{P}_j : \mathbf{D}^h \longrightarrow \mathbf{D}_j$ associated with the bilinear form $a(\cdot\,,\cdot)$ by

$$a(\mathbf{P}_j\mathbf{u},\, \mathbf{w}) = a(\mathbf{u},\, \mathbf{w}) \quad \forall\, \mathbf{w} \in \mathbf{D}_j$$

for $j \in \{0,\, 1, \ldots, J\}$.

ALGORITHM 3.1 (Additive domain decomposition).

1. *For $j = 0, 1, \ldots, J$, compute $\mathbf{v}_j \in \tilde{\mathbf{V}}_j$ by solving problems* (3.4). *Then set*

$$\mathbf{v}_I^h = \mathbf{v}_0 + \mathbf{v}_1 + \cdots + \mathbf{v}_J.$$

2. *Compute an approximation $\mathbf{v}_D$ of $\mathbf{v}_D^h \in \mathbf{D}^h$ by applying conjugate gradient iteration to*

$$\mathbf{P}\mathbf{v}_D = \mathbf{F}, \tag{3.5}$$

where $\mathbf{P} = \mathbf{P}_0 + \mathbf{P}_1 + \cdots + \mathbf{P}_J$, $\mathbf{F} = \mathbf{F}_0 + \mathbf{F}_1 + \cdots + \mathbf{F}_J$, and $\mathbf{F}_j = \mathbf{P}_j\mathbf{v}_D^h$.

3. *Set*

$$\mathbf{v}^h = \mathbf{v}_D + \mathbf{v}_I^h.$$

Remark 3.1. The right-hand side $\mathbf{F}$ in (3.5) can be computed by solving the coarse-grid problem and local subproblems. Specifically, for each $j \in \{0, 1, \ldots, J\}$, $\mathbf{F}_j$ is the solution of the following problem:

$$a(\mathbf{F}_j,\, \mathbf{w}) = a(\mathbf{P}_j \mathbf{v}_D^h,\, \mathbf{w}) = -a(\mathbf{v}_I^h,\, \mathbf{w}) \quad \forall\, \mathbf{w} \in \mathbf{D}_j. \tag{3.6}$$

For $\mathbf{F}_j$ given in (3.6), we can see as follows that (3.5) is equivalent to problem (3.3). Given (3.3), define $\mathbf{F}_j$ as in (3.6), and $\mathbf{P}$ and $\mathbf{F}$ as above. Then $a(\mathbf{F}_j, w) = a(v_D^h, w)$ $\forall w \in \mathbf{D}_j$, so that $\mathbf{F}_j = \mathbf{P}_j \mathbf{v}_D^h$, and summing on j yields $\mathbf{P}\mathbf{v}_D^h = \mathbf{F}$. To complete the equivalence, we claim that $\mathbf{v}_D^h$ is the only solution of (3.5). It suffices to show that $\mathbf{P}\mathbf{u} = \mathbf{0}$ implies that $\mathbf{u} = \mathbf{0}$ for $\mathbf{u} \in \mathbf{D}^h$. If $\mathbf{P}\mathbf{u} = \mathbf{0}$, then

$$0 = a(\mathbf{P}\mathbf{u}, \mathbf{u}) = \sum_j a(\mathbf{P}_j\mathbf{u}, \mathbf{u}) = \sum_j a(\mathbf{P}_j\mathbf{u}, \mathbf{P}_j\mathbf{u}),$$

so that $a(\mathbf{P}_j\mathbf{u}, \mathbf{P}_j\mathbf{u}) = 0$ $\forall j$; hence $\mathbf{P}_j\mathbf{u} = \mathbf{0}$ $\forall j$. In Lemma 5.1 below, we prove that $\mathbf{u}$ has a decomposition $\mathbf{u} = \mathbf{u}_0 + \mathbf{u}_1 + \cdots + \mathbf{u}_J$, where $\mathbf{u}_j \in \mathbf{D}_j$. With this,

$$a(\mathbf{u}, \mathbf{u}) = \sum_j a(\mathbf{u}_j, \mathbf{u}) = \sum_j a(\mathbf{u}_j, \mathbf{P}_j\mathbf{u}) = 0,$$

and hence $\mathbf{u} = \mathbf{0}$, as claimed.

At each iteration of the conjugate gradient method applied to (3.5), we need to compute the action of the projection operator $\mathbf{P}_j$ on a given $\mathbf{u} \in \mathbf{D}^h$, which may be obtained by solving the following problem:

$$a(\mathbf{P}_j\mathbf{u},\, \mathbf{w}) = a(\mathbf{u},\, \mathbf{w}) \quad \forall\, \mathbf{w} \in \mathbf{D}_j. \tag{3.7}$$

When analyzing the preconditioned conjugate gradient method for a system of linear equations, the crucial issue is to estimate the condition number of the preconditioned operator. In section 5, we will establish a uniform estimate of the condition number for $\mathbf{P}$ and find a basis for $\mathbf{D}^h$ that allows for efficient computations.

Algorithm 3.2 (Multiplicative domain decomposition).

1. *Compute* $\mathbf{v}_I^h$ *as in the first step of* Algorithm 3.1.
2. *Given an approximation* $\mathbf{v}_D^l \in \mathbf{D}^h$ *to the solution* $\mathbf{v}_D^h$ *of* (3.3)*, define the next approximation* $\mathbf{v}_D^{l+1} \in \mathbf{D}^h$ *as follows:*
   (a) *Set* $W_{-1} = \mathbf{v}_D^l$.
   (b) *For* $j = 0, 1, \ldots, J$ *in turn, define* W_j *by*
   $$W_j = W_{j-1} + \omega \mathbf{P}_j(\mathbf{v}_D^h - W_{j-1}),$$
   where the parameter $\omega \in (0,\, 2)$.
   (c) *Set* $\mathbf{v}_D^{l+1} = W_J$.
3. *Set*
   $$\mathbf{v}^h = \mathbf{v}_I^h + \mathbf{v}_D^L.$$

Remark 3.2. $\mathbf{P}_j(\mathbf{v}_D^h - W_{j-1})$ can be computed by solving the following problem:

$$a(\mathbf{P}_j(\mathbf{v}_D^h - W_{j-1}),\, \mathbf{w}) = -a(\mathbf{v}_I^h + W_{j-1},\, \mathbf{w}) \quad \forall\, \mathbf{w} \in \mathbf{D}_j. \tag{3.8}$$

A simple computation implies that the error propagation operator of multiplicative domain decomposition at the second step of Algorithm 3.2 has the form of

$$\mathbf{E} = (\mathbf{I} - \mathbf{P}_J)(\mathbf{I} - \mathbf{P}_{J-1}) \cdots (\mathbf{I} - \mathbf{P}_0). \tag{3.9}$$

Define a norm associated with the bilinear form $a(\cdot,\, \cdot)$ by

$$\|\mathbf{u}\|_a = a(\mathbf{u},\, \mathbf{u})^{1/2} \quad \forall\, \mathbf{u} \in \mathbf{D}^h.$$

We shall show in the last section that $\|\mathbf{E}\|_a$ is bounded by a constant which is less than one and independent of the mesh size h and the number of subdomains.

4. Construction of a divergence-free basis. Since the technique of the mixed method leads to a saddle-point problem, which causes the final system to be indefinite, many well-established efficient linear system solvers cannot be applied. As we mentioned earlier, (2.5) could be symmetric and positive definite if we discretize it in the discrete divergence-free subspace $\mathbf{D}^h$. The construction of a basis for $\mathbf{D}^h$ is essential.

In this section, we will construct a computationally convenient basis for $\mathbf{D}^h$—the divergence-free subspace of $\mathbf{V}^h$. We will do this by first constructing a vector potential space $\mathbf{U}^h$ such that

$$\mathbf{D}^h = \mathbf{curl}\,\mathbf{U}^h. \tag{4.1}$$

Next, we will find a basis for $\mathbf{U}^h$, and we will define a basis for $\mathbf{D}^h$ by simply taking the curls of the vector potential basis functions.

Denote the mesh on $\Omega = (0,1)^3$ by $0 = x_0 < \cdots < x_i < \cdots < x_n = 1$, and similarly with y_j and z_k, $0 \le j,k \le n$. The assumption of the same number n of intervals in each direction is merely for convenience and is not necessary for the construction to follow. Let

$$\phi^x_{i,j,k}(x,y,z) = \chi_i(x)\psi_j(y)\psi_k(z), \qquad 1 \le i \le n, \quad 1 \le j,k \le n-1,$$

where χ_i is the characteristic function of (x_{i-1}, x_i), ψ_j is the standard hat function supported on (y_{j-1}, y_{j+1}), and similarly ψ_k is supported on (z_{k-1}, z_{k+1}). Then $\phi^x_{i,j,k}$ is the standard bilinear nodal basis function on $(y_{j-1}, y_{j+1}) \times (z_{k-1}, z_{k+1})$, extended as a constant in the x-direction in the ith slice only, zero in the other slices. For economy of notation, write $\phi_i(y,z)$ for $\phi^x_{i,j,k}(x,y,z)$, where the single index i, $1 \le i \le n(n-1)^2$, runs through the triples (i,j,k) lexicographically (k varying most rapidly). The support of a typical $\phi_i(y,z)$ consists of a $1 \times 2 \times 2$ set of 4 cells and is shown in Figure 4.1. Similarly, let

$$\begin{aligned} \phi^y_{j,i,k}(x,y,z) &= \chi_j(y)\psi_i(x)\psi_k(z) = \phi_j(x,z), \quad 1 \le j \le n,\ 1 \le i,k \le n-1, \\ \phi^z_{k,i,j}(x,y,z) &= \chi_k(z)\psi_i(x)\psi_j(y) = \phi_k(x,y), \quad 1 \le k \le n,\ 1 \le i,j \le n-1, \end{aligned}$$

where j, k run lexicographically through (j,i,k) and (k,i,j), respectively. Finally, let $\mathbf{U}^h$ be defined as follows:

$$\mathbf{U}^h = \text{span}\left\{ \begin{pmatrix} \phi_i(y,z) \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ \phi_j(x,z) \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ \phi_k(x,y) \end{pmatrix} \right\}, \tag{4.2}$$

where $1 \le i \le (n-1)^2$ (thus, only the first yz-slice is included) and $1 \le j,k \le n(n-1)^2$ (all xz- and xy-slices are included). Note that the number of excluded ϕ_i's is $(n-1)^3$. If the number of intervals in the x-, y-, and z-directions were ℓ, m, and n, respectively, the number excluded would be $(\ell-1)(m-1)(n-1)$ and would be the same if all but one xz- or xy-slice were excluded instead of all but one yz-slice.

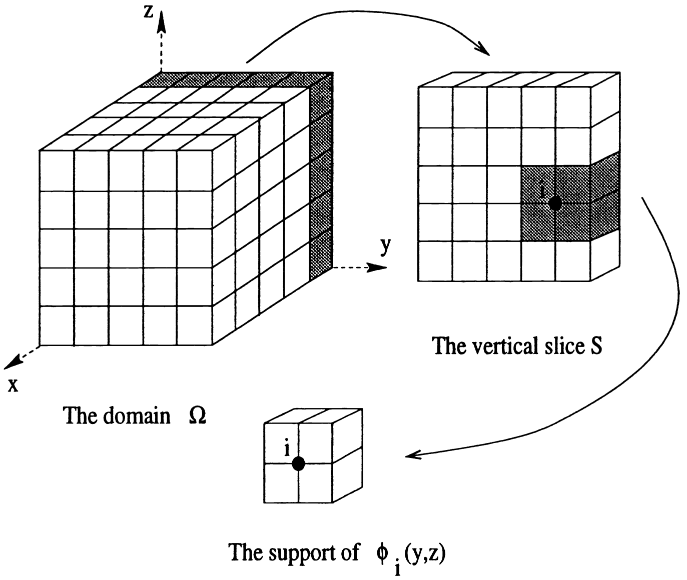


FIG. 4.1. *The support of a typical potential basis function.*

Next, we list some properties of $\mathbf{U}^h$ which follow directly from the definition of the potential space.

Remark 4.1. $\mathbf{U}^h \not\subset H(div;\Omega)$ (because, e.g., $\phi_i(y,z)$ is discontinuous in x), and hence, $\mathbf{U}^h \not\subset H^1(\Omega)^3$.

Remark 4.2. Every $\mathbf{\Phi} \in \mathbf{U}^h$ satisfies $\mathbf{\Phi} \times \mathbf{n} = \mathbf{0}$ on $\partial\Omega$ (because, e.g., $\phi_j(0,z)$ and $\phi_k(0,y)$ are identically zero).

Remark 4.3. $\mathbf{U}^h$ is locally divergence-free, i.e., $\nabla\cdot\mathbf{\Phi} = 0$ on each element $K \in \mathcal{T}^h$ for every $\mathbf{\Phi} \in \mathbf{U}^h$.

Remark 4.4. $\mathbf{U}^h \subset H(curl;\Omega)$, and hence $\mathbf{curl}\,\mathbf{U}^h \subset \mathbf{V}^h$. To see this, consider as a typical case the vector function $(\phi_i(y,z),0,0) \in \mathbf{U}^h$, whose curl is $(0, \partial\phi_i/\partial z, -\partial\phi_i/\partial y)$. Because ϕ_i is discontinuous only in the x-direction and no x-derivatives appear in the curl, we have $(\phi_i,0,0) \in H(curl;\Omega)$. Further, the y-component of $\mathbf{curl}(\phi_i,0,0)$ is $\partial\phi_i/\partial z = \chi_i(x)\psi_j(y)\psi_k'(z)$, which is continuous piecewise linear in y and discontinuous piecewise constant in x and z; similarly, the other components have the correct form to yield $\mathbf{curl}(\phi_i,0,0) \in \mathbf{V}^h$.

Since $\operatorname{div}\mathbf{curl} \equiv 0$, we have $\mathbf{curl}\,\mathbf{U}^h \subset \mathbf{D}^h$. Counting dimensions,

$$\dim \mathbf{U}^h = (2n+1)(n-1)^2 = 2n^3 - 3n^2 + 1.$$

Also, $\operatorname{div}\mathbf{V}^h$ consists of those piecewise constants with integral zero over Ω, and hence has dimension $n^3 - 1$, and we obtain

$$\dim \mathbf{D}^h = \dim \mathbf{V}^h - \dim \operatorname{div} \mathbf{V}^h = 3(n-1)n^2 - (n^3 - 1) = 2n^3 - 3n^2 + 1.$$

We claim that the curls of the vectors in (4.2) are linearly independent, so that

$$\dim \mathbf{D}^h = \dim \mathbf{curl}\,\mathbf{U}^h = \dim \mathbf{U}^h = 2n^3 - 3n^2 + 1,$$

which implies that for every divergence-free vector $\mathbf{v} \in \mathbf{D}^h$ there exists a unique potential vector $\mathbf{\Phi} \in \mathbf{U}^h$ such that

$$\mathbf{v} = \mathbf{curl}\,\mathbf{\Phi}.$$

To prove linear independence, first note that vectors in $\mathbf{V}^h$ can be characterized in terms of normal fluxes across the $3(n-1)n^2$ interior faces between elements. For example, some calculations will show that $\mathbf{curl}\,(\phi_1(y,z),0,0) = \mathbf{curl}\,(\phi^x_{1,1,1}(x,y,z),0,0)$ has y-component 1 on face $(1,3/2,1) = (x_0,x_1) \times \{y_1\} \times (z_0,z_1)$ and -1 on face $(1,3/2,2) = (x_0,x_1) \times \{y_1\} \times (z_1,z_2)$, and z-component 1 on $(1,2,3/2)$ and -1 on $(1,1,3/2)$, where the four fluxes have been scaled to unit magnitude. This is shown in Figure 4.2. We denote this particular curl by $+1(1,3/2,1)-1(1,3/2,2)+1(1,2,3/2)-1(1,1,3/2)$.

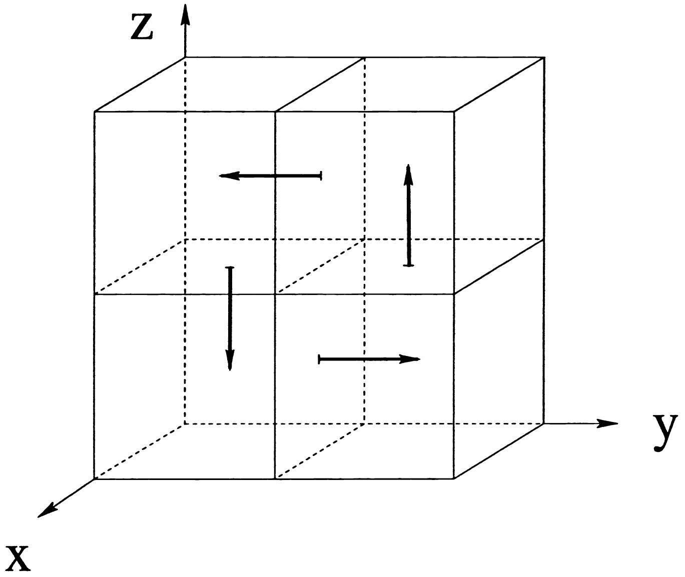


FIG. 4.2. *The curl of a typical potential basis function.*

Now consider $\mathbf{curl}\,(0,\phi_j(x,z),0)$, $1 \le j \le n(n-1)^2$. Put $\phi_j(x,z) = \phi^y_{j,i,k}(x,y,z)$ in lexicographic order, noting that $\mathbf{curl}\,(0,\phi^y_{j,i,k},0) = +1(i+1/2,j,k+1) - 1(i+1/2,j,k) + 1(i,j,k+1/2) - 1(i+1,j,k+1/2)$ and that face $(i+1/2,j,k+1)$ appears for the first time in $\mathbf{curl}\,(0,\phi^y_{j,i,k},0)$. Since each curl introduces a nonzero flux on a new face, the curls of $(0,\phi_j(x,z),0)$ are linearly independent. Next, we have $\mathbf{curl}\,(0,0,\phi_k(x,y)) = \mathbf{curl}\,(0,0,\phi^z_{k,i,j}) = +1(i+1,j+1/2,k) - 1(i,j+1/2,k) + 1(i+1/2,j,k) - 1(i+1/2,j+1,k)$, and face $(i+1,j+1/2,k)$ appears for the first time in $\mathbf{curl}\,(0,0,\phi^z_{k,i,j})$. Thus, the curls of $(0,\phi_j(x,z),0)$ and $(0,0,\phi_k(x,y))$, $1 \le j,k \le n(n-1)^2$, are all linearly independent.

Finally, consider a linear combination

$$\sum_i \alpha_i \mathbf{curl}\,(\phi_i(y,z),0,0) = \sum_{j=1}^{n-1}\sum_{k=1}^{n-1} \alpha_{jk}\mathbf{curl}\,(\phi^x_{1,j,k},0,0)$$

(terms from first yz-slice only). Because the curls of $(\phi_i,0,0)$ are linearly independent by the argument applied above to the curls of $(0,\phi_j,0)$, it suffices to show that this linear combination is independent of the curls of $(0, \phi_j, 0)$ and $(0, 0, \phi_k)$. We have $\mathbf{curl}\,(\phi^x_{1,j,k}, 0, 0) = +1(1, j + 1/2, k) - 1(1, j + 1/2, k + 1) + 1(1, j + 1, k + 1/2) - 1(1, j, k + 1/2)$. Each of these four terms occurs exactly once in the curls of $(0, \phi_j(x, z), 0)$ and $(0, 0, \phi_k(x, y))$, namely in $-\mathbf{curl}\,(0, 0, \phi^z_{k,1,j})$, $+\mathbf{curl}\,(0, 0, \phi^z_{k+1,1,j})$, $+\mathbf{curl}\,(0, \phi^y_{j+1,1,k}, 0)$, $-\mathbf{curl}\,(0, \phi^y_{j,1,k}, 0)$, respectively. Hence, a dependency relationship for $\mathbf{curl}\,(\phi^x_{1,j,k}, 0, 0)$ in terms of the preceding curls must involve these four curls, and when they are combined we get $\mathbf{curl}\,(\phi^x_{1,j,k}, 0, 0) - \mathbf{curl}\,(\phi^x_{2,j,k}, 0, 0)$. Applying this fact to each term of the linear combination, we have $\sum_{j=1}^{n-1} \sum_{k=1}^{n-1} \alpha_{jk}(\mathbf{curl}\,(\phi^x_{1,j,k}, 0, 0) - \mathbf{curl}\,(\phi^x_{2,j,k}, 0, 0))$. To cancel $\sum_{j,k} \alpha_{jk}(-\mathbf{curl}\,(\phi^x_{2,j,k}, 0, 0))$ with the preceding curls, the forced combination yields $\sum_{j,k} \alpha_{jk}(\mathbf{curl}\,(\phi^x_{2,j,k}, 0, 0) - \mathbf{curl}\,(\phi^x_{3,j,k}, 0, 0))$, and so on until $\sum_{j,k} \alpha_{jk}(-\mathbf{curl}\,(\phi^x_{n,j,k}, 0, 0))$ remains, and it is not possible to cancel it. It follows that no dependency relationship exists, so that the curls of the vectors in (4.2) are indeed linearly independent.

The vector functions in (4.2) constitute one choice of a basis for $\mathbf{U}^h$. As noted above, this choice includes all $2n(n-1)^2$ vectors of the forms $(0, \phi_j, 0)$ and $(0, 0, \phi_k)$, but only $(n-1)^2$ vectors of the type $(\phi_i, 0, 0)$ with support contained in one vertical slice S of Ω (say, the shaded one in Figure 4.1).

Remark 4.5. The above-defined basis for $\mathbf{U}^h$ (and hence for $\mathbf{D}^h$) consists of vector functions with minimal possible support. (A moment's reflection shows that a nontrivial divergence-free vector function must be supported on at least four elements, as in the pattern in Figure 4.2.)

Now we need to prove the following Poincaré-type inequality.

LEMMA 4.1. *There exists a constant $C(\Omega) > 0$, independent of the quasi-uniform mesh size h, such that for all $\mathbf{\Phi} \in \mathbf{U}^h$ we have*

$$\|\mathbf{\Phi}\|_{L^2(\Omega)^3} \le C(\Omega)\, \|\mathbf{curl}\,\mathbf{\Phi}\|_{L^2(\Omega)^3}. \tag{4.3}$$

(Since the vector potential space $\mathbf{U}^h \not\subset H^1(\Omega)^3$, inequality (4.3) does not follow from the standard Poincaré inequality.)

Proof. Keeping in mind our choice for a basis in $\mathbf{U}^h$, we have

$$\mathbf{\Phi} = (\Phi_x, \Phi_y, \Phi_z)^T,$$

and since Φ_x vanishes outside the vertical slice S we have

$$\begin{aligned}\|\mathbf{\Phi}\|^2_{L^2(\Omega)^3} &= \|\Phi_x\|^2_{L^2(S)} + \|\Phi_y\|^2_{L^2(S)} + \|\Phi_z\|^2_{L^2(S)} \\ &\quad + \|\Phi_y\|^2_{L^2(\Omega\setminus S)} + \|\Phi_z\|^2_{L^2(\Omega\setminus S)}.\end{aligned}$$

Let us estimate the last term first. Noting that $\Phi_z(x, y, z)$ is a continuous function of x and vanishes for $x = 0$ and $x = 1$, we can write

$$\Phi_z(x, y, z) = \int_1^x \frac{\partial \Phi_z(x', y, z)}{\partial x} dx'.$$

After squaring both sides of the above identity, then using the Cauchy–Schwarz inequality on the right-hand side, and finally integrating both sides over $\Omega\setminus S$, we obtain

$$\|\Phi_z\|_{L^2(\Omega\setminus S)} \le C(\Omega)\, \|\mathbf{curl}\,\mathbf{\Phi}\|_{L^2(\Omega)^3},$$

where we have used the fact that on $\Omega\setminus S$ we have $\mathbf{curl}\,\mathbf{\Phi} = (*, -\frac{\partial \Phi_z}{\partial x}, \frac{\partial \Phi_y}{\partial x})^T$. Exactly in the same manner we get $\|\Phi_y\|_{L^2(\Omega\setminus S)} \le C(\Omega)\, \|\mathbf{curl}\,\mathbf{\Phi}\|_{L^2(\Omega)^3}$.

The next step of the proof will be estimating $\|\Phi_z\|_{L^2(S)}$ in terms of $\|\mathbf{curl}\,\mathbf{\Phi}\|_{L^2(\Omega)^3}$. Since Φ_z is a piecewise constant function in the z-direction, let us denote by $\Phi_z^k(x,y)$ the restriction of Φ_z to the kth horizontal slice of Ω. Note that Φ_z^k is linear in x within S and vanishes when $x=0$. Then

$$\begin{aligned}\|\Phi_z\|^2_{L^2(S)} &= \sum_{k=1}^n h\int_0^1\int_0^h \left[\Phi_z^k(x,y)\right]^2 dx\,dy\\ &\le \sum_{k=1}^n h\int_0^1\int_0^h \left[\Phi_z^k(h,y)\right]^2 dx\,dy\\ &= h^2\sum_{k=1}^n \int_0^1 \left[\Phi_z^k(h,y)\right]^2 dy \le h^2\,C(\Omega)\,\|\mathbf{curl}\,\mathbf{\Phi}\|^2_{L^2(\Omega)^3},\end{aligned}$$

where to obtain the last inequality we have again integrated $\frac{\partial\Phi_z}{\partial x}$ over $\Omega\backslash S$. The term $\|\Phi_y\|_{L^2(S)}$ is estimated in an analogous manner.

Finally, consider the identity on S,

$$\Phi_x(y,z) = \int_0^z \left[\frac{\partial\Phi_x(y,z')}{\partial z} - \frac{\partial\Phi_z(x,y,z')}{\partial x}\right] dz' + \int_0^z \frac{\partial\Phi_z(x,y,z')}{\partial x} dz'.$$

Again, after we square both sides, apply the Cauchy–Schwarz inequality on the right-hand side, integrate both sides over S, and note that $\frac{\partial\Phi_x}{\partial z} - \frac{\partial\Phi_z}{\partial x}$ is a component of $\mathbf{curl}\,\mathbf{\Phi}$, we get

$$\|\Phi_x\|^2_{L^2(S)} \le C(\Omega)\left\{\|\mathbf{curl}\,\mathbf{\Phi}\|^2_{L^2(S)^3} + \left\|\frac{\partial\Phi_z}{\partial x}\right\|^2_{L^2(S)}\right\}.$$

Now we complete the proof by applying an inverse inequality on the last term and using the estimate for $\|\Phi_z\|^2_{L^2(S)}$ that was obtained earlier. □

COROLLARY 4.2. *The linear system* (3.3) *to be solved in* $\mathbf{D}^h$ *has a symmetric and positive definite matrix with condition number of order* $O(h^{-2})$.

5. Convergence analysis. In this section, we provide a uniform upper bound for the condition number of the preconditioned operator $\mathbf{P}$ which indicates that the conjugate gradient iteration for problem (3.5) converges uniformly with respect to the mesh size h and the number of subdomains J. We also establish the uniform convergence of the multiplicative domain decomposition proposed in the second step of Algorithm 3.2. These convergence rates do depend on the factor ζ_1 in the minimum overlap $\zeta_1 H$, where H is the coarse-grid mesh size.

Here and henceforth, we shall use C with or without a subscript to denote a generic positive constant independent of the mesh size h and the number of subdomains J. The next lemma plays an essential role in estimating the minimum eigenvalue of the preconditioned operator $\mathbf{P}$.

LEMMA 5.1. *For any* $\mathbf{v}\in\mathbf{D}^h$, *there exists a decomposition of the form*

$$\mathbf{v} = \mathbf{v}_0 + \mathbf{v}_1 + \cdots + \mathbf{v}_J \quad \textit{with} \quad \mathbf{v}_j \in \mathbf{D}_j$$

and

$$\sum_{j=0}^J a(\mathbf{v}_j,\,\mathbf{v}_j) \le C\,a(\mathbf{v},\,\mathbf{v}), \tag{5.1}$$

where the positive constant C is independent of the mesh size h and the number of subdomains J (but depends on the factor ζ_1 in the minimum overlap $\zeta_1 H$).

Proof. For any $\mathbf{v} \in \mathbf{D}^h$, there exists a vector potential (cf. [7]) $\mathbf{\Phi} \in H^1(\Omega)^3$ such that $\mathbf{v} = \mathbf{curl}\,\mathbf{\Phi}$, $\mathbf{\Phi} \times \mathbf{n} = \mathbf{0}$ on $\partial\Omega$, and

$$\|\mathbf{\Phi}\|_{L^2(\Omega)^3} \le C\,\|\mathbf{curl}\,\mathbf{\Phi}\|_{L^2(\Omega)^3} \quad \text{and} \quad \|\nabla\mathbf{\Phi}\|_{L^2(\Omega)^3} \le C\,\|\mathbf{curl}\,\mathbf{\Phi}\|_{L^2(\Omega)^3}. \tag{5.2}$$

Let $\mathbf{U}^H$, associated with the coarse triangulation $\mathcal{T}^H$, be defined similarly as in the previous section and $\mathbf{Q}^H$ be the standard L^2 projection operator onto $\mathbf{U}^H$. Let $\mathbf{\Psi} = \mathbf{\Phi} - \mathbf{Q}^H\mathbf{\Phi}$; then it is easy to check (see [2]) that

$$\|\mathbf{\Psi}\|_{L^2(\Omega)^3} \le C\,H\|\nabla\,\mathbf{\Phi}\|_{L^2(\Omega)^3} \quad \text{and} \quad \|\mathbf{curl}\,(\mathbf{Q}^H\mathbf{\Phi})\|_{L^2(\Omega)^3} \le C\,\|\nabla\,\mathbf{\Phi}\|_{L^2(\Omega)^3}. \tag{5.3}$$

Define $\mathbf{v}_0 = \mathbf{curl}\,(\mathbf{Q}^H\mathbf{\Phi})$; then $\mathbf{v}_0 \in \mathbf{D}_0$. By using inequalities (5.3) and (5.2), we have that

$$\begin{aligned} \|\mathbf{v}_0\|_{L^2(\Omega)^3} &= \|\mathbf{curl}\,(\mathbf{Q}^H\mathbf{\Phi})\|_{L^2(\Omega)^3} \le C\,\|\nabla\,\mathbf{\Phi}\|_{L^2(\Omega)^3} \\ &\le C\,\|\mathbf{curl}\,\mathbf{\Phi}\|_{L^2(\Omega)^3} = C\,\|\mathbf{v}\|_{L^2(\Omega)^3}. \end{aligned} \tag{5.4}$$

Now, let $\theta_j \in C_0^\infty(\Omega_j)$, $j = 1, \ldots, J$, be a partition of unity such that

$$|\nabla\theta_j| \le C\,\zeta_1^{-1}\,H^{-1}, \tag{5.5}$$

and let

$$\mathbf{v}_j = \mathbf{\Pi}_h\mathbf{curl}(\theta_j\mathbf{\Psi}) \in \mathbf{D}_j.$$

Note that $\mathbf{v} = \mathbf{curl}\,\mathbf{\Phi} = \mathbf{v}_0 + \mathbf{curl}\,\mathbf{\Psi}$ and $\mathbf{\Pi}_h\mathbf{v} = \mathbf{v}$. Then linearity of $\mathbf{\Pi}_h$ and $\mathbf{curl}$ imply that $\mathbf{v}$ has a decomposition of the form

$$\mathbf{v} = \mathbf{v}_0 + \mathbf{v}_1 + \cdots + \mathbf{v}_J.$$

Since

$$\mathbf{curl}(\theta_j\mathbf{\Psi}) = \mathbf{\Psi} \times \nabla\theta_j + \theta_j\mathbf{curl}\,\mathbf{\Psi},$$

it follows from inequalities (2.3) and (2.10), the Cauchy–Schwarz inequality, and inequality (5.5) that for $j \in \{1, 2, \ldots, J\}$

$$\begin{aligned} a(\mathbf{v}_j,\, \mathbf{v}_j) &\le C\,\|\mathbf{v}_j\|^2_{L^2(\Omega)^3} \\ &\le C\,\|\mathbf{curl}(\theta_j\mathbf{\Psi})\|^2_{L^2(\Omega)^3} \\ &\le C\int_{\Omega_j}\left(|\nabla\theta_j|^2|\mathbf{\Psi}|^2 + \theta_j^2|\mathbf{curl}\,\mathbf{\Psi}|^2\right) \\ &\le C\,\zeta_1^{-2}\,H^{-2}\int_{\Omega_j}|\mathbf{\Psi}|^2 + C\int_{\Omega_j}|\mathbf{curl}\,\mathbf{\Psi}|^2. \end{aligned}$$

By summing the above inequality over j, it follows from the fact that the maximum number of subdomain overlaps is bounded, and from inequalities (5.2), (5.3), (5.4), and (2.3), that

$$\begin{aligned} \sum_{j=0}^{J} a(\mathbf{v}_j,\, \mathbf{v}_j) &\le C\,\|\mathbf{v}_0\|^2_{L^2(\Omega)^3} + C\,\zeta_1^{-2}\,H^{-2}\int_\Omega|\mathbf{\Psi}|^2 + C\int_\Omega|\mathbf{curl}\,\mathbf{\Psi}|^2 \\ &\le C\,\|\mathbf{v}_0\|^2_{L^2(\Omega)^3} + C\,\zeta_1^{-2}\,\|\mathbf{v}\|^2_{L^2(\Omega)^3} \\ &\le C\,(1+\zeta_1^{-2})\,\|\mathbf{v}\|^2_{L^2(\Omega)^3} \\ &\le C\,a(\mathbf{v},\, \mathbf{v}). \end{aligned}$$

This completes the proof of the lemma. □

Now, the standard argument provides the condition number estimate for $\mathbf{P}$.

THEOREM 5.1. *For any vector* $\mathbf{v} \in \mathbf{D}^h$, *we have*

$$C_1 a(\mathbf{v}, \mathbf{v}) \leq a(\mathbf{P}\mathbf{v}, \mathbf{v}) \leq C_2 a(\mathbf{v}, \mathbf{v}), \tag{5.6}$$

where the constants C_1 *and* C_2 *are independent of* h *and* J. *(*C_1 *contains the factor* $(1+\zeta_1^{-2})^{-1}$.*)*

Proof. The proof of the right-hand inequality follows from the boundedness of $\mathbf{P}_j$ and the maximum number of subdomain overlaps. The left-hand inequality follows from Lemma 5.1 and Lions' lemma [9]. □

Remark 5.1. In 2-D, a special Poincaré-type lemma (see [5, Lemma 3.1]), together with a bound of $\|\nabla\phi\|$ in terms of $\|\mathbf{curl}\,\phi\|$, allows an argument from Chapter 5 of [14] to prove a condition-number bound involving $1+\zeta_1^{-1}$ instead of $1+\zeta_1^{-2}$. It is not clear whether the analogous bound holds in 3-D.

To analyze the convergence of the multiplicative domain decomposition method defined at the second step in Algorithm 3.2, we note that for any $\mathbf{w} \in \mathbf{D}$ we have by the definition of the projection operators $\mathbf{P}_j$

$$a(\omega\mathbf{P}_j\mathbf{w}, \omega\mathbf{P}_j\mathbf{w}) = \omega\, a(\omega\mathbf{P}_j\mathbf{w}, \mathbf{w}). \tag{5.7}$$

And Lemma 5.1, the Cauchy–Schwarz inequality, and the bound on the number of subdomain overlaps give that

$$\begin{aligned} a(\mathbf{v}, \mathbf{v}) &= \sum_{j=0}^{J} a(\mathbf{v}, \mathbf{v}_j) = \sum_{j=0}^{J} a(\mathbf{P}_j\mathbf{v}, \mathbf{v}_j) \\ &\leq \left(\sum_{j=0}^{J} a(\mathbf{P}_j\mathbf{v}, \mathbf{P}_j\mathbf{v})\right)^{1/2} \left(\sum_{j=0}^{J} a(\mathbf{v}_j, \mathbf{v}_j)\right)^{1/2} \\ &\leq C \left(\sum_{j=0}^{J} a(\omega\mathbf{P}_j\mathbf{v}, \omega\mathbf{P}_j\mathbf{v})\right)^{1/2} a(\mathbf{v}, \mathbf{v})^{1/2}, \end{aligned}$$

which implies that

$$a(\mathbf{v}, \mathbf{v}) \leq C \sum_{j=0}^{J} a(\omega\mathbf{P}_j\mathbf{v}, \omega\mathbf{P}_j\mathbf{v}). \tag{5.8}$$

Hence, a straightforward consequence of [1] (see also Remark 2.2 in [3]) gives the following result.

THEOREM 5.2. *The iterative method defined at the second step in Algorithm* 3.2 *is uniformly convergent; i.e.,*

$$\|\mathbf{E}\|_a \leq \gamma < 1, \tag{5.9}$$

where γ *is a constant that does not depend on the number of subdomains and the mesh size. (*γ *does depend on* ζ_1.*)*

6. Numerical results. We briefly summarize some computations [15] that will be presented in more detail elsewhere. The additive preconditioner has been implemented and run on a variety of test problems. Corollary 4.2 was confirmed, as the smallest and largest eigenvalues of the system matrix varied as $O(h)$ and $O(h^{-1})$, respectively. For coarse grids ranging from $H = 1/4$ (thus $4 \times 4 \times 4$) to $H = 1/32$, with fine grids $h = H/4$ and overlaps $\zeta_1 H = h$, the iteration counts needed to reduce the preconditioned residual by 10 orders of magnitude were 31 to 32 for constant k (Poisson's equation), 31 for $k = 10^{-5}$ in $(1/4,\ 3/4)^3$ and $k = 1$ elsewhere, and 32 to 36 for $k = 10^{-5}$ in randomly-distributed coarse-grid blocks and $k = 1$ elsewhere. These results correspond to norm reductions of 0.47 (31 iterations) to 0.52 (36 iterations) per iteration. When random heterogeneity was combined with random anisotropy (k a diagonal tensor, three random entries of 10^{-5}, 10^{-4}, $\ldots$, 10^0 in each coarse-grid block), so that there was an increasing number of random blocks on finer grids, norm reductions were significantly worse (0.79 to 0.91) and worsened on finer grids. The theory of this paper does not address the dependence of iteration counts on jumps in coefficients, but it appears that this dependence is substantial only when heterogeneity and anisotropy are intertwined.

REFERENCES

[1] J. H. Bramble, J. E. Pasciak, J. Wang, and J. Xu, *Convergence estimates for product iterative methods with applications to domain decomposition and multigrid*, Math. Comp., 57 (1991), pp. 1–21.

[2] J. H. Bramble and J. Xu, *Some estimates for a weighted L^2 projection*, Math. Comp., 56 (1991), pp. 463–476.

[3] Z. Cai, *Norm estimates of product operators with application to domain decomposition*, Appl. Math. Comp., 53 (1993), pp. 251–276.

[4] L. C. Cowsar, J. Mandel, and M. F. Wheeler, *Balancing domain decomposition for mixed finite elements*, Math. Comp., 64 (1995), pp. 989–1015.

[5] M. Dryja and O. B. Widlund, *Domain decomposition algorithms with small overlap*, SIAM J. Sci. Comput., 15 (1994), pp. 604–620.

[6] R. E. Ewing and J. Wang, *Analysis of the Schwarz algorithm for mixed finite element methods*, RAIRO Math. Modél. Anal. Numér., 26 (1992), pp. 739–756.

[7] V. Girault and P. A. Raviart, *Finite Element Methods for Navier–Stokes Equations: Theory and Algorithms*, Springer-Verlag, New York, 1986.

[8] R. Glowinski and M. F. Wheeler, *Domain decomposition and mixed finite element methods for elliptic problems*, in Proceedings of the 1st International Symposium on Domain Decomposition Methods for Partial Differential Equations, R. Glowinski, G. H. Golub, G. A. Meurant, and J. Périaux, eds., SIAM, Philadelphia, 1987, pp. 144–172.

[9] P. L. Lions, *On the Schwarz alternating method* I, in Proceedings of the 1st International Symposium on Domain Decomposition Methods for Partial Differential Equations, R. Glowinski, G. H. Golub, G. A. Meurant, and J. Périaux, eds., SIAM, Philadelphia, 1987, pp. 1–42.

[10] T. F. Mathew, *Domain Decomposition and Iterative Refinement Methods for Mixed Finite Element Discretizations of Elliptic Problems*, Ph.D. Thesis, Courant Institute, New York, 1989.

[11] R. Mosé, P. Siegel, P. Ackerer, and G. Chavent, *Application of the mixed hybrid finite element approximation in a groundwater flow model: Luxury or necessity?*, Water Resources Res., 30 (1994), pp. 3001–3012.

[12] J. C. Nedelec, *Mixed finite elements in $\mathbb{R}^3$*, Numer. Math., 35 (1980), pp. 315–341.

[13] P. A. Raviart and J. M. Thomas, *A mixed finite element method for 2nd order elliptic problems*, in Mathematical Aspects of Finite Element Methods, Lecture Notes in Math. 606, I. Galligani and E. Magenes, eds., Springer-Verlag, New York, 1977, pp. 292–315.

[14] B. Smith, P. Bjørstad, and W. Gropp, *Domain Decomposition*, Cambridge University Press, London, Cambridge, 1996.

[15] J. D. Wilson, *Efficient Solver for Mixed and Control-Volume Mixed Finite Element Methods in Three Dimensions*, Ph.D. Thesis, University of Colorado at Denver, Denver, CO, 2001; also available online at http://www-math.cudenver.edu/graduate/thesis/jwthesis.ps.gz.

SIAM J. Numer. Anal.
Vol. 41, No. 1, pp. 195–207

WELL POSEDNESS OF THE INITIAL VALUE PROBLEM FOR VERTICALLY DISCRETIZED HYDROSTATIC EQUATIONS[*]

ANDREI BOURCHTEIN[†] AND VLADIMIR KADYCHNIKOV[‡]

Abstract. Vertically discretized linearized hydrostatic equations in hybrid coordinates are considered. The matrix of vertical structure, which depends on vertical discretization and determines the classification of the obtained system of time-dependent partial differential equations, is derived. The main theorem about oscillatory properties of the matrix of vertical structure is proved. This result ensures the well posedness of the initial value problem for vertically discretized primitive equations.

Key words. numerical weather prediction, initial value problems, well posedness, oscillatory matrices

AMS subject classifications. 65M99, 86A10, 15A48

PII. S0036142901398313

1. Introduction.

Semi-implicit schemes based on Eulerian and semi-Lagrangian treatments of advection have been intensively used in numerical weather forecast models in the last three decades. These two approaches have been developed primarily by Robert [31, 32] for barotropic equations and are currently the most popular techniques in all areas of atmospheric modeling. It is sufficient to mention some hydrostatic and nonhydrostatic atmospheric models with different areas of application (e.g., mesoscale, regional and global forecasting, climate simulation, and air quality monitoring) based on finite-difference, spectral, and finite-element numerical schemes and designed for operational forecasts as well as research simulations. Numerous bibliographic sources indicated in classical books [17, 21, 27] can be completed by abundant reference lists in recent papers [14, 28, 37].

Construction and solution of the implicit coupled equations connected to linear terms of the primitive equations are important parts of all semi-implicit schemes. This implicit part depends on vertical discretization and the choice of reference temperature profile. Due to its linearity, these discretized equations can be vertically decoupled, which reduces the three-dimensional space problem to the set of two-dimensional systems in the form of shallow water equations. The classification of these two-dimensional systems is determined by the matrix of the vertical structure. The properties of this matrix have been studied and discussed in the context of design of numerical weather prediction models [6, 8, 30, 15], forecast accuracy [4, 5], normal mode initialization [38, 39], and four-dimensional data assimilation [9, 12], among many other papers and applications.

Usually, the authors' analysis of the implicit linear equations (including stability and convergence analysis) is based on the supposition that this matrix has positive distinct eigenvalues, which guarantees a complete decoupling of the primitive linearized equations and the well posedness of initial value problems for all obtained shallow water systems [6, 8, 26, 30, 33]. Although this supposition has been confirmed by numerical computations in different models, it has not been demonstrated analytically

[*] Received by the editors November 16, 2001; accepted for publication (in revised form) October 3, 2002; published electronically March 19, 2003.
http://www.siam.org/journals/sinum/41-1/39831.html

[†] Rua Anchieta 4715 bloco K, Ap. 304, 96020-250 Pelotas, Brazil (burstein@terra.com.br).

[‡] Department of Mathematics, Pelotas State University, Campus Universitario da UFPel, 96010-900, Pelotas, Brazil (kadychnikov@terra.com.br).

for any numerical scheme. The goal of the present work is verifying the well posedness of the initial value problem for a family of linearized vertically discretized hydrostatic equations to provide additional theoretical justification for using the semi-implicit Eulerian and semi-Lagrangian schemes. Our analysis is for a large set of vertical grids, including, in particular, the optimal vertical grids constructed in [4, 5] to improve forecast accuracy.

This paper is divided into two main sections. In section 2, we present primitive hydrostatic equations, which we linearize and vertically discretize using one of the commonly used vertical grids. The matrix of vertical structure is derived, and its influence on vertical decoupling is discussed. Section 3 is devoted to the study of the properties of the matrix of vertical structure. The principal result consists of the proof of the oscillatory properties of the considered vertical matrix. Some concluding remarks are presented in the final section.

2. Deriving the matrix of vertical structure. We set out the continuous equations in Cartesian coordinates x, y of a conformal mapping projection and hybrid vertical coordinate $\eta(p, p_s)$, which is a monotonic function of the pressure p and also depends on the surface pressure p_s in such a way that

$$\eta\left(0, p_s\right) = 0, \qquad \eta\left(p_s, p_s\right) = 1.$$

Following Kasahara [23] and Simmons and Burridge [36], the primitive hydrostatic equations can be written as follows:

horizontal momentum equations

$$\frac{du}{dt} = -\frac{u^2+v^2}{2} m_x^2 + fv - RTP_x - \Phi_x, \tag{1}$$

$$\frac{dv}{dt} = -\frac{u^2+v^2}{2} m_y^2 - fu - RTP_y - \Phi_y, \tag{2}$$

hydrostatic equation

$$\Phi_\eta = -RTP_\eta, \tag{3}$$

thermodynamic equation

$$\frac{dT}{dt} = \frac{RT}{c_p}\frac{dP}{dt}, \tag{4}$$

continuity equation

$$\frac{dp_\eta}{dt} = -p_\eta m^2 \left(u_x + v_y\right) - p_\eta \dot{\eta}_\eta, \tag{5}$$

pressure equation

$$P_t + \dot{\eta} P_\eta = -\frac{m^2}{p}\int_0^\eta \left(up_\eta\right)_x + \left(vp_\eta\right)_y d\eta, \tag{6}$$

surface pressure equation

$$\left(P_s\right)_t = -\frac{m^2}{p_s}\int_0^1 \left(up_\eta\right)_x + \left(vp_\eta\right)_y d\eta. \tag{7}$$

The last two equations are the consequence of the continuity equation and the upper and lower boundary conditions

$$\dot{\eta} = 0 \quad \text{at} \quad \eta = 0; 1.$$

Here

$$\frac{d\varphi}{dt} = \varphi_t + m^2 \left(u\varphi_x + v\varphi_y \right) + \dot{\eta}\varphi_\eta$$

is the three-dimensional individual derivative, u and v are the horizontal projection velocity components, and the following common denotations are used: t is the time, $P = \ln p$, $P_s = \ln p_s$, m is the mapping factor, f is the Coriolis parameter, T is the temperature, $\Phi = gz$ is the geopotential, z is the height, g is the gravitational acceleration, R is the gas constant, c_p is the specific heat at constant pressure, and the subscripts t, x, y, η denote the partial derivatives with respect to the indicated variable. (Note that other subscripts do not mean differentiation.)

The first step to obtain the vertical structure matrix is linearization of this system about a state of rest:

$$u_t = fv - RT_0 P_x - \Phi_x, \tag{8}$$

$$v_t = -fu - RT_0 P_y - \Phi_y, \tag{9}$$

$$\Phi_\eta = -RT_0 P_\eta - RTS_\eta, \tag{10}$$

$$T_t = \frac{RT_0}{c_p} \cdot \left(P_t + S_\eta \dot{\eta} \right), \tag{11}$$

$$P_t + S_\eta \dot{\eta} = -\frac{1}{\sigma} \int_0^\eta \sigma_\eta D d\eta, \tag{12}$$

$$\left(P_s \right)_t = -\frac{1}{\sigma_s} \int_0^1 \sigma_\eta D d\eta. \tag{13}$$

Here $T_0 =$ const is the reference temperature profile, $p_0 = \sigma(\eta)$ is the reference pressure profile ($\sigma(\eta)$ is the positive strictly increasing function), $\sigma_s = \sigma(1)$, $S = \ln \sigma$, and $D = m^2(u_x + v_y)$ is the horizontal divergence.

Substituting the right-hand side of (12) in (11) and introducing a new unknown function $G = \Phi + RT_0 P$, the system (8)–(11), (13) can be written in the following form:

$$u_t = fv - G_x, \tag{14}$$

$$v_t = -fu - G_y, \tag{15}$$

$$G_\eta = -RTS_\eta, \tag{16}$$

$$T_t = -\frac{RT_0}{c_p \sigma} \int_0^\eta \sigma_\eta D d\eta, \tag{17}$$

$$(G_s)_t = -\frac{RT_0}{\sigma_s}\int_0^1 \sigma_\eta D d\eta. \tag{18}$$

Equation (12), which can be considered an equation for the vertical velocity component, is decoupled from the last system and has no effect upon the linear analysis of the primitive equations. Note that the analogous system (with or without Coriolis terms) can be obtained for the implicit part of semi-implicit Eulerian or semi-Lagrangian schemes [6, 8, 26, 30, 33, 34].

Of course, studying the continuous problem one can eliminate T from system (14)–(18) and, seeking separable solutions, can obtain the differential vertical structure equation [13, 18]. However, to construct the vertical structure matrix used in numerical schemes we have to follow the procedure of vertical discretization used in numerical models. Thus, the second step consists of choosing the type of vertical grid and vertical discretization. We choose a rather popular Lorenz vertical grid which carries horizontal velocities, temperature, and geopotential at the same model levels that represent model layers, while vertical velocity is carried at the interfaces of these layers (see Figure 1).

This grid divides the considered atmosphere in K vertical layers. The boundaries $\eta_{k+1/2}$, $k = 0, \ldots, K$ (solid lines) of these layers can be chosen arbitrarily, and the inner levels η_k, $k = 1, \ldots, K$ (dashed lines) satisfy the natural inequalities

$$\eta_{k-1/2} < \eta_k < \eta_{k+1/2}, \qquad k = 1, \ldots, K. \tag{19}$$

In [6, 8, 11], the values η_k are defined as midlines of the layers; that is,

$$\eta_k = \frac{\eta_{k+1/2} + \eta_{k-1/2}}{2}, \qquad k = 1, \ldots, K,$$

but it is not essential for our analysis. The restrictions $\eta_{1/2} = 0$, $\eta_{K+1/2} = 1$, and (19) are uniquely used in subsequent reasoning. Therefore, the vertical grid construction is sufficiently flexible, and it involves a wide set of grids.

This type of grid was introduced by Lorenz [25] to construct a vertically discrete balanced model, which conserves the total energy, the mean potential temperature, and the variance of the potential temperature under adiabatic and frictionless processes. Following Lorenz's approach, Arakawa and Lamb [1] and Arakawa and Suarez [2] constructed vertically discrete models based on the Lorenz grid satisfying the most integral constraints of the continuous system. Thus, the distribution of variables on the Lorenz grid is more straightforward for keeping conservation properties of the primitive equations. For this reason, it seems to be the most widespread vertical grid used for hydrostatic equation models [3]. We can mention the NASA numerical weather prediction and climate models [6, 11, 35], NCEP global and mesoscale models [7, 22], and the Pennsylvania State University MM5 model [16], among others, which use this vertical approximation.

All the variables u, v, T, G of the linearized system (14)–(18) are defined at the levels η_k. Since variable $\dot{\eta}$ is not used in linear analysis, the Lorenz grid coincides with the completely unstaggered vertical grid. Therefore, our analysis is also applied to this last grid used, for example, in [24]. Note that the reference pressure $\sigma(\eta)$ is a given function, and then it can be calculated at both boundaries and inner levels.

The approximation of the hydrostatic equation (16) is realized by formulas

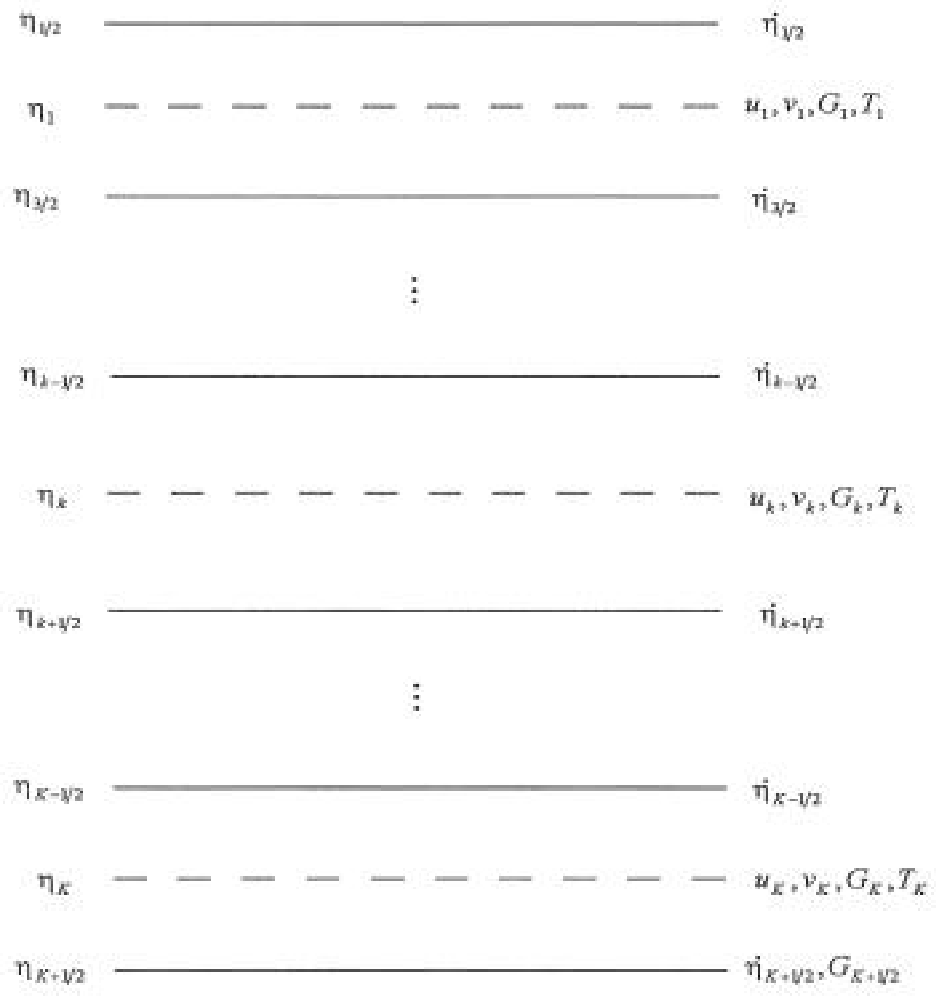


FIG. 1. *K layer Lorenz vertical grid.*

$$G_k = G_{K+1} + R\left[\sum_{i=k+1}^{K} \ln\frac{\sigma_{i+1/2}}{\sigma_{i-1/2}}\cdot T_i + \ln\frac{\sigma_{k+1/2}}{\sigma_k}\cdot T_k\right], \qquad k = 1,\dots,K,$$

$$G_{K+1} \equiv G_s.$$

The thermodynamic and continuity equations (17)–(18) are approximated as

$$T_{k_t} = -\frac{RT_0}{c_p}\cdot\left[\sum_{i=1}^{k-1}\frac{\sigma_{i+1/2}-\sigma_{i-1/2}}{\sigma_k}D_i + \frac{\sigma_{k+1/2}-\sigma_{k-1/2}}{2\sigma_k}D_k\right], \qquad k = 1,\dots,K,$$

$$(G_s)_t = -\frac{RT_0}{\sigma_{K+1/2}}\sum_{i=1}^{K}(\sigma_{i+1/2}-\sigma_{i-1/2})\,D_i.$$

In the above, a summation is defined to be zero if the lower limit of the summation index exceeds the upper limit.

Introducing vector columns

$$\mathbf{u}=(u_1,\dots,u_K)^T,\mathbf{v}=(v_1,\dots,v_K)^T,\mathbf{D}=(D_1,\dots,D_K)^T,\mathbf{G}=(G_1,\dots,G_K)^T,$$

$$\mathbf{T}=(T_1,\dots,T_K,G_s)^T,$$

we can write discrete analogues of (14)–(18) in the form

$$\mathbf{u}_t=f\mathbf{v}-\mathbf{G}_x, \tag{20}$$

$$\mathbf{v}_t=-f\mathbf{u}-\mathbf{G}_y, \tag{21}$$

$$\mathbf{G}=R\mathbf{A}_T\cdot\mathbf{T}, \tag{22}$$

$$\mathbf{T}_t=-\frac{RT_0}{c_p}\mathbf{A}_D\cdot\mathbf{D}. \tag{23}$$

Here $\mathbf{A}_T$ and $\mathbf{A}_D$ are the $K\times(K+1)$ and $(K+1)\times K$ matrices, respectively, determined as follows:

$$\mathbf{A}_T=\begin{pmatrix}
\ln\frac{\sigma_{3/2}}{\sigma_1} & \ln\frac{\sigma_{5/2}}{\sigma_{3/2}} & \cdots & \ln\frac{\sigma_{k+1/2}}{\sigma_{k-1/2}} & \cdots & \ln\frac{\sigma_{K-1/2}}{\sigma_{K-3/2}} & \ln\frac{\sigma_{K+1/2}}{\sigma_{K-1/2}} & 1\\
0 & \ln\frac{\sigma_{5/2}}{\sigma_2} & \cdots & \ln\frac{\sigma_{k+1/2}}{\sigma_{k-1/2}} & \cdots & \ln\frac{\sigma_{K-1/2}}{\sigma_{K-3/2}} & \ln\frac{\sigma_{K+1/2}}{\sigma_{K-1/2}} & 1\\
\vdots & \vdots & \ddots & \vdots & & \vdots & \vdots & \vdots\\
0 & 0 & \cdots & \ln\frac{\sigma_{k+1/2}}{\sigma_k} & \cdots & \ln\frac{\sigma_{K-1/2}}{\sigma_{K-3/2}} & \ln\frac{\sigma_{K+1/2}}{\sigma_{K-1/2}} & 1\\
\vdots & \vdots & & \vdots & \ddots & \vdots & \vdots & \vdots\\
0 & 0 & \cdots & 0 & \cdots & \ln\frac{\sigma_{K-1/2}}{\sigma_{K-1}} & \ln\frac{\sigma_{K+1/2}}{\sigma_{K-1/2}} & 1\\
0 & 0 & \cdots & 0 & \cdots & 0 & \ln\frac{\sigma_{K+1/2}}{\sigma_K} & 1
\end{pmatrix}$$

and

$$\mathbf{A}_D=$$

$$\begin{pmatrix}
\frac{\sigma_{3/2}-\sigma_{1/2}}{2\sigma_1} & 0 & \cdots & 0 & \cdots & 0 & 0\\
\frac{\sigma_{3/2}-\sigma_{1/2}}{\sigma_2} & \frac{\sigma_{5/2}-\sigma_{3/2}}{2\sigma_2} & \cdots & 0 & \cdots & 0 & 0\\
\vdots & \vdots & \ddots & \vdots & & \vdots & \vdots\\
\frac{\sigma_{3/2}-\sigma_{1/2}}{\sigma_k} & \frac{\sigma_{5/2}-\sigma_{3/2}}{\sigma_k} & \cdots & \frac{\sigma_{k+1/2}-\sigma_{k-1/2}}{2\sigma_k} & \cdots & 0 & 0\\
\vdots & \vdots & & \vdots & \ddots & \vdots & \vdots\\
\frac{\sigma_{3/2}-\sigma_{1/2}}{\sigma_{K-1}} & \frac{\sigma_{5/2}-\sigma_{3/2}}{\sigma_{K-1}} & \cdots & \frac{\sigma_{k+1/2}-\sigma_{k-1/2}}{\sigma_{K-1}} & \cdots & \frac{\sigma_{K-1/2}-\sigma_{K-3/2}}{2\sigma_{K-1}} & 0\\
\frac{\sigma_{3/2}-\sigma_{1/2}}{\sigma_K} & \frac{\sigma_{5/2}-\sigma_{3/2}}{\sigma_K} & \cdots & \frac{\sigma_{k+1/2}-\sigma_{k-1/2}}{\sigma_K} & \cdots & \frac{\sigma_{K-1/2}-\sigma_{K-3/2}}{\sigma_K} & \frac{\sigma_{K+1/2}-\sigma_{K-1/2}}{2\sigma_K}\\
\frac{\sigma_{3/2}-\sigma_{1/2}}{\sigma_{K+1/2}/c_p} & \frac{\sigma_{5/2}-\sigma_{3/2}}{\sigma_{K+1/2}/c_p} & \cdots & \frac{\sigma_{k+1/2}-\sigma_{k-1/2}}{\sigma_{K+1/2}/c_p} & \cdots & \frac{\sigma_{K-1/2}-\sigma_{K-3/2}}{\sigma_{K+1/2}/c_p} & \frac{\sigma_{K+1/2}-\sigma_{K-1/2}}{\sigma_{K+1/2}/c_p}
\end{pmatrix}.$$

Finally, differentiating hydrostatic equations with respect to t and eliminating $\mathbf{T}$, we obtain

$$\mathbf{G}_t = R\mathbf{A}_T T_t = R\mathbf{A}_T\left(-\frac{RT_0}{c_p}\mathbf{A}_D\mathbf{D}\right) = -\frac{R}{c_p}RT_0\mathbf{A}_T\mathbf{A}_D\mathbf{D} = -\frac{R}{c_p}RT_0\mathbf{A}\mathbf{D},$$

where $\mathbf{A} = \mathbf{A}_T\mathbf{A}_D$ is the $K \times K$ matrix of the vertical structure. If all eigenvalues of matrix $\mathbf{A}$ are positive and respective eigenvectors are linearly independent (that is, these form a basis in the space R^K), then the system

$$\mathbf{u}_t = f\mathbf{v} - \mathbf{G}_x, \tag{24}$$

$$\mathbf{v}_t = -f\mathbf{u} - \mathbf{G}_y, \tag{25}$$

$$\mathbf{G}_t = -\frac{R}{c_p}RT_0\mathbf{A}\mathbf{D} \tag{26}$$

can be vertically decoupled in the K linearized shallow water systems with different gravitational wave speeds. To make this, one can represent the vertical structure matrix $\mathbf{A}$ in the spectral form $\mathbf{A} = \mathbf{S}\Lambda\mathbf{S}^{-1}$, where Λ is the diagonal eigenvalue matrix and $\mathbf{S}$ is the matrix of eigenvectors (that is, vertical normal modes) of the $\mathbf{A}$, and, multiplying (24)–(26) on the left by $\mathbf{S}^{-1}$, one can obtain

$$\tilde{\mathbf{u}}_t = f\tilde{\mathbf{v}} - \tilde{\mathbf{G}}_x, \tag{27}$$

$$\tilde{\mathbf{v}}_t = -f\tilde{\mathbf{u}} - \tilde{\mathbf{G}}_y, \tag{28}$$

$$\tilde{\mathbf{G}}_t = -\frac{R}{c_p}RT_0\mathbf{\Lambda}\tilde{\mathbf{D}}, \tag{29}$$

where coefficients of vertical mode expansion are introduced by formula

$$\tilde{\varphi} = \mathbf{S}^{-1}\varphi, \qquad \varphi = \mathbf{u}, \mathbf{v}, \mathbf{G}.$$

If at least one of the eigenvalues of matrix $\mathbf{A}$ is negative, then the corresponding shallow water system in (27)–(29) is elliptic, and the initial value problem is not well posed for this system, which results in an incorrect problem for system (20)–(23) with any initial conditions. This leads to an ill posed initial value problem for vertically discretized primitive equations (1)–(7), and, therefore, their numerical solution has no meaning.

Thus, the properties of vertical structure matrix $\mathbf{A}$ are crucial for the well posedness of vertically discretized equations (20)–(23), which is a necessary condition for the well posedness of the numerical scheme. In the following section, we verify some properties of matrix $\mathbf{A}$, which guarantee the positivity of all its eigenvalues and the completeness of the set of eigenvectors.

3. Oscillatory properties of matrix A. A matrix is called totally positive if all its minors are nonnegative. A strictly totally positive matrix is a matrix whose minors are all positive. A $K \times K$ matrix is called oscillatory if it is totally positive and its certain power is strictly totally positive.

We use the following criterion for oscillatory matrices given in [19].

Theorem G1. *Matrix* $\mathbf{A}$ *is oscillatory if*

(1) $\mathbf{A}$ *is totally positive;*

(2) $\det \mathbf{A} > 0$*;*

(3) $a_{ij} > 0$ *for* $|i-j| \le 1$*; that is, all the entries on the principal diagonal, first superdiagonal, and first subdiagonal are positive.*

The following fundamental theorem is true for oscillatory matrices [19].

Theorem G2. *An oscillatory* $K \times K$ *matrix* $\mathbf{A}$ *has* K *distinct positive eigenvalues* $\lambda_1 > \lambda_2 > \cdots > \lambda_K > 0$*, and eigenvector* $\mathbf{s}_k (k = 1, \ldots, K)$*, which corresponds to eigenvalue* λ_k*, has* $k-1$ *variations of sign.*

Thus, if the oscillatoriness of matrix $\mathbf{A}$ is demonstrated, it will guarantee the well posedness of (20)–(23). According to Theorem G1, we have to show three properties of matrix $\mathbf{A}$ to guarantee its oscillatoriness. We start with the study of the total positivity of this matrix.

A direct consequence of the well-known Cauchy–Binet identity for determinants is that the product of totally positive matrices is again a totally positive matrix. Thus, to show the total positivity of matrix $\mathbf{A} = \mathbf{A}_T \mathbf{A}_D$ it is sufficient to prove that each matrix $\mathbf{A}_T$ and $\mathbf{A}_D$ is totally positive.

Let us consider matrix $\mathbf{A}_T$. We can factor it into the form

$$\mathbf{A}_T = \mathbf{B}\mathbf{C},$$

where $\mathbf{B}$ is a $K \times (K+1)$ matrix

$$\mathbf{B} = \begin{pmatrix} b_1 & 1 & \cdots & 1 & \cdots & 1 & 1 & 1 \\ 0 & b_2 & \cdots & 1 & \cdots & 1 & 1 & 1 \\ \vdots & \vdots & \ddots & \vdots & & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & b_k & \cdots & 1 & 1 & 1 \\ \vdots & \vdots & & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & \cdots & b_{K-1} & 1 & 1 \\ 0 & 0 & \cdots & 0 & \cdots & 0 & b_K & 1 \end{pmatrix}$$

with elements

$$b_1 = \ln \frac{\sigma_{3/2}}{\sigma_1}, \quad b_k = \frac{\ln(\sigma_{k+1/2}/\sigma_k)}{\ln(\sigma_{k+1/2}/\sigma_{k-1/2})}, \quad k = 2, \ldots, K$$

and $\mathbf{C}$ is a $(K+1) \times (K+1)$ diagonal matrix $\mathbf{C} = \mathbf{diag}[1, c_2, \ldots, c_K, 1]$ with elements

$$c_k = \ln \frac{\sigma_{k+1/2}}{\sigma_{k-1/2}}, \qquad k = 2, \ldots, K.$$

Since (19) holds and $\sigma(\eta)$ is a positive strictly increasing function, the elements b_k, $k = 1, \ldots, K$, and c_k, $k = 2, \ldots, K$, are positive and $b_k < 1$ for $k = 2, \ldots, K$.

Obviously, matrix $\mathbf{C}$ is totally positive. The principal property of matrix $\mathbf{B}$ is established in the following theorem. Although this result can be derived from the test for total positivity provided in [20], for the sake of completeness we include a direct proof.

Theorem 1. *Matrix* $\mathbf{B}$ *is totally positive.*

Proof. To prove this result we use the determinantal criterion given in [10] and [29].

Criterion C1. $\mathbf{A}$ is an $m \times n$ matrix, $n \geq m$, such that $\det \mathbf{A}[1,\dots,k|1,\dots,k] > 0$ for all $1 \leq k \leq m$. Then $\mathbf{A}$ is totally positive if and only if all minors with initial consecutive columns or initial consecutive rows are nonnegative.

Hereinafter, $\mathbf{A}[\alpha|\beta]$ denotes the $k \times l$ submatrix of $m \times n$ matrix $\mathbf{A}$ ($k \leq m$ and $l \leq n$), containing rows numbered by $\alpha = [\alpha_1,\dots,\alpha_k]$ and columns numbered by $\beta = [\beta_1,\dots,\beta_l]$, where α and β are increasing sequences of k and l natural numbers α_i, $i = 1,\dots,k$, and β_j, $j = 1,\dots,l$, respectively.

First, note that $\det \mathbf{B}[1,\dots,k|1,\dots,k] = b_1 \times \cdots \times b_k > 0$ for any $k = 1,\dots,K$. Second, $\det \mathbf{B}[\alpha|1,\dots,k] = 0$ for any $k = 1,\dots,K$ and $\alpha \neq [1,\dots,k]$. Therefore, we can simplify Criterion C1 for matrix $\mathbf{B}$ to the following form.

Criterion C2. $\mathbf{B}$ is totally positive if and only if all minors with initial consecutive rows are nonnegative; that is, $\det \mathbf{B}[1,\dots,k|\beta] \geq 0$ for any k and for any increasing set of k indices $\beta \neq [1,\dots,k]$.

Let us consider two cases.

(1) If $\beta_{k-1} > k$, then the columns with indices β_{k-1} and β_k contain all their entries above the first principal diagonal (above the diagonal with entries $b_1,\dots,b_K$). Hence, these two columns coincide and $\det \mathbf{B}[1,\dots,k|\beta] = 0$.

(2) If $\beta_{k-1} \leq k$, then, taking into account that $\beta_k > k$, we have

$$\det \mathbf{B}[1,\dots,k|\beta] = \begin{vmatrix} b_{1,\beta_1} & \dots & b_{1,\beta_{k-1}} & b_{1,\beta_k} \\ \vdots & \ddots & \vdots & \vdots \\ b_{k-1,\beta_1} & \dots & b_{k-1,\beta_{k-1}} & b_{k-1,\beta_k} \\ b_{k,\beta_1} & \dots & b_{k,\beta_{k-1}} & b_{k,\beta_k} \end{vmatrix}$$

$$= \begin{vmatrix} b_{1,\beta_1} & \dots & b_{1,\beta_{k-1}} & 1 \\ \vdots & \ddots & \vdots & \vdots \\ b_{k-1,\beta_1} & \dots & b_{k-1,\beta_{k-1}} & 1 \\ b_{k,\beta_1} & \dots & b_{k,\beta_{k-1}} & 1 \end{vmatrix} = \begin{vmatrix} b_{1,\beta_1} - b_{k,\beta_1} & \dots & b_{1,\beta_{k-1}} - b_{k,\beta_{k-1}} & 0 \\ \vdots & \ddots & \vdots & \vdots \\ b_{k-1,\beta_1} - b_{k,\beta_1} & \dots & b_{k-1,\beta_{k-1}} - b_{k,\beta_{k-1}} & 0 \\ b_{k,\beta_1} & \dots & b_{k,\beta_{k-1}} & 1 \end{vmatrix}. \tag{30}$$

Expanding in cofactors along the last column and observing that $b_{k,\beta_i} = 0$ for all $i = 1,\dots,k-2$ because $\beta_i < \beta_{k-1} \leq k$, we obtain the $k-1$ order determinant

$$\det \mathbf{B}[1,\dots,k|\beta] = \begin{vmatrix} b_{1,\beta_1} & \dots & b_{1,\beta_{k-2}} & b_{1,\beta_{k-1}} - b_{k,\beta_{k-1}} \\ \vdots & \ddots & \vdots & \vdots \\ b_{k-2,\beta_1} & \dots & b_{k-2,\beta_{k-2}} & b_{k-2,\beta_{k-1}} - b_{k,\beta_{k-1}} \\ b_{k-1,\beta_1} & \dots & b_{k-1,\beta_{k-2}} & b_{k-1,\beta_{k-1}} - b_{k,\beta_{k-1}} \end{vmatrix}. \tag{31}$$

Now two situations can occur. The first, when $\beta_{k-1} = k-1$, is the simplest. In this case, $\beta_i = i$ for $i = 1,\dots,k-2$, $b_{k,\beta_{k-1}} = 0$, and, consequently,

$$\det \mathbf{B}[1,\dots,k|\beta] = \det \mathbf{B}[1,\dots,k-1|1,\dots,k-1] = b_1 \times \cdots \times b_{k-1} > 0.$$

In the second case, $\beta_{k-1} = k$ and $b_{k,\beta_{k-1}} = b_{k,k} = b_k$, $b_{k-1,\beta_{k-1}} = b_{k-1,k} = 1$. Therefore, we simplify (31) to

$$\det \mathbf{B}[1,\dots,k|\beta] = \begin{vmatrix} b_{1,\beta_1} & \dots & b_{1,\beta_{k-2}} & 1 - b_k \\ \vdots & \ddots & \vdots & \vdots \\ b_{k-2,\beta_1} & \dots & b_{k-2,\beta_{k-2}} & 1 - b_k \\ b_{k-1,\beta_1} & \dots & b_{k-1,\beta_{k-2}} & 1 - b_k \end{vmatrix}, \tag{32}$$

and we continue to reduce the order of determinant

$$(33)\qquad = (1-b_k)\begin{vmatrix} b_{1,\beta_1} & \dots & b_{1,\beta_{k-3}} & b_{1,\beta_{k-2}} - b_{k-1,\beta_{k-2}} \\ \vdots & \ddots & \vdots & \vdots \\ b_{k-3,\beta_1} & \dots & b_{k-3,\beta_{k-3}} & b_{k-3,\beta_{k-2}} - b_{k-1,\beta_{k-2}} \\ b_{k-2,\beta_1} & \dots & b_{k-2,\beta_{k-3}} & b_{k-2,\beta_{k-2}} - b_{k-1,\beta_{k-2}} \end{vmatrix}.$$

Obtained determinant (33) is similar to (31), but it is one order less. Again we have two choices. If $\beta_{k-2} = k-2$, then $\beta_i = i$ for $i = 1, \dots, k-3$ and $b_{k-1,\beta_{k-2}} = 0$, which gives the following result:

$$\begin{aligned} \det \mathbf{B}[1,\dots,k|\beta] &= (1-b_k)\det \mathbf{B}[1,\dots,k-2|1,\dots,k-2] \\ &= b_1 \times \cdots \times b_{k-2} \times (1-b_k) > 0. \end{aligned}$$

In the opposite case, $\beta_{k-2} = k-1$, $b_{k-1,\beta_{k-2}} = b_{k-1}$, $b_{k-2,\beta_{k-2}} = 1$, and we get the determinant in the form (32), but its order is one less than the order of (32).

We can continue this procedure until a determinant of the first order is obtained. The final result of the calculus of the determinant (30) can be represented in the form

$$\det \mathbf{B}[1,\dots,k|\beta] = d_1 \times \cdots \times d_{k-1},$$

where

$$d_i = b_i \text{ or } (1-b_{i+1}), \qquad i = 1, \dots, k-1,$$

which depends on the choice of the index set β. In either case, all $d_i > 0$, and, consequently,

$$(34)\qquad \det \mathbf{B}[1,\dots,k|\beta] = d_1 \times \cdots \times d_{k-1} > 0.$$

Since we consider the arbitrary value of k, Criterion C2 is proved. Therefore, matrix **B** is totally positive and Theorem 1 is proved.

Hence, matrix $\mathbf{A}_T$ is totally positive as the product of two totally positive matrices **B** and **C**.

To prove the total positivity of matrix $\mathbf{A}_D$ we use factorization in the form

$$\mathbf{A}_D = \mathbf{FGH},$$

where **G** is $(K+1) \times K$ matrix

$$\mathbf{G} = \begin{pmatrix} 1/2 & 0 & \cdots & 0 & \cdots & 0 & 0 \\ 1 & 1/2 & \cdots & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & & \vdots & \vdots \\ 1 & 1 & \cdots & 1/2 & \cdots & 0 & 0 \\ \vdots & \vdots & & \vdots & \ddots & \vdots & \vdots \\ 1 & 1 & \cdots & 1 & \cdots & 1/2 & 0 \\ 1 & 1 & \cdots & 1 & \cdots & 1 & 1/2 \\ 1 & 1 & \cdots & 1 & \cdots & 1 & 1 \end{pmatrix},$$

F is $(K+1) \times (K+1)$ diagonal matrix $\mathbf{F} = \mathbf{diag}[f_1, \dots, f_K, f_{K+1}]$ with elements

$$f_k = \frac{1}{\sigma_k}, \quad k = 1, \dots, K, \quad f_{K+1} = \frac{c_p}{\sigma_{K+1/2}},$$

and $\mathbf{H}$ is $K \times K$ diagonal matrix $\mathbf{H} = \mathbf{diag}[h_1, \ldots, h_K]$ with elements

$$h_k = \sigma_{k+1/2} - \sigma_{k-1/2}, \qquad k = 1, \ldots, K.$$

It is evident that matrices $\mathbf{F}$ and $\mathbf{H}$ are totally positive. Matrix $\mathbf{G}$ is totally positive because it is a particular case of matrix $\mathbf{B}^T$. Therefore, matrix $\mathbf{A}_D$ is totally positive. Finally, due to the total positivity of matrices $\mathbf{A}_T$ and $\mathbf{A}_D$, matrix $\mathbf{A} = \mathbf{A}_T\mathbf{A}_D$ is totally positive too.

It remains to verify the second and third conditions of Theorem G1. To show the second condition we observe that evaluation (34) considered for $k = K$ means that all K order minors of matrix $\mathbf{B}$ are positive. Obviously, all K order minors of matrix $\mathbf{C}$ are nonnegative and some of these are positive. Therefore, by Cauchy–Binet formulas, all K order minors of matrix $\mathbf{A}_T$ are positive. Analogously, we can show that all K order minors of matrix $\mathbf{A}_D$ are positive. Applying the Cauchy–Binet identity to the determinant of $K \times K$ matrix $\mathbf{A}$ we demonstrate its positivity. Finally, the positivity of the elements $a_{ij} > 0$ for $|i - j| \leq 1$ follows directly from a structure of matrices $\mathbf{A}_T$ and $\mathbf{A}_D$.

Thus, all three conditions of Theorem G1 are satisfied, and, therefore, matrix $\mathbf{A}$ is oscillatory. According to Theorem G2, this implies that all eigenvalues are positive and distinct, which guarantees the strict hyperbolicity of all shallow water systems (27)–(29). Therefore, the initial value problem is well posed for discretized equations (20)–(23).

4. Conclusion. An analytical study of the vertical structure matrix of discretized hydrostatic equations is performed. The principal statement about the oscillatoriness of this matrix is proved using the determinantal criterion for total positivity and the oscillatory matrix test. The properties of the oscillatory matrix ensure that all the decoupled linearized shallow water equations are strictly hyperbolic. Therefore, the initial value problem is well posed for linearized primitive equations, which is the necessary condition for the convergence of semi-implicit schemes.

REFERENCES

[1] A. Arakawa and V. Lamb, *Computational design of the basic dynamical processes of the UCLA general circulation model*, Methods Comput. Phys. 17, J. Chang, ed., Academic Press, New York, 1977, pp. 173–265.

[2] A. Arakawa and M.J. Suarez, *Vertical differencing of the primitive equations in sigma coordinates*, Mon. Wea. Rev., 111 (1983), pp. 34–45.

[3] A. Arakawa and C.S. Konor, *Vertical differencing of the primitive equations based on the Charney-Phillips grid in hybrid $\sigma - p$ vertical coordinates*, Mon. Wea. Rev., 124 (1996), pp. 511–528.

[4] F. Baer and M. Ji, *Optimal vertical discretization for atmospheric models*, Mon. Wea. Rev., 117 (1989), pp. 391–406.

[5] F. Baer and Y. Zhu, *Forecast accuracy with optimum vertical model truncation*, Mon. Wea. Rev., 120 (1992), pp. 2579–2591.

[6] J.R. Bates, S. Moorthi, and R.W. Higgins, *A global multilevel atmospheric model using a vector semi-Lagrangian finite-difference scheme. Part* I*: Adiabatic formulation*, Mon. Wea. Rev., 121 (1993), pp. 244–263.

[7] T.L. Black, *The new NMC mesoscale ETA model: Description and forecast examples*, Wea. Forecasting, 9 (1994), pp. 265–278.

[8] A. Bourchtein, *Semi-Lagrangian semi-implicit space splitting regional baroclinic atmospheric model*, Appl. Numer. Math., 40 (2002), pp. 307–326.

[9] G. Brunet, *Empirical normal-mode analysis of atmospheric data*, J. Atmospheric. Sci., 51 (1994), pp. 932–952.

[10] J.M. Carnicer, J.M. Pena, and R.A. Zalik, *Strictly totally positive systems*, J. Approx. Theory, 92 (1998), pp. 411–441.

[11] M. Chen and J.R. Bates, *Forecast experiments with a global finite-difference semi-Lagrangian model*, Mon. Wea. Rev., 124 (1996), pp. 1992–2007.

[12] S.E. Cohn and D.P. Dee, *Observability of discretized partial differential equations*, SIAM J. Numer. Anal., 25 (1988), pp. 586–617.

[13] S.E. Cohn and D.P. Dee, *An analysis of the vertical structure equation for arbitrary thermal profiles*, Q. J. R. Meteor. Soc., 115 (1989), pp. 143–171.

[14] J. Cote, S. Gravel, A. Methot, A. Patoine, M. Roch, and A. Staniforth, *The operational CMC-MRB global environmental multiscale (GEM) model. Part* I*: Design considerations and formulation*, Mon. Wea. Rev., 126 (1998), pp. 1373–1395.

[15] R. Daley, *The development of efficient time integration schemes using model normal modes*, Mon. Wea. Rev., 108 (1980), pp. 100–110.

[16] J. Dudhia, *A nonhydrostatic version of the Penn State–NCAR mesoscale model: Validation tests and simulation of an Atlantic cyclone and cold front*, Mon. Wea. Rev., 121 (1993), pp. 1493–1513.

[17] D.R. Durran, *Numerical Methods for Wave Equations in Geophysical Fluid Dynamics.*, Springer-Verlag, New York, 1999.

[18] C. Eckart, *Hydrodynamics of Oceans and Atmospheres*, Pergamon Press, New York, 1960.

[19] F.R. Gantmacher, *Applications of the Theory of Matrices*, Interscience, New York, 1959.

[20] M. Gasca and J.M. Pena, *Total positivity and Neville elimination*, Linear Algebra Appl., 165 (1992), pp. 25–44.

[21] G.J. Haltiner and R.T. Williams, *Numerical Prediction and Dynamic Meteorology*, John Wiley and Sons, New York, 1980.

[22] E. Kalnay, M. Kanamitsu, and W.E. Baker, *Global numerical weather prediction at the National Meteorological Center*, Bull. Amer. Meteor. Soc., 71 (1990), pp. 1410–1428.

[23] A. Kasahara, *Various vertical coordinate systems used for numerical weather prediction*, Mon. Wea. Rev., 102 (1974), pp. 509–522.

[24] L.M. Leslie and R.J. Purser, *Three-dimensional mass-conservating semi-Lagrangian scheme employing forward trajectories*, Mon. Wea. Rev., 123 (1995), pp. 2551–2566.

[25] E.N. Lorenz, *Energy and numerical weather prediction*, Tellus, 12 (1960), pp. 364–373.

[26] A. McDonald and J. Haugen, *A two-time-level, three-dimensional semi-Lagrangian, semi-implicit, limited-area gridpoint model of the primitive equations*, Mon. Wea. Rev., 120 (1992), pp. 2603–2621.

[27] F. Mesinger and A. Arakawa, *Numerical Methods Used in Atmospheric Models*, GARP Publications Series, WMO/ICSU Joint Organizing Committee, Geneva, Switzerland, 1976.

[28] F. Mesinger, *Dynamics of limited-area models: Formulation and numerical methods*, Meteor. Atmos. Phys., 63 (1997), pp. 3–14.

[29] J.M. Pena, *Determinantal criteria for total positivity*, Linear Algebra Appl., 332/334 (2001), pp. 131–137.

[30] H. Ritchie, C. Temperton, A. Simmons, M. Hortal, T. Davies, D. Dent, and M. Hamrud, *Implementation of the semi-Lagrangian method in a high-resolution version of the ECMWF forecast model*, Mon. Wea. Rev., 123 (1995), pp. 489–514.

[31] A. Robert, *The integration of a spectral model of the atmosphere by the implicit method*, in Proceedings of the WMO/IUGG Symposium on Numerical Weather Prediction, Tokyo, 1969, pp. VII-19–VII-24.

[32] A. Robert, *A stable numerical integration scheme for the primitive meteorological equations*, Atmos.-Ocean, 19 (1981), pp. 35–46.

[33] A. Robert, J. Henderson, and C. Turnbull, *An implicit time integration scheme for baroclinic models of the atmosphere*, Mon. Wea. Rev., 100 (1972), pp. 329–335.

[34] A. Robert, T.L. Yee, and H. Ritchie, *A semi-Lagrangian and semi-implicit numerical integration scheme for multilevel atmospheric models*, Mon. Wea. Rev., 113 (1985), pp. 388–394.

[35] S. Schubert, M. Suarez, C.K. Park, and S. Moorthi, *GCM simulations on interseasonal variability in Pacific/North American region*, J. Atmospheric Sci., 50 (1993), pp. 1991–2007.

[36] A.J. Simmons and D.M. Burridge, *An energy and angular-momentum conserving vertical finite-difference scheme and hybrid vertical coordinates*, Mon. Wea. Rev., 109 (1981), pp. 758–766.

[37] A. Staniforth and J. Côté, *Semi-Lagrangian integration schemes for atmospheric models—A review*, Mon. Wea. Rev., 119 (1991), pp. 2206–2223.

[38] C. Temperton and D.L. Williamson, *Normal mode initialization for a multilevel grid-point model. Part* I*: Linear aspects*, Mon. Wea. Rev., 109 (1981), pp. 729–743.

[39] C. Temperton and M. Roch, *Implicit normal mode initialization for an operational regional model*, Mon. Wea. Rev., 119 (1991), pp. 667–677.

SIAM J. NUMER. ANAL.
Vol. 41, No. 1, pp. 208–234

APPROXIMATION OF A THIN PLATE SPLINE SMOOTHER USING CONTINUOUS PIECEWISE POLYNOMIAL FUNCTIONS[*]

STEPHEN ROBERTS[†], MARKUS HEGLAND[‡], AND IRFAN ALTAS[§]

Abstract. A new smoothing method is proposed which can be viewed as a finite element thin plate spline. This approach combines the favorable properties of finite element surface fitting with those of thin plate splines. The method is based on first order techniques similar to mixed finite element techniques for the biharmonic equation. The existence of a solution to our smoothing problem is demonstrated, and the approximation theory for uniformly spread data is presented in the case of both exact and noisy data. This convergence analysis seems to be the first for a discrete smoothing spline with data perturbed by white noise. Numerical results are presented which verify our theoretical results and demonstrate our method on a large real life data set.

Key words. thin plate splines, finite element methods, surface smoothing

AMS subject classifications. Primary, 65D15; Secondary, 41A15

PII. S0036142901383296

1. Introduction.

We propose a new finite element approximation of the thin plate spline for surface fitting in $\mathbb{R}^2$. We call our new smoother the TPSFEM smoother. This new approach combines the favorable properties of finite element surface fitting with those of thin plate splines. In particular, the method can deal with very large data sets consisting of tens of millions of predictor and response observations. Our smoothing function is a finite element approximation. As a consequence, the computational problem can be broken into two stages:

1. Forming the finite element matrices and vectors, which entails only a single scan of the data.
2. Solving for the finite element solution. This entails solving a sparse matrix system with size dependent on the discretization of a finite element mesh.

The size of the finite element structures can be chosen to be independent of the size of the data. This enables us to deal with very large data sets (tens of millions). For instance, in section 4 we present the results of smoothing geomagnetic data [6] with 735,700 data points.

Our method uses techniques typical of mixed finite element methods (see Brezzi and Fortin [11]) and of first order systems least squares (FOSLS) methods (see Cai, Manteuffel, and McCormick [14, 12, 13] and Cai et al. [15]). Auxiliary functions representing the gradient of the smoother are introduced to lower the order of smoothness of the resulting equations from $H^2(\Omega)$ to $H^1(\Omega)$. This allows for the use of simpler finite elements spaces and results in a better-conditioned problem.

The combined use of smoothing techniques and mixed finite element techniques provides a smoothing method which can deal with large data sets but still retains

[*] Received by the editors January 8, 2001; accepted for publication (in revised form) July 19, 2002; published electronically March 19, 2003.
http://www.siam.org/journals/sinum/41-1/38329.html

[†] Mathematical Research Institute, Australian National University, Canberra, ACT 0200, Australia (Stephen.Roberts@anu.edu.au).

[‡] Computer Sciences Laboratory, The Research School of Information Sciences and Engineering (RSISE), Australian National University, Canberra, ACT 0200, Australia (Markus.Hegland@anu.edu.au).

[§] School of Information Studies, Charles Sturt University, Wagga Wagga, NSW 2678, Australia (ialtas@golum.riv.csu.edu.au).

the good smoothing properties of the thin plate spline. An initial discussion of the method can be found in [23] and [24], and a discussion of the implementation and solution of the resulting discrete equations in [16] and [17].

The rest of the paper is organized as follows. The next section provides a detailed description and motivation for our method. In section 3 we describe the main convergence result, Theorem 3.2 on page 215, which justifies our claim that the TPSFEM method has approximation properties similar to those of a discrete thin plate spline. In section 4 we describe the numerical method used to solve our smoothing problem and present some numerical results for test problems and a real life large data set. The rest of the paper involves proving various properties of our method. The TPSFEM smoothing method produces a unique solution, provided that the values of the predictor variable are not collinear. This is shown in section 5. In section 6 we provide a number of useful results that will be used in section 7 to prove convergence of our method, that is, to prove Theorem 3.2. Finally in section 8 we conclude our discussion.

2. Description of the method.

2.1. Sobolev spaces. Before proceeding, let us introduce some notation. The $L^2(\Omega)$ inner-product and norm are given, respectively, by

$$(v,w)_{L^2(\Omega)} = \int_\Omega vw\,d\mathbf{x} \quad \text{and} \quad \|v\|^2_{L^2(\Omega)} = (v,v)_{L^2(\Omega)}.$$

Standard $H^1(\Omega)$ and $H^2(\Omega)$ Sobolev semi-inner-products are given by

$$(v,w)_{H^1(\Omega)} = \int_\Omega \left(\frac{\partial v}{\partial x_1}\frac{\partial w}{\partial x_1} + \frac{\partial v}{\partial x_2}\frac{\partial w}{\partial x_2}\right) d\mathbf{x}$$

and

$$(v,w)_{H^2(\Omega)} = \int_\Omega \left(\frac{\partial^2 v}{\partial^2 x_1}\frac{\partial^2 w}{\partial^2 x_1} + 2\frac{\partial^2 v}{\partial x_1\partial x_2}\frac{\partial^2 w}{\partial x_1\partial x_2} + \frac{\partial^2 v}{\partial^2 x_2}\frac{\partial^2 w}{\partial^2 x_2}\right) d\mathbf{x}.$$

The corresponding seminorms are given by

$$|v|^2_{H^1(\Omega)} = (v,v)_{H^1(\Omega)} \quad \text{and} \quad |v|^2_{H^2(\Omega)} = (v,v)_{H^2(\Omega)},$$

and the associated norms by

$$\|v\|^2_{H^1(\Omega)} = \|v\|^2_{L^2(\Omega)} + |v|^2_{H^1(\Omega)},$$
$$\|v\|^2_{H^2(\Omega)} = \|v\|^2_{L^2(\Omega)} + |v|^2_{H^2(\Omega)}.$$

Note that we use single bars $|\cdot|$ to denote seminorms, and double bars $\|\cdot\|$ to denote norms. For vector functions $\mathbf{u} = (u_1,u_2), \mathbf{v} = (v_1,v_2) \in H^1(\Omega)^2$, we define the (semi-)inner products and (semi-)norms as follows:

$$\begin{aligned}(\mathbf{u},\mathbf{v})_{L^2(\Omega)^2} &= (u_1,v_1)_{L^2(\Omega)} + (u_2,v_2)_{L^2(\Omega)},\\ \|\mathbf{u}\|^2_{L^2(\Omega)^2} &= (\mathbf{u},\mathbf{u})_{L^2(\Omega)^2},\\ (\mathbf{u},\mathbf{v})_{H^1(\Omega)^2} &= (u_1,v_1)_{H^1(\Omega)} + (u_2,v_2)_{H^1(\Omega)},\\ |\mathbf{u}|^2_{H^1(\Omega)^2} &= (\mathbf{u},\mathbf{u})_{H^1(\Omega)^2}.\end{aligned}$$

The dual norm $H^{-1}(\Omega)$ is given by

$$\|v\|_{H^{-1}(\Omega)} = \sup_{u\in H^1_0(\Omega)} \frac{\int_\Omega v(\mathbf{x})u(\mathbf{x})\,d\mathbf{x}}{\|u\|_{H^1(\Omega)}}.$$

2.2. Overview: Thin plate spline. To describe our method we first recall the definition of the standard thin plate spline, which provides a smoother that predicts a real response variable y, given the value of a predictor variable $\mathbf{x} \in \mathbb{R}^2$. The input to the construction of this smoother is an array of response values

$$\mathbf{y} = [y^{(1)} \ \cdots \ y^{(n)}]^T \in \mathbb{R}^n$$

and a corresponding array of predictor values

$$P = [\mathbf{x}^{(1)} \ \cdots \ \mathbf{x}^{(n)}]^T \in \mathbb{R}^{n\times 2}.$$

We suppose that the predictor values all lie in a convex bounded domain $\Omega \subset \mathbb{R}^2$.

The thin plate spline is the solution of the following minimization problem.

PROBLEM 1 (Thin plate spline). *For a fixed response vector* $\mathbf{y}$, *find the minimizer* $\bar{s}_\alpha(\mathbf{y})$ *of the functional*

$$\begin{aligned}\bar{J}_\alpha(s, \mathbf{y}) &= n^{-1}\sum_{i=1}^{n}(s(\mathbf{x}^{(i)}) - y^{(i)})^2 + \alpha|s|^2_{H^2(\Omega)} \\ &= \|\boldsymbol{\rho}^n s - \mathbf{y}\|_n^2 + \alpha|s|^2_{H^2(\Omega)}\end{aligned}$$

over all $s \in H^2(\Omega)$.

Here we have introduced the notation

$$\boldsymbol{\rho}^n s = [s(\mathbf{x}^{(1)}) \ \cdots \ s(\mathbf{x}^{(n)})]^T$$

and a norm on $\mathbb{R}^n$ defined by

$$\langle \mathbf{z}, \mathbf{w}\rangle_n = n^{-1}\mathbf{z}^T\mathbf{w}, \qquad \|\mathbf{z}\|_n^2 = \langle \mathbf{z}, \mathbf{z}\rangle_n.$$

We have absorbed the n factor into the definition of this norm, so that if s is a "smooth" function, then $\|\boldsymbol{\rho}^n s\|_n$ provides an estimate of $\|s\|_{L^2(\Omega)}$. (This is quantified in Lemma 6.2.)

Note that we have defined the thin plate spline with respect to a domain Ω. Indeed it is more common to use $\mathbb{R}^2$ than Ω. In that case the smoother is rotationally invariant. We are using a finite domain since we will be using simple finite element spaces.

The functional $\bar{J}_\alpha$ also depends on a smoothing parameter α. An appropriate choice of α depends on the size of the data errors and the number of data points. In many cases an appropriate choice of α can be made automatically using generalized cross-validation (see Craven and Wahba [19] or Wahba [29, Chapter 4]). But in this paper we will generally consider α as being fixed.

It is very reasonable to assume that the predictor variables $\mathbf{x}^{(i)}$ are not collinear (i.e., they do not all lie on a line in $\mathbb{R}^2$). When the domain is $\mathbb{R}^2$, it has been shown by Duchon [20] that Problem 1 has a unique solution and that the solution has an explicit representation as a sum of radial basis functions. This approach requires the solution of a symmetric indefinite dense linear system of equations that has a size proportional to the number of data records, n. Although this initial approach was improved later [7, 8, 9, 10, 26], these techniques require complex data structures and algorithms and an $O(n)$ workspace. Thus, it is a challenge to use standard thin plate splines for applications that have very large data sets.

Standard finite element discretizations of a problem like Problem 1 involve minimizing the functional $\bar{J}_\alpha$ over finite dimensional, finite element subspaces of $H^2(\Omega)$.

An analysis of these standard methods can be found in the work of Arcangéli and his coworkers [2, 3, 4, 5, 27]. There are reasonably simple $H^2(\Omega)$ finite element spaces (for instance, the Bogner–Fox–Schmidt rectangle [18, pp. 76–77]), but these are necessarily more complicated than the simpler $H^1(\Omega)$ finite element spaces. In particular, these smoother spaces lead to more dense stiffness matrices. For instance, the Bogner–Fox–Schmidt rectangle space will lead to sparse matrices with 36 nonzeros per variable, as compared to 9 for the space of bilinear functions. In addition, the conditioning of the $H^2(\Omega)$ matrix equations is typically $O(h^{-4})$, where h is a measure of the "mesh size" of the finite element space. Hence for problems which do not need the added smoothness it is more efficient to use the method described in this paper.

2.3. The $\boldsymbol{H^1(\Omega)}$ method. Our aim is to smooth very large data sets, which often necessitates fine meshes (typically $h <$ diameter$(\Omega)/500$ in many of our applications). As such, it is convenient to concentrate on methods in which the underlying finite element space is as simple as possible. This is very similar to the philosophy of using mixed finite element methods to solve the Navier–Stokes equation in fluid mechanics. To this end we reformulate Problem 1 so that only first order derivatives occur. In this case simple finite element spaces in $H^1(\Omega)$ can be utilized to discretize the problem. The conditioning of the resulting matrix problems are $O(h^{-2})$ and are readily solved using multigrid preconditioners.

As in many mixed finite element methods (see [11]), a new vector variable $\mathbf{u} = (u_1, u_2)$ is introduced. This variable represents the gradient of the function s sought in Problem 1. Our method formulates the minimization in terms of this new variable $\mathbf{u}$.

Suppose that $\mathbf{u}$ is the gradient of s, the minimizer of Problem 1. Then

$$\boldsymbol{\nabla} s = \mathbf{u}, \tag{2.1}$$

which determines s up to a constant. This constant can be determined by noting that a necessary condition for s to be the minimizer of Problem 1 is that

$$\langle \boldsymbol{\rho}^n s, \mathbf{e}\rangle_n = \langle \mathbf{y}, \mathbf{e}\rangle_n, \tag{2.2}$$

where $\mathbf{e}$ is the vector of all ones, $\mathbf{e} = [1 \ \cdots \ 1]^T$.

On the other hand, for general $\mathbf{u}$ we cannot expect to find an s satisfying (2.1) and (2.2). But for a given $\mathbf{u} \in H^1(\Omega)^2$ it is always possible to find an $s \in H^2(\Omega)$ satisfying

$$(\boldsymbol{\nabla} s, \boldsymbol{\nabla} v)_{L^2(\Omega)^2} = (\mathbf{u}, \boldsymbol{\nabla} v)_{L^2(\Omega)^2} \tag{2.3}$$

for all $v \in H^1(\Omega)$. In fact, there will be a unique such s satisfying (2.2) and (2.3), which we denote $\Phi(\mathbf{u}, \mathbf{y})$. Observe that $\Phi : H^1(\Omega)^2 \times \mathbb{R}^n \to H^2(\Omega)$.

The function $s = \Phi(\mathbf{u}, \mathbf{y})$ also satisfies a Neumann boundary value problem

$$\begin{aligned} \Delta\, s &= \boldsymbol{\nabla} \cdot \mathbf{u} \quad \text{in } \Omega, \\ \boldsymbol{\nabla} s \cdot \mathbf{n} &= \mathbf{u} \cdot \mathbf{n} \quad \text{on } \partial\Omega, \end{aligned}$$

together with the normalization (2.2).

Now consider the associated operator $\Psi : H^1(\Omega)^2 \to H^2(\Omega)$, where $g = \Psi(\mathbf{u})$ satisfies the Neumann boundary value problem

$$\Delta\, g = \boldsymbol{\nabla} \cdot \mathbf{u} \quad \text{in } \Omega, \tag{2.4}$$

$$\boldsymbol{\nabla} g \cdot \mathbf{n} = \mathbf{u} \cdot \mathbf{n} \quad \text{on } \partial\Omega, \tag{2.5}$$

$$\langle \boldsymbol{\rho}^n g, \mathbf{e}\rangle_n = 0. \tag{2.6}$$

It is easy to see that $\Phi(\mathbf{u}, \mathbf{y})$ and $\Psi(\mathbf{u})$ differ by a constant since they satisfy the same Neumann boundary value problem, and indeed

$$\Phi(\mathbf{u}, \mathbf{y}) = \Psi(\mathbf{u}) + \langle \mathbf{y}, \mathbf{e} \rangle_n.$$

We are thus led to the following problem.

Problem 2 (H^1 smoother). *For a fixed response vector $\mathbf{y}$, find $\mathbf{u}_\alpha(\mathbf{y}) \in H^1(\Omega)^2$ which minimizes the functional*

$$J_\alpha(\mathbf{u}, \mathbf{y}) = \|\boldsymbol{\rho}^n \Phi(\mathbf{u}, \mathbf{y}) - \mathbf{y}\|_n^2 + \alpha |\mathbf{u}|_{H^1(\Omega)^2}^2$$

over all $\mathbf{u} \in H^1(\Omega)^2$. The function $\Phi(\mathbf{u}_\alpha(\mathbf{y}), \mathbf{y})$ will correspond to our H^1 smoother and will be denoted $s_\alpha(\mathbf{y})$.

Problems 1 and 2 are not equivalent. For instance, there will be $\mathbf{u}$'s which are not the gradient of any s. Indeed, to force equivalence we would have to ensure that we minimize over those $\mathbf{u}$'s which are the gradient of some function, and this can be shown to be equivalent to $\mathbf{u}$ having zero curl ($\partial u_1/\partial x_2 - \partial u_2/\partial x_1 = 0$). While this can be done, and is the subject of many methods associated with the solution of the Navier–Stokes equation (see [21]), we instead advocate that the zero curl condition be dropped completely. The resulting Problem 2 is much easier to solve than one in which a zero curl condition is enforced. In addition, it is our claim that the smoothing function provided by Problem 2 produces a fit to the data (at least near the data points) that is similar to the standard thin plate spline. The precise statement of this result is our main convergence result, Theorem 3.2. The main advantages of this new minimization problem are that we can work with simple subspaces of $H^1(\Omega)$ and the problem becomes essentially an $H^1(\Omega)$ minimization problem, for which there are efficient solvers (for example, multigrid solvers).

2.4. The TPSFEM method. To discretize our problem we can minimize J_α over any simple finite dimensional subspace of $H^1(\Omega)^2$. In particular, we use simple continuous piecewise polynomial spaces $\mathbb{V}^h \subset H^1(\Omega)$ parameterized by h, the mesh size. Associated with each finite element space $\mathbb{V}^h$ is a mesh $\mathcal{T}^h$, consisting of a set of "elements" (usually triangles or quadrilaterals) $K \in \mathcal{T}^h$. The mesh size is given by

$$h = \max_{K \in \mathcal{T}^h} \text{diameter}(K).$$

We assume that the finite element spaces $\mathbb{V}^h$, and in particular the associated meshes $\mathcal{T}^h$, satisfy the following assumption.

Assumption 1 (Quasi-uniform meshes). *We will assume that the meshes $\mathcal{T}^h$ are "quasi-uniform"; that is,*

$$\max_{K \in \mathcal{T}^h} \frac{\text{diameter}(K)}{\rho(K)}$$

and

$$\frac{\max_{K \in \mathcal{T}^h} \text{diameter}(K)}{\min_{K \in \mathcal{T}^h} \text{diameter}(K)}$$

are bounded uniformly in h. Here $\rho(K)$ is the maximum diameter of any ball contained within K. Hence the elements of $\mathcal{T}^h$ cannot get too small relative to h, and they cannot get too "long and thin."

We also assume that the family of finite element spaces $\mathbb{V}^h$ satisfies the following approximation properties.

ASSUMPTION 2 (Standard approximation). *For all $\mathbb{V}^h$ there exists a linear operator $Q^h : H^1(\Omega) \to \mathbb{V}^h$ and a constant $C > 0$ such that for all $v \in H^2(\Omega)$*

$$\begin{aligned} \|v - Q^h v\|_{L^2(\Omega)} &\le Ch^2 |v|_{H^2(\Omega)}, \\ |v - Q^h v|_{H^1(\Omega)} &\le Ch|v|_{H^2(\Omega)}, \\ \sum_{K \in \mathcal{T}^h} \|v - Q^h v\|^2_{H^2(K)} &\le C|v|^2_{H^2(\Omega)}. \end{aligned}$$

In what follows, we will also need to control the $H^{-1}(\Omega)$ norm of the finite element approximation. As such, we also assume that our family of finite element spaces satisfies the following assumption.

ASSUMPTION 3 ($H^{-1}(\Omega)$ approximation). *For all $v \in H^1(\Omega)$ the linear operator $Q^h : H^1(\Omega) \to \mathbb{V}^h$ introduced in Assumption 2 satisfies*

$$\begin{aligned} \|v - Q^h v\|_{H^{-1}(\Omega)} &\le Ch^2 |v|_{H^1(\Omega)}, \\ \|v - Q^h v\|_{L^2(\Omega)} &\le Ch|v|_{H^1(\Omega)}, \\ |v - Q^h v|_{H^1(\Omega)} &\le C|v|_{H^1(\Omega)} \end{aligned}$$

for some constant $C > 0$.

The spaces of continuous piecewise linear functions associated with quasi-uniform triangulations of Ω or quadrilateral grids with bilinear functions satisfy Assumptions 2 and 3, with Q^h given by the $L^2(\Omega)$ projection.

The evaluation of $J_\alpha(\mathbf{u}, \mathbf{y})$ involves the calculation of $\Psi(\mathbf{u})$. An approximation of $g = \Psi(\mathbf{u})$ in $\mathbb{V}^h$ is given by $g^h = \Psi^h(\mathbf{u})$, where

$$(\boldsymbol{\nabla} g^h, \boldsymbol{\nabla} v)_{L^2(\Omega)^2} = (\mathbf{u}, \boldsymbol{\nabla} v)_{L^2(\Omega)^2} \tag{2.7}$$

for all $v \in \mathbb{V}^h$, also subject to the condition (2.6). We then let

$$\Phi^h(\mathbf{u}, \mathbf{y}) = \Psi^h(\mathbf{u}) + \langle \mathbf{y}, \mathbf{e} \rangle_n.$$

Our final discrete problem becomes the following.

PROBLEM 3 (TPSFEM smoother). *For a given response vector $\mathbf{y}$, find $\mathbf{u}^h_\alpha(\mathbf{y}) \in \mathbb{V}^h \times \mathbb{V}^h$, which minimizes the functional*

$$J^h_\alpha(\mathbf{u}, \mathbf{y}) = \|\boldsymbol{\rho}^n \Phi^h(\mathbf{u}, \mathbf{y}) - \mathbf{y}\|^2_n + \alpha |\mathbf{u}|^2_{H^1(\Omega)^2}$$

over the space $\mathbb{V}^h \times \mathbb{V}^h$. The function $\Phi^h(\mathbf{u}^h_\alpha(\mathbf{y}), \mathbf{y})$ will correspond to our TPSFEM smoothing function and is denoted $s^h_\alpha(\mathbf{y})$.

3. Approximation properties of the TPSFEM smoother. To obtain some quantitative information about the approximation properties of our smoother, we follow Utreras [28] and make the following assumptions about the form of the response variable observations and the values of the predictor variables. First the response variable.

ASSUMPTION 4 (Data model). *We suppose that our response variable is modelled by a smooth function $f \in H^2(\Omega)$ plus error. That is, the value of the response variable at a data point satisfies*

$$y^{(i)} = f(\mathbf{x}^{(i)}) + \nu^{(i)},$$

where $\nu^{(i)}$ denotes measurement error. We suppose that the $\nu^{(i)}$ are independent identically distributed random variables with mean zero and variance σ^2.

We also make an assumption about the spread of the values of the predictor variable. Let

$$d = \sup_{\mathbf{x}\in\Omega} \text{distance}(\mathbf{x}, \{\mathbf{x}^{(i)}\}) \tag{3.1}$$

be the maximum distance of any point in Ω from a data point. Following the problem formulation in [28], we assume that the predictor variables are "uniformly spread" throughout Ω in the following sense.

ASSUMPTION 5 (Uniformly spread data). *We assume that d, as a function of the number of data points n, is controlled by the minimum distance between data points. That is, there exists a constant $C > 1$ such that*

$$d < C \min_{i\neq j} |\mathbf{x}^{(i)} - \mathbf{x}^{(j)}|.$$

This implies that any ball of radius, say, $2d$ contained in Ω will have at least three data points, with the number uniformly bounded above. By filling larger balls $B \subset \Omega$ with balls of radius $2d$, we conclude that the number of predictor points in a larger ball is proportional to the area of the ball. In particular we have that

$$d^2 n_B \preceq \text{area}(B) \preceq d^2 n_B. \tag{3.2}$$

Here $n_B = \text{card}\{\mathbf{x}^{(i)} : \mathbf{x}^{(i)} \in B\}$, the number of predictor points in B. We will use this notation for any set B.

REMARK 1. *Note that we have introduced the notation*

$$f(d,h,\alpha) \preceq g(d,h,\alpha),$$

which will signify that

$$f(d,h,\alpha) \leq Cg(d,h,\alpha)$$

for some constant $C > 0$ uniformly for all choices of d, h, and α over the specified range of values. This notation will be used extensively in what follows.

REMARK 2. *Since our set Ω is convex, we can assume that it satisfies an interior cone condition and so conclude that* (3.2) *holds for all balls of radius greater than* $2d$ *with center contained in Ω (i.e., the balls may extend over the boundary of Ω).*

Combining the assumptions on the data and the finite element spaces, we conclude a similar point density estimate for elements of our meshes. Namely, we claim the following.

PROPOSITION 3.1. *Provided that the finite element spaces satisfy Assumption* 1 *and the data satisfies Assumption* 5, *then there exists a constant $C_1 > 1$ such that, for any mesh $\mathcal{T}^h$ with $h > C_1 d$, the elements $K \in \bigcup_{h>C_1 d} \mathcal{T}^h$ satisfy*

$$d^2 n_K \preceq \text{area}(K) \preceq d^2 n_K. \tag{3.3}$$

Proof. Assumption 1 implies that for each $K \in \mathcal{T}^h$ there exist balls $\underline{B}_K$ and $\overline{B}_K$ such that $\underline{B}_K \subset K \subset \overline{B}_K$ and $\text{area}(\overline{B}_K) \preceq \text{area}(K) \preceq \text{area}(\underline{B}_K)$. The elements

K cannot get too long and thin. Provided the elements are large enough ($h > C_1\, d$ for some C_1), we can assume that for all elements $K \in \bigcup_{h>C_1\, d} \mathcal{T}^h$ the inscribed balls satisfy radius$(\underline{B}_K) > 2d$. Hence (3.2) implies that area$(\underline{B}_K) \preceq d^2 n_{\underline{B}_K}$ and $d^2 n_{\overline{B}_K} \preceq \text{area}(\overline{B}_K)$. Combining these estimates leads to

$$d^2 n_K \preceq d^2 n_{\overline{B}_K} \preceq \text{area}(\overline{B}_K) \preceq \text{area}(K) \preceq \text{area}(\underline{B}_K) \preceq d^2 n_{\underline{B}_K} \preceq d^2 n_K,$$

which implies (3.3). □

REMARK 3. *By summing* (3.3) *over all K in the mesh, we conclude that*

$$d^2 n \preceq \text{area}(\Omega) \preceq d^2 n. \tag{3.4}$$

Now we can state the main convergence property of our smoother when applied to data which is uniformly spread.

THEOREM 3.2 (Main convergence result). *Suppose that our finite element spaces satisfy Assumptions* 1, 2, *and* 3 *and our data satisfies Assumptions* 4 *and* 5. *Then there exist constants $C_1 > 1$ and $\alpha_0 > 0$ such that for all $f \in H^2(\Omega)$ the expected errors of the TPSFEM smoother satisfy*

$$E\|s_\alpha^h(\mathbf{y}) - f\|^2_{L^2(\Omega)} \preceq (\alpha + h^4 + d^4)\|f\|^2_{H^2(\Omega)} + \frac{\sigma^2}{\alpha^{1/2}} \frac{(h^4 + d^4)}{h^2}, \tag{3.5}$$

$$E|s_\alpha^h(\mathbf{y}) - f|^2_{H^1(\Omega)} \preceq \frac{1}{\alpha^{1/2}}(\alpha + h^4 + d^4)\|f\|^2_{H^2(\Omega)} + \frac{\sigma^2}{\alpha} \frac{(h^4 + d^4)}{h^2}, \tag{3.6}$$

$$E\|\mathbf{u}_\alpha^h(\mathbf{y})\|^2_{H^1(\Omega)^2} \preceq \frac{1}{\alpha}(\alpha + h^4 + d^4)\|f\|^2_{H^2(\Omega)} + \frac{\sigma^2}{\alpha^{3/2}} \frac{(h^4 + d^4)}{h^2}, \tag{3.7}$$

provided that h and α satisfy $h > C_1\, d$ and $d^4 + h^4 < \alpha < \alpha_0$.

Hence, under standard assumptions on the data, our smoothing method satisfies smoothing properties very similar to those of the standard thin plate spline (see [28, Theorem 1.1]).

To help qualify the contributions of the approximations in our method, we have elected to present the convergence result with the explicit α, d, and h dependencies retained. We see that the errors naturally divide into a bias and a variance term. In the bias term, the quantity $\alpha + d^4 + h^4$ arises from the

1. smoothing error (α),
2. error in approximating $L^2(\Omega)$ norms with pointwise norms (d^4), and
3. finite element approximation error (h^4).

The variance term is essentially of the form $\sigma^2 d^2/\alpha^{1/2}$ for small h, which is the bound obtained for the standard thin plate spline. Actually we conjecture that a more sophisticated analysis will show that the variance term is of the form $\sigma^2 d^2/(\alpha^{1/2}+h^2)$, which would demonstrate the smoothing influence of coarser grids.

In section 7 we prove this theorem using an argument very similar to that found in Utreras [28]. First, in Theorem 7.3 we bound the norms of bias of the method $s_\alpha^h(\boldsymbol{\rho}^n f) - f$. Then in Theorem 7.8 the norms of the variance $s_\alpha^h(\boldsymbol{\nu})$ are shown to be bounded by the analogous bounds for the standard thin plate spline. This convergence analysis seems to be the first for a discrete smoothing spline with data perturbed by white noise. It should be possible to apply our methodology to proving convergence for the standard discrete smoothing splines.

It should be noted that, given the requirements that $h > C_1\, d$ and $d^4 + h^4 < \alpha < \alpha_0$, we can obtain the somewhat "cleaner" estimates

$$\begin{aligned} E\|s^h_\alpha(\mathbf{y}) - f\|^2_{L^2(\Omega)} &\preceq \alpha\|f\|^2_{H^2(\Omega)} + h^2\frac{\sigma^2}{\alpha^{1/2}}, \\ E|s^h_\alpha(\mathbf{y}) - f|^2_{H^1(\Omega)} &\preceq \alpha^{1/2}\|f\|^2_{H^2(\Omega)} + h^2\frac{\sigma^2}{\alpha}, \\ E\|\mathbf{u}^h_\alpha(\mathbf{y})\|^2_{H^1(\Omega)^2} &\preceq \|f\|^2_{H^2(\Omega)} + h^2\frac{\sigma^2}{\alpha^{3/2}}. \end{aligned}$$

4. The numerical method.

In this section we will outline the method we use to calculate our smoother. In subsection 4.2 we describe the generalized cross-validation method that we often use to calculate an appropriate value for the smoothing parameter α. Finally in subsection 4.3 we provide two numerical experiments to compare our method with Duchon's method and to demonstrate our method on a large data fitting problem.

4.1. The discrete equations.

The TPSFEM method provides a gradient function $\mathbf{u} = \mathbf{u}^h_\alpha(\mathbf{y})$ and an associated smoothing function $s = \Phi^h_\alpha(\mathbf{u}, \mathbf{y})$. To obtain an explicit representation of these functions in $\mathbb{V}^h$ we must calculate their expansion in terms of a basis for $\mathbb{V}^h$. As such, let

$$\mathbf{h}(\mathbf{x}) = \begin{bmatrix} h_1(\mathbf{x}) \\ \vdots \\ h_m(\mathbf{x}) \end{bmatrix}$$

denote a vector of basis functions for the finite element space $\mathbb{V}^h$. The values of the basis functions at the data points are encapsulated by the matrix

$$\boldsymbol{H}^T = \boldsymbol{\rho}^n \mathbf{h}(\mathbf{x})^T = \begin{bmatrix} h_1(\mathbf{x}^{(1)}) & \cdots & h_m(\mathbf{x}^{(1)}) \\ \vdots & \vdots & \vdots \\ h_1(\mathbf{x}^{(n)}) & \cdots & h_m(\mathbf{x}^{(n)}) \end{bmatrix}.$$

Then s and $\mathbf{u} = [u_1, u_2]^T$ will be of the form

$$s(\mathbf{x}) = \mathbf{h}(\mathbf{x})^T\mathbf{c}, \quad u_1(\mathbf{x}) = \mathbf{h}(\mathbf{x})^T\mathbf{g}_1, \quad u_2(\mathbf{x}) = \mathbf{h}(\mathbf{x})^T\mathbf{g}_2, \tag{4.1}$$

where the vectors $\mathbf{c}$, $\mathbf{g}_1$, and $\mathbf{g}_2$ represent the linear combination coefficients in the basis $\mathbf{h}$.

Relation (2.3) between s and $\mathbf{u}$ can be written as

$$\boldsymbol{L}\mathbf{c} = \boldsymbol{G}_1\mathbf{g}_1 + \boldsymbol{G}_2\mathbf{g}_2, \tag{4.2}$$

where $\boldsymbol{L} = (\boldsymbol{\nabla}\mathbf{h}, \boldsymbol{\nabla}\mathbf{h}^T)_{L^2(\Omega)}$ is a matrix approximation to the operator $-\Delta$, and $\boldsymbol{G}_1 = (\partial_{x_1}\mathbf{h}, \mathbf{h}^T)_{L^2(\Omega)}$ and $\boldsymbol{G}_2 = (\partial_{x_2}\mathbf{h}, \mathbf{h}^T)_{L^2(\Omega)}$ are matrix approximations to the operators $-\partial_{x_1}$ and $-\partial_{x_2}$, respectively. Consequently

$$\mathbf{c} = \boldsymbol{L}^+(\boldsymbol{G}_1\mathbf{g}_1 + \boldsymbol{G}_2\mathbf{g}_2),$$

where $\boldsymbol{L}^+$ is the pseudoinverse of $\boldsymbol{L}$ which satisfies $L^+ H\mathbf{e} = 0$. In our implementation all calculations involving $\boldsymbol{L}^+$ are provided by a multigrid Poisson solver.

Our discrete functional is now equivalent to

$$J^h_\alpha(\mathbf{u},\mathbf{y}) = \tilde{J}^h_\alpha(\mathbf{g},\mathbf{y}) = \|\mathcal{K}\mathbf{g} + \langle \mathbf{y},\mathbf{e}\rangle_n \mathbf{e} - \mathbf{y}\|^2_n + \alpha \mathbf{g}^T \mathcal{L}\mathbf{g}, \tag{4.3}$$

where $\mathbf{g} = [\mathbf{g}_1^T \quad \mathbf{g}_2^T]^T$, $\mathcal{K} = \boldsymbol{H}^T L^+ [\boldsymbol{G}_1 \quad \boldsymbol{G}_2]$, and

$$\mathcal{L} = \begin{bmatrix} \boldsymbol{L} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{L} \end{bmatrix}.$$

Our smoothing problem consists of minimizing this functional over all vectors $\mathbf{g} \in \mathbb{R}^{2m}$.

The minimizer satisfies the equation

$$\left[\mathcal{K}^T\mathcal{K} + n\alpha\mathcal{L}\right]\mathbf{g} = \mathcal{K}^T\tilde{\mathbf{y}}, \tag{4.4}$$

where $\tilde{\mathbf{y}} = \mathbf{y} - \langle \mathbf{y},\mathbf{e}\rangle_n \mathbf{e}$. In terms of the components $\mathbf{g}_1$ and $\mathbf{g}_2$, this becomes

$$\begin{bmatrix} \boldsymbol{G}_1^T\boldsymbol{Z}\boldsymbol{G}_1 + n\alpha\boldsymbol{L} & \boldsymbol{G}_1^T\boldsymbol{Z}\boldsymbol{G}_2 \\ \boldsymbol{G}_2^T\boldsymbol{Z}\boldsymbol{G}_1 & \boldsymbol{G}_2^T\boldsymbol{Z}\boldsymbol{G}_2 + n\alpha\boldsymbol{L} \end{bmatrix} \begin{bmatrix} \mathbf{g}_1 \\ \mathbf{g}_2 \end{bmatrix} = \begin{bmatrix} \boldsymbol{G}_1^T\boldsymbol{L}^+\boldsymbol{z} \\ \boldsymbol{G}_2^T\boldsymbol{L}^+\boldsymbol{z} \end{bmatrix}, \tag{4.5}$$

where $\boldsymbol{Z} = \boldsymbol{L}^+\boldsymbol{H}\boldsymbol{H}^T\boldsymbol{L}^+$ and $\boldsymbol{z} = \boldsymbol{H}^T\tilde{\mathbf{y}}$. The matrix $\boldsymbol{H}\boldsymbol{H}^T$ and the vector $\boldsymbol{z}$ depend on the data and are evaluated by reading the data arrays from disk and computing the contribution due to each data point. This is then accumulated into global data structures. Thus, the formation of $\boldsymbol{H}\boldsymbol{H}^T$ and $\boldsymbol{z}$ is scalable with respect to the data size. These operations parallelize well because the different segments of the data can be processed independently.

The matrices $\boldsymbol{L}$, $\boldsymbol{G}_1$, and $\boldsymbol{G}_2$ depend on the specified mesh size, which in many cases can be chosen independently of the data. It can be seen that the system (4.5) is symmetric positive definite, and so we use the conjugate gradient method to solve the equation. In fact, we use the preconditioned conjugate gradient (PCG) method, with preconditioner matrix

$$\boldsymbol{Q} = \begin{bmatrix} \boldsymbol{L}^+ & \mathbf{0} \\ \mathbf{0} & \boldsymbol{L}^+ \end{bmatrix}.$$

Each iteration of our PCG solver involves four applications of $\boldsymbol{L}^+$. For smooth problems we need on the order of twenty iterations of our PCG method to reduce the residual of the equation by a factor of 10^{-10}.

4.2. Estimation of optimal $\boldsymbol{\alpha}$ and $\boldsymbol{h}$. The parameters α and h provide a measure of the smoothing effect of our method. We usually determine "optimal" values for these parameters in a two-step process. For a given h we use generalized cross-validation (GCV) to determine α. Then we decrease h until the value of α obtained by the GCV method stabilizes at a fixed value.

4.2.1. The generalized cross-validation method. For a fixed h we obtain an "optimal" value for α using the GCV method. The GCV function

$$V(\alpha) = \frac{\|(I - \mathcal{K}[\mathcal{K}^T\mathcal{K} + n\alpha\mathcal{L}]^{-1}\mathcal{K}^T)\tilde{\mathbf{y}}\|^2_n}{[tr(I - \mathcal{K}[\mathcal{K}^T\mathcal{K} + n\alpha\mathcal{L}]^{-1}\mathcal{K}^T)]^2}$$

provides a cumulative measure of the error of estimating the response value at a point by a smoother in which individual data points have been "left out." This function provides a measure of how well the smoother is fitting the data, and also how well "new" data is approximated by the smoother. The reader is referred to Wahba's book [29] for a description of cross-validation and GCV methods. In the GCV method the "optimal" smoothing parameter α is the minimizer of the GCV function [29, p. 43]. For us, the problem is how to calculate the function $V(\alpha)$ efficiently. The most difficult term is the trace term, as we do not want to assemble the full matrix $[\mathcal{K}^T\mathcal{K}+n\alpha\mathcal{L}]^{-1}$. We use an approximation due to Hutchinson [25], in which the trace term is approximated by an unbiased stochastic estimator. In particular, we use the function

$$\bar{V}(\alpha) = \frac{\|(I - \mathcal{K}[\mathcal{K}^T\mathcal{K} + n\alpha\mathcal{L}]^{-1}\mathcal{K}^T)\tilde{\mathbf{y}}\|_n^2}{[tr(I) - \mathbf{u}^T\mathcal{K}[\mathcal{K}^T\mathcal{K} + n\alpha\mathcal{L}]^{-1}\mathcal{K}^T\mathbf{u}]^2}$$

as an approximation of the standard GCV function. Here $\mathbf{u}$ is a random vector whose entries take on the values 1 and -1 with equal probability $\frac{1}{2}$. To calculate $\bar{V}(\alpha)$ for any fixed α, we must solve two smoothing problems (4.4). We have found that it is necessary to obtain the optimal α only to within an order of magnitude, and so we generally need to calculate $\bar{V}(\alpha)$ for only four to five values of α.

4.3. Numerical experiments. We now provide two numerical experiments. The first verifies that the TPSFEM method and the standard thin plate spline smoother provide similar results for a simple smoothing problem. Our second experiment demonstrates our method on a reasonably large data set.

First the comparison with the thin plate spline smoother. For clarity, we will apply our method to a small problem in order to *accentuate* the difference between the two smoothers. Our data is of the form

$$\mathbf{y} = \boldsymbol{\rho}^n f + \boldsymbol{\nu},$$

where the random variables $\boldsymbol{\nu}$ are normally distributed with expectation 0 and standard deviation 0.2. The components of the data points $\mathbf{x}^{(i)}$ are independently normally distributed with expectation 0.5 and standard deviation 0.25. For f the "peaks" function from MATLAB has been chosen. The peaks function is formed as a linear combination of several scaled and translated Gaussian distributions. The smoothing parameter is $\alpha = 10^{-6}$. In this case we use bilinear elements on a uniform 32 by 32 grid.

As can be seen from Figure 4.1, the two smoothers approximate the data well in the region of high density and show slightly different behavior on the boundary where the two smoothers satisfy different conditions and the data points are less dense. Generally the qualitative behavior is similar.

As another example of our method we present results in Figure 4.2 from smoothing a survey of magnetic field strengths known as the Ebagoola magnetic data set [6]. The survey comprises 735,700 data points consisting of latitude and longitude and magnetic field strength obtained from an aerial survey of the Ebagoola region of Cape York Peninsula in northern Australia. The region surveyed is approximately 61 by 63 kilometers. The distance between data points is very small in the direction of flight (about 10 meters), whereas the distance between adjacent flight paths is longer, on the order of 500 meters.

The smoother was obtained in 7.5 minutes using a MATLAB implementation on a 135 MHz SUN Sparc Station. Reading in and parsing the 17MB ASCII data set takes 5 minutes on the same workstation. It should be noted the TPSFEM smoother

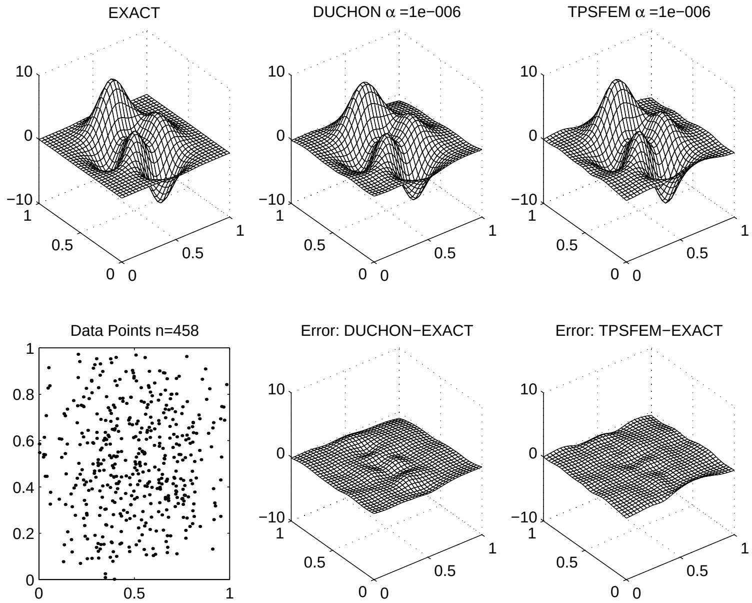


Fig. 4.1. *Comparison of standard thin plate spline (DUCHON) and the TPSFEM smoother; an example fitting data with* 458 *points. Approximation properties are similar except near the boundaries, where there are only a small number of data points.*

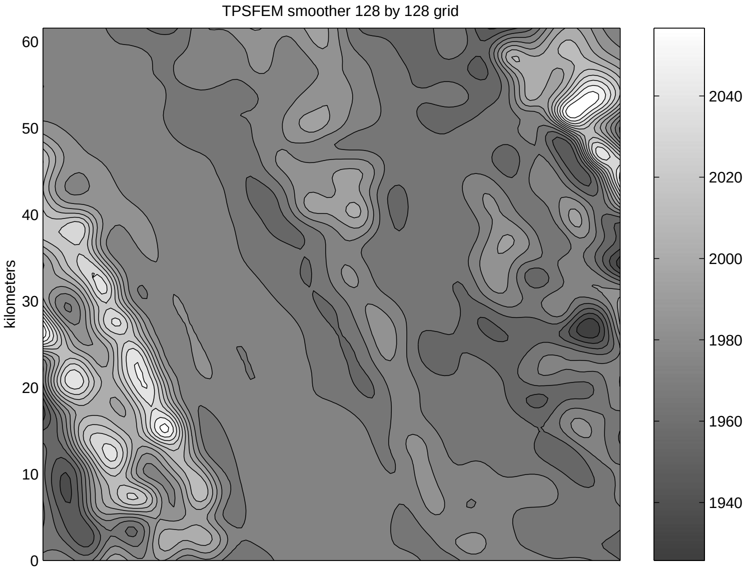


Fig. 4.2. *The TPSFEM smoother for the Ebagoola magnetic data set of* 735,700 *points with a finite element grid of* 128 *by* 128.

provides an extreme compression of the data. The grid size has been chosen to match the data spread between flight paths. This is provided by using bilinear elements on a uniform 128 by 128 grid. Even so, it not guaranteed that each element will have data points. The use of the TPSFEM provides a robust smoother which deals with the large data size and is also guaranteed to have a solution for essentially any distribution of data points (as opposed to a simple finite element least squares fit, which will fail if any finite element contains no data points).

5. Existence and uniqueness of the minimizers. We will now show that if the values of the predictor variables $\mathbf{x}^{(i)}$ are not collinear, then the unconstrained minimization problems 2 and 3 will have unique solutions. This will be shown by proving that equivalent associated variational equations satisfy standard continuity and ellipticity properties.

First we note that the thin plate spline problem, Problem 1 on page 210, is equivalent to the following variational problem.

PROBLEM 4 (Thin plate spline smoother). *For fixed* $\mathbf{y} \in \mathbb{R}^n$, *find* $\bar{s}_\alpha(\mathbf{y}) \in H^2(\Omega)$ *such that*

$$\bar{a}_\alpha(s_\alpha(\mathbf{y}), v) = \bar{F}(v, \mathbf{y}) \qquad \textit{for all } v \in H^2(\Omega), \tag{5.1}$$

where

$$\bar{a}_\alpha(w, v) = \langle \boldsymbol{\rho}^n w, \boldsymbol{\rho}^n v \rangle_n + \alpha (w, v)_{H^2(\Omega)}$$

and

$$\bar{F}(v, \mathbf{y}) = \langle \mathbf{y}, \boldsymbol{\rho}^n v \rangle_n.$$

The existence and uniqueness of the solution of this problem is provided by Duchon [20].

The H^1 smoothing problem, Problem 2 on page 212, is equivalent to the following variational problem.

PROBLEM 5 (H^1 smoother). *For fixed* $\mathbf{y} \in \mathbb{R}^n$, *find* $\mathbf{u}_\alpha(\mathbf{y}) \in H^1(\Omega)^2$ *such that*

$$a_\alpha(\mathbf{u}_\alpha(\mathbf{y}), \mathbf{v}) = F(\mathbf{v}, \mathbf{y}) \qquad \textit{for all } \mathbf{v} \in H^1(\Omega)^2, \tag{5.2}$$

where

$$a_\alpha(\mathbf{u}, \mathbf{v}) = \langle \boldsymbol{\rho}^n \Psi(\mathbf{u}), \boldsymbol{\rho}^n \Psi(\mathbf{v}) \rangle_n + \alpha (\mathbf{u}, \mathbf{v})_{H^1(\Omega)^2}$$

and

$$F(\mathbf{v}, \mathbf{y}) = \langle \mathbf{y}, \boldsymbol{\rho}^n \Psi(\mathbf{v}) \rangle_n.$$

The equivalence of Problems 2 and 5 follows from the fact that

$$\begin{aligned} J_\alpha(\mathbf{u}, \mathbf{y}) &= \|\boldsymbol{\rho}^n \Phi(\mathbf{u}, \mathbf{y}) - \mathbf{y}\|_n^2 + \alpha |\mathbf{u}|_{H^1(\Omega)^2}^2 \\ &= \|\boldsymbol{\rho}^n \Psi(\mathbf{u}) - \mathbf{y} + \langle \mathbf{y}, \mathbf{e} \rangle_n \mathbf{e}\|_n^2 + \alpha |\mathbf{u}|_{H^1(\Omega)^2}^2. \end{aligned}$$

It is elementary that the minimization problem is equivalent to the variational equation (5.2), where a_α is given above and

$$F(\mathbf{v}, \mathbf{y}) = \langle \mathbf{y} - \langle \mathbf{y}, \mathbf{e} \rangle_n \mathbf{e}, \boldsymbol{\rho}^n \Psi(\mathbf{v}) \rangle_n.$$

Now $\langle \boldsymbol{\rho}^n \Psi(\mathbf{v}), \mathbf{e}\rangle_n = 0$, and so

$$F(\mathbf{v}, \mathbf{y}) = \langle \mathbf{y}, \boldsymbol{\rho}^n \Psi(\mathbf{v})\rangle_n,$$

as required.

The TPSFEM variational problem is given by the following.

PROBLEM 6 (TPSFEM smoother). *For fixed* $\mathbf{y} \in \mathbb{R}^n$*, find* $\mathbf{u}_\alpha^h(\mathbf{y}) \in \mathbb{V}^h \times \mathbb{V}^h$ *such that*

$$a_\alpha^h(\mathbf{u}_\alpha^h(\mathbf{y}), \mathbf{v}) = F^h(\mathbf{v}, \mathbf{y}) \qquad \textit{for all } \mathbf{v} \in \mathbb{V}^h \times \mathbb{V}^h, \tag{5.3}$$

where

$$a_\alpha^h(\mathbf{u}_\alpha^h, \mathbf{v}) = \langle \boldsymbol{\rho}^n \Psi^h(\mathbf{u}), \boldsymbol{\rho}^n \Psi^h(\mathbf{v})\rangle_n + \alpha(\mathbf{u}, \mathbf{v})_{H^1(\Omega)^2}$$

and

$$F^h(\mathbf{v}, \mathbf{y}) = \langle \mathbf{y}, \boldsymbol{\rho}^n \Psi^h(\mathbf{v})\rangle_n.$$

The equivalence of Problem 3 on page 213 and Problem 6 follows in exactly the same way as the previous equivalence statement.

To prove the existence and uniqueness of solutions of these variational equations, we must show continuity and ellipticity of the operators. First we show the continuity of bilinear forms $a_\alpha(\cdot,\cdot)$ and $a_\alpha^h(\cdot,\cdot)$ and the linear functionals $F(\cdot, \mathbf{y})$ and $F^h(\cdot, \mathbf{y})$.

PROPOSITION 5.1 (Continuity). *The bilinear forms* $a_\alpha(\cdot,\cdot)$ *and* $a_\alpha^h(\cdot,\cdot)$ *and the linear forms* $F(\cdot, \mathbf{y})$ *and* $F^h(\cdot, \mathbf{y})$ *are continuous on* $H^1(\Omega)^2$*.*

Proof. The continuity of these linear operators follows from the fact that the pointwise evaluations of the functions $\Psi(\mathbf{v})$ and $\Psi^h(\mathbf{v})$ are continuous operations when $\mathbf{v} \in H^1(\Omega)^2$. Recall that the function $\Psi(\mathbf{v})$ satisfies (2.4). Now $\mathbf{v} \in H^1(\Omega)^2$, and so $\boldsymbol{\nabla} \cdot \mathbf{v} \in L^2(\Omega)$ and $\mathbf{v} \cdot \mathbf{n} \in H^{1/2}(\partial\Omega)$. The standard regularity theory for second order elliptic equations on convex domains and the trace theorem (see [22]) imply that

$$|\Psi(\mathbf{v})|_{H^1(\Omega)} + |\Psi(\mathbf{v})|_{H^2(\Omega)} \preceq \|\boldsymbol{\nabla} \cdot \mathbf{v}\|_{L^2(\Omega)} + \|\mathbf{v} \cdot \mathbf{n}\|_{H^{1/2}(\partial\Omega)} \preceq \|\mathbf{v}\|_{H^1(\Omega)^2}.$$

The condition $\langle \boldsymbol{\rho}^n \Psi(\mathbf{v}), \mathbf{e}\rangle_n = 0$ implies that a Poincaré inequality holds for $\Psi(\mathbf{v})$, namely,

$$\|\Psi(\mathbf{v})\|_{L^2(\Omega)} \preceq |\Psi(\mathbf{v})|_{H^1(\Omega)},$$

and so

$$\|\Psi(\mathbf{v})\|_{H^2(\Omega)} \preceq \|\mathbf{v}\|_{H^1(\Omega)^2}. \tag{5.4}$$

The Sobolev inequality implies that the pointwise norm $\|\cdot\|_n$ is bounded by the $H^2(\Omega)$ norm, and so we conclude that there exists a constant $C > 0$ (perhaps dependent on the data distribution) such that

$$\|\boldsymbol{\rho}^n \Psi(\mathbf{v})\|_n \leq C\|\mathbf{v}\|_{H^1(\Omega)^2}. \tag{5.5}$$

A similar estimate is necessary for $\Psi^h(\mathbf{v})$, the finite element approximation of $\Psi(\mathbf{v})$. It is useful to observe that since the finite element space $\mathbb{V}^h$ satisfies Assumption 2 on page 213, standard error estimates imply that

$$\|\Psi^h(\mathbf{v}) - \Psi(\mathbf{v})\|^2_{L^2(\Omega)} \preceq h^4 \|\Psi(\mathbf{v})\|^2_{H^2(\Omega)}, \tag{5.6}$$

$$|\Psi^h(\mathbf{v}) - \Psi(\mathbf{v})|^2_{H^1(\Omega)} \preceq h^2 \|\Psi(\mathbf{v})\|^2_{H^2(\Omega)}, \tag{5.7}$$

$$\sum_{K \in \mathcal{T}^h} |\Psi^h(\mathbf{v}) - \Psi(\mathbf{v})|^2_{H^2(K)} \preceq \|\Psi(\mathbf{v})\|^2_{H^2(\Omega)}. \tag{5.8}$$

We cannot assume that $\Psi^h(\mathbf{v}) \in H^2(\Omega)$, but we can assume that $\Psi^h(\mathbf{v})$ restricted to each element $K \in \mathcal{T}^h$ is in $H^2(K)$. Hence there exists a constant $C > 0$ such that

$$\|\boldsymbol{\rho}^n \Psi^h(\mathbf{v})\|^2_n \leq C \sum_{K \in \mathcal{T}^h} \|\Psi^h(\mathbf{v})\|^2_{H^2(K)}.$$

Equation (5.8) implies that

$$\begin{aligned} &\sum_{K \in \mathcal{T}^h} \|\Psi^h(\mathbf{v})\|^2_{H^2(K)} \\ &\preceq \sum_{K \in \mathcal{T}^h} \|\Psi^h(\mathbf{v}) - \Psi(\mathbf{v})\|^2_{H^2(K)} + \sum_{K \in \mathcal{T}^h} \|\Psi(\mathbf{v})\|^2_{H^2(K)} \preceq \|\Psi(\mathbf{v})\|^2_{H^2(\Omega)}. \end{aligned}$$

By (5.4) we conclude that there exists a constant $C > 0$ such that

$$\|\boldsymbol{\rho}^n \Psi^h(\mathbf{v})\|_n \leq C \|\mathbf{v}\|_{H^1(\Omega)^2}. \tag{5.9}$$

Equations (5.5) and (5.9) then imply that the bilinear functionals $a_\alpha(\cdot,\cdot)$ and $a^h_\alpha(\cdot,\cdot)$ and the linear functionals $F(\cdot,\mathbf{y})$ and $F^h(\cdot,\mathbf{y})$ are continuous. $\square$

PROPOSITION 5.2 (Ellipticity). *Suppose that the values of the predictor variable are not collinear. Then $a_\alpha(\cdot,\cdot)$ and $a^h_\alpha(\cdot,\cdot)$ are $H^1(\Omega)^2$-elliptic.*

Proof. First observe that for any constants c_1 and c_2

$$\Psi((c_1,c_2)) = \Psi^h((c_1,c_2)) = c_1(\mathrm{x}_1 - \langle \boldsymbol{\rho}^n \mathrm{x}_1, \mathbf{e}\rangle_n) + c_2(\mathrm{x}_2 - \langle \boldsymbol{\rho}^n \mathrm{x}_2, \mathbf{e}\rangle_n)$$

since both $H^1(\Omega)$ and $\mathbb{V}^h$ contain the linear functions. Without loss of generality we may suppose that the origin of the $\mathbf{x}$ variable has been chosen so that $\langle \boldsymbol{\rho}^n \mathrm{x}_1, \mathbf{e}\rangle_n = 0$ and $\langle \boldsymbol{\rho}^n \mathrm{x}_2, \mathbf{e}\rangle_n = 0$. In this case we have

$$\Psi((c_1,c_2)) = \Psi^h((c_1,c_2)) = c_1 \mathrm{x}_1 + c_2 \mathrm{x}_2. \tag{5.10}$$

Thus constants are mapped to linear functions via both $\Psi(\cdot)$ and $\Psi^h(\cdot)$.

For $\mathbf{v} = (v_1, v_2) \in H^1(\Omega)^2$ let $c_1 = \int_\Omega v_1 \, d\mathbf{x}$ and $c_2 = \int_\Omega v_2 \, d\mathbf{x}$. Then the Poincaré inequality implies that

$$\|\mathbf{v} - (c_1,c_2)\|^2_{L^2(\Omega)^2} \leq C|\mathbf{v}|^2_{H^1(\Omega)^2} \quad \text{and} \quad \|\mathbf{v}\|^2_{H^1(\Omega)^2} \leq C[c_1^2 + c_2^2 + |\mathbf{v}|^2_{H^1(\Omega)^2}].$$

Let $P = [\mathbf{x}^{(1)} \quad \cdots \quad \mathbf{x}^{(n)}]^T$. Since the points $\mathbf{x}^{(i)}$ are not collinear, the matrix P has full rank (of two), and the matrix $P^T P$ will be positive definite. Hence

$$c_1^2 + c_2^2 \leq C \frac{1}{n} [c_1 \ c_2] P^T P [c_1 \ c_2]^T = C \|\boldsymbol{\rho}^n \Psi((c_1,c_2))\|^2_n,$$

and so

$$\|\mathbf{v}\|^2_{H^1(\Omega)^2} \leq C[\|\boldsymbol{\rho}^n \Psi((c_1,c_2))\|^2_n + |\mathbf{v}|^2_{H^1(\Omega)^2}]$$

for some constant C (which can depend on h and n). Finally note that by the triangle inequality and the Poincaré inequality

$$\begin{aligned} \|\boldsymbol{\rho}^n\Psi((c_1,c_2))\|_n^2 &= \|\boldsymbol{\rho}^n\Psi((c_1,c_2)-\mathbf{v})+\boldsymbol{\rho}^n\Psi(\mathbf{v})\|_n^2 \\ &\le 2\|\boldsymbol{\rho}^n\Psi((c_1,c_2)-\mathbf{v})\|_n^2+2\|\boldsymbol{\rho}^n\Psi(\mathbf{v})\|_n^2 \\ &= 2\|(c_1,c_2)-\mathbf{v}\|_{H^1(\Omega)^2}^2+2\|\boldsymbol{\rho}^n\Psi(\mathbf{v})\|_n^2 \\ &\le C[|\mathbf{v}|_{H^1(\Omega)^2}^2+\|\boldsymbol{\rho}^n\Psi(\mathbf{v})\|_n^2], \end{aligned} \tag{5.11}$$

and consequently

$$\|\mathbf{v}\|_{H^1(\Omega)^2}^2 \le C[\|\boldsymbol{\rho}^n\Psi(\mathbf{v})\|_n^2+\alpha|\mathbf{v}|_{H^1(\Omega)^2}^2] = Ca_\alpha(\mathbf{v},\mathbf{v}).$$

The proof for a_α^h is identical. □

The following theorem is a direct application of a standard existence/uniqueness result for abstract variation problems; see [18, pp. 2–3].

THEOREM 5.3 (Existence-uniqueness). *Suppose that the values of the predictor variable are not collinear. Then there is a unique* $\mathbf{u}\in H^1(\Omega)^2$ *which solves Problems* 2 *and* 5. *Similarly, there is a unique* $\mathbf{u}^h\in\mathbb{V}^h\times\mathbb{V}^h$ *which solves Problems* 3 *and* 6.

Hence our H^1 and TPSFEM smoothing problems have unique solutions. They are not the standard thin plate splines, but, as we will prove in section 7, they do possess smoothing properties similar to the standard smoothing splines.

6. Auxilary results. We will now quote and prove a number of results which allow us to estimate the error due to approximating $\Psi(\mathbf{v})$ with $\Psi^h(\mathbf{v})$, and estimating the $L^2(\Omega)$ norms with $\mathbb{R}^n$ norms (and visa versa).

We quote an important result which follows from Utreras [28, Theorems 3.3 and 3.4] when applied to each $K\in\mathcal{T}^h$.

LEMMA 6.1 (Element norms). *Suppose that our finite element spaces satisfy Assumption* 1 *and our data satisfies Assumption* 5. *Then there exists a constant* $C_1>1$ *such that for any* $K\in\mathcal{T}^h$ *where* $h>C_1\,d$, *and for any* $f\in H^2(K)$, *the following bounds hold:*

$$d^2\sum_{\mathbf{x}^{(i)}\in K} f(\mathbf{x}^{(i)})^2 \preceq \|f\|_{L^2(K)}^2+d^4|f|_{H^2(K)}^2$$

and

$$\|f\|_{L^2(K)}^2 \preceq d^2\sum_{\mathbf{x}^{(i)}\in K} f(\mathbf{x}^{(i)})^2+d^4|f|_{H^2(K)}^2.$$

Proof. Results of this type, where the element K is replaced by small balls of radius $2d$, are shown in Utreras [28, Theorems 3.3 and 3.4]. Just as in Proposition 3.1, we can cover each element K with balls of radius $2d$ and sum to obtain our result. □

We are working with functions in $H^1(\Omega)$, which are not in $H^2(\Omega)$ but are controlled in $H^2(K)$ for each element K of a mesh. Hence we need the following generalization of Utreras' result.

LEMMA 6.2 (Pointwise norm equivalence). *Suppose that our finite element spaces satisfy Assumption* 1 *and our data satisfies Assumption* 5. *Then there exists a constant* $C_1>1$ *such that for any* $K\in\mathcal{T}^h$ *where* $h>C_1\,d$, *and for* $f\in H^1(\Omega)$ *such that* f *restricted to each* K *has bounded* $H^2(K)$ *norm, we have that*

$$\|\boldsymbol{\rho}^n f\|_n^2 \preceq \|f\|_{L^2(\Omega)}^2+d^4\sum_{K\in\mathcal{T}^h}|f|_{H^2(K)}^2$$

and

$$\|f\|^2_{L^2(\Omega)} \preceq \|\boldsymbol{\rho}^n f\|^2_n + d^4 \sum_{K\in\mathcal{T}^h} |f|^2_{H^2(K)}.$$

Proof. This result follows by summing the inequalities from Lemma 6.1 over all $K \in \mathcal{T}^h$ and observing that, since the data points are "uniformly spread," d^2 is comparable to $1/n$. □

With this lemma we can prove the following useful results.

LEMMA 6.3 (Ψ^h consistency). *Suppose that our finite element spaces satisfy Assumption* 1 *and our data satisfies Assumption* 5. *Then there exists a constant* $C_1 > 1$ *such that for* $h > C_1\, d$ *and for* $\mathbf{v} \in H^1(\Omega)^2$

$$\|\boldsymbol{\rho}^n\Psi^h(\mathbf{v}) - \boldsymbol{\rho}^n\Psi(\mathbf{v})\|^2_n \preceq (h^4 + d^4)\|\mathbf{v}\|^2_{H^1(\Omega)^2}.$$

Proof. By Lemma 6.2

$$\begin{aligned}&\|\boldsymbol{\rho}^n\Psi^h(\mathbf{v}) - \boldsymbol{\rho}^n\Psi(\mathbf{v})\|^2_n\\ &\preceq \|\Psi^h(\mathbf{v}) - \Psi(\mathbf{v})\|^2_{L^2(\Omega)} + d^4 \sum_{K\in\mathcal{T}^h} |\Psi^h(\mathbf{v}) - \Psi(\mathbf{v})|^2_{H^2(K)}.\end{aligned}$$

The first term is bounded by (5.6) and the second term by (5.8). Consequently

$$\|\boldsymbol{\rho}^n\Psi^h(\mathbf{v}) - \boldsymbol{\rho}^n\Psi(\mathbf{v})\|^2_n \preceq (h^4 + d^4)\|\Psi(\mathbf{v})\|^2_{H^2(\Omega)}.$$

By (5.4), $\|\Psi(\mathbf{v})\|_{H^2(\Omega)}$ is bounded by $\|\mathbf{v}\|^2_{H^1(\Omega)^2}$ so that

$$\|\boldsymbol{\rho}^n\Psi^h(\mathbf{v}) - \boldsymbol{\rho}^n\Psi(\mathbf{v})\|^2_n \preceq (h^4 + d^4)\|\mathbf{v}\|^2_{H^1(\Omega)^2}. \qquad \square$$

LEMMA 6.4 (Ψ stability). *Suppose that our finite element spaces satisfy Assumption* 1 *and our data satisfies Assumption* 5. *Then there exists a constant* $C_1 > 1$ *such that for* $h > C_1\, d$ *and for* $\mathbf{v} \in H^1(\Omega)^2$

$$\|\boldsymbol{\rho}^n\Psi(\mathbf{v}^h) - \boldsymbol{\rho}^n\Psi(\mathbf{v})\|^2_n \preceq (h^4 + d^4)\|\mathbf{v}\|^2_{H^1(\Omega)^2},$$

where $\mathbf{v}^h = Q^h\mathbf{v}$*, the projection operator referred to in Assumptions* 2 *and* 3, *has been applied to each component of* $\mathbf{v}$.

Proof. By Lemma 6.2

$$\begin{aligned}&\|\boldsymbol{\rho}^n\Psi(\mathbf{v}^h) - \boldsymbol{\rho}^n\Psi(\mathbf{v})\|^2_n\\ &\preceq \|\Psi(\mathbf{v}^h - \mathbf{v})\|^2_{L^2(\Omega)} + d^4|\Psi(\mathbf{v}^h - \mathbf{v})|^2_{H^2(\Omega)}.\end{aligned}$$

Standard regularity results for second order elliptic equations provide the bounds

$$\begin{aligned}\|\Psi(\mathbf{v}^h - \mathbf{v})\|^2_{L^2(\Omega)} &\preceq \|\mathbf{v}^h - \mathbf{v}\|^2_{H^{-1}(\Omega)^2},\\ |\Psi(\mathbf{v}^h - \mathbf{v})|^2_{H^2(\Omega)} &\preceq \|\mathbf{v}^h - \mathbf{v}\|^2_{H^1(\Omega)^2}.\end{aligned}$$

Assumption 3 on page 213 implies that

$$\begin{aligned}\|\mathbf{v}^h - \mathbf{v}\|^2_{H^{-1}(\Omega)^2} &\preceq h^4\|\mathbf{v}\|^2_{H^1(\Omega)^2},\\ \|\mathbf{v}^h - \mathbf{v}\|^2_{H^1(\Omega)^2} &\preceq \|\mathbf{v}\|^2_{H^1(\Omega)^2}.\end{aligned}$$

Consequently

$$\|\boldsymbol{\rho}^n\Psi(\mathbf{v}^h) - \boldsymbol{\rho}^n\Psi(\mathbf{v})\|_n^2 \preceq (h^4 + d^4)\|\mathbf{v}\|_{H^1(\Omega)^2}^2. \qquad \square$$

Lemma 6.5 ($H^1(\Omega)^2$ norm equivalence). *Suppose that our finite element spaces satisfy Assumption* 1 *and our data satisfies Assumption* 5. *Then there exists a constant* $C_1 > 1$ *such that for* $h > C_1\, d$ *and for* $\mathbf{v} \in H^1(\Omega)^2$

$$\|\mathbf{v}\|_{H^1(\Omega)^2}^2 \preceq \|\boldsymbol{\rho}^n\Psi^h(\mathbf{v})\|_n^2 + |\mathbf{v}|_{H^1(\Omega)^2}^2,$$

where the implied constant is independent of the data points.

Proof. In the previous section we proved this result for a fixed choice of data points. Now we want to show that the constant is independent of the data points, provided that they satisfy our "uniformly spread" assumption, Assumption 5.

We know that

$$\|\mathbf{v}\|_{H^1(\Omega)^2}^2 \preceq c_1^2 + c_2^2 + |\mathbf{v}|_{H^1(\Omega)^2}^2,$$

where $c_1 = \int_\Omega v_1 \, d\mathbf{x}$ and $c_2 = \int_\Omega v_2 \, d\mathbf{x}$. By explicit calculation,

$$c_1^2 + c_2^2 \preceq \int_\Omega (c_1\mathrm{x}_1 + c_2\mathrm{x}_2)^2 d\mathbf{x}.$$

Lemma 6.2 implies that

$$\int_\Omega (c_1\mathrm{x}_1 + c_2\mathrm{x}_2)^2 d\mathbf{x} = \|c_1\mathrm{x}_1 + c_2\mathrm{x}_2\|_{L^2(\Omega)}^2 \preceq \|\boldsymbol{\rho}^n\Psi^h((c_1, c_2))\|_n^2$$

since the function $c_1\mathrm{x}_1 + c_2\mathrm{x}_2$ is linear and so $\|c_1\mathrm{x}_1 + c_2\mathrm{x}_2\|_{H^2(\Omega)} = 0$. Using a similar argument as in (5.11), we conclude that

$$\|\boldsymbol{\rho}^n\Psi^h((c_1, c_2))\|_n^2 \preceq \|\boldsymbol{\rho}^n\Psi^h(\mathbf{v})\|_n^2 + |\mathbf{v}|_{H^1(\Omega)^2}^2.$$

Consequently

$$c_1^2 + c_2^2 \preceq \|\boldsymbol{\rho}^n\Psi^h(\mathbf{v})\|_n^2 + |\mathbf{v}|_{H^1(\Omega)^2}^2,$$

and so

$$\|\mathbf{v}\|_{H^1(\Omega)^2}^2 \preceq \|\boldsymbol{\rho}^n\Psi^h(\mathbf{v})\|_n^2 + |\mathbf{v}|_{H^1(\Omega)^2}^2. \qquad \square$$

Lemma 6.6 (Energy orthogonality). *For any* $\mathbf{y} \in \mathbb{R}^n$,

1. *For all* $s \in H^2(\Omega)$

$$\bar{J}_\alpha(s, \mathbf{y}) = \bar{J}_\alpha(\bar{s}_\alpha(\mathbf{y}), \mathbf{y}) + \bar{a}_\alpha(\bar{s}_\alpha(\mathbf{y}) - s, \bar{s}_\alpha(\mathbf{y}) - s).$$

2. *For all* $\mathbf{v} \in H^1(\Omega)^2$

$$J_\alpha(\mathbf{v}, \mathbf{y}) = J_\alpha(\mathbf{u}_\alpha(\mathbf{y}), \mathbf{y}) + a_\alpha(\mathbf{u}_\alpha(\mathbf{y}) - \mathbf{v}, \mathbf{u}_\alpha(\mathbf{y}) - \mathbf{v}).$$

3. *For all* $\mathbf{v}^h \in \mathbb{V}^h \times \mathbb{V}^h$

$$J_\alpha^h(\mathbf{v}^h, \mathbf{y}) = J_\alpha^h(\mathbf{u}_\alpha^h(\mathbf{y}), \mathbf{y}) + a_\alpha^h(\mathbf{u}_\alpha^h(\mathbf{y}) - \mathbf{v}^h, \mathbf{u}_\alpha^h(\mathbf{y}) - \mathbf{v}^h).$$

Proof. Expand $\bar{J}_\alpha(s - \bar{s}_\alpha(\mathbf{y}) + \bar{s}_\alpha(\mathbf{y}), \mathbf{y})$, $J_\alpha(\mathbf{v} - \mathbf{u}_\alpha(\mathbf{y}) + \mathbf{u}_\alpha(\mathbf{y}), \mathbf{y})$, and $J_\alpha^h(\mathbf{v}^h - \mathbf{u}_\alpha^h(\mathbf{y}) + \mathbf{u}_\alpha^h(\mathbf{y}), \mathbf{y})$, and use (5.1), (5.2), and (5.3), respectively, to eliminate cross terms. $\square$

7. Convergence of the method. In this section we will prove the main convergence result, Theorem 3.2. For a function $f \in H^2(\Omega)$, we will use the notation $\mathbf{u}_\alpha(f)$ to denote $\mathbf{u}_\alpha(\boldsymbol{\rho}^n f)$. Similarly we define $\mathbf{u}_\alpha^h(f)$, $s_\alpha(f)$, and $s_\alpha^h(f)$.

Recall that we assume that the response values satisfy Assumption 4 on page 213, namely,

$$\mathbf{y} = \boldsymbol{\rho}^n f + \boldsymbol{\nu},$$

where $\boldsymbol{\nu} = [\nu^{(1)} \ \cdots \ \nu^{(n)}]^T$ and the $\nu^{(i)}$'s are independent identically distributed random variables with zero mean and variance σ^2.

It is clear that $\mathbf{u}_\alpha^h$ and s_α^h are linear operators, and so

$$\begin{aligned} \mathbf{u}_\alpha^h(\mathbf{y}) &= \mathbf{u}_\alpha^h(f) + \mathbf{u}_\alpha^h(\boldsymbol{\nu}), \\ s_\alpha^h(\mathbf{y}) &= s_\alpha^h(f) + s_\alpha^h(\boldsymbol{\nu}). \end{aligned}$$

We want to measure how well $s_\alpha^h(\mathbf{y})$ approximates f. Using the linearity of the operators, we can separate the errors into those due to bias and those due to variance. In particular we have

$$\|s_\alpha^h(\mathbf{y}) - f\|_{L^2(\Omega)} \le \|s_\alpha^h(f) - f\|_{L^2(\Omega)} + \|s_\alpha^h(\boldsymbol{\nu})\|_{L^2(\Omega)}, \tag{7.1}$$

$$|s_\alpha^h(\mathbf{y}) - f|_{H^1(\Omega)} \le |s_\alpha^h(f) - f|_{H^1(\Omega)} + |s_\alpha^h(\boldsymbol{\nu})|_{H^1(\Omega)}, \tag{7.2}$$

$$\|\mathbf{u}_\alpha^h(\mathbf{y})\|_{H^1(\Omega)} \le \|\mathbf{u}_\alpha^h(f)\|_{H^1(\Omega)} + \|\mathbf{u}_\alpha^h(\boldsymbol{\nu})\|_{H^1(\Omega)}. \tag{7.3}$$

To produce our estimates we will use the following result proved by Utreras for the exact thin plate spline.

THEOREM 7.1 (see Utreras [28, equations (6.2) and (6.4)]). *Suppose that the data satisfies Assumptions* 4 *and* 5. *Then there exists a constant* $\alpha_0 > 0$ *such that*

$$E\bar{a}_\alpha(\bar{s}(\boldsymbol{\nu}), \bar{s}(\boldsymbol{\nu})) \preceq \sigma^2 \frac{d^2}{\alpha^{1/2}} \tag{7.4}$$

for $d^4 < \alpha < \alpha_0$.

Note that we have written the bound in terms of d instead of n by using the result $n^{-1} \preceq d^2$.

Throughout this section we will be making the same basic assumption about the allowed range of values for h and α. We quantify this assumption as follows.

ASSUMPTION 6 (Constraints on h and α). *Let* C_1 *and* α_0 *be the constants introduced in Proposition* 3.1 *and Theorem* 7.1, *respectively. We assume that* h *and* α *satisfy* $h > C_1\, d$ *and* $d^4 + h^4 < \alpha < \alpha_0$.

We can now prove convergence of our method by first looking at exact data (the bias term) and then estimating the variance.

7.1. Exact response variable data. We deal first with the error for exact data. Using an argument almost identical to that found in Utreras [28, Theorem 4.1], we will show that the value of the functional $J_\alpha^h(\mathbf{u}_\alpha^h(f), \boldsymbol{\rho}^n f)$ can be controlled by the smoothness of the underlying model function f.

PROPOSITION 7.2 (Energy bound). *Let the finite element spaces satisfy Assumptions* 1, 2, *and* 3, *the data satisfy Assumptions* 4 *and* 5, *and* h *and* α *satisfy Assumption* 6. *Then for all* $f \in H^2(\Omega)$

$$J_\alpha^h(\mathbf{u}_\alpha^h(f), \boldsymbol{\rho}^n f) \preceq (\alpha + h^4 + d^4)\|f\|_{H^2(\Omega)}.$$

Proof. Let $\mathbf{v} = \boldsymbol{\nabla} f$. Then $f = \Phi(\mathbf{v}, \boldsymbol{\rho}^n f)$. Let $\mathbf{v}^h = Q^h \mathbf{v}$. As $\mathbf{u}_\alpha^h(f)$ is the minimizer of $J_\alpha^h(\cdot, \boldsymbol{\rho}^n f)$, we conclude that

$$J_\alpha^h(\mathbf{u}_\alpha^h(f), \boldsymbol{\rho}^n f) \le J_\alpha^h(\mathbf{v}^h, \boldsymbol{\rho}^n f).$$

Now

$$\begin{aligned} J_\alpha^h(\mathbf{v}^h, \boldsymbol{\rho}^n f) &= \|\boldsymbol{\rho}^n \Phi^h(\mathbf{v}^h, \boldsymbol{\rho}^n f) - \boldsymbol{\rho}^n f\|_n^2 + \alpha |\mathbf{v}^h|_{H^1(\Omega)^2}^2 \\ &\preceq \|\boldsymbol{\rho}^n \Phi^h(\mathbf{v}^h, \boldsymbol{\rho}^n f) - \boldsymbol{\rho}^n \Phi(\mathbf{v}, \boldsymbol{\rho}^n f)\|_n^2 \\ &\quad + \|\boldsymbol{\rho}^n \Phi(\mathbf{v}, \boldsymbol{\rho}^n f) - \boldsymbol{\rho}^n f\|_n^2 + \alpha |\mathbf{v}^h|_{H^1(\Omega)^2}^2 \\ &\preceq \|\boldsymbol{\rho}^n \Psi^h(\mathbf{v}^h) - \boldsymbol{\rho}^n \Psi(\mathbf{v})\|_n^2 \\ &\quad + \|\boldsymbol{\rho}^n \Phi(\mathbf{v}, \boldsymbol{\rho}^n f) - \boldsymbol{\rho}^n f\|_n^2 + \alpha |\mathbf{v}^h|_{H^1(\Omega)^2}^2. \end{aligned}$$

Using Lemmas 6.3 and 6.4, a simple triangle inequality argument bounds the first term, namely,

$$\begin{aligned} \|\boldsymbol{\rho}^n \Psi^h(\mathbf{v}^h) - \boldsymbol{\rho}^n \Psi(\mathbf{v})\|_n^2 &\preceq \|\boldsymbol{\rho}^n \Psi^h(\mathbf{v}^h) - \boldsymbol{\rho}^n \Psi(\mathbf{v}^h)\|_n^2 + \|\boldsymbol{\rho}^n \Psi(\mathbf{v}^h) - \boldsymbol{\rho}^n \Psi(\mathbf{v})\|_n^2 \\ &\preceq (h^4 + d^4)[\|\mathbf{v}^h\|_{H^1(\Omega)^2}^2 + \|\mathbf{v}\|_{H^1(\Omega)^2}^2] \\ &\preceq (h^4 + d^4)\|\mathbf{v}\|_{H^1(\Omega)^2}^2. \end{aligned}$$

The term $\|\boldsymbol{\rho}^n \Phi(\mathbf{v}, \boldsymbol{\rho}^n f) - \boldsymbol{\rho}^n f\|_n$ is zero since $\Phi(\mathbf{v}, \boldsymbol{\rho}^n f) = f$. Taking this together, we conclude that

$$\begin{aligned} J_\alpha^h(\mathbf{v}^h, \boldsymbol{\rho}^n f) &\preceq (h^4 + d^4)\|\mathbf{v}\|_{H^1(\Omega)^2}^2 + \alpha |\mathbf{v}^h|_{H^1(\Omega)^2}^2 \\ &\preceq (\alpha + h^4 + d^4)\|f\|_{H^2(\Omega)}^2, \end{aligned}$$

since Assumption 2 and the triangle inequality imply that

$$\|\mathbf{v}^h\|_{H^1(\Omega)^2} \preceq \|\mathbf{v}\|_{H^1(\Omega)^2},$$

and, by the definition of $\mathbf{v}$,

$$\|\mathbf{v}\|_{H^1(\Omega)^2} = \|\boldsymbol{\nabla} f\|_{H^1(\Omega)^2} \le \|f\|_{H^2(\Omega)}. \qquad \square$$

We have proved that

$$\|\boldsymbol{\rho}^n s_\alpha^h(f) - \boldsymbol{\rho}^n f\|_n^2 + \alpha |\mathbf{u}_\alpha^h(f)|_{H^1(\Omega)^2}^2 \preceq (\alpha + h^4 + d^4)\|f\|_{H^2(\Omega)}^2.$$

Now we can obtain $L^2(\Omega)$ and $H^1(\Omega)$ norm estimates for the error for exact data.

THEOREM 7.3 (Convergence: exact data). *We suppose that our finite element spaces satisfy Assumptions* 1, 2, *and* 3, *our data satisfies Assumptions* 4 *and* 5, *and* h *and* α *satisfy Assumption* 6. *Then for all* $f \in H^2(\Omega)$

$$\|s_\alpha^h(f) - f\|_{L^2(\Omega)}^2 \preceq (\alpha + h^4 + d^4)\|f\|_{H^2(\Omega)}^2, \tag{7.5}$$

$$|s_\alpha^h(f) - f|_{H^1(\Omega)}^2 \preceq \frac{1}{\alpha^{1/2}}(\alpha + h^4 + d^4)\|f\|_{H^2(\Omega)}^2, \tag{7.6}$$

$$\|\mathbf{u}_\alpha^h(f)\|_{H^1(\Omega)^2}^2 \preceq \frac{1}{\alpha}(\alpha + h^4 + d^4)\|f\|_{H^2(\Omega)}^2. \tag{7.7}$$

Proof. First we prove (7.7). Proposition 7.2 implies that

$$|\mathbf{u}_\alpha^h(f)|_{H^1(\Omega)^2}^2 \preceq \frac{1}{\alpha}(\alpha + h^4 + d^4)\|f\|_{H^2(\Omega)}^2 \tag{7.8}$$

and

$$\|\boldsymbol{\rho}^n \Phi^h(\mathbf{u}^h_\alpha(f), \boldsymbol{\rho}^n f) - \boldsymbol{\rho}^n f\|_n^2 \preceq (\alpha + h^4 + d^4)\|f\|^2_{H^2(\Omega)}. \tag{7.9}$$

Lemma 6.5, together with (7.8), provides the bound

$$\begin{aligned}\|\mathbf{u}^h_\alpha(f)\|^2_{H^1(\Omega)^2} &\preceq \|\boldsymbol{\rho}^n \Psi^h(\mathbf{u}^h(f))\|_n^2 + |\mathbf{u}^h_\alpha(f)|^2_{H^1(\Omega)^2}\\ &\preceq \|\boldsymbol{\rho}^n \Psi^h(\mathbf{u}^h(f))\|_n^2 + \frac{1}{\alpha}(\alpha + h^4 + d^4)\|f\|^2_{H^2(\Omega)}.\end{aligned}$$

Using the bound (7.9), the Sobolev inequality (which implies that the pointwise norm $\|\boldsymbol{\rho}^n \cdot\|_n$ is bounded by the norm $\|\cdot\|_{H^2(\Omega)}$) and a triangle inequality estimate give us

$$\begin{aligned}\|\boldsymbol{\rho}^n \Psi^h(\mathbf{u}^h(f))\|_n^2 &\preceq \|\boldsymbol{\rho}^n \Phi^h(\mathbf{u}^h(f), \boldsymbol{\rho}^n f) - \boldsymbol{\rho}^n f - \langle \boldsymbol{\rho}^n f, \mathbf{e}\rangle_n \mathbf{e} + \boldsymbol{\rho}^n f\|_n^2\\ &\preceq \|\boldsymbol{\rho}^n \Phi^h(\mathbf{u}^h(f), \boldsymbol{\rho}^n f) - \boldsymbol{\rho}^n f\|_n^2 + \|\langle \boldsymbol{\rho}^n f, \mathbf{e}\rangle_n \mathbf{e}\|_n^2 + \|\boldsymbol{\rho}^n f\|_n^2\\ &\preceq (\alpha + h^4 + d^4)\|f\|^2_{H^2(\Omega)} + \|\boldsymbol{\rho}^n f\|_n^2 \preceq \frac{1}{\alpha}(\alpha + h^4 + d^4)\|f\|^2_{H^2(\Omega)}.\end{aligned}$$

Consequently

$$\|\mathbf{u}^h_\alpha(f)\|^2_{H^1(\Omega)^2} \preceq \frac{1}{\alpha}(\alpha + h^4 + d^4)\|f\|^2_{H^2(\Omega)}.$$

Equation (7.5) follows from Lemma 6.2,

$$\begin{aligned}&\|s^h_\alpha(f) - f\|^2_{L^2(\Omega)} = \|\Phi^h(\mathbf{u}^h_\alpha(f), \boldsymbol{\rho}^n f) - f\|^2_{L^2(\Omega)}\\ &\qquad\preceq \|\boldsymbol{\rho}^n \Phi^h(\mathbf{u}^h_\alpha(f), \boldsymbol{\rho}^n f) - \boldsymbol{\rho}^n f\|_n^2 + d^4 \sum_{K \in \mathcal{T}^h} |\Phi^h(\mathbf{u}^h_\alpha(f), \boldsymbol{\rho}^n f) - f|^2_{H^2(K)},\end{aligned}$$

which by (7.9) and the triangle inequality implies

$$\begin{aligned}&\preceq (\alpha + h^4 + d^4)\alpha\|f\|^2_{H^2(\Omega)} + d^4 \sum_{K \in \mathcal{T}^h} \Big[|\Phi^h(\mathbf{u}^h_\alpha(f), \boldsymbol{\rho}^n f)|^2_{H^2(K)} + |f|^2_{H^2(K)}\Big]\\ &\preceq (\alpha + h^4 + d^4)\|f\|^2_{H^2(\Omega)},\end{aligned}$$

since by (5.8)

$$\sum_{K \in \mathcal{T}^h} |\Psi^h(\mathbf{u}^h_\alpha(f))|^2_{H^2(K)} \preceq \|\Psi(\mathbf{u}^h_\alpha(f))\|^2_{H^2(\Omega)} \preceq \|f\|^2_{H^2(\Omega)}.$$

Finally, (7.6) will follow from an interpolation result which estimates the H^1 norm of $s^h_\alpha(f) - f$ by use of the preceding L^2 estimate and an H^2 estimate. Unfortunately $s^h_\alpha(f) = \Phi^h(\mathbf{u}^h_\alpha(f), \boldsymbol{\rho}^n f)$ is not in $H^2(\Omega)$. On the other hand, $\Phi(\mathbf{u}^h_\alpha(f), \boldsymbol{\rho}^n f)$ is in $H^2(\Omega)$. We will first use an interpolation result to estimate $|\Phi(\mathbf{u}^h_\alpha(f), \boldsymbol{\rho}^n f) - f|^2_{H^1(\Omega)}$ in terms of $\|\Phi(\mathbf{u}^h_\alpha(f), \boldsymbol{\rho}^n f) - f\|^2_{L^2(\Omega)}$ and $|\Phi(\mathbf{u}^h_\alpha(f), \boldsymbol{\rho}^n f) - f|^2_{H^2(\Omega)}$. Now

$$\begin{aligned}&\|\Phi(\mathbf{u}^h_\alpha(f), \boldsymbol{\rho}^n f) - f\|^2_{L^2(\Omega)}\\ &\preceq \|\Psi(\mathbf{u}^h_\alpha(f)) - \Psi^h(\mathbf{u}^h_\alpha(f))\|^2_{L^2(\Omega)} + \|s^h_\alpha(f) - f\|^2_{L^2(\Omega)}\\ &\preceq (\alpha + h^4 + d^4)\|f\|^2_{H^2(\Omega)}\end{aligned}$$

and

$$|\Phi(\mathbf{u}_\alpha^h(f), \boldsymbol{\rho}^n f) - f|^2_{H^2(\Omega)} \preceq \frac{1}{\alpha}(\alpha + h^4 + d^4)\|f\|^2_{H^2(\Omega)}.$$

An interpolation estimate (see, for example, [1, Theorem 4.14]) then implies that

$$|\Phi(\mathbf{u}_\alpha^h(f), \boldsymbol{\rho}^n f) - f|^2_{H^1(\Omega)} \preceq \frac{1}{\alpha^{1/2}}(\alpha + h^4 + d^4)\|f\|^2_{H^2(\Omega)}.$$

Now we can estimate the quantity $|\Phi^h(\mathbf{u}_\alpha^h(f), \boldsymbol{\rho}^n f) - f|^2_{H^1(\Omega)}$. Equation (5.7), together with the previous estimate, implies that

$$\begin{aligned} |s_\alpha^h(f) - f|^2_{H^1(\Omega)} &= |\Phi^h(\mathbf{u}_\alpha^h(f), \boldsymbol{\rho}^n f) - f|^2_{H^1(\Omega)} \\ &\preceq |\Psi^h(\mathbf{u}_\alpha^h(f)) - \Psi(\mathbf{u}_\alpha^h(f))|^2_{H^1(\Omega)} + |\Phi(\mathbf{u}_\alpha^h(f), \boldsymbol{\rho}^n f) - f|^2_{H^1(\Omega)} \\ &\preceq \left(h^2 + \frac{1}{\alpha^{1/2}}(\alpha + h^4 + d^4)\right)\|f\|^2_{H^2(\Omega)} \\ &\preceq \frac{1}{\alpha^{1/2}}(\alpha + h^4 + d^4)\|f\|^2_{H^2(\Omega)} \end{aligned}$$

(since $h^2 < \alpha^{1/2}$), which provides the required $H^1(\Omega)$ seminorm estimate for $s_\alpha^h(f) - f$. □

7.2. Variance of the error. We will now bound

$$E|s^h(\mathbf{u}^h(\boldsymbol{\nu}))|^2_{L^2(\Omega)}, \quad E|s^h(\mathbf{u}^h(\boldsymbol{\nu}))|^2_{H^1(\Omega)}, \quad \text{and} \quad E|\mathbf{u}^h(\boldsymbol{\nu})|^2_{H^1(\Omega)},$$

the expected error due to the random part of the response variable data.

The idea is to use Utreras' result for the thin plate spline, namely Theorem 7.1, and in particular the bound (7.4), to bound $Ea_\alpha^h(\mathbf{u}^h(\boldsymbol{\nu}), \mathbf{u}^h(\boldsymbol{\nu}))$. Observe that $\boldsymbol{\nu} = \sum_{k=1}^n \nu^{(k)}\mathbf{e}_k$, where $\mathbf{e}_k$ is the kth coordinate vector. Using this expansion, we observe that

$$\begin{aligned} a_\alpha^h(\mathbf{u}_\alpha^h(\boldsymbol{\nu}), \mathbf{u}_\alpha^h(\boldsymbol{\nu})) &= a^h\left(\sum_{k=1}^n \nu^{(k)}\mathbf{u}_\alpha^h(\mathbf{e}_k), \sum_{j=1}^n \nu^{(j)}\mathbf{u}_\alpha^h(\mathbf{e}_k)\right) \\ &= \sum_{k,j=1}^n \nu^{(k)}\nu^{(j)} a^h(\mathbf{u}_\alpha^h(\mathbf{e}_k), \mathbf{u}_\alpha^h(\mathbf{e}_j)). \end{aligned}$$

Hence

$$\begin{aligned} Ea_\alpha^h(\mathbf{u}_\alpha^h(\boldsymbol{\nu}), \mathbf{u}_\alpha^h(\boldsymbol{\nu})) &= \sum_{k,j=1}^n E(\nu^{(k)}\nu^{(j)}) a_\alpha^h(\mathbf{u}_\alpha^h(\mathbf{e}_k), \mathbf{u}_\alpha^h(\mathbf{e}_j)) \\ &= \sum_{k=1}^n \sigma^2 a_\alpha^h(\mathbf{u}_\alpha^h(\mathbf{e}_k), \mathbf{u}_\alpha^h(\mathbf{e}_k)) \end{aligned} \tag{7.10}$$

since the $\nu^{(k)}$ are independent random variables with zero mean and variance σ^2. Similarly the thin plate spline function $\bar{s}_\alpha(\boldsymbol{\nu})$ satisfies

$$E\bar{a}_\alpha(\bar{s}_\alpha(\boldsymbol{\nu}), \bar{s}_\alpha(\boldsymbol{\nu})) = \sum_k^n \sigma^2 \bar{a}_\alpha(\bar{s}_\alpha(\mathbf{e}_k), \bar{s}_\alpha(\mathbf{e}_k)). \tag{7.11}$$

We will bound $a^h_\alpha(\mathbf{u}^h_\alpha(\mathbf{e}_k), \mathbf{u}^h_\alpha(\mathbf{e}_k))$ in terms of $\bar{a}_\alpha(\bar{s}_\alpha(\mathbf{e}_k), \bar{s}_\alpha(\mathbf{e}_k))$. To do this, note that

$$\begin{aligned} \frac{1}{n} - \frac{1}{n^2} &= \|\mathbf{e}_k - \langle \mathbf{e}_k, \mathbf{e}\rangle_n \mathbf{e}\|_n^2 \\ &= \bar{J}(\bar{s}_\alpha(\mathbf{e}_k), \mathbf{e}_k) + \bar{a}_\alpha(\bar{s}_\alpha(\mathbf{e}_k), \bar{s}_\alpha(\mathbf{e}_k)) \\ &= J_\alpha(\mathbf{u}_\alpha(\mathbf{e}_k), \mathbf{e}_k) + a_\alpha(\mathbf{u}_\alpha(\mathbf{e}_k), \mathbf{u}_\alpha(\mathbf{e}_k)) \\ &= J^h_\alpha(\mathbf{u}^h_\alpha(\mathbf{e}_k), \mathbf{e}_k) + a^h_\alpha(\mathbf{u}^h_\alpha(\mathbf{e}_k), \mathbf{u}^h_\alpha(\mathbf{e}_k)). \end{aligned} \tag{7.12}$$

This follows from Lemma 6.6 with $s = \langle \mathbf{e}_k, \mathbf{e}\rangle_n$ and $\mathbf{v} = \mathbf{v}^h = 0$. Essentially (7.12) shows that a result like

$$a^h_\alpha(\mathbf{u}^h_\alpha(\mathbf{e}_k), \mathbf{u}^h_\alpha(\mathbf{e}_k)) \le \bar{a}_\alpha(\bar{s}_\alpha(\mathbf{e}_k), \bar{s}_\alpha(\mathbf{e}_k))$$

holds if an opposite inequality of the form

$$\bar{J}_\alpha(\bar{s}_\alpha(\mathbf{e}_k), \mathbf{e}_k) \le J^h_\alpha(\mathbf{u}^h_\alpha(\mathbf{e}_k), \mathbf{e}_k)$$

holds. In fact we will be able to prove the following slightly weaker results.

LEMMA 7.4 (Energy bound). *We suppose that our finite element spaces satisfy Assumptions* 1, 2, *and* 3, *our data satisfies Assumptions* 4 *and* 5, *and h and α satisfy Assumption* 6. *Then*

$$J_\alpha(\mathbf{u}_\alpha(\mathbf{e}_k), \mathbf{e}_k) \le \left(1 + \frac{h^2}{\alpha^{1/2}}\right) J^h_\alpha(\mathbf{u}^h_\alpha(\mathbf{e}_k), \mathbf{e}_k) + \frac{C_1}{n\alpha^{1/2}} \frac{(h^4 + d^4)}{h^2}$$

for some constant C_1 independent of h, d, and α.

LEMMA 7.5 (Thin plate spline lower bound). *We suppose that our finite element spaces satisfy Assumptions* 1, 2, *and* 3, *our data satisfies Assumptions* 4 *and* 5, *and h and α satisfy Assumption* 6. *Then there exists a constant $C_2 > 0$ such that for $\alpha' = C_2\alpha$*

$$\bar{J}_{\alpha'}(\bar{s}_{\alpha'}(\mathbf{e}_k), \mathbf{e}_k) \le J_\alpha(\mathbf{u}_\alpha(\mathbf{e}_k), \mathbf{e}_k).$$

Before proving these lemmas, let us use them to show the following.

PROPOSITION 7.6 (Thin plate spline bound). *We suppose that our finite element spaces satisfy Assumptions* 1, 2, *and* 3, *our data satisfies Assumptions* 4 *and* 5, *and h and α satisfy Assumption* 6. *Then there exists a constant $C_2 > 0$ such that for $\alpha' = C_2\alpha$*

$$a^h_\alpha(\mathbf{u}^h_\alpha(\mathbf{e}_k), \mathbf{u}^h_\alpha(\mathbf{e}_k)) \le \bar{a}_{\alpha'}(\bar{s}_{\alpha'}(\mathbf{e}_k), \bar{s}_{\alpha'}(\mathbf{e}_k)) + \frac{C_3}{n\alpha^{1/2}} \frac{(h^4 + d^4)}{h^2}$$

for some constant C_3.

Proof. Combining (7.12), Lemma 7.4, and Lemma 7.5, we conclude that

$$\begin{aligned} &\left(1 + \frac{h^2}{\alpha^{1/2}}\right) a^h_\alpha(\mathbf{u}^h_\alpha(\mathbf{e}_k), \mathbf{u}^h_\alpha(\mathbf{e}_k)) \\ &\le \bar{a}_{\alpha'}(\bar{s}_{\alpha'}(\mathbf{e}_k), \bar{s}_{\alpha'}(\mathbf{e}_k)) + \left(\frac{1}{n} - \frac{1}{n^2}\right)\left(\frac{h^2}{\alpha^{1/2}}\right) + \frac{C_1}{n\alpha^{1/2}} \frac{(h^4 + d^4)}{h^2}, \end{aligned}$$

where C_1 is defined in Lemma 7.4. Our assumption on h ensures that $h > d$. So we conclude that for $C_3 = C_1 + 1$,

$$a_\alpha^h(\mathbf{u}_\alpha^h(\mathbf{e}_k), \mathbf{u}_\alpha^h(\mathbf{e}_k)) \leq \bar{a}_{\alpha'}(\bar{s}_{\alpha'}(\mathbf{e}_k), \bar{s}_{\alpha'}(\mathbf{e}_k)) + \frac{C_3}{n\alpha^{1/2}}\frac{(h^4+d^4)}{h^2},$$

as required. □

Now we can estimate $Ea_\alpha^h(\mathbf{u}_\alpha^h(\boldsymbol{\nu}), \mathbf{u}_\alpha^h(\boldsymbol{\nu}))$.

Theorem 7.7 (Energy norm convergence: random data). *We suppose that our finite element spaces satisfy Assumptions* 1, 2, *and* 3, *our data satisfies Assumptions* 4 *and* 5, *and h and α satisfy Assumption* 6. *Then*

$$Ea_\alpha^h(\mathbf{u}_\alpha^h(\boldsymbol{\nu}), \mathbf{u}_\alpha^h(\boldsymbol{\nu})) \preceq \frac{\sigma^2}{\alpha^{1/2}}\frac{(h^4+d^4)}{h^2}.$$

Proof. By summing the inequalities in the statement of Proposition 7.6 over k and noting (7.10) and (7.11), we arrive at the conclusion that

$$Ea_\alpha^h(\mathbf{u}_\alpha^h(\boldsymbol{\nu}), \mathbf{u}_\alpha^h(\boldsymbol{\nu})) \leq E\bar{a}_{\alpha'}(\bar{s}_{\alpha'}(\mathbf{e}_k), \bar{s}_{\alpha'}(\mathbf{e}_k)) + \frac{C_3\sigma^2}{\alpha^{1/2}}\frac{(h^4+d^4)}{h^2}.$$

Equation (7.4) then implies that there exists a constant C_4 such that

$$E\bar{a}_{\alpha'}(\bar{s}_{\alpha'}(\mathbf{e}_k), \bar{s}_{\alpha'}(\mathbf{e}_k)) \leq C_4\sigma^2\frac{d^2}{\alpha'^{1/2}} \leq \frac{C_4\sigma^2}{C_2^{1/2}\alpha^{1/2}}d^2 \preceq \frac{\sigma^2}{\alpha^{1/2}}\frac{(h^4+d^4)}{h^2}.$$

Putting this together, the required result is obtained. □

Note that we have proved that

$$E(\|\boldsymbol{\rho}^n\Psi^h(\mathbf{u}_\alpha^h(\boldsymbol{\nu}))\|_n^2 + \alpha|\mathbf{u}_\alpha^h(\boldsymbol{\nu})|_{H^1(\Omega)^2}^2) \preceq \frac{\sigma^2}{\alpha^{1/2}}\frac{(h^4+d^4)}{h^2}.$$

We can now obtain $L^2(\Omega)$ and $H^1(\Omega)$ estimates of the error.

Theorem 7.8 (Convergence: random data). *We suppose that our finite element spaces satisfy Assumptions* 1, 2, *and* 3, *our data satisfies Assumptions* 4 *and* 5, *and h and α satisfy Assumption* 6. *Then*

$$\begin{aligned} E\|s_\alpha^h(\boldsymbol{\nu})\|_{L^2(\Omega)}^2 &\preceq \frac{\sigma^2}{\alpha^{1/2}}\frac{(h^4+d^4)}{h^2}, \\ E|s_\alpha^h(\boldsymbol{\nu})|_{H^1(\Omega)}^2 &\preceq \frac{\sigma^2}{\alpha}\frac{(h^4+d^4)}{h^2}, \\ E\|\mathbf{u}_\alpha^h(\boldsymbol{\nu})\|_{H^1(\Omega)^2}^2 &\preceq \frac{\sigma^2}{\alpha^{3/2}}\frac{(h^4+d^4)}{h^2}. \end{aligned}$$

Proof. Just as in the proof of Theorem 7.3, we can use Lemma 6.2 to obtain an $L^2(\Omega)$ estimate from the pointwise error provided by Theorem 7.7. The $H^1(\Omega)$ estimate of $\mathbf{u}_\alpha^h(\boldsymbol{\nu})$ also follows from Theorem 7.7. Finally the $H^1(\Omega)$ estimate of $s_\alpha^h(\boldsymbol{\nu})$ follows from an interpolation result based on the $L^2(\Omega)$ and an $H^2(\Omega)$ estimate of $s(\mathbf{u}_\alpha^h(\boldsymbol{\nu}), \boldsymbol{\nu})$. See the proof of (7.6) in Theorem 7.3 for a similar argument. □

The proof of Theorem 3.2 now follows simply from Theorems 7.3 and 7.8 and the bounds provided by (7.1), (7.2), and (7.3).

7.3. Proof of Lemma 7.4. Finally we return to the proof of Lemma 7.4, which quantifies the difference between $J_\alpha(\cdot, \mathbf{e}_k)$ and $J_\alpha^h(\cdot, \mathbf{e}_k)$ at their respective minimizers.

Now

$$\begin{aligned} J_\alpha(\mathbf{u}_\alpha(\mathbf{e}_k), \mathbf{e}_k) &\le J_\alpha(\mathbf{u}_\alpha^h(\mathbf{e}_k), \mathbf{e}_k) \\ &\le J_\alpha^h(\mathbf{u}_\alpha^h(\mathbf{e}_k), \mathbf{e}_k) + \|\boldsymbol{\rho}^n \Phi(\mathbf{u}_\alpha^h(\mathbf{e}_k), \mathbf{e}_k) - \mathbf{e}_k\|_n^2 \\ &\quad - \|\boldsymbol{\rho}^n \Phi^h(\mathbf{u}_\alpha^h(\mathbf{e}_k), \mathbf{e}_k) - \mathbf{e}_k\|_n^2. \end{aligned}$$

Elementary manipulations and the Cauchy-Schwarz inequality imply that

$$\begin{aligned} &\|\boldsymbol{\rho}^n \Phi(\mathbf{u}_\alpha^h(\mathbf{e}_k), \mathbf{e}_k) - \mathbf{e}_k\|_n^2 - \|\boldsymbol{\rho}^n \Phi^h(\mathbf{u}_\alpha^h(\mathbf{e}_k), \mathbf{e}_k) - \mathbf{e}_k\|_n^2 \\ &= \|\boldsymbol{\rho}^n \Phi(\mathbf{u}_\alpha^h(\mathbf{e}_k), \mathbf{e}_k) - \boldsymbol{\rho}^n \Phi^h(\mathbf{u}_\alpha^h(\mathbf{e}_k), \mathbf{e}_k)\|_n^2 \\ &\quad - 2\langle \boldsymbol{\rho}^n \Phi(\mathbf{u}_\alpha^h(\mathbf{e}_k), \mathbf{e}_k) - \boldsymbol{\rho}^n \Phi^h(\mathbf{u}_\alpha^h(\mathbf{e}_k), \mathbf{e}_k), \boldsymbol{\rho}^n \Phi^h(\mathbf{u}_\alpha^h(\mathbf{e}_k), \mathbf{e}_k) - \mathbf{e}_k\rangle_n. \end{aligned}$$

The geometric, arithmetic mean inequality implies that for any $0 < \epsilon < 1$

$$\langle \mathbf{v}, \mathbf{w}\rangle_n \le \|\mathbf{v}\|_n \|\mathbf{w}\|_n \le \frac{\epsilon^{-1}}{2}\|\mathbf{v}\|_n^2 + \frac{\epsilon}{2}\|\mathbf{w}\|_n^2.$$

Consequently

$$\begin{aligned} &\|\boldsymbol{\rho}^n \Phi(\mathbf{u}_\alpha^h(\mathbf{e}_k), \mathbf{e}_k) - \mathbf{e}_k\|_n^2 - \|\boldsymbol{\rho}^n \Phi^h(\mathbf{u}_\alpha^h(\mathbf{e}_k), \mathbf{e}_k) - \mathbf{e}_k\|_n^2 \\ &\le (1+\epsilon^{-1})\|\boldsymbol{\rho}^n \Phi(\mathbf{u}_\alpha^h(\mathbf{e}_k), \mathbf{e}_k) - \boldsymbol{\rho}^n \Phi^h(\mathbf{u}_\alpha^h(\mathbf{e}_k), \mathbf{e}_k)\|_n^2 \\ &\quad + \epsilon\|\boldsymbol{\rho}^n \Phi^h(\mathbf{u}_\alpha^h(\mathbf{e}_k), \mathbf{e}_k) - \mathbf{e}_k\|_n^2. \end{aligned}$$

Since $\|\boldsymbol{\rho}^n \Phi^h(\mathbf{u}_\alpha^h(\mathbf{e}_k), \mathbf{e}_k) - \mathbf{e}_k\|_n^2 \le J_\alpha^h(\mathbf{u}_\alpha^h(\mathbf{e}_k), \mathbf{e}_k)$, we conclude that

$$\begin{aligned} J_\alpha(\mathbf{u}_\alpha(\mathbf{e}_k), \mathbf{e}_k) &\le (1+\epsilon) J_\alpha^h(\mathbf{u}_\alpha^h(\mathbf{e}_k), \mathbf{e}_k) \\ &\quad + (1+\epsilon^{-1})\|\boldsymbol{\rho}^n \Phi(\mathbf{u}_\alpha^h(\mathbf{e}_k), \mathbf{e}_k) - \boldsymbol{\rho}^n \Phi^h(\mathbf{u}_\alpha^h(\mathbf{e}_k), \mathbf{e}_k)\|_n^2 \\ &\le (1+\epsilon) J_\alpha^h(\mathbf{u}_\alpha^h(\mathbf{e}_k), \mathbf{e}_k) + 2\epsilon^{-1}(h^4 + d^4)\|\mathbf{u}_\alpha^h(\mathbf{e}_k)\|_{H^1(\Omega)^2}^2 \\ &\le (1+\epsilon) J_\alpha^h(\mathbf{u}_\alpha^h(\mathbf{e}_k), \mathbf{e}_k) + 2\epsilon^{-1}\frac{(h^4 + d^4)}{n\alpha}. \end{aligned}$$

The last inequality follows from (7.12) since

$$\alpha\|\mathbf{u}_\alpha^h(\mathbf{e}_k)\|_{H^1(\Omega)^2}^2 \le a_\alpha^h(\mathbf{u}_\alpha^h(\mathbf{e}_k), \mathbf{u}_\alpha^h(\mathbf{e}_k)) \le \frac{1}{n}.$$

Finally we choose $\epsilon = h^2/(\alpha^{1/2})$ to obtain the required estimate.

7.4. Proof of Lemma 7.5. Next we compare the functionals $\bar{J}_\alpha$ and J_α. First consider the following decomposition of a function in $H^1(\Omega)^2$. A $\mathbf{u} \in H^1(\Omega)^2$ can be decomposed into gradient and curl components via

$$\mathbf{u} = V(\mathbf{u}) + W(\mathbf{u}),$$

where $\mathbf{v} = V(\mathbf{u})$ satisfies

$$\begin{aligned} \boldsymbol{\nabla} \cdot \mathbf{v} &= \boldsymbol{\nabla} \cdot \mathbf{u} && \text{in } \Omega, \\ \boldsymbol{\nabla} \times \mathbf{v} &= 0 && \text{in } \Omega, \\ \mathbf{v} \cdot \mathbf{n} &= \mathbf{u} \cdot \mathbf{n} && \text{on } \partial\Omega, \end{aligned}$$

and $\mathbf{w} = W(\mathbf{u})$ satisfies

$$\begin{aligned} \boldsymbol{\nabla} \times \mathbf{w} &= \boldsymbol{\nabla} \times \mathbf{u} && \text{in } \Omega, \\ \boldsymbol{\nabla} \cdot \mathbf{w} &= 0 && \text{in } \Omega, \\ \mathbf{w} \cdot \mathbf{n} &= 0 && \text{on } \partial\Omega. \end{aligned}$$

It is possible to show that there exists a constant $C_2 > 0$ such that

$$|V(\mathbf{u})|^2_{H^1(\Omega)^2} + |W(\mathbf{u})|^2_{H^1(\Omega)^2} \leq \frac{1}{C_2} |\mathbf{u}|^2_{H^1(\Omega)^2}.$$

If we let $\alpha' = C_2 \alpha$, then

$$\bar{J}_{\alpha'}(\bar{s}_{\alpha'}(\mathbf{e}_k), \mathbf{e}_k) \leq J_\alpha(\mathbf{u}_\alpha(\mathbf{e}_k), \mathbf{e}_k). \tag{7.13}$$

Let $s' = \bar{s}_{\alpha'}(\mathbf{e}_k)$ and $\mathbf{u}' = \boldsymbol{\nabla} s'$. Note that $V(\mathbf{u}') = \boldsymbol{\nabla} s'$ and $W(\mathbf{u}') = 0$ and $\Phi(\mathbf{u}', \mathbf{e}_k) = s'$. Now (7.13) follows from

$$\begin{aligned} \bar{J}_{\alpha'}(s', \mathbf{e}_k) &= \|\boldsymbol{\rho}^n s' - \mathbf{e}_k\|^2_n + \alpha' |s'|^2_{H^2(\Omega)} \\ &= \|\boldsymbol{\rho}^n \Phi(\mathbf{u}', \mathbf{e}_k) - \mathbf{e}_k\|^2_n + \alpha' |V(\mathbf{u}')|^2_{H^1(\Omega)^2} + \alpha' |W(\mathbf{u}')|^2_{H^1(\Omega)^2} \\ &\leq \|\boldsymbol{\rho}^n \Phi(\mathbf{u}_\alpha(\mathbf{e}_k), \mathbf{e}_k) - \mathbf{e}_k\|^2_n + \alpha' |V(\mathbf{u}_\alpha(\mathbf{e}_k))|^2_{H^1(\Omega)^2} + \alpha' |W(\mathbf{u}_\alpha(\mathbf{e}_k))|^2_{H^1(\Omega)^2} \\ &\leq \|\boldsymbol{\rho}^n \Phi(\mathbf{u}_\alpha(\mathbf{e}_k), \mathbf{e}_k) - \mathbf{e}_k\|^2_n + \alpha |\mathbf{u}_\alpha(\mathbf{e}_k)|^2_{H^1(\Omega)^2} \\ &= J_\alpha(\mathbf{u}_\alpha(\mathbf{e}_k), \mathbf{e}_k). \end{aligned}$$

Here we have used the fact that $\mathbf{u}' = \boldsymbol{\nabla} s'$ will be the minimizer of the functional

$$M_{\alpha'}(\mathbf{u}, \mathbf{e}_k) = \|\boldsymbol{\rho}^n \Phi(\mathbf{u}, \mathbf{e}_k) - \mathbf{e}_k\|^2_n + \alpha' |V(\mathbf{u})|^2_{H^1(\Omega)^2} + \alpha' |W(\mathbf{u})|^2_{H^1(\Omega)^2}$$

over $H^1(\Omega)^2$.

It should be noted that the functional $M_\alpha(\cdot, \mathbf{y})$ provides an alternative method for approximating a thin plate spline, the disadvantage being the need to have an auxiliary vector function $\mathbf{w}(\mathbf{u})$ (or $\mathbf{v}(\mathbf{u})$). On the other hand, we do intend to analyze this alternative method in a future paper and compare it to the method discussed here.

8. Conclusions. We have introduced a finite element method which provides a flexible and practical tool for smoothing as well as surface fitting. It is a scalable method in the sense that the size of the data set does not affect the overall size of the system obtained from the discretization of the equations. The assembly of the finite element matrices involves only a single scan of the data set. We have demonstrated that the method can be used to provide data fits of very large data sets, providing results similar to those obtained by standard thin plate splines. Additionally, the TPSFEM method as described can easily be generalized to three dimensions.

REFERENCES

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

[2] D. Apprato, R. Arcangéli, and R. Manzanilla, *Sur la construction de surfaces de classe C^k à partir d'un grand nombre de données de Lagrange*, Math. Model. Numer. Anal., 21 (1987), pp. 529–555.

[3] R. Arcangéli, *D^m-splines sur un domaine borné de $\mathbb{R}^n$*, Publication URA CNRS 1204 n° 86/2, University of Pau, Pau, France, 1986.

[4] R. Arcangéli, *Some applications of discrete D^m-splines*, in Mathematical Methods in Computer Aided Geometric Design, T. Lyche and L. L. Schumaker, eds., Academic Press, New York, 1989, pp. 35–44.

[5] R. Arcangéli, R. Manzanilla, and J. J. Torrens, *Approximation spline de surfaces de type explicite comportant des failles*, Math. Model. Numer. Anal., 31 (1997), pp. 643–676.

[6] Australian Society of Exploration Geophysicists, *Magnet Data for the Ebagoola Region of Queensland, Australia*, http://www.aseg.org.au/asegeduc/fr26.htm.

[7] R. K. Beatson, G. Goodsell, and M. J. D. Powell, *On multigrid techniques for thin plate spline interpolation in two dimensions*, in The Mathematics of Numerical Analysis, Lectures in Appl. Math. 32, AMS, Providence, RI, 1996, pp. 77–97.

[8] R. K. Beatson and W. A. Light, *Fast evaluation of radial basis functions: Methods for two-dimensional polyharmonic splines*, IMA J. Numer. Anal., 17 (1997), pp. 343–372.

[9] R. K. Beatson and G. N. Newsam, *Fast evaluation of radial basis functions*, Comput. Math. Appl., 24 (1992), pp. 7–19.

[10] R. K. Beatson and M. J. D. Powell, *An iterative method for thin plate spline interpolation that employs approximations to Lagrange functions*, in Numerical Analysis 1993 (Dundee, 1993), Pitman Res. Notes Math. Ser. 303, Longman Scientific and Technical, Harlow, UK, 1994, pp. 17–39.

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

[12] Z. Cai, T. A. Manteuffel, and S. F. McCormick, *First-order system least squares for velocity-vorticity-pressure form of the Stokes equations, with application to linear elasticity*, Electron. Trans. Numer. Anal., 3 (1995), pp. 150–159 (electronic).

[13] Z. Cai, T. A. Manteuffel, and S. F. McCormick, *First-order system least squares for second-order partial differential equations: Part* II, SIAM J. Numer. Anal., 34 (1997), pp. 425–454.

[14] Z. Cai, T. A. Manteuffel, and S. F. McCormick, *First-order system least squares for the Stokes equations, with application to linear elasticity*, SIAM J. Numer. Anal., 34 (1997), pp. 1727–1741.

[15] Z. Cai, T. A. Manteuffel, S. F. McCormick, and S. V. Parter, *First-order system least squares (FOSLS) for planar linear elasticity: Pure traction problem*, SIAM J. Numer. Anal., 35 (1998), pp. 320–335.

[16] P. Christen, I. Altas, M. Hegland, S. Roberts, K. Burrage, and R. Sidje, *Parallelization of a finite element surface fitting algorithm for data mining*, ANZIAM J., 42 (2000), pp. C385–C399.

[17] P. Christen, M. Hegland, O. Nielsen, S. Roberts, P. Strazdins, and I. Altas, *Scalable parallel algorithms for surface fitting and data mining*, Parallel Computing, 27 (2001), pp. 941–961.

[18] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North–Holland, Amsterdam, 1978.

[19] P. Craven and G. Wahba, *Smoothing noisy data with spline functions. Estimating the correct degree of smoothing by the method of generalized cross-validation*, Numer. Math., 31 (1978/79), pp. 377–403.

[20] J. Duchon, *Splines minimizing rotation-invariant semi-norms in Sobolev spaces*, in Constructive Theory of Functions of Several Variables, Lecture Notes in Math. 571, 1977, pp. 85–100.

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

[22] P. Grisvard, *Elliptic Problems in Nonsmooth Domains*, Pitman, Boston, 1985.

[23] M. Hegland, S. Roberts, and I. Altas, *Finite element thin plate splines for surface fitting*, in Computational Techniques and Applications: CTAC97 (Singapore, 1997), B. Noye, M. Teubner, and A. Gill, eds., World Scientific, River Edge, NJ, pp. 289–296.

[24] M. Hegland, S. Roberts, and I. Altas, *Finite element thin plate splines for data mining applications*, in Mathematical Methods for Curves and Surfaces, II (Lillehammer, 1997), Vanderbilt University Press, Nashville, TN, 1998, pp. 245–252.

[25] M. F. Hutchinson, *A stochastic estimator of the trace of the influence matrix for Laplacian smoothing splines*, Comm. Statist. Simulation Comput., 19 (1990), pp. 433–450.

[26] R. Sibson and G. Stone, *Computation of thin-plate splines*, SIAM J. Sci. Statist. Comput., 12 (1991), pp. 1304–1313.

[27] J. J. Torrens, *Discrete smoothing D^m-splines: Applications to surface fitting*, in Mathematical Methods for Curves and Surfaces, II (Lillehammer, 1997), Vanderbilt University Press, Nashville, TN, 1998, pp. 477–484.

[28] F. I. Utreras, *Convergence rates for multivariate smoothing spline functions*, J. Approx. Theory, 52 (1988), pp. 1–27.

[29] G. Wahba, *Spline Models for Observational Data*, CBMS-NSF Regional Conf. Ser. in Appl. Math. 59, SIAM, Philadelphia, 1990.

SIAM J. Numer. Anal.
Vol. 41, No. 1, pp. 235–257

VARIATIONAL MESH ADAPTATION METHODS FOR AXISYMMETRICAL PROBLEMS*

WEIMING CAO†, RICARDO CARRETERO-GONZÁLEZ‡, WEIZHANG HUANG§, AND ROBERT D. RUSSELL¶

Abstract. We study variational mesh adaptation for axially symmetric solutions to two-dimensional problems. The study is focused on the relationship between the mesh density distribution and the monitor function and is carried out for a traditional functional that includes several widely used variational methods as special cases and a recently proposed functional that allows for a weighting between mesh isotropy (or regularity) and global equidistribution of the monitor function. The main results are stated in Theorems 4.1 and 4.2. For axially symmetric problems, it is natural to choose axially symmetric mesh adaptation. To this end, it is reasonable to use the monitor function in the form $G = \lambda_1(r)\boldsymbol{e}_r\boldsymbol{e}_r^T + \lambda_2(r)\boldsymbol{e}_\theta\boldsymbol{e}_\theta^T$, where $\boldsymbol{e}_r$ and $\boldsymbol{e}_\theta$ are the radial and angular unit vectors.

It is shown that when higher mesh concentration at the origin is desired, a choice of λ_1 and λ_2 satisfying $\lambda_1(0) < \lambda_2(0)$ will make the mesh denser at $r = 0$ than in the surrounding area whether or not λ_1 has a maximum value at $r = 0$. The purpose can also be served by choosing λ_1 to have a local maximum at $r = 0$ when a Winslow-type monitor function with $\lambda_1(r) = \lambda_2(r)$ is employed. On the other hand, it is shown that the traditional functional provides little control over mesh concentration around a ring $r = r_\lambda > 0$ by choosing λ_1 and λ_2.

In contrast, numerical results show that the new functional provides better control of the mesh concentration through the monitor function. Two-dimensional numerical results are presented to support the analysis.

Key words. mesh adaptation, variational method, mesh regularity, equidistribution

AMS subject classifications. 65M50, 65M60

PII. S0036142902401591

1. Introduction.

Mesh adaptation has become an indispensable tool for use in the numerical solution of PDEs. One of the most widely used approaches for generating adaptive meshes is a variational method. With such a method, meshes are generated as images of a reference mesh through a coordinate transformation between the physical and computational (or logical) domains. The transformation is determined as the minimizer of a functional formulated to measure difficulties in the numerical approximation of the physical solution, typically through a so-called monitor function prescribed by the user to control the mesh adaptation. A variational method often results in an elliptic (PDE) mesh generation system. Such a system generates smooth meshes, allows for full specification of mesh behavior at the boundary, does not propagate boundary singularities into the domain, has less danger of producing mesh overlappings, and can be solved efficiently using many well-developed

*Received by the editors January 28, 2002; accepted for publication (in revised form) September 23, 2002; published electronically March 19, 2003. This work was supported in part by NSF (USA) grant DMS-0074240, and NSERC (Canada) grant OGP-0008781.

http://www.siam.org/journals/sinum/41-1/40159.html

†Department of Mathematics, The University of Texas at San Antonio, San Antonio, TX 78249 (wcao@math.utsa.edu).

‡Department of Mathematics, Simon Fraser University, Burnaby, BC V5A 1S6, Canada. Current address: Nonlinear Dynamical Systems Group, Department of Mathematics and Statistics, San Diego State University, San Diego, CA 92182 (carreter@math.sdsu.edu).

§Department of Mathematics, the University of Kansas, Lawrence, KS 66045 (huang@math.ukans.edu).

¶Department of Mathematics, Simon Fraser University, Burnaby, BC V5A 1S6, Canada (rdr@cs.sfu.ca).

algorithms. Moreover, the equidistribution principle, a concept which has been used successfully in one-dimensional mesh adaptation [3], can be naturally extended to multidimensions in the variational framework. Finally, many mesh features, such as orthogonality, smoothness, and concentration, can be incorporated explicitly into the mesh adaptation functional.

A number of variational methods have been developed in the past. For example, Winslow [15] proposes the variable diffusion method for which the mesh lines play the role of equipotentials of a potential problem [14]. Brackbill and Saltzman [1] develop a popular method combining mesh concentration, smoothness, and orthogonality. Several functionals are formulated by Steinberg and Roache [13] to control mesh properties such as the spacing of the points, areas or volumes of the cells, and the angles between mesh lines. Dvinsky [4] uses the energy of harmonic mappings as his mesh adaptation functional. Knupp [9, 10] and Knupp and Robidoux [11] develop functionals based on the idea of conditioning the Jacobian matrix of the coordinate transformation. A functional balancing mesh regularity and adaptivity is proposed by Huang [6].

Some theoretical work has been devoted to better understanding the existing methods. Cao, Huang, and Russell [2] study the qualitative effect of monitor functions on the resulting mesh for a general class of variational methods that includes Winslow's method [15] and the method using harmonic mappings [4] as special cases. In the recent work of Huang and Sun [8], the monitor function for the functional of [6] is defined based on interpolation error estimates, and asymptotic error bounds are obtained for interpolation on the resulting adaptive meshes satisfying the so-called isotropy and equidistribution conditions. The ability of the resulting method to generate adaptive meshes satisfying these conditions is also demonstrated numerically. Nevertheless, more work remains to be done on better understanding the existing variational methods, especially on precisely how the monitor function controls the concentration of the generated mesh.

In this paper we present such a study for two functionals, the traditional one studied in [2] and the new one proposed in [6], for the simple but important case of two-dimensional problems with axisymmetrical solutions. These types of problems arise in many practical situations, particularly for problems with blowup or quenching solutions. There has been considerable recent interest in solving higher-dimensional blowup problems such as the Schrödinger equation, and this work was motivated by the observation that the standard moving mesh procedures generally perform inadequately on such problems (e.g., see [2, 12]).

Let (x,y) be the coordinates in the physical domain Ω, and let (ξ,η) be the coordinates in the computational domain Ω_c. The traditional functional is

$$I_{trad}[\xi,\eta] = \int_{\Omega} \left(\nabla\xi^T G^{-1}\nabla\xi + \nabla\eta^T G^{-1}\nabla\eta\right) dxdy \tag{1.1}$$

and the new functional in [6] has the form

$$\begin{aligned} I_{new}[\xi,\eta] \quad &= \quad \gamma\int_{\Omega}\sqrt{g}\left(\nabla\xi^T G^{-1}\nabla\xi + \nabla\eta^T G^{-1}\nabla\eta\right)^q dxdy \\ &\quad + (1-2\gamma)2^q\int_{\Omega}\frac{\sqrt{g}}{(J\sqrt{g})^q}dxdy, \end{aligned} \tag{1.2}$$

where $J = x_\xi y_\eta - x_\eta y_\xi$ is the Jacobian of the coordinate transformation, G is the (matrix) monitor function with determinant g, and $q \geq 1$ and $\gamma \in (0, 1/2]$ are parameters. Here, $q \geq 1$ is required in order for the first integral of (1.2) to be convex. The features of these functionals and the roles of the parameters will be discussed in sections 2 and 3.

Axisymmetrical problems. For simplicity, we assume that the physical domain is $\Omega = \{(x, y) \mid x^2 + y^2 < 1\}$ and the computational domain is $\Omega_c = \{(\xi, \eta) \mid \xi^2 + \eta^2 < 1\}$. Let the polar coordinate systems for the physical and computational domains be

$$\begin{cases} x = r\cos\theta, \\ y = r\sin\theta, \end{cases} \qquad \begin{cases} \xi = R\cos\Theta, \\ \eta = R\sin\Theta. \end{cases}$$

Consider the case where the solution $u(x, y)$ is axially symmetric; i.e., u is invariant under rotation about the center $(0, 0)$. It is natural to choose an axially symmetric coordinate transformation

$$R = R(r), \qquad \Theta = \theta \tag{1.3}$$

for mesh adaptation. To this end, it is reasonable to use the monitor function in the form

$$G = \lambda_1(r)\boldsymbol{e}_r\boldsymbol{e}_r^T + \lambda_2(r)\boldsymbol{e}_\theta\boldsymbol{e}_\theta^T, \tag{1.4}$$

where $\boldsymbol{e}_r$ and $\boldsymbol{e}_\theta$ are unit vectors in the radial and angular directions, respectively. Thus, G is determined by its radial and angular components $\lambda_1 > 0$ and $\lambda_2 > 0$.

We are interested in the relationship between the monitor function and the mesh distribution. In particular, we focus on the mesh density $D(r)$. The Jacobian of the coordinate transformation J is easily seen to satisfy

$$\frac{1}{J} \equiv \det\left(\frac{\partial(\xi, \eta)}{\partial(x, y)}\right) = \frac{R}{r}\frac{dR}{dr}, \tag{1.5}$$

and thus the mesh density is given by

$$D(r) = \frac{R}{r}\frac{dR}{dr}. \tag{1.6}$$

The central aim of this paper is to gain insight into how much control one has on the mesh density $D(r)$ by appropriately choosing λ_1 and λ_2. In order for the variational method to be successful one needs that the solution to the variational problem gives a mesh distribution compatible with the chosen monitor function. For example, it is natural to choose one or more of the eigenvalues of the monitor function to have a higher value (a maximum) in the region where a physical solution needs a high concentration of mesh points; e.g., see [2]. It will become clear below that this is not always achievable and that if one is not careful in choosing the appropriate relation between λ_1 and λ_2 it is possible for the mesh density maximum to occur at a different location than that of the maximum of the eigenvalue. This can in turn lead to a large error in the numerical approximation of the physical solution.

An outline of the paper is as follows. In sections 2 and 3 basic properties of the traditional and new functionals for radially symmetric problems are presented. In section 4 we carry out an in-depth analysis on the control of the mesh density via the monitor function. In particular, we find that the relationship between the radial (λ_1) and the angular (λ_2) components of the monitor function is crucial for a good control of the mesh density. Section 5 presents some two-dimensional numerical results highlighting in part the lack of control of the mesh concentration for a wide choice of monitor functions. A brief analysis is given in section 6 for the traditional functional applied to spherically symmetric problems in three dimensions. Finally, section 7 contains conclusions and comments.

2. The traditional functional. In this section we consider the traditional functional (1.1) for axisymmetrical problems and give some of its basic properties. It is a generalization of the functionals for Winslow's method and Dvinsky's method of harmonic mappings. The monitor function G can be defined by arbitrarily choosing λ_1 and λ_2. However, it is worth pointing out that a number of commonly used monitor functions can be obtained through the interdependent relationship

$$\lambda_2 = \lambda_1^p \tag{2.1}$$

for some power p. For example, we have

$$\begin{array}{lll} \text{(HM)} & p=-1: & \text{harmonic mapping monitor function;} \\ \text{(Al)} & p=0: & \text{arclength monitor function;} \\ \text{(Ws)} & p=1: & \text{Winslow's monitor function;} \\ \text{(St)} & p=2: & \text{strong concentration monitor function.} \end{array} \tag{2.2}$$

In polar coordinates the gradient operator reads as

$$\nabla = \boldsymbol{e}_r \frac{\partial}{\partial r} + \frac{\boldsymbol{e}_\theta}{r}\frac{\partial}{\partial \theta},$$

and it follows from (1.3) and (1.4) that

$$\nabla\xi^T G^{-1}\nabla\xi + \nabla\eta^T G^{-1}\nabla\eta = \frac{1}{\lambda_1}\left(\frac{dR}{dr}\right)^2 + \frac{1}{\lambda_2}\left(\frac{R}{r}\right)^2. \tag{2.3}$$

Substituting (2.3) into (1.1) gives

$$I_{trad}[R] = 2\pi\int_0^1 \left[\frac{1}{\lambda_1}\left(\frac{dR}{dr}\right)^2 + \frac{1}{\lambda_2}\left(\frac{R}{r}\right)^2\right] rdr,$$

and its Euler–Lagrange equation is

$$-\frac{d}{dr}\left(\frac{r}{\lambda_1}\frac{dR}{dr}\right) + \frac{R}{r\lambda_2} = 0. \tag{2.4}$$

This equation is supplemented with the boundary conditions

$$R(0) = 0, \qquad R(1) = 1. \tag{2.5}$$

For a given monitor function (i.e., for given λ_1 and λ_2), solving (2.4) determines the resulting mesh transformation $R(r)$.

2.1. Nonnegativeness and mesh crossing. We have $R(r) \geq 0$ for $r \in (0,1)$. To see this, we note that the minimum of $R(r)$ occurs at the left end and/or an interior point due to the boundary conditions (2.5). If $R(0) = \min R(r)$, then we have $R(r) \geq 0$ from (2.5). If instead the minimum point is $r_0 \in (0,1)$, then $R'(r_0) = 0$ and $R''(r_0) \geq 0$. From (2.4)

$$\frac{1}{r_0\lambda_2}R(r_0) = \frac{d}{dr}\left(\frac{r}{\lambda_1}\right)\left(\frac{dR}{dr}\right)\bigg|_{r_0} + \frac{r}{\lambda_1}\frac{d^2R}{dr^2}\bigg|_{r_0} \geq 0.$$

Hence, in either case $R(r) \geq R(r_0) \geq 0$. Furthermore, (2.4) gives

$$\frac{dR}{dr} = \frac{\lambda_1}{r}\int_0^r \frac{R(x)}{x\lambda_2(x)}\,dx,$$

so it follows that $\frac{dR}{dr} > 0$ for $r \in (0,1)$; i.e., the mesh transformation is guaranteed to be nonsingular and produce no mesh crossing.

2.2. Mesh transformation for harmonic mappings. For the case of harmonic mappings ($p = -1$ or $\lambda_2 = 1/\lambda_1$) it is possible to find an analytical form for the mesh transformation $R(r)$. We explicitly construct $R(r)$ here since it then serves as the basis of study for other cases. Using the change of coordinates

$$s(r) = \int_1^r \frac{\lambda_1(x)}{x} dx, \tag{2.6}$$

(2.4) reads

$$-\frac{d^2R}{ds^2} + R = 0.$$

Its solution satisfying the boundary conditions (2.5) is $R(r) = e^s$. In section 4.1, using a transformation based on (2.6), we study more general monitor functions (including arclength and Winslow) in detail.

3. The new functional. The formulation of the new functional (1.2) is based on the so-called isotropy (or regularity) and equidistribution (or adaptation) requirements for an error distribution [6]. Specifically, the first integral term corresponds to the regularity requirement, while the second is associated with equidistribution. These two requirements are balanced by adjusting the value of the parameter γ. When $q = 1$ or $\gamma = 1/2$, the second integral becomes constant or simply vanishes, and only the isotropy plays a role. When $q = 1$ the functional gives rise to the energy functional of a harmonic mapping. The relation between the new and traditional functionals will be addressed later in section 3.3.

From (1.4) the determinant of G is $g = \det(G) = \lambda_1\lambda_2$. Let

$$\Lambda = \sqrt{\lambda_1\lambda_2}, \tag{3.1}$$

$$\mu_1(r) = \frac{\lambda_1}{\Lambda^{1/q}}, \qquad \mu_2(r) = \frac{\lambda_2}{\Lambda^{1/q}}. \tag{3.2}$$

Using the symmetry assumption, we can rewrite (1.2) as

$$I_{new}[R] = \gamma \int_0^1 \left[\frac{1}{\mu_1}\left(\frac{dR}{dr}\right)^2 + \frac{1}{\mu_2}\left(\frac{R}{r}\right)^2 \right]^q r dr + (1-2\gamma)2^q \int_0^1 \left[\frac{RR'}{r\sqrt{g}}\right]^q r\sqrt{g} dr.$$

Its Euler–Lagrange equation is given by

$$\begin{aligned} &-\frac{\gamma}{\beta^{q-1}}\frac{d}{dr}\left(\frac{r\beta^{q-1}R'}{\mu_1}\right) + \frac{\gamma R}{r\mu_2} \\ &-\frac{(1-2\gamma)2^{q-1}(q-1)R}{\beta^{q-1}}\left(\frac{RR'}{r\sqrt{g}}\right)^{q-2}\frac{d}{dr}\left(\frac{RR'}{r\sqrt{g}}\right) = 0, \end{aligned} \tag{3.3}$$

where

$$\beta = \frac{1}{\mu_1}\left(\frac{dR}{dr}\right)^2 + \frac{1}{\mu_2}\left(\frac{R}{r}\right)^2.$$

The highly nonlinear form of the new functional does not lend itself to a straightforward analytical treatment of its basic properties. Nonetheless, we devote the rest of this section to the study of several special cases of (3.3) subject to the boundary conditions (2.5). These cases are important because they help to better understand the functional and link it to the traditional one.

3.1. The exact equidistribution case ($\gamma = 0$). We first consider the case $\gamma = 0$ which corresponds to exact equidistribution. Assuming that $R(r) > 0$ for $r \in (0,1)$, (3.3) implies

$$\frac{1}{\sqrt{g}} \frac{R}{r} \frac{dR}{dr} = \alpha, \tag{3.4}$$

where α is a constant. From (1.5), this is equivalent to

$$\frac{1}{J\sqrt{g}} = \alpha,$$

which is a multidimensional generalization of the well-known equidistribution principle in one dimension. This equation guarantees that J, the Jacobian of the coordinate transformation, does not change sign in the domain.

3.2. The pure isotropy case ($\gamma = 1/2$). For $\gamma = 1/2$ the mesh equation (3.3) reduces to

$$\frac{d}{dr}\left[\frac{\beta^{q-1} r\, R'}{\mu_1}\right] = \frac{\beta^{q-1} R}{r\mu_2}.$$

As in section 2.1, it is easy to show that $R(r) \geq 0$ and $R'(r) \geq 0$. Thus, for this case the mesh is also guaranteed not to cross.

In Figure 1 we depict $R'(r)$ for the traditional functional and the new functional with $\gamma = 1/2$ and several values of q. As can be seen, the mesh transformation for the new functional with different values of q is quite similar to the traditional functional. This is not surprising since for $\gamma = 1/2$ the new functional shifts all the weight towards isotropy and thus resembles the traditional functional.

3.3. The case $q = 1$. When $q = 1$, the second integral in (1.2) becomes constant. From (3.1) the mesh equation (3.3) reduces to

$$-\frac{d}{dr}\left(\frac{\Lambda r}{\lambda_1}\frac{dR}{dr}\right) + \frac{\Lambda R}{r\lambda_2} = 0. \tag{3.5}$$

Once again it is easy to prove that mesh crossing will not occur. Note that the mesh equation (3.5) is independent of the parameter γ and very similar to (2.4) for the traditional functional. In fact, for the harmonic mapping case where $\Lambda = 1$, the mesh equations (3.5) and (2.4) are identical.

For the Winslow monitor function case ($\Lambda = \lambda_1 = \lambda_2$) the mesh equation is

$$\frac{d}{dr}\left(r\frac{dR}{dr}\right) = \frac{R}{r}. \tag{3.6}$$

The solution of (3.6) compatible with the boundary conditions is $R(r) = r$. Therefore, the case $q = 1$ of the new functional method gives a trivial coordinate transformation $R = r$ and does not allow for any control of the mesh concentration when a Winslow-type monitor function is used.

Finally, for the arclength monitor function ($\lambda_2 = 1$) the mesh equation is

$$-\frac{d}{dr}\left(\frac{r}{\sqrt{\lambda_1}}\frac{dR}{dr}\right) + \frac{R}{r\sqrt{\lambda_1}} = 0.$$

This mesh equation is equivalent to that for the traditional functional (2.4) using a Winslow-type monitor function with $\sqrt{\lambda_1}$ instead of λ_1.

In summary, except for the Winslow case, for $q = 1$ the new functional corresponds to the traditional functional with a suitable choice of the monitor function.

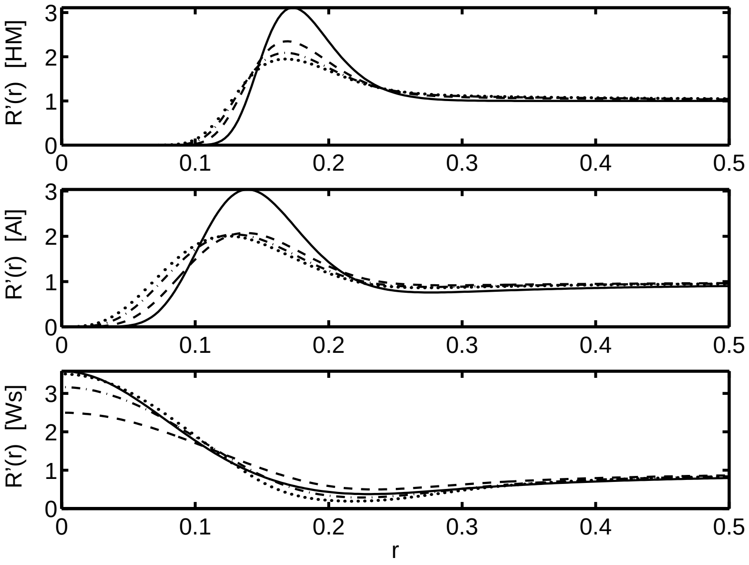


Fig. 1. *Comparison of $R'(r)$ for the traditional functional (solid line) and the new functional with $\gamma = 1/2$ ($q = 2$: dashed, $q = 3$: dotted-dashed, $q = 4$: dotted). The plots correspond to the three popular choices of monitor function (harmonic mapping, arclength, and Winslow) with $\lambda_1(r) = 1 + \exp(-r^2/a)/a$ ($a = 0.01$).*

4. Control of mesh density via $\boldsymbol{\lambda_1}$ and $\boldsymbol{\lambda_2}$.

4.1. The traditional functional. The Euler–Lagrange equation (2.4) for the traditional functional relates the coordinate transformation to the monitor function for a given choice of λ_1 and λ_2. The purpose of this section is to use this to show that precise control of the mesh density $D(r)$ cannot be achieved from the choice of λ_1 and λ_2. In fact, we prove that the maximum for the mesh concentration does not occur at the maximum of λ_1, resulting in misplacement of mesh concentration.

Let us then take (2.4) and solve for the mesh density $D(r)$ in (1.6). Motivated by the transformation (2.6) leading to the exact solution of (2.4) for the harmonic mapping monitor function ($\lambda_2 = 1/\lambda_1$), we consider the change of dependent variable

$$R(r) = e^{s(r)} \quad \text{with} \quad s(r) = \int_1^r \frac{\lambda_1(x)\, v(x)}{x} dx \tag{4.1}$$

for a to-be-determined and bounded function v. Substituting this into (2.4) yields the ODE for v

$$\frac{dv}{dr} = \frac{\lambda_1}{r}\left(\frac{1}{\Lambda^2} - v^2\right). \tag{4.2}$$

It satisfies

$$v(0) = \frac{1}{\Lambda(0)}, \tag{4.3}$$

since any other initial value (at $r = 0$) produces an unbounded solution v. The choice (4.3) is compatible with the special case of the harmonic mapping where $v(r) = 1$.

LEMMA 4.1. *$v(r) > 0$ for all $r \in [0,1]$.*

Proof. This is an immediate result of the initial condition $v(0) = 1/\Lambda(0) > 0$ and the fact that $v' > 0$ on the line $v = 0$. □

The overall behavior of v is determined by the nullcline

$$v_{\text{null}}(r) = \frac{1}{\Lambda(r)}. \tag{4.4}$$

LEMMA 4.2. *$v_{\min} \le v(r) \le v_{\max}$ for all $r \in [0,1]$, where $v_{\min} = \min_r\{1/\Lambda(r)\}$ and $v_{\max} = \max_r\{1/\Lambda(r)\}$. Thus, the solution $v(r)$ is bounded by the minimum and maximum of the nullcline.*

Proof. Note that $v' > 0$ below the nullcline and $v' < 0$ above it. Since $v_{\min} \le v_{\text{null}}(r)$, we have $v' \ge 0$. This and $v(0) = 1/\Lambda(0) \ge v_{\min}$ imply that $v(r) \ge v_{\min}$. Similarly, we have $v(r) \le v_{\max}$. □

Define r_λ as the point where λ_1 attains its maximum, i.e.,

$$\lambda_1(r_\lambda) = \max_{r\in[0,1]} \lambda_1(r).$$

We have the following lemma.

LEMMA 4.3. *Let r_λ be a strict maximum point of λ_1 (so $\lambda_1^{''}(r_\lambda) < 0$), and let $\lambda_2 = c\lambda_1^p$ for some power $p > -1$ and some constant $c > 0$. Then, $v(r_\lambda) > \frac{1}{\Lambda(r_\lambda)}$.*

Proof. For this particular choice of λ_2, we have

$$\Lambda'(r_\lambda) = 0, \quad \Lambda''(r_\lambda) \ne 0, \quad v_{\min} = \frac{1}{\Lambda(r_\lambda)}.$$

We prove the lemma by contradiction. From Lemma 4.2, we can assume only $v(r_\lambda) = 1/\Lambda(r_\lambda)$. By differentiating (4.2) twice and using the fact that $\lambda_1^{'}(r_\lambda) = 0$, we get

$$\begin{aligned} v'(r_\lambda) &= v''(r_\lambda) = 0,\\ v'''(r_\lambda) &= -\frac{2\lambda_1(r_\lambda)\Lambda''(r_\lambda)}{r_\lambda\Lambda(r_\lambda)^3} \ne 0. \end{aligned}$$

This implies that

$$\begin{aligned} v(r) &= v(r_\lambda) + \frac{(r-r_\lambda)^3}{6}v'''(r_\lambda) + O((r-r_\lambda)^4)\\ &= v_{\min} + \frac{(r-r_\lambda)^3}{6}v'''(r_\lambda) + O((r-r_\lambda)^4). \end{aligned}$$

Hence, $v(r) < v_{\min}$ at some points in the neighborhood of r_λ, which contradicts Lemma 4.2. □

Figure 2 shows a typical vector field for v.

To study the mesh density, note that in terms of s,

$$D(r) = \frac{R}{r}\frac{dR}{dr} = \frac{e^s}{r}s'e^s = \frac{s'e^{2s}}{r},$$

and its rate of change

$$\begin{aligned} \frac{dD}{dr} &= \frac{d}{dr}\left(\frac{s'e^{2s}}{r}\right)\\ &= \frac{e^{2s}}{r}\left(s'' + 2s'^2 - \frac{s'}{r}\right)\\ &= \frac{e^{2s}}{r^2}\left(\lambda_1'v + \lambda_1 v' - 2\frac{\lambda_1 v}{r} + 2\frac{\lambda_1^2 v^2}{r}\right). \end{aligned} \tag{4.5}$$

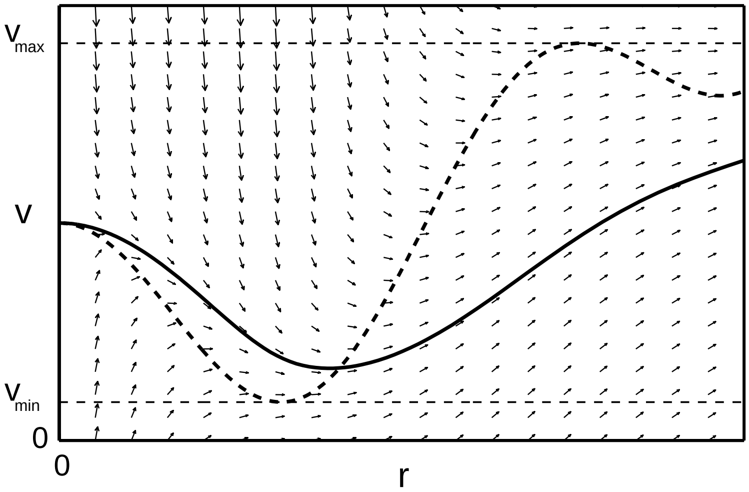


FIG. 2. *Typical vector field for* (4.2). *The total variation* $[v_{\min}, v_{\max}]$ *of the nullcline* (4.4) *(dashed line) bounds the behavior of the solution with* $v(0) = 1/\Lambda(0)$ *(solid line).*

Using (4.2) one obtains

$$\frac{dD}{dr} = \frac{e^{2s}}{r^2}\left\{\lambda_1' v + \frac{\lambda_1^2}{r}\left[\left(v - \frac{1}{\lambda_1}\right)^2 + \left(\frac{1}{\Lambda^2} - \frac{1}{\lambda_1^2}\right)\right]\right\}. \tag{4.6}$$

Equation (4.6) determines where the mesh density reaches an extremum in terms of λ_1 and λ_2. In general, it is desired that the mesh has a higher concentration of points at the maximum location of λ_1 so that the mesh concentration can be controlled by choosing λ_1. We first consider mesh concentration at the origin $r = 0$.

THEOREM 4.1. (i) *If* $\lambda_1(0) - \lambda_2(0) \neq 0$, *then* $D'(0)$ *has the same sign as* $\lambda_1(0) - \lambda_2(0)$ *whether* λ_1 *has a maximum at* $r = 0$ *or not. Specifically, if* $\lambda_1(0) > \lambda_2(0)$, *then* $D'(0) > 0$ *(i.e., the mesh at the origin is coarser than in the surrounding area), and if* $\lambda_1(0) < \lambda_2(0)$, *then* $D'(0) < 0$ *(i.e., the mesh at the origin is denser than in the surrounding area).*

(ii) *Let* $\lambda_2(r) = \lambda_1(r)$. *If* $\lambda_1^{'}(0) \neq 0$, *then* $D'(0)\lambda_1^{'}(0) > 0$. *If* $\lambda_1^{'}(0) = 0$ *but* $\lambda_1^{''}(0) \neq 0$, *then* $D'(0)\lambda_1^{''}(0) > 0$.

Proof. Let $y(r) = \lambda_1(r)v(r)$. Note that $y(0) = \lambda_1(0)/\Lambda(0) = \sqrt{\lambda_1(0)/\lambda_2(0)}$. Equation (4.5) can be rewritten as

$$\frac{dD}{dr} = \frac{e^{2s}}{r^2}\left(y' - \frac{2y}{r} + \frac{2y^2}{r}\right).$$

Expanding the bracketed terms on the right-hand side about $r = 0$, we get

$$\begin{aligned}\frac{dD}{dr} &= \frac{e^{2s}}{r^2}\left\{\frac{2}{r}y(0)\left(y(0) - 1\right) + y'(0)\left(4y(0) - 1\right) + O(r)\right\} \\ &= \frac{e^{2s}}{r^2}\left\{\frac{2}{r}\sqrt{\frac{\lambda_1(0)}{\lambda_2(0)}}\left(\sqrt{\frac{\lambda_1(0)}{\lambda_2(0)}} - 1\right) + y'(0)\left(4\sqrt{\frac{\lambda_1(0)}{\lambda_2(0)}} - 1\right) + O(r)\right\}. \qquad (4.7)\end{aligned}$$

Thus, if $\lambda_1(0) \neq \lambda_2(0)$, the first term in the bracket dominates. In this case, $D'(r)$ has the same sign as $\lambda_1(0) - \lambda_2(0)$. The result in (i) follows.

We now prove part (ii) using (4.6). This can also be done through (4.5), but higher order terms must be used. Using the assumption $\lambda_1(r) = \lambda_2(r)$ and expanding the bracketed terms of (4.6) about $r = 0$, we get

$$\frac{dD}{dr} = \frac{e^{2s}}{r^2}\{v(0)\lambda_1^{'}(0) + r[\lambda_1(0)v'(0) + \lambda_1^{'}(0)(1 + v'(0)) + \lambda_1^{''}(0)v(0)] + O(r^2)\}.$$

The results in (ii) follow since $\lambda_1^{'}(0) = 0$ implies $v'(0) = 0$. □

We now consider the case where mesh concentration away from the origin is desired, i.e., when $r_\lambda > 0$. Let

$$r_D : \qquad D(r_D) = \max_{r \in [0,1]} D(r).$$

The following theorem shows the relative positioning of r_D with respect to r_λ.

THEOREM 4.2. *Let* $r_\lambda > 0$.

(i) *If* $\lambda_1(r_\lambda) > \lambda_2(r_\lambda)$, *then* $D'(r_\lambda) > 0$ *and thus* $r_D > r_\lambda$.

(ii) *Further, if we assume that* $\lambda_2(r) = \lambda_1(r)$ *(Winslow's method) and* r_λ *is a strict maximum point of* λ_1 *(i.e.,* $\lambda_1{''}(r_\lambda) < 0$*), then* $D'(r_\lambda) > 0$ *or again* $r_D > r_\lambda$.

Proof. (i) The result is an immediate consequence of (4.6) and the assumptions.

(ii) When $\lambda_2(r) = \lambda_1(r)$, we have $\Lambda(r) = \lambda_1(r)$. Lemma 4.3, the fact that $\lambda_1(r_\lambda) = 0$, and (4.6) imply that $D'(r_\lambda) > 0$. □

We note that a result can be obtained for the general choice of λ_2 satisfying $\lambda_2(r) = \lambda_1(r)$ only at $r = r_\lambda$. Moreover, numerical experiments (see below) show that the mismatch between r_D and r_λ for the Winslow monitor function is relatively small.

The situation with $\lambda_1(r_\lambda) < \lambda_2(r_\lambda)$ is much more complex. Note that the last term in (4.6) is now negative. In order to determine the relative positions of r_D and r_λ it is necessary to compare all the terms on the right-hand side of (4.6). It is possible for r_D and r_λ to coincide. However, numerical results (see section 4.1.4) also show that for $\lambda_2 = \lambda_1^p$ with $p > 1$, r_D can be located on either side of r_λ.

It is emphasized that part (i) of both Theorems 4.1 and 4.2 requires no explicit relationship between λ_1 and λ_2, although we typically apply them to the monitor functions defined in (2.2).

4.1.1. The harmonic mapping case ($p = -1$). In this case, $\lambda_2 = 1/\lambda_1$. Assuming that $\lambda_1(r_\lambda) > 1$, we have $\lambda_1(r_\lambda) > \lambda_2(r_\lambda)$. Theorem 4.2 implies that $r_D > r_\lambda$, or the location of maximum mesh density is to the right of the maximum for λ_1.

When a (local) higher mesh concentration at the origin is desired, Theorem 4.1 implies that if (a) $\lambda_1(0) > 1$, the mesh at the origin is coarser than in the surrounding area, whether λ_1 has a maximum at $r = 0$ or not. This effect is clearly depicted in Figure 3 (left column, second plot) where $r_D > r_\lambda = 0$ implies a failure to concentrate the points at the origin. If instead (b) $0 < \lambda_1(0) \leq 1$, then $\lambda_1(0) \leq \lambda_2(0)$ and the mesh will be denser in the center than in the surrounding area.

We now consider conditions under which $r_\lambda > 0$ and r_D coincide. There is a tight restriction on the choice of λ_1 since $D' = 0$ must hold where $\lambda_1^{'} = 0$. Notice that we have $\Lambda = 1$ for the current case. Equation (4.6) implies that this can be achieved if and only if $\lambda_1(r_\lambda) = 1$. However, this cannot hold in general unless the high mesh concentration is desired only at the global maximum point and $\lambda_1(r_\lambda) = 1$ can be achieved by rescaling.

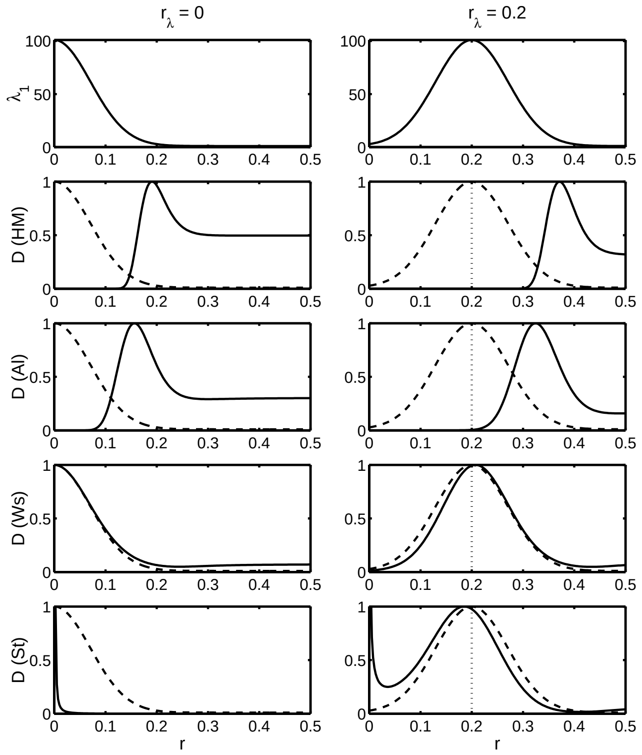


Fig. 3. *Normalized mesh densities obtained with the traditional functional for different monitor functions defined in* (2.2) *and with* $\lambda_1(r) = 1+\exp(-(r-r_\lambda)^2/a)/a$ $(a = 0.01)$ *that has its maximum located at* $r_\lambda = 0$ *(left column) and* $r_\lambda = 0.2$ *(right column). The top plot depicts* $\lambda_1(r)$*. For guidance, we plot, along with the normalized densities (solid lines), the normalized curve for* $\lambda_1(r)$ *(dashed lines).*

From the above analysis we see that if $\lambda_1(r_\lambda) > 1$, r_D will be located to the right of r_λ. This failure to place the higher concentration of points in the desired area is depicted in Figure 3 (right column, second plot).

4.1.2. The arclength case ($p = 0$). For the arclength case $\lambda_2(r) = 1$, a similar analysis as the one for the harmonic mapping case can be carried out. We assume $\lambda_1(r_\lambda) > 1$ since this is the one commonly used in the literature.

If $r_\lambda > 0$, Theorem 4.2 and $\lambda_1(r_\lambda) > 1 = \lambda_2$ imply that the maximum of the mesh density occurs at a location to the right of that of the maximum of λ_1. This mismatch is illustrated in Figure 3 (right column, third plot).

The argument for $r_\lambda = 0$ is similar, and there is again a mismatch (to the right) between the locations of the maxima of the mesh density and λ_1 (see Figure 3, left column, third plot.

4.1.3. The Winslow case ($p = 1$). If a high mesh concentration is desired at a strict maximum point $r_\lambda > 0$ of λ_1, Theorem 4.2 implies that r_D will be located to the right of r_λ. Nevertheless, as Figure 3 (right column, fourth plot) shows, the mismatch between the maxima for D and λ_1 can be very small (compared with the other cases).

On the other hand, Theorem 4.1 implies that the mesh has higher concentration

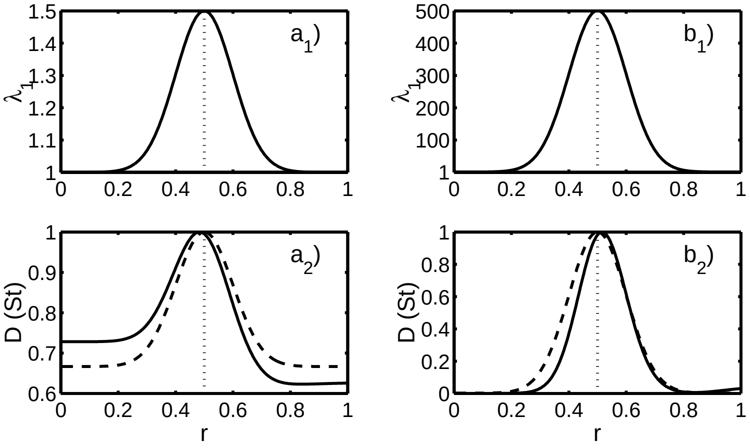


FIG. 4. *Normalized mesh densities obtained with the traditional functional for the strong concentration case ($p = 2$) with $\lambda_1(r) = 1 + A\exp(-(r - r_\lambda)^2/a)/a$ ($a = 0.02$, $r_\lambda = 0.5$) and* (a) *$A = 0.01$ and* (b) *$A = 10$. The top plots show $\lambda_1(r)$. For guidance, in a_2 and b_2 we plot the normalized curve for $\lambda_1(r)$ (dashed lines) along with the normalized densities (solid lines).*

at $r = 0$ if either $\lambda_1'(0) < 0$ or $r = 0$ is a local maximum point of λ_1. Figure 3 (left column, fourth plot) shows good agreement between the shape of λ_1 and the mesh density.

4.1.4. The strong concentration case ($p = 2$). Consider first the case where a higher mesh concentration at the origin is desired. Theorem 4.1 implies that the maximum for the mesh density is located at the origin if $\lambda_1(0) > 1$ (see the last plot in Figure 3, left column). However, the rate of change of the density may be a very large negative value—proportional to $\lim_{r\to 0} e^{2s}/r^3$. This effect is observed in Figure 3 (last plot, left column) where the mesh density is very steep at the origin, giving an overconcentration of points at $r = 0$. Incidentally, our use of the term "strong concentration" for the $p = 2$ case reflects this behavior.

For $r_\lambda > 0$, the current situation is more complex than the previous cases and Theorem 4.2 does not apply if $\lambda_1(r_\lambda) > 1$. Figure 4 shows that r_D can be located to either side of r_λ. Since in this case we have $\lambda_1(0) < \lambda_2(0)$, Theorem 4.1 implies that the mesh concentration has a maximum at the origin. Thus, it is possible for the mesh concentration to have two (or more) maxima, one near the desired location r_λ and a spurious (and steep) maximum at $r = 0$ (see the last plot in Figure 3, right column).

4.2. The new functional. The Euler–Lagrange equation (3.3) corresponding to the new functional is too complex to carry out an analysis similar to the one for the traditional functional, and we instead perform a numerical study of the relation between the monitor function (λ_1 and λ_2) and the mesh density $D(r)$. In particular, we show that by appropriate control of the weighting γ between isotropy and equidistribution it is possible to reduce the mismatch between the location of the maximum for the monitor function and that of the maximum for $D(r)$. As for the traditional functional, we concentrate our attention on monitor functions of the type (2.1) and use the same notation as in (2.2) to designate the most popular choices of p. Note that for the harmonic mapping monitor function $g = 1$, and equidistribution reads as $J =$ constant, giving no mesh control in the new functional. As a result, it is expected

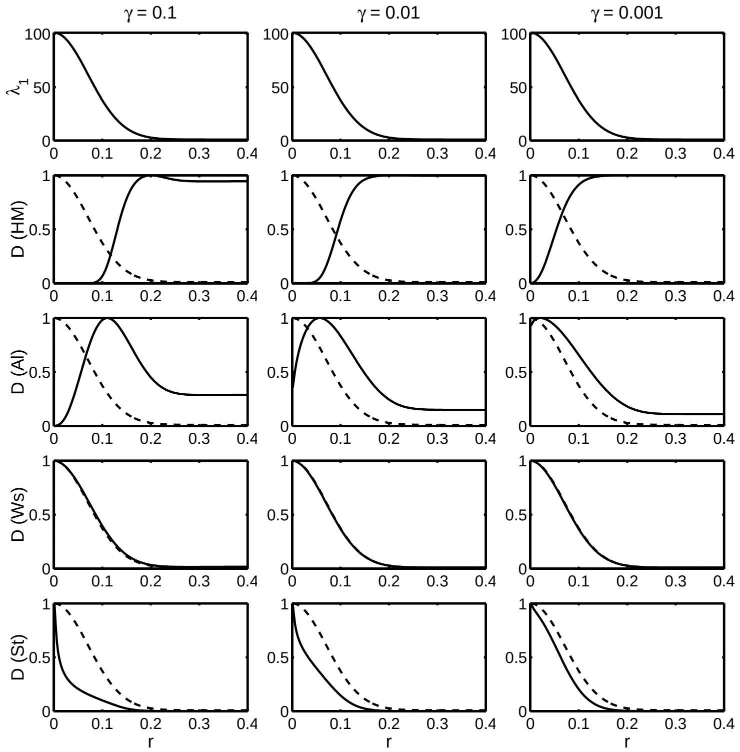


FIG. 5. *Normalized mesh densities (solid lines) obtained with the new functional for a monitor function (dashed lines) such that* $\lambda_1(r) = 1 + \exp(-(r - r_\lambda)^2/a)/a$ *(*$a = 0.01$, $r_\lambda = 0$*) for different choices of* λ_2 *and* γ *(*$q = 2$*).*

that the new functional combined with the harmonic mapping monitor function gives no better results than those with the traditional functional, even for a small value of γ.

4.2.1. Concentration at $r = 0$. For $r_\lambda = 0$, Figure 5 shows the monitor function and the mesh density for the various choices of monitor function (2.2) and weighting between isotropy and equidistribution. For large γ (close to 1/2), the new functional tends to emphasize isotropy, giving similar results to those for the traditional functional. For $\gamma = 0.1$ (first column in Figure 5), the harmonic mapping and the arclength monitor functions tend to misplace the position of the maximum for the density as before. For the Winslow and strong concentration cases, $D(r)$ achieves its maximum at $r = 0$.

Decreasing γ puts more weight on equidistribution, allowing for a better distribution of the mesh density. In fact, by decreasing γ (second and third columns in Figure 5) the maximum for the mesh density is pulled towards the correct position $r = 0$. As pointed out above, the harmonic mapping case fails to have its maximum at $r = 0$ even for very small γ. For the other cases ($p \geq 0$), as γ tends to zero, not only the mesh density has its maximum placed correctly, but its shape tends to mimic the shape of λ_1. This suggests that for small γ it is possible to control the position of the maximum mesh concentration as well as the shape of the mesh density from the choice of monitor function. Interestingly, the Winslow case provides the best control

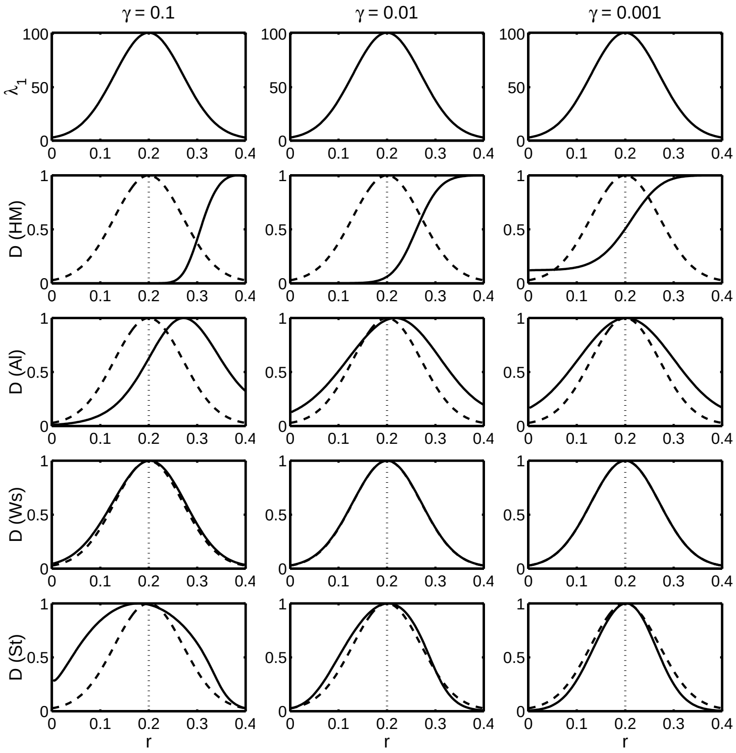


Fig. 6. *Normalized mesh densities (solid lines) obtained with the new functional for a monitor function (dashed lines) such that* $\lambda_1(r) = 1 + \exp(-(r - r_\lambda)^2/a)/a$ *(*$a = 0.001$*,* $r_\lambda = 0.2$*) for different choices of* λ_2 *and* γ *(*$q = 2$*).*

on the mesh density, and for small γ ($\gamma < 0.01$) $D(r)$ is almost indistinguishable from $\lambda_1(r)$.

4.2.2. Concentration at $r > 0$. For $r_\lambda > 0$ we obtain similar results to those for the traditional functional when using a large value of γ (see left column in Figure 6). In particular, the position of the mesh density maximum does not coincide with r_λ except in the Winslow case. As we decrease γ, the new functional weights more towards equidistribution, and the location of the maximum for $D(r)$ tends to approach r_λ, again reinforcing the observation that for small γ and $p \geq 0$ it is possible to have a good control on the mesh density (maximum and shape) from the monitor function.

5. Numerical results. In this section we present some numerical results obtained with the functionals (1.1) and (1.2). For simplicity, square physical and computational domains and structured meshes are used in the computation. As a consequence, axially symmetric meshes are not generated. Nevertheless, the numerical results are sufficient to support the analysis of the previous sections and highlight the level of control of mesh concentration through the monitor functions.

The (two-dimensional) Euler–Lagrange equations for functionals (1.1) and (1.2) are discretized with central finite differences and solved using the moving mesh PDE approach [5, 7]. With this approach, a derivative $(\partial \boldsymbol{x})/(\partial t)$ (where $\boldsymbol{x} = (x, y)^T$) with respect to pseudotime t is added to the Euler–Lagrange equation, and the resulting parabolic system is integrated using a modified backward Euler scheme with which the coefficients of terms $(\partial \boldsymbol{x})/(\partial \xi^i)$ and $(\partial^2 \boldsymbol{x})/(\partial \xi^i \partial \xi^j)$ are calculated at the previous time level. The linear algebraic system is solved using a preconditioned conjugate gradient method. The converged mesh is obtained when the root-mean-square norm of the residual is less than 10^{-4}. All computations start with a uniform mesh of size 41×41 and use a uniform boundary correspondence between Ω and Ω_c. We use $q = 2$ in all cases and, following the common practice, choose λ_1 to be greater than 1.

Example 5.1. The first example is to generate adaptive meshes for the monitor function (1.4) with

$$\lambda_1 = 1 + \frac{1}{a} e^{-(r-0.2)^2/a}, \tag{5.1}$$

where $r = \sqrt{x^2 + y^2}$ and $a = 0.01$. In the (x, y) coordinate system, G has the form

$$G = \frac{\lambda_1}{x^2 + y^2} \begin{pmatrix} x^2 & xy \\ xy & y^2 \end{pmatrix} + \frac{\lambda_2}{x^2 + y^2} \begin{pmatrix} y^2 & -xy \\ -xy & x^2 \end{pmatrix}. \tag{5.2}$$

The goal is to generate meshes with higher point concentration around the circle $x^2 + y^2 = 0.2^2$.

The meshes obtained are shown in Figures 7 and 8. The first row corresponds to the traditional functional, while the second, third, and fourth rows are for the new functional with $\gamma = 0.5$, 0.1 and 0.01, respectively. Each column is associated with a given monitor function.

The left column of Figure 7 shows that the mesh concentration is badly misplaced for both the traditional and new functionals using the harmonic mapping monitor function ($p = -1$). In this case the traditional functional gives exactly the harmonic mapping method used by Dvinsky [4]. Note that the new functional does not work well, as expected, since $g = 1$ and $J =$ constant, giving no control of mesh concentration.

For the arclength monitor function ($p = 0$, the right column of Figure 7), the traditional functional still produces the mismatched concentration. However, since $g = \lambda$ and the equidistribution becomes $J\sqrt{\lambda} =$ constant, the new functional bears the feature of equidistribution and leads to the correct concentration when a small value of γ is used.

Interestingly, with the Winslow-type monitor function, both the traditional and new functionals generate correct mesh concentration—see the left column of Figure 8. For the case of strong concentration with $p = 2$ (see the right column of Figure 8), the new functional produces the correct results, whereas the traditional one seems to overconcentrate mesh points inside the circle $x^2 + y^2 = 0.2^2$, although there is also concentration around the circle.

From these two figures one can also see that the new functional with $\gamma = 0.5$ leads to results similar to but slightly less adaptive than those obtained with the traditional functional.

Example 5.2. The second example is to generate adaptive meshes for the monitor function (1.4) with

$$\lambda_1 = 1 + \frac{1}{a} e^{-r^2/a}, \qquad a = 0.01. \tag{5.3}$$

The goal is now to generate adaptive meshes with higher point concentration at the origin.

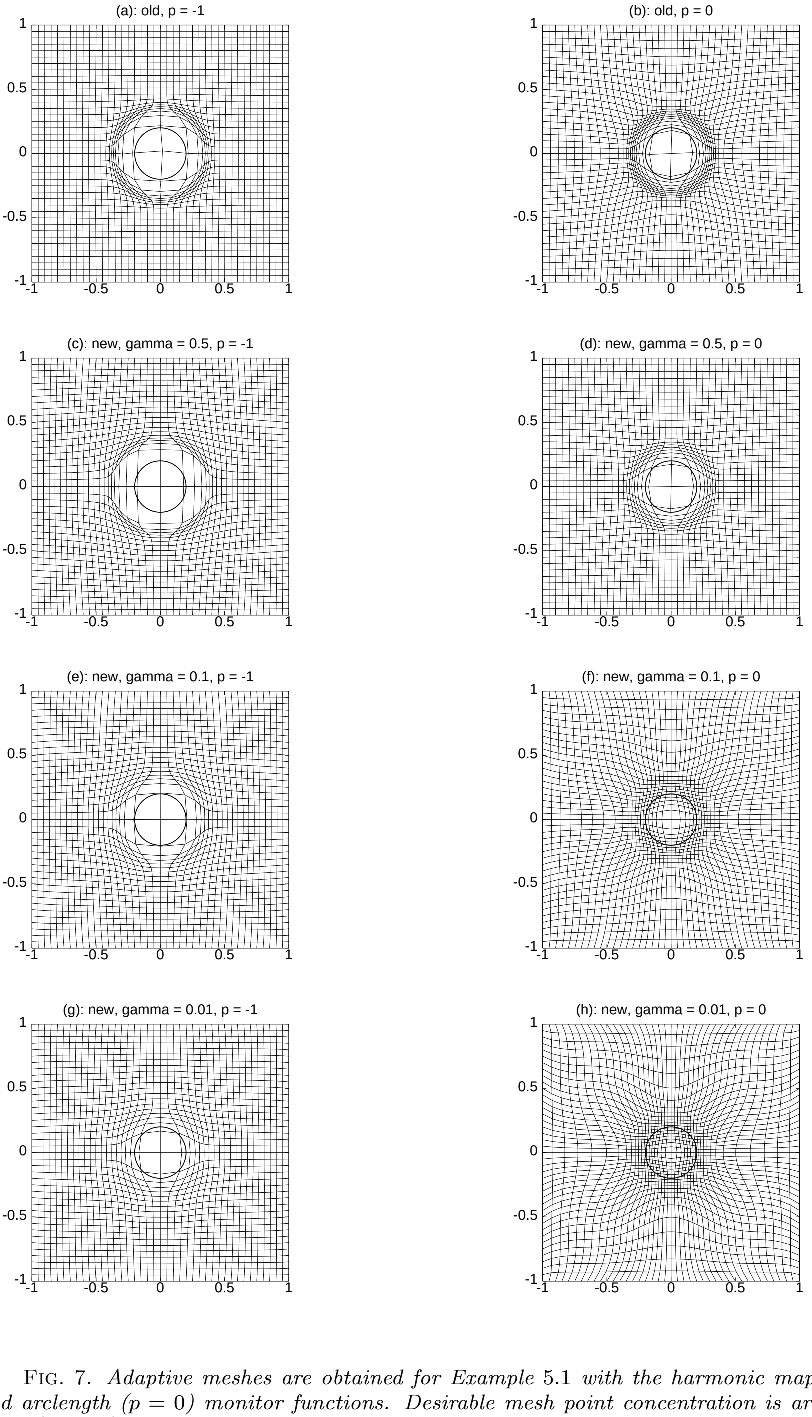


FIG. 7. *Adaptive meshes are obtained for Example* 5.1 *with the harmonic mapping* ($p = -1$) *and arclength* ($p = 0$) *monitor functions. Desirable mesh point concentration is around the circle* $x^2 + y^2 = 0.2^2$ *(the bold solid circle).*

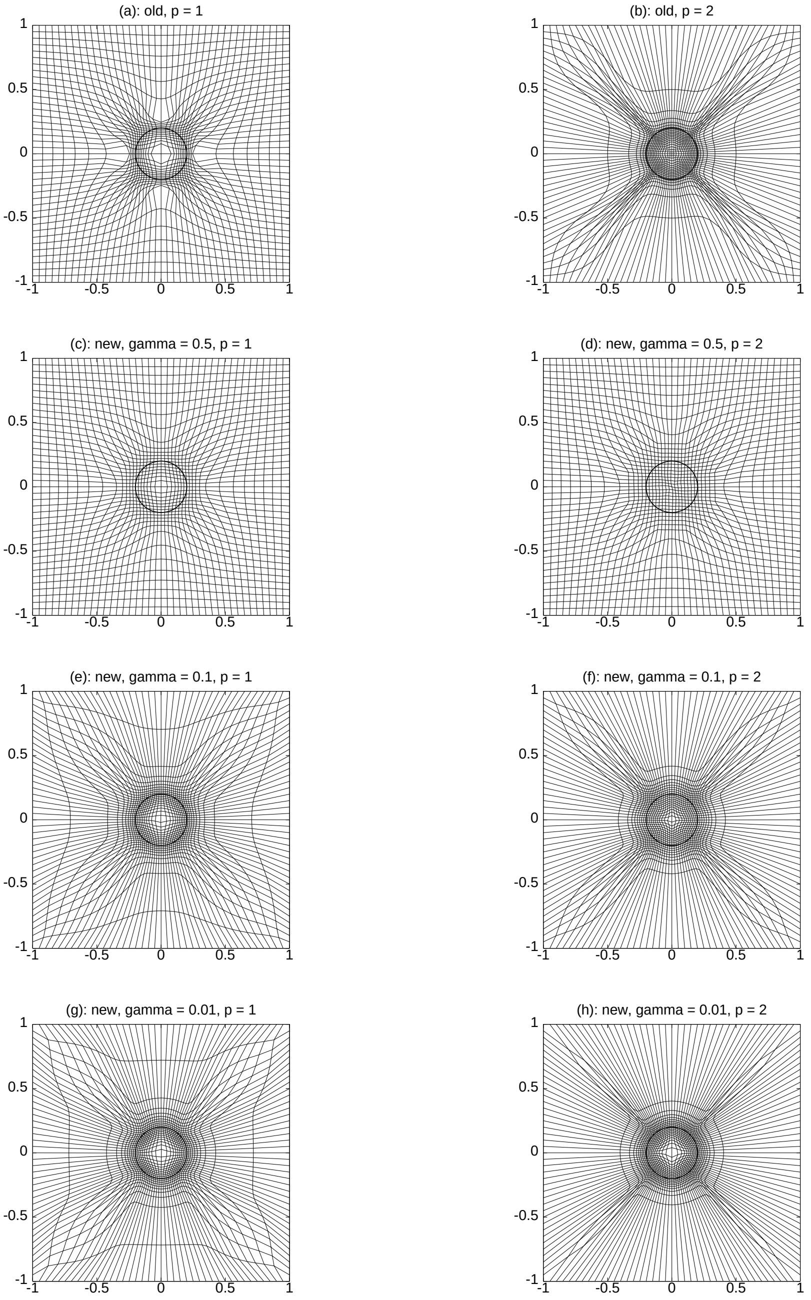

(a): old, p = 1
(b): old, p = 2
(c): new, gamma = 0.5, p = 1
(d): new, gamma = 0.5, p = 2
(e): new, gamma = 0.1, p = 1
(f): new, gamma = 0.1, p = 2
(g): new, gamma = 0.01, p = 1
(h): new, gamma = 0.01, p = 2
1
0.5
0
-0.5
-1
-1
-0.5
0
0.5
1

FIG. 8. *Adaptive meshes are obtained for Example* 5.1 *with the Winslow-type* ($p = 1$) *and strong concentration* ($p = 2$) *monitor functions. Desirable mesh point concentration is around the circle* $x^2 + y^2 = 0.2^2$ *(the bold solid circle).*

The meshes obtained are shown in Figures 9 and 10. The results confirm the observations made in Example 5.1 and the analysis given in the preceding sections. That is, the traditional functional misplaces meshes for the harmonic mapping and arclength monitor functions and correctly places them for the Winslow-type and strong concentration monitor functions; the new functional with $\gamma = 0.5$ leads to meshes similar to but slightly less adaptive than those obtained with the traditional functional; and the new functional with a small value of γ leads to meshes with correct concentration when the arclength, Winslow-type, or strong concentration monitor function is used.

6. The traditional functional for spherically symmetric problems. A similar analysis can be carried out for the traditional functional applied to spherically symmetric problems in three dimensions. Consider

$$I_{trad}[\xi,\eta,\zeta] = \int_\Omega \left(\nabla\xi^T G^{-1}\nabla\xi + \nabla\eta^T G^{-1}\nabla\eta + \nabla\zeta^T G^{-1}\nabla\zeta\right) dxdydz, \tag{6.1}$$

where $\Omega = \{(x,y,z) \,|\, x^2+y^2+z^2 < 1\}$. Take $\Omega_c = \{(\xi,\eta,\zeta) \,|\, \xi^2+\eta^2+\zeta^2 < 1\}$, and let the spherical coordinates for Ω and Ω_c be

$$\begin{cases} x = r\sin(\theta)\cos(\phi), \\ y = r\sin(\theta)\sin(\phi), \\ z = r\cos(\theta), \end{cases} \qquad \begin{cases} \xi = R\sin(\Theta)\cos(\Phi), \\ \eta = R\sin(\Theta)\sin(\Phi), \\ \zeta = R\cos(\Theta). \end{cases}$$

Consider the case where the physical solution is spherically symmetric. Assume that the corresponding mesh adaptation is also spherically symmetric, i.e.,

$$R = R(r), \quad \Theta = \theta, \quad \Phi = \phi. \tag{6.2}$$

Then it is reasonable to use the monitor function in the form

$$G = \lambda_1(r)\boldsymbol{e}_r\boldsymbol{e}_r^T + \lambda_2(r)\boldsymbol{e}_\theta\boldsymbol{e}_\theta^T + \lambda_3(r)\boldsymbol{e}_\phi\boldsymbol{e}_\phi^T, \tag{6.3}$$

where $\boldsymbol{e}_r$, $\boldsymbol{e}_\theta$, and $\boldsymbol{e}_\phi$ are the unit vectors in the radial, latitudinal, and longitudinal axes. Under the symmetry assumption, (6.1) reduces to

$$I_{trad}[R] = \int_0^1 \left[\frac{1}{\lambda_1}\left(\frac{dR}{dr}\right)^2 + \frac{2}{\lambda_{23}}\left(\frac{R}{r}\right)^2\right] r^2 dr,$$

where λ_{23} is defined as

$$\frac{2}{\lambda_{23}} = \frac{1}{\lambda_2} + \frac{1}{\lambda_3}.$$

The corresponding boundary value problem is given by

$$\begin{aligned} &-\frac{d}{dr}\left(\frac{r^2}{\lambda_1}\frac{dR}{dr}\right) + \frac{2}{\lambda_{23}}R = 0, \\ &R(0) = 0, \qquad R(1) = 1. \end{aligned} \tag{6.4}$$

The transformation (4.1) can be used for analyzing (6.4). We obtain the equation for v

$$v' = \frac{\lambda_1}{r}\left(\frac{2}{\lambda_1\lambda_{23}} - \frac{v}{\lambda_1} - v^2\right) \tag{6.5}$$

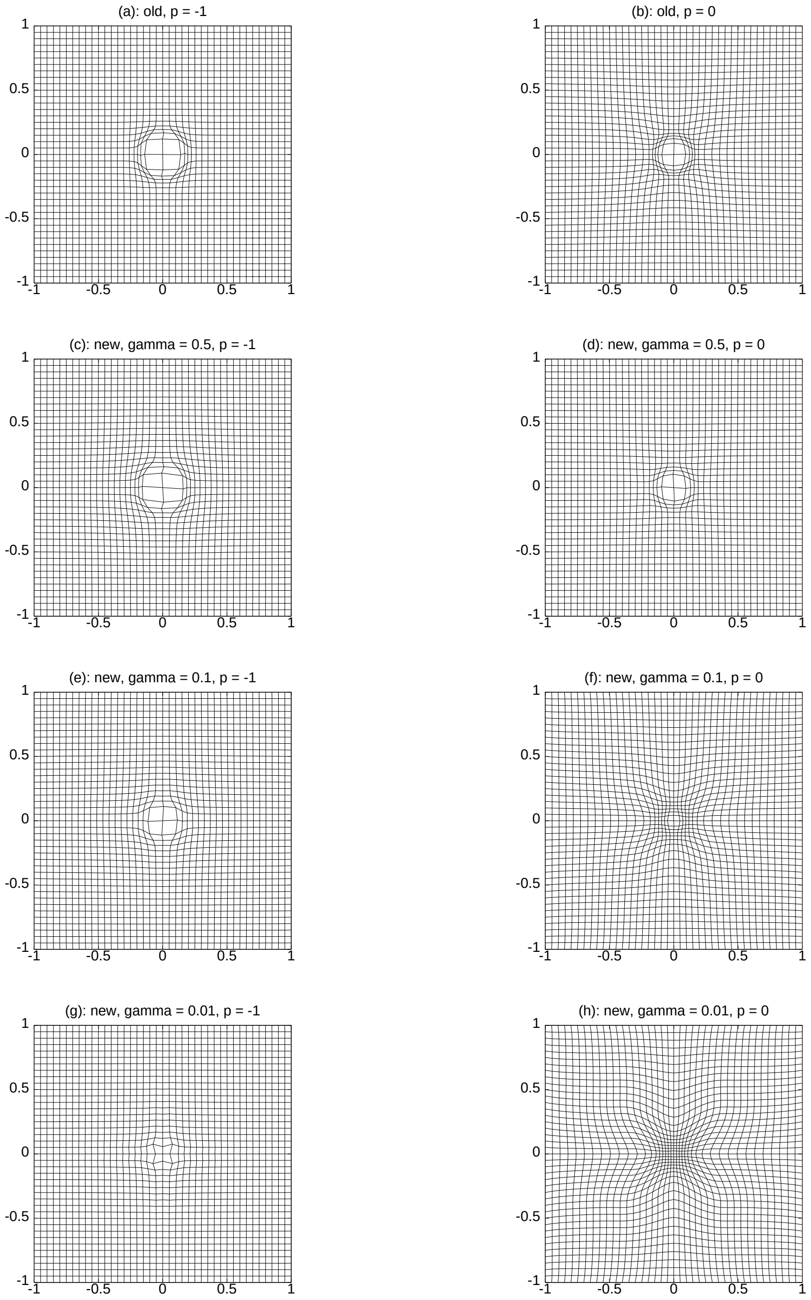


Fig. 9. *Adaptive meshes are obtained for Example* 5.2 *with the harmonic mapping* ($p = -1$) *and arclength* ($p = 0$) *monitor functions. Desirable mesh point concentration is near the origin.*

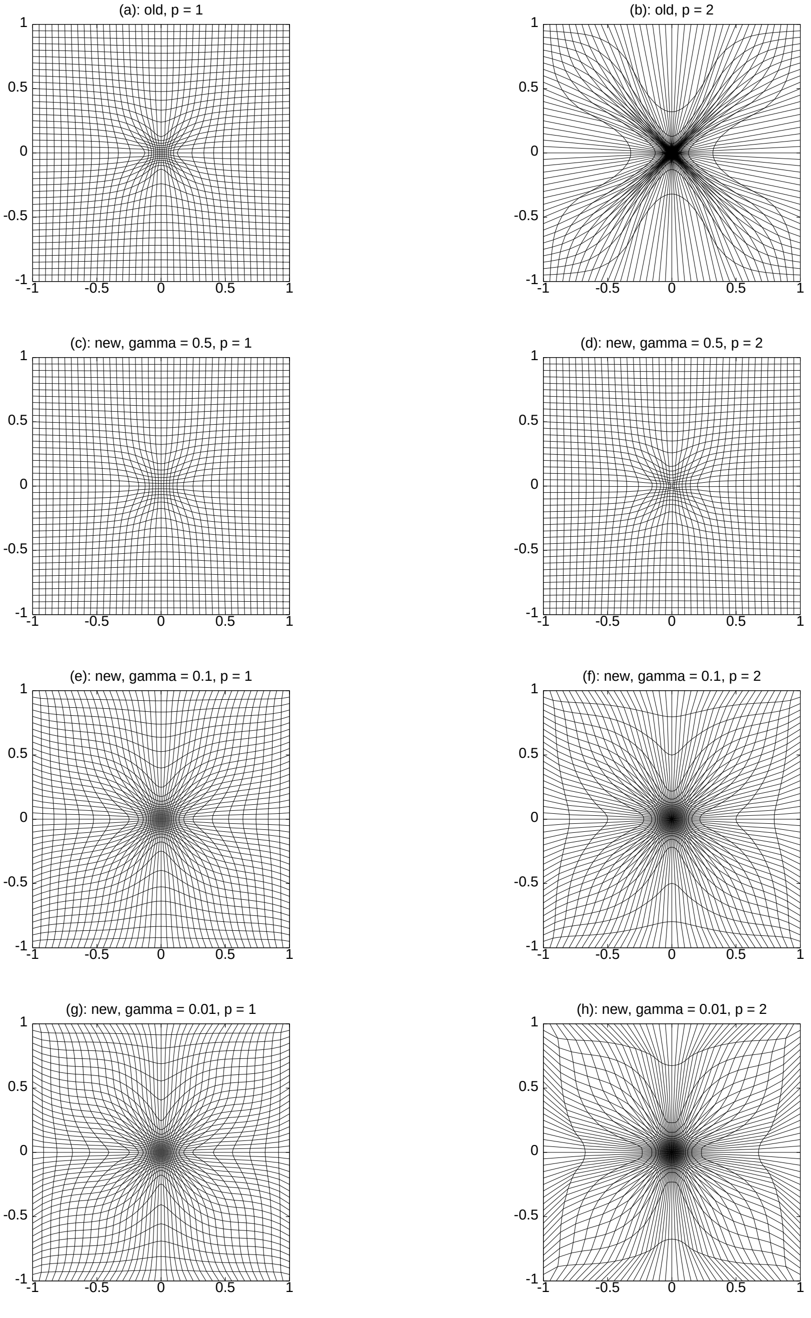


FIG. 10. *Adaptive meshes are obtained for Example* 5.2 *with the Winslow-type* ($p = 1$) *and strong concentration* ($p = 2$) *monitor functions. Desirable mesh point concentration is near the origin.*

which is subject to the initial condition

$$v(0) = \frac{1}{2}\sqrt{\frac{1}{\lambda_1(0)^2} + \frac{8}{\lambda_1(0)\lambda_{23}(0)}} - \frac{1}{2\lambda_1(0)}. \tag{6.6}$$

It is straightforward to show that the solution v of (6.5) has the properties stated in Lemmas 4.1–4.3.

In the current situation, the mesh density has the form

$$D(r) = \frac{R^2}{r^2}\frac{dR}{dr}.$$

Its rate of change reads as

$$\begin{aligned} D'(r) &= \frac{e^{3s}}{r^3}\left[(\lambda_1 v)' - \frac{3(\lambda_1 v)}{r} + \frac{3(\lambda_1 v)^2}{r}\right] \\ &= \frac{e^{3s}}{r^3}\left[\lambda_1^{'} v + \frac{2\lambda_1^2}{r}\left(\left(v - \frac{1}{\lambda_1}\right)^2 + \left(\frac{1}{\lambda_1\lambda_{23}} - \frac{1}{\lambda_1^2}\right)\right)\right]. \end{aligned} \tag{6.7}$$

We have the following theorems which are basically identical to Theorems 4.1 and 4.2. One may notice that in this three-dimensional case, the relation between λ_1 and λ_{23}, rather than those between λ_1 and each of λ_2 and λ_3, plays a role in affecting the corresponding mesh adaptation.

THEOREM 6.1. (i) *If $\lambda_1(0) - \lambda_{23}(0) \neq 0$, then $D'(0)$ has the same sign as $\lambda_1(0) - \lambda_{23}(0)$, whether $r = 0$ is a maximum point of λ_1 or not.*

(ii) *Let $\lambda_{23}(r) = \lambda_1(r)$. If $\lambda_1^{'}(0) \neq 0$, then $D'(0)\lambda_1^{'}(0) > 0$. If $\lambda_1^{'}(0) = 0$ but $\lambda_1^{''}(0) \neq 0$, then $D'(0)\lambda_1^{''}(0) > 0$.*

THEOREM 6.2. *Let $r_\lambda > 0$.*

(i) *If $\lambda_1(r_\lambda) > \lambda_2(r_\lambda)$, then $D'(r_\lambda) > 0$ and thus $r_D > r_\lambda$.*

(ii) *Further, if we assume that $\lambda_{23}(r) = \lambda_1(r)$ and r_λ is a strict maximum point of λ_1 (i.e., $\lambda_1''(r_\lambda) < 0$), then $D'(r_\lambda) > 0$ or $r_D > r_\lambda$.*

7. Conclusions and comments. The question of how variational grid generators behave when solving problems with axisymmetric solutions has been investigated. Specifically, two functionals have been analyzed in the previous sections for their abilities to precisely control the mesh concentration via monitor functions. One is the traditional functional (1.1) which includes Winslow's method and Dvinsky's method of harmonic mappings as special cases. The other is the new functional (1.2) proposed by Huang in [6] which explicitly includes the isotropy (or regularity) and equidistribution features. The analysis is primarily done for axisymmetrical problems in two dimensions. For axially symmetric mesh adaptation, it is reasonable to use a monitor function of the form in (1.4).

Theoretical results for the traditional functional are given in Theorems 4.1 and 4.2. Specifically, when higher mesh concentration at the origin is desired, a choice of the radial and angular components λ_1 and λ_2 of the monitor function satisfying $\lambda_1(0) < \lambda_2(0)$ will make the mesh denser at $r = 0$ than in the surrounding area whether or not λ_1 has a maximum value at $r = 0$. The purpose can also be served by choosing λ_1 to have a local maximum at $r = 0$ when a Winslow-type monitor function with $\lambda_1(r) = \lambda_2(r)$ is employed. Unfortunately, the choice $\lambda_2(r) = \lambda_1(r)^p$ with $p < 0$, which includes Dvinsky's method of harmonic mappings and the arclength monitor function as special cases, will not satisfy the condition $\lambda_1(0) < \lambda_2(0)$ if $\lambda_1(0) > 1$ (as commonly taken in the literature) and leads to a mesh with coarser concentration of points in the center than in the surrounding area.

On the other hand, when higher mesh concentration around a ring $r = r_\lambda > 0$ is desired, the traditional functional provides far less control by choosing λ_1 and λ_2. Indeed, Theorem 4.2 shows that there surely is a mismatch between the position r_λ of the maximum of λ_1 and the location r_D of the maximum of the mesh density if either (a) $\lambda_1(r_\lambda) > \lambda_2(r_\lambda)$ (which is the case for the harmonic mapping or the arclength monitor function with $\lambda_1(r_\lambda) > 1$) or (b) a Winslow-type monitor function is used and r_λ is a strict maximum point of λ_1. Moreover, a mismatch between r_D and r_λ is also possible for the case $\lambda_1(r_\lambda) < \lambda_2(r_\lambda)$. Indeed, the numerical results show that r_D can be located to either side of r_λ when $\lambda_1(r_\lambda) > 1$ and λ_2 is taken as $\lambda_2 = \lambda_1^p$ for $p > 1$. Nevertheless, the numerical results suggest that $|r_D - r_\lambda|$ is relatively small for the Winslow case $\lambda_2 = \lambda_1$. The analysis also shows that for the harmonic mapping case $\lambda_2 = 1/\lambda_1$, r_D can be made to agree with r_λ by rescaling λ_1 such that $\lambda_1(r_\lambda) = 1$. However, this can be done if the mesh concentration is needed only at the location of the (global) maximum of λ_1.

For axially symmetric problems, the new functional leads to a nonlinear mesh equation too complex to permit an analysis like that for the traditional functional. Nevertheless, numerical results presented in sections 4 and 5 show that the new functional offers explicit control for mesh concentration by adjusting the value of γ that weights the isotropy and equidistribution. Specifically, when using a large value of γ (close to 1/2) we obtain similar results to those for the traditional functional cases. However, as we decrease γ, the new functional weights more towards equidistribution, and both the location of the maximum and the profile of the mesh density tend to coincide with those of λ_1 for a monitor function with a nonconstant determinant. For the case of the harmonic mapping monitor function, the determinant is $g = 1$ and equidistribution becomes $J =$ constant so no control of mesh concentration is possible by choosing λ_1. Thus, as expected, the new functional does not work in this case even when a small value of γ is used.

Analysis has also been carried out for the traditional functional applied to spherically symmetric problems in three dimensions. The results are stated in Theorems 6.1 and 6.2.

In the future we intend to investigate a number of higher-dimensional axisymmetrical problems arising in physical applications and show the practicability of the methods which have performed well here.

Acknowledgment. The authors are grateful to the referees for their valuable comments.

REFERENCES

[1] J. U. Brackbill and J. S. Saltzman, *Adaptive zoning for singular problems in two dimensions*, J. Comput. Phys., 46 (1982), pp. 342–368.

[2] W. Cao, W. Huang, and R. D. Russell, *A study of monitor functions for two-dimensional adaptive mesh generation*, SIAM J. Sci. Comput., 20 (1999), pp. 1978–1994.

[3] C. de Boor, *Good approximation by splines with variable knots* II, in Proceedings of the Conference on the Numerical Solution of Differential Equations, Dundee, Scotland, 1973, Lecture Notes in Math. 363, G. A. Watson, ed., Springer-Verlag, Berlin, 1974, pp. 12–20.

[4] A. S. Dvinsky, *Adaptive grid generation from harmonic maps on Riemannian manifolds*, J. Comput. Phys., 95 (1991), pp. 450–476.

[5] W. Huang, *Practical aspects of formulation and solution of moving mesh partial differential equations*, J. Comput. Phys., 171 (2001), pp. 753–775.

[6] W. Huang, *Variational mesh adaptation: Isotropy and equidistribution*, J. Comput. Phys., 174 (2001), pp. 903–924.

[7] W. Huang and R. D. Russell, *A high dimensional moving mesh strategy*, Appl. Numer. Math., 26 (1997), pp. 63–76.

[8] W. Huang and W. Sun, *Variational mesh adaptation* II: *Error estimates and monitor functions*, J. Comput. Phys., to appear.

[9] P. Knupp, *Mesh generation using vector-fields*, J. Comput. Phys., 119 (1995), pp. 142–148.

[10] P. M. Knupp, *Jacobian-weighted elliptic grid generation*, SIAM J. Sci. Comput., 17 (1996), pp. 1475–1490.

[11] P. M. Knupp and N. Robidoux, *A framework for variational grid generation: Conditioning the Jacobian matrix with matrix norms*, SIAM J. Sci. Comput., 21 (2000), pp. 2029–2047.

[12] W. Ren and X.-P. Wang, *An iterative grid redistribution method for singular problems in multiple dimensions*, J. Comput. Phys., 159 (2000), pp. 246–273.

[13] S. Steinberg and P. J. Roache, *Variational grid generation*, Numer. Methods Partial Differential Equations, 2 (1986), pp. 71–96.

[14] A. Winslow, *Numerical solution of the quasi-linear Poisson equation in a nonuniform triangle mesh*, J. Comput. Phys., 1 (1967), pp. 149–172.

[15] A. M. Winslow, *Adaptive Mesh Zoning by the Equipotential Method*, Technical report UCID-19062, Lawrence Livemore Laboratory, Livermore, CA, 1981.

SIAM J. NUMER. ANAL.
Vol. 41, No. 1, pp. 258–271

CONVERGENCE ESTIMATES FOR THE GENERALIZED DAVIDSON METHOD FOR SYMMETRIC EIGENVALUE PROBLEMS I: THE PRECONDITIONING ASPECT*

E. OVTCHINNIKOV†

Abstract. The generalized Davidson (GD) method can be viewed as a generalization of the preconditioned steepest descent (PSD) method for solving symmetric eigenvalue problems. There are two aspects of this generalization. The most obvious one is that in the GD method the new approximation is sought in a larger subspace, namely the one that spans all the previous approximate eigenvectors, in addition to the current one and the preconditioned residual thereof. Another aspect relates to the preconditioning. Most of the available results for the PSD method are associated with the same view on preconditioning as in the case of linear systems. Consequently, they fail to detect the superlinear convergence for certain "ideal" preconditioners, such as the one corresponding to the "exact" version of the Jacobi–Davidson method—one of the most familiar instances of the GD method. Focusing on the preconditioning aspect, this paper advocates an alternative approach to measuring the quality of preconditioning for eigenvalue problems and presents corresponding nonasymptotic convergence estimates for the GD method in general and Jacobi–Davidson method in particular that correctly detect known cases of the superlinear convergence.

Key words. iterative methods for eigenvalue problems, preconditioning, generalized Davidson method, convergence estimates, superlinear convergence

AMS subject classifications. 65F15, 65N12, 65N22, 65N30

PII. S0036142902411756

1. Introduction.

This paper presents convergence estimates for a class of iterative methods for finding the smallest eigenvalue of the generalized eigenvalue problem

$$Lu = \lambda M u \tag{1}$$

or, equivalently, finding the global minimum of the Rayleigh quotient functional

$$\lambda(u) \equiv \frac{(Lu, u)}{(Mu, u)}, \tag{2}$$

where L and M are, respectively, a symmetric and a symmetric positive definite linear operator in a Euclidean space $\mathcal{E}$, and $(\cdot, \cdot)$ is the scalar product in $\mathcal{E}$. The class of methods in question embraces methods based on the following iterative scheme for calculating λ_0 numerically: *given an arbitrary nontrivial vector u^0 calculate a sequence λ^n of approximations to λ_0 using the following recurrent formula:*

$$\begin{aligned} u^n &= \arg \min_{u \in \mathcal{D}^n, (Mu,u)=1} \lambda(u), \quad \lambda^n = \lambda(u^n), \\ \mathcal{D}^{n+1} &= \operatorname{span}\{u^0, \ldots, u^n, K_n(L - \lambda^n M)u^n\}, \end{aligned} \tag{3}$$

where K_n are some linear operators in $\mathcal{E}$.

Two familiar methods based on the iterative scheme (3) are the Davidson method (see, e.g., [4, 3]) and the Jacobi–Davidson method (see, e.g., [29, 27, 9, 28] and the

[*] Received by the editors January 11, 2002; accepted for publication (in revised form) September 16, 2002; published electronically March 26, 2003.
http://www.siam.org/journals/sinum/41-1/41175.html

[†] Harrow School of Computer Science, University of Westminster, Watford Road, London HA1 3TP, United Kingdom (eeo@wmin.ac.uk).

relevant chapters in [1]). In the former, L is a matrix, $M = I$ (the unit matrix), and $K_n = (D_L - \lambda^n I)^{-1}$, where D_L is the diagonal of L (i.e., a diagonal matrix of the same size and with the same diagonal entries as in L). In the Jacobi–Davidson method $v = K_n r$ is calculated by approximately solving the linear system

$$(1 - \pi_n)^T (L - \lambda^n M)(1 - \pi_n)v = (1 - \pi_n)^T r, \quad \pi_n v = 0, \tag{4}$$

where $\pi_n w = (Mw, u^n)u^n$ (this approach is apparently closely related to the so-called shift-and-invert technique—see, e.g., [1, 24]). The general case is sometimes referred to as the generalized Davidson (GD) method [1, 25, 22], which is the term adopted in the present paper.

Nowadays, extensive numerical evidence of the efficiency of the GD method, especially in the Jacobi–Davidson case, is available in the literature (see, e.g., the above references). The present paper is one of a few that focus on *theoretical* convergence results for the GD method. Accordingly, the GD method is presented here in an abstract mathematical form (3), which is convenient for the convergence analysis, but not for its practical implementation, which should follow, e.g., the guidelines in [1].

The GD method can be viewed as a generalization of the preconditioned steepest descent (PSD) iterations

$$u^{n+1} = u^n - \tau_n g^n, \quad g^n = K_n(L - \lambda(u^n)M)u^n, \tag{5}$$

where τ_n are suitably chosen parameters, e.g.,

$$\tau_n = \arg\min_\tau \lambda(u^n - \tau g^n), \tag{6}$$

and K_n is a symmetric positive definite operator usually referred to as *preconditioner*. Preconditioned eigensolvers based on (5) are among the oldest and best-studied to date (see, e.g., [26, 7, 5, 6, 11, 12, 17, 18, 19]; for the related historical overview, see [2, 13]). An obvious aspect of the generalization of (5)–(6) into the GD method is the minimization of $\lambda(u)$ over a larger subspace, and this aspect is addressed by Part II of this paper [23], whereas here we address a less obvious, yet important aspect of the preconditioning.

Formally speaking, both (3) and (5) can be used with the same preconditioners[1] K_n. However, the convergence results available to date for (5) (see, e.g., the references of the previous paragraph; cf. also section 3 below) are associated with essentially the same view on preconditioning as that for linear systems, which is certainly not the case with Davidson and Jacobi–Davidson methods. The "ideal" preconditioner for the (nondegenerate) linear system $Lu = f$ is obviously $K = L^{-1}$—indeed, with this preconditioner the convergence can be achieved in just one iteration. Accordingly, in the general case the quality of a preconditioner K is measured by its closeness to L^{-1}, which can be expressed quantitatively via the spectral condition number (the ratio of the largest and the smallest eigenvalue) of KL, the smaller the better. The convergence results for (5) mentioned above are formulated in terms of the same ratio (or related quantities) and thereby suggest the same "ideal" preconditioner. However, for eigenvalue problems $K = L^{-1}$ is obviously not the best choice, since it fails to deliver a superlinear convergence achievable with some other preconditioners.

It is argued in [13] (cf. also [29]) that an "ideal" preconditioner for (5) would be the pseudoinverse of $L_0 \equiv L - \lambda_0 M$. Indeed, with this preconditioner the convergence

[1] In the PSD method, K_n usually does not depend on n, but this fact is of little importance for the convergence analysis of this method.

of (5), and hence (3), is superlinear (in fact, it is cubic—cf. section 5). Accordingly, the natural measure of the quality of preconditioning for eigenvalue problems should be based rather on how far the preconditioner K_n is from the pseudoinverse of L_0. An example of a convergence estimate for (5) in such terms is an asymptotic estimate in [26], which was later reproduced for the GD method in [22]. The major disadvantage of both is that they are asymptotic, i.e., contain unknown "small enough" quantities. In the important case of a parameter-dependent problem (e.g., a discretized differential one, which involves discretization parameters), such estimates are insufficient for comprehensive analysis of the convergence of iterative methods and its dependence on the parameters of the problem. In the present paper a similar but *nonasymptotic* estimate is derived (Theorem 5.2), alongside an analogous estimate related to the Jacobi–Davidson-type preconditioning (Theorem 4.2).

The outline of the paper is as follows. After introducing the notation in section 2, we discuss in section 3 the available convergence estimates for the iterations (5) that are applicable to the GD method. It should be noted that the discussion is restricted to the most advanced results, leaving out quite a few historically important ones (cf. [2, 13, 16] for a more detailed review). In section 4 we study the convergence of the Jacobi–Davidson method. We show that the approximate solution of (4) is equivalent to the use of preconditioners K_n (in the GD method) that satisfy the assumption (22), which measures the quality of preconditioning against the "ideal" case of solving (4) exactly. Based on this assumption, we present a new convergence estimate that correctly predicts the cubic convergence in this "ideal" case. In section 5 we discuss the relation between the two "ideal" cases mentioned above and present a new nonasymptotic convergence estimate for the GD method that detects the cubic convergence in the case when K_n is the pseudoinverse of L_0.

2. Notation. In this paper we use the standard notation $A > 0$ ($A \geq 0$) to declare that a symmetric linear operator A is positive definite (resp., semidefinite). Accordingly, $A \geq B$ stands for $A - B \geq 0$, etc. For $A \geq 0$ we denote (Au, v) by $(u, v)_A$, and $\|u\|_A$ stands for $\sqrt{(u, u)_A}$.

The minimal eigenvalue of (1) is denoted λ_0, and λ_1 denotes the second smallest *distinct* eigenvalue. The invariant subspace of $M^{-1}L$ corresponding to λ_0 is denoted by $\mathcal{I}_0$, and π denotes the $(\cdot,\cdot)_M$-orthogonal projection onto $\mathcal{I}_0$. The orthogonal complement to $\mathcal{I}_0$ in $\mathcal{E}$ is denoted by $\bar{\mathcal{I}}_0$. For any projection π' we denote $\bar{\pi}' \equiv 1 - \pi'$.

The projection π has the following simple properties:

$$L\pi = \pi^T L = \lambda_0 M\pi, \quad M\pi = \pi^T M, \quad L_0 \equiv L - \lambda_0 M = L_0\bar{\pi} = \bar{\pi}^T L_0,$$

the last one implying that

$$\|\bar{\pi}u\|^2_{L_0} = \|u\|^2_{L_0} = \|u\|^2_L - \lambda_0\|u\|^2_M = (\lambda(u) - \lambda_0)\|u\|^2_M.$$

Further,

$$\begin{aligned}\lambda(u)\|u\|^2_M = \|u\|^2_L = \|\pi u\|^2_L + \|\bar{\pi}u\|^2_L = \lambda_0\|\pi u\|^2_M + \lambda(\bar{\pi}u)\|\bar{\pi}u\|^2_M \\ \geq \lambda_0\|\pi u\|^2_M + \lambda_1\|\bar{\pi}u\|^2_M = \lambda_0\|u\|^2_M + (\lambda_1 - \lambda_0)\|\bar{\pi}u\|^2_M.\end{aligned}$$

Thus, for any vector u

$$\|\bar{\pi}u\|^2_M \leq \delta(\lambda(u))\|u\|^2_M, \quad \|\bar{\pi}u\|^2_{L_0} = (\lambda(u) - \lambda_0)\|u\|^2_M, \tag{7}$$

where

$$\delta(\lambda) = \frac{\lambda - \lambda_0}{\lambda_1 - \lambda_0}.$$

3. Available convergence estimates. From the minimax principle for eigenvalues it follows that $\lambda(u_0(\mathcal{D}^{n+1})) \le \lambda(u^n - \tau g^n)$ for any τ. Hence, convergence results for any method based on (5) in terms of eigenvalues[2] apply to (3) with the same preconditioners K_n. In view of this, let us start with a brief account of (the most advanced) available convergence results for the PSD method (5), paying special attention to the assumptions on K_n.

As mentioned in the introduction, in the convergence analysis for methods based on (5) the following assumption on K_n is standard:

$$aK_n^{-1} \le L \le bK_n^{-1}, \tag{8}$$

where a and b are some positive constants, or, equivalently,

$$aL^{-1} \le K_n \le bL^{-1} \tag{9}$$

or else

$$(1-\gamma)L \le \tau_n L K_n L \le (1+\gamma)L, \tag{10}$$

where $0 \le \gamma < 1$ and $\tau_n > 0$. The equivalence between (8) and (10) is established by the following relationships:

$$\tau_n = \frac{2}{a+b}, \quad \gamma = \frac{b-a}{b+a}. \tag{11}$$

One of the best available convergence results for the preconditioned iterations (5) under the assumption (10) is given in [19] (see also [14, 15, 16] and the related results in [17, 18]):

$$\Delta(\lambda^{n+1}) \le q(\gamma)^2 \Delta(\lambda^n), \quad \Delta(\mu) = \frac{\mu - \lambda_0}{\lambda_1 - \mu}, \quad q(\gamma) = \gamma + (1-\gamma)\frac{\lambda_0}{\lambda_1}, \tag{12}$$

where $\lambda^n \equiv \lambda(u^n)$. The above estimate is sharp for iterations (5) with fixed $\tau_n = \tau$. For variable τ_n one can find in [12] an asymptotically better estimate: assuming (8) one has

$$\lambda^{n+1} - \lambda_0 \le (\tilde{\gamma} + \epsilon(\lambda^n - \lambda_0))^2 (\lambda^n - \lambda_0), \tag{13}$$

where $\epsilon(t) = \mathcal{O}(\sqrt{t})$ and

$$\tilde{\gamma} = \frac{1-\xi}{1+\xi}, \quad \xi = \frac{a}{b}\left(1 - \frac{\lambda_0}{\lambda_1}\right).$$

The same result for the corresponding variant of the GD method was obtained independently in [22].

The main disadvantage of the above results is their failure to detect cubic convergence of (5) achievable with, e.g., $K = L_0^{-1}\bar{\pi}^T$ (here and below L_0^{-1} is the inverse of the restriction of $L_0 = L - \lambda_0 M$ onto $\bar{\mathcal{I}}_0$). To overcome this limitation, which is apparently closely related to the form of the assumptions (8) and (10), [22] considers the assumption

$$(1-\gamma)L_0 \le \tau_n L_0 K_n L_0 \le (1+\gamma)L_0, \tag{14}$$

[2] It is important to emphasize that the convergence results for (5) in other terms (e.g., those in [8, 20, 30]) may not be straightforwardly applicable to (3).

leading to the following asymptotic estimate:

$$\lambda^{n+1} - \lambda_0 \le (\gamma + \epsilon_n)^2(\lambda^n - \lambda_0), \tag{15}$$

where $\epsilon_n = \mathcal{O}(\sqrt{\lambda^n - \lambda_0})$. It should be noted that [22] addresses the GD method. As far as the PSD method is concerned, the above estimate was actually available as early as in 1958 [26] and represents the first convergence result for preconditioned eigensolvers. We observe that in the "ideal" case $\gamma = 0$ the estimate (15) predicts quadratic convergence for (5) (in reality, it is cubic—cf. section 5).

As mentioned in the introduction, the main drawback of the estimate (15) is the presence of an unknown (albeit asymptotically insignificant) term ϵ_n. Furthermore, it is not clear how to apply this estimate to some practical implementations of the GD method, such as the Jacobi–Davidson method. In section 4 below we show that the approximate solution of (4) in the Jacobi–Davidson method is equivalent to using the GD method with preconditioners K_n satisfying the assumption (22). In subsequent section 5 we show how this new assumption is related to (a generalization of) (14), thereby illuminating the relevance of the latter to the Jacobi–Davidson method and, furthermore, present a nonasymptotic analogue of (15).

4. A new convergence estimate for the Jacobi–Davidson method.

Let us rewrite (4) in the following equivalent form:

$$\tilde{L}_n v = \bar{\pi}_n^T r, \tag{16}$$

where

$$\tilde{L}_n = \omega \pi_n^T M \pi_n + \bar{\pi}_n^T (L - \lambda(u^n)M)\bar{\pi}_n \tag{17}$$

and ω is a positive constant. The auxiliary result below shows that (16) is correctly posed provided that λ_0 is simple and $\lambda(u^n)$ is "close enough" to λ_0.

LEMMA 4.1. *If λ_0 is a simple eigenvalue and $\lambda^n \equiv \lambda(u^n) < (\lambda_0 + \lambda_1)/2$, then*

$$\tilde{L}_n \ge c_n M, \quad c_n = \min\{\lambda_0 + \lambda_1 - 2\lambda^n, \omega\}.$$

Proof. See Appendix A.2 (cf. also [20]). □

If (16) is solved exactly, then the Jacobi–Davidson method is known to have cubic convergence [24, 29]. If the size of the problem makes the exact solution impractical, then one has to resort to solving (16) iteratively (cf. inexact shift-and-invert iterations—see, e.g., [30]). We note that the iterative solution of (4) by the preconditioned conjugate gradient method was studied in [20]. Just like this paper, [20] focuses on the preconditioning aspect and thereby considers actually the iterations (5) (referred to as "simplified GD method" there) rather than (3). However, the convergence results in [20], unlike those below, are given in terms of residuals, and hence they do not directly apply to the GD method.[3]

Below we assume that (16) is solved by some iterative method with the error propagation operators T_i (e.g., Chebyshev semi-iterative method [10]), i.e.,

$$v^i - v = T_i(v^0 - v). \tag{18}$$

[3] A recent report [21], which was not available to the author at the time of submission, has removed this limitation.

We assume that $\tilde{L}_n T_i$ are symmetric and that we stop (18) after k iterations, where k is such that

$$-\delta_k \tilde{L}_n \le \tilde{L}_n T_k \le \delta_k \tilde{L}_n, \quad \delta_k < 1, \tag{19}$$

which implies the reduction of the $\tilde{L}_n$-norm of the error by a factor of δ_k. If we start with the zero initial guess v^0, then

$$v^k = (1 - T_k)v = (1 - T_k)\tilde{L}_n^{-1}\bar{\pi}_n^T r \equiv K_n \bar{\pi}_n^T r = K_n r \tag{20}$$

and from (19) we have

$$(1 - \delta_k)\tilde{L}_n \le \tilde{L}_n K_n \tilde{L}_n \le (1 + \delta_k)\tilde{L}_n, \tag{21}$$

which shows that in this Jacobi–Davidson implementation of the GD method K_n can be viewed as a preconditioner for $\tilde{L}_n$. Accordingly, Theorem 4.2 below presents a convergence estimate for (5) with the following assumption on K_n that generalizes (21) (cf. (9) and (10)):

$$a^0 \tilde{L}_n^{-1} \le K_n \le b^0 \tilde{L}_n^{-1}. \tag{22}$$

THEOREM 4.2. *Assume that λ_0 is a simple eigenvalue of* (1). *If $\lambda^{n_0} < \frac{\lambda_0 + \lambda_1}{2}$ for some n_0 and the preconditioner K_n satisfies* (22) *for $n \ge n_0$, then the convergence of the iterations* (5) *for $n \ge n_0$ is described by the following estimate:*

$$0 \le \lambda^{n+1} - \lambda_0 \le \tilde{q}(\kappa^0, \nu, \delta(\lambda^n))^2 (\lambda^n - \lambda_0), \tag{23}$$

where

$$\kappa^0 = \frac{b^0}{a^0}, \quad \nu = \frac{\omega}{\lambda_1 - \lambda_0},$$

$$\tilde{q}(u, v, w) = \frac{u\tilde{\rho}(v, w) - 1}{u\tilde{\rho}(v, w) + 1}, \quad \tilde{\rho}(v, w) = \frac{1}{1 - 2w} \frac{(1 + w)^2 + vw}{(1 - w)^2}.$$

Proof. See Appendix A.3. □

COROLLARY 4.3. *The convergence estimate of Theorem* 4.2 *applies to the iterations* (3).

We observe that in the above Jacobi–Davidson implementation of the GD method $a^0 = 1 - \delta_k$ and $b^0 = 1 + \delta_k$ (cf. (21)), and hence $\tilde{q}(\kappa^0, \nu, \delta(\lambda^n)) = \delta_k + \mathcal{O}(\delta(\lambda^n))$. Thus, in the "ideal" case $\delta_k = 0$ (i.e., the system (4) solved exactly) the convergence is indeed cubic.

Finally, we note that Theorem 4.2 remains valid for $\omega = 0$, provided that $\tilde{L}_n^{-1}$ in (22) is the pseudoinverse of $\tilde{L}_n$: this case obviously corresponds to solving (4) by an iterative method in the subspace of vectors orthogonal to u^n in the scalar product $(\cdot, \cdot)_M$.

5. A new convergence estimate for the GD method. The aim of this section is to present a nonasymptotic convergence estimate for the GD method that would detect cubic convergence in another "ideal" case mentioned above, namely the case when K_n is the pseudoinverse of L_0. The form of the operator in the left-hand side of (4) suggests that this case is closely related to the case of solving (4) (or, equivalently, (16)) exactly. Below we express this relationship quantitatively based on the following lemma.

Lemma 5.1. *Let P be the $(\cdot,\cdot)_M$-orthogonal projection onto a subspace of $\mathcal{E}$ of the same dimension as $\mathcal{I}_0$, and denote*

$$L_{\alpha,P} = \alpha P^T M P + \bar{P}^T (L - \lambda_P M)\bar{P}, \tag{24}$$

where $\lambda_P = \max_u \lambda(Pu)$ and $\bar{P} = 1 - P$. If $\delta_P \equiv \delta(\lambda_P) < \delta_0$, where

$$\delta_0 = \frac{\min\{\alpha, \lambda_1 - \lambda_0\}}{\alpha + \lambda_1 - \lambda_0},$$

then

$$\left(1 - 2\sqrt{\frac{\delta_P}{\delta_0}} - \frac{\delta_P}{\delta_1}\right) L_{\alpha,\pi} \le L_{\alpha,P} \le \left(1 + \sqrt{\frac{\delta_P}{\delta_0}}\right)^2 L_{\alpha,\pi}, \tag{25}$$

where

$$\delta_1 = \min\left\{1, \frac{\alpha}{\lambda_1 - \lambda_0}\right\}.$$

Proof. See Appendix A.1. □

From Lemma 5.1 it follows that if λ_0 is simple, then

$$\tilde{a}_n \tilde{L} \le \tilde{L}_n \le \tilde{b}_n \tilde{L},$$

where $\tilde{a}_n = 1 - \mathcal{O}(\sqrt{\delta(\lambda^n)})$, $\tilde{b}_n = 1 + \mathcal{O}(\sqrt{\delta(\lambda^n)})$, and $\lambda^n = \lambda(u^n)$. Therefore, the operators $\tilde{L}_n$ defined by (17) converge to the operator

$$\tilde{L} = \omega\, \pi^T M \pi + \bar{\pi}^T L_0 \bar{\pi}, \tag{26}$$

and the assumption (22) is asymptotically equivalent to

$$a^0 \tilde{L}^{-1} \le K_n \le b^0 \tilde{L}^{-1}. \tag{27}$$

In the convergence result below the above assumption is slightly weakened, for the sake of generality (and in order to cover the case when K is the pseudoinverse of L_0), to become

$$a_0 \bar{\pi} L_0^{-1} \bar{\pi}^T \le K_n \le b_0 \left(\alpha \pi M \pi^T + \bar{\pi} L_0 \bar{\pi}^T\right)^{-1}, \tag{28}$$

where a_0, b_0, and α are positive constants.[4] Apart from its relation to (22), the assumption (28) can also be viewed as a generalization of (8) in the following sense: if K_n satisfies (8), then it also satisfies (28) with

$$a_0 = a\frac{\lambda_1 - \lambda_0}{\lambda_1}, \quad b_0 = b, \quad \alpha = \lambda_0. \tag{29}$$

It is also easy to see that (28) implies (14) with

$$\tau_n = \frac{2}{a_0 + b_0}, \quad \gamma = \frac{b_0 - a_0}{b_0 + a_0}. \tag{30}$$

[4] The use of different letters indicates that (28) does not need the same constants as (22).

The main convergence result of this section combines the advantages of the results discussed in section 3 and is free from their limitations: it is nonasymptotic, it produces the same asymptotic convergence factor as in (15) (the smallest one for (5) available to date), and it adequately predicts the cubic convergence in the "ideal" case $a_0 = b_0$.

THEOREM 5.2. *Assuming that* $\lambda^{n_0} < \lambda_1$ *for some* n_0 *and that the condition* (28) *is satisfied for* $n \geq n_0$, *the convergence of the iterations* (5) *for* $n \geq n_0$ *is described by the following estimate:*

$$0 \leq \lambda^{n+1} - \lambda_0 \leq q(\kappa_0, \sigma, \delta(\lambda^n))^2 (\lambda^n - \lambda_0), \tag{31}$$

where

$$\kappa_0 = \frac{b_0}{a_0}, \quad \sigma = \frac{\lambda_1 - \lambda_0}{\alpha},$$

$$q(u, v, w) = \frac{u\rho(v, w) - 1}{u\rho(v, w) + 1}, \quad \rho(v, w) = \frac{(1 + w)^2 + vw}{(1 - w)^2}.$$

Proof. See Appendix A.4. □

COROLLARY 5.3. *If* K_n *satisfies* (8), *then* (13) *is valid with* $\epsilon(t) = \mathcal{O}(t)$.

COROLLARY 5.4. *If* K_n *satisfies* (28), *then* (15) *is valid with* $\epsilon_n = \mathcal{O}(\lambda^n - \lambda_0)$.

COROLLARY 5.5. *The convergence estimate of Theorem* 5.2 *applies to the GD method.*

We observe that $q(\kappa_0, \sigma, \delta(\lambda^n)) < 1$ for $\lambda^n < \lambda_1$, and, by Corollary 5.4, the asymptotically insignificant term in (31) is smaller than that in (15). Furthermore, in the "ideal" case $\kappa_0 = 1$ one has $q(\kappa_0, \sigma, \delta(\lambda^n)) = \mathcal{O}(\delta(\lambda^n))$ and, thus, $\delta(\lambda^{n+1}) = \mathcal{O}(\delta(\lambda^n)^3)$ (cubic convergence). The last relationship, together with the discussion at the beginning of this section, advocates the use of the assumption (28) for measuring the quality of preconditioning for eigenvalue problems, instead of the standard assumptions (8) or (10), which are certainly more simple but do not guarantee the cubic convergence for any values of a, b, and γ.

Appendix A. Proofs and auxiliary results.

LEMMA A.1. *Assuming that*

$$a_u L_0 \leq L_0 \bar{\pi}_u K \bar{\pi}_u^T L_0 \leq b_u L_0, \tag{32}$$

where $K = K^T$,

$$\pi_u v = \frac{(Mv, u)}{(Mu, u)} u,$$

and $0 \neq u \in \mathcal{E}$, *the following inequalities are valid:*

$$0 \leq \min_\tau \lambda(u - \tau K r(u)) - \lambda_0 \leq \left(\frac{b_u - a_u}{b_u + a_u} \right)^2 (\lambda(u) - \lambda_0). \tag{33}$$

Proof. We have

$$\operatorname{span}\{u, Kr(u)\} = \operatorname{span}\{u, \bar{\pi}_u K r(u)\},$$

and hence

$$\min_\tau \lambda(u - \tau K r(u)) = \min_\tau \lambda(u - \bar{\pi}_u \tau K r(u)).$$

It is easy to verify that

$$\pi_u^T M = M\pi_u = \pi_u^T M \pi_u$$

and

$$\pi_u^T r(u) = 0, \quad r(u) = \bar{\pi}_u^T r(u) = \bar{\pi}_u^T L_0 u.$$

Therefore,

$$u(\tau) \equiv u - \tau\bar{\pi}_u K r(u) = u - \tau\bar{\pi}_u K \bar{\pi}_u^T r(u) = (1 - \tau K_u L_0)u,$$

where $K_u \equiv \bar{\pi}_u K \bar{\pi}_u^T$. Since $L_0 = \bar{\pi}^T L_0 \bar{\pi}$ we have

$$\bar{\pi} u(\tau) = (1 - \tau\bar{\pi} K_u \bar{\pi}^T L_0)\bar{\pi} u \equiv T_\tau \bar{\pi} u.$$

The operator $T_\tau : \bar{\mathcal{I}}_0 \to \bar{\mathcal{I}}_0$ is symmetric in the scalar product $(\cdot,\cdot)_{L_0}$. Furthermore, from (32) it follows that for

$$\tau = \tau_{opt} \equiv \frac{2}{a_u + b_u}$$

its spectral radius $\rho(T_\tau)$ is

$$\rho(T_{\tau_{opt}}) = \rho_{opt} \equiv \frac{b_u - a_u}{b_u + a_u}.$$

Hence,

$$\|u(\tau_{opt})\|_{L_0} \le \rho_{opt} \|u\|_{L_0}.$$

Since $\|u\|_{L_0}^2 = (\lambda(u) - \lambda_0)\|u\|_M^2$, $\|u(\tau_{opt})\|_{L_0}^2 = (\lambda(u(\tau_{opt})) - \lambda_0)\|u(\tau_{opt})\|_M^2$, and $\|u(\tau_{opt})\|_M^2 \ge \|u\|_M^2$ we finally obtain

$$\min_\tau \lambda(u - \tau K r(u)) - \lambda_0 = \min_\tau (\lambda(u(\tau)) - \lambda_0) \le \rho_{opt}^2 (\lambda(u) - \lambda_0).$$

The right-hand side inequality in (33) follows from the minimax principle. □

A.1. Proof of Lemma 5.1. From (7) it follows that if $\tilde{\pi}$ is a projection in the scalar product $(\cdot,\cdot)_M$ onto a subspace $\tilde{\mathcal{I}}_0$ of the same dimension as $\mathcal{I}_0$, then

$$\|(\pi - \tilde{\pi})u\|_M^2 \le \delta(\lambda^0)\|u\|_M^2, \quad \lambda^0 = \max_{v \in \tilde{\mathcal{I}}_0} \lambda(v). \tag{34}$$

Since $L_{\alpha,\pi} \ge \alpha_0 M$, where $\alpha_0 = \min\{\alpha, \lambda_1 - \lambda_0\}$, using (34) we obtain

$$\|(P - \pi)u\|_M^2 \le \delta_P \|u\|_M^2 \le \frac{\delta_P}{\alpha_0}\|u\|_{L_{\alpha,\pi}}^2.$$

Further,

$$\|(P - \pi)u\|_{L_0}^2 = \|\bar{\pi} P u\|_{L_0}^2 \le (\lambda_P - \lambda_0)\|u\|_M^2 \le \frac{\lambda_P - \lambda_0}{\alpha_0}\|u\|_{L_{\alpha,\pi}}^2.$$

Thus, in the norm $[\cdot]$ given by $[u]^2 = \alpha\|u\|_M^2 + \|u\|_{L_0}^2$ we have

$$[(P - \pi)u]^2 \le \frac{\alpha\delta_P + \lambda_P - \lambda_0}{\alpha_0}\|u\|_{L_{\alpha,\pi}}^2 = \frac{\delta_P}{\delta_0}\|u\|_{L_{\alpha,\pi}}^2. \tag{35}$$

Since $\delta_0 \le \frac{1}{2}$, the same calculations as in the proof of Lemma 4.1 show that $L_{\alpha,P} > 0$. Further,

$$\|u\|^2_{L_{\alpha,P}} = \alpha\|Pu\|^2_M + \|\bar{P}u\|^2_{L_0} - (\lambda_P - \lambda_0)\|\bar{P}u\|^2_M.$$

Using the relationships $Pu = \pi u + (P - \pi)u$ and $\bar{P}u = \bar{\pi}u - (P - \pi)u$ we obtain

$$\begin{aligned}\alpha\|Pu\|^2_M + \|\bar{P}u\|^2_{L_0} &= \alpha(\|\pi u\|^2_M + 2(\pi u, (P-\pi)u)_M + \|(P-\pi)u\|^2_M) \\ &\quad + \|\bar{\pi}u\|^2_{L_0} - 2(\bar{\pi}u, (P-\pi)u)_{L_0} + \|(P-\pi)u\|^2_{L_0} \\ &= \|u\|^2_{L_{\alpha,\pi}} + [(P-\pi)u]^2 + 2(\alpha(\pi u, (P-\pi)u)_M - (\bar{\pi}u, (P-\pi)u)_{L_0}).\end{aligned}$$

We have

$$\begin{aligned}&|\alpha(\pi u, (P-\pi)u)_M - (\bar{\pi}u, (P-\pi)u)_{L_0}| \\ &\le \alpha\|\pi u\|_M\|(P-\pi)u\|_M + \|\bar{\pi}u\|_{L_0}\|(P-\pi)u\|_{L_0} \\ &\le \|u\|_{L_{\alpha,\pi}}[(P-\pi)u] \le \sqrt{\frac{\delta_P}{\delta_0}}\|u\|^2_{L_{\alpha,\pi}}\end{aligned}$$

and

$$(\lambda_P - \lambda_0)\|\bar{P}u\|^2_M \le (\lambda_P - \lambda_0)\|u\|^2_M \le \frac{\lambda_P - \lambda_0}{\alpha_0}\|u\|^2_{L_{\alpha,\pi}} = \frac{\delta_P}{\delta_1}\|u\|^2_{L_{\alpha,\pi}}.$$

Thus,

$$\|u\|^2_{L_{\alpha,P}} \le \left(1 + \sqrt{\frac{\delta_P}{\delta_0}}\right)^2 \|u\|^2_{L_{\alpha,\pi}}$$

and

$$\|u\|^2_{L_{\alpha,P}} \ge \left(1 - 2\sqrt{\frac{\delta_P}{\delta_0}} - \frac{\delta_P}{\delta_1}\right)\|u\|^2_{L_{\alpha,\pi}}.$$

In what follows we assume that $\|u^n\|_M = 1$, and we denote $\pi_n v \equiv \pi_{u^n}$ and $\tilde{K}_n \equiv \bar{\pi}_n K_n \bar{\pi}_n^T$.

A.2. Proof of Lemma 4.1. We have

$$\begin{aligned}(\tilde{L}_n v, v) &= ((L - \lambda^n M)\bar{\pi}_n v, \bar{\pi}_n v) + \omega\|\pi_n v\|^2_M, \\ ((L - \lambda^n M)\bar{\pi}_n v, \bar{\pi}_n v) &= \|\bar{\pi}\bar{\pi}_n v\|^2_{L-\lambda^n M} - (\lambda^n - \lambda_0)\|\pi\bar{\pi}_n v\|^2_M \\ &\ge (\lambda_1 - \lambda^n)\|\bar{\pi}\bar{\pi}_n v\|^2_M - (\lambda^n - \lambda_0)\|\pi\bar{\pi}_n v\|^2_M \\ &= (\lambda_1 - \lambda^n)(\|\bar{\pi}\bar{\pi}_n v\|^2_M - \tilde{\delta}(\lambda^n)\|\pi\bar{\pi}_n v\|^2_M),\end{aligned}$$

where

$$\tilde{\delta}(\mu) = \frac{\mu - \lambda_0}{\lambda_1 - \mu} = \frac{1}{1 - \delta(\mu)} - 1, \quad \lambda_0 \le \mu < \lambda_1.$$

From (7) it follows that

$$\|\pi\bar{\pi}_n v\|^2_M = \|(\pi - \pi_n)\bar{\pi}_n v\|^2_M \le \delta(\lambda^n)\|\bar{\pi}_n v\|^2_M$$

and, thus,

$$\|\bar\pi\bar\pi_n v\|_{M}^2 = \|\bar\pi_n v\|_{M}^2 - \|\pi\bar\pi_n v\|_{M}^2 \ge (1-\delta(\lambda^n))\|\bar\pi_n v\|_{M}^2.$$

Therefore,

$$\begin{aligned}(\tilde L_n v, v) &\ge (\lambda_1-\lambda^n)(1-\delta(\lambda^n)-\delta(\lambda^n)\tilde\delta(\lambda^n))\|\bar\pi_n v\|_{M}^2 + \omega\|\pi_n v\|_{M}^2\\ &= (\lambda_1-\lambda^n)(1-\tilde\delta(\lambda^n))\|\bar\pi_n v\|_{M}^2 + \omega\|\pi_n v\|_{M}^2\\ &= (\lambda_0+\lambda_1-2\lambda^n)\|\bar\pi_n v\|_{M}^2 + \omega\|\pi_n v\|_{M}^2\\ &\ge \min\{\lambda_0+\lambda_1-2\lambda^n, \omega\}\,\|v\|_{M}^2.\end{aligned}$$

A.3. Proof of Theorem 4.2. Denote $B_n = K_n^{-1}$ and consider an auxiliary linear system

$$B_n v = \bar\pi_n^T L_0 u. \tag{36}$$

We have

$$\bar\pi_n^T L_0 u = L_0 u - (L_0 u, u^n) M u^n.$$

Hence, multiplying (36) by $\bar\pi u$ (in the scalar product in $\mathcal{E}$) we obtain

$$\|u\|_{L_0}^2 = (Mu^n, \bar\pi u)(L_0 u, u^n) + (B_n v, \bar\pi u).$$

The first term in the right-hand side can be estimated as follows:

$$\begin{gathered}|(Mu^n, \bar\pi u)(L_0 u, u^n)| \le \|\bar\pi u^n\|_{M}\|\bar\pi u\|_{M}\|u^n\|_{L_0}\|u\|_{L_0},\\ \|\bar\pi u^n\|_{M}^2 \le \frac{\lambda^n-\lambda_0}{\lambda_1-\lambda_0}, \quad \|u^n\|_{L_0}^2 = \lambda^n-\lambda_0, \quad \|\bar\pi u\|_{M}^2 \le \frac{1}{\lambda_1-\lambda_0}\|u\|_{L_0}^2,\\ |(Mu^n, \bar\pi u)(L_0 u, u^n)| \le \delta(\lambda^n)\|u\|_{L_0}^2.\end{gathered}$$

For the second term we have

$$|(B_n v, \bar\pi u)| \le \|v\|_{B_n}\|\bar\pi u\|_{B_n}$$

and, thus,

$$(1-\delta(\lambda^n))\|u\|_{L_0}^2 \le \|v\|_{B_n}\|\bar\pi u\|_{B_n}.$$

The second factor in the right-hand side can be estimated as follows:

$$\begin{gathered}\|\bar\pi u\|_{B_n}^2 \le \frac{1}{a^0}\|\bar\pi u\|_{\tilde L_n}^2 = \frac{1}{a^0}\left(\nu(\lambda_1-\lambda_0)\|\pi_n\bar\pi u\|_{M}^2 + ((L-\lambda^n M)\bar\pi_n\bar\pi u, \bar\pi_n\bar\pi u)\right),\\ \|\pi_n\bar\pi u\|_{M}^2 = (M\bar\pi u, u^n)^2 \le \frac{\delta(\lambda^n)}{\lambda_1-\lambda_0}\|u\|_{L_0}^2,\\ ((L-\lambda^n M)\bar\pi_n\bar\pi u, \bar\pi_n\bar\pi u) \le \|\bar\pi_n\bar\pi u\|_{L_0}^2,\\ \begin{aligned}\|\bar\pi_n\bar\pi u\|_{L_0} = \|\bar\pi u - (M\bar\pi u, u^n)u^n\|_{L_0} &\le \|u\|_{L_0} + |(M\bar\pi u, u^n)|\|u^n\|_{L_0}\\ &\le (1+\delta(\lambda^n))\|u\|_{L_0},\end{aligned}\\ \|\bar\pi u\|_{B_n}^2 \le \frac{1}{a^0}\left(\nu\delta(\lambda^n) + (1+\delta(\lambda^n))^2\right)\|u\|_{L_0}^2,\end{gathered}$$

which leads to the following estimate:

$$\|u\|^2_{L_0} \le \frac{1}{a^0} \frac{\nu\delta(\lambda^n) + (1+\delta(\lambda^n))^2}{(1-\delta(\lambda^n))^2} \|v\|^2_{B_n} = \frac{1}{a^0_n} \|v\|^2_{B_n}, \tag{37}$$

where

$$a^0_n = a^0 \frac{(1-\delta(\lambda^n))^2}{(1+\delta(\lambda^n))^2 + \nu\delta(\lambda^n)}.$$

Now, multiplying (36) by v we have

$$\begin{aligned}(B_n v, v) = (L_0 u, \bar{\pi}_n v) \le \|u\|_{L_0} \|\bar{\pi}_n v\|_{L_0} &\le \frac{\lambda_1 - \lambda_0}{\lambda_1 - \lambda^n} \|u\|_{L_0} \|\bar{\pi}\bar{\pi}_n v\|_{L-\lambda^n M} \\ &= (1+\tilde{\delta}(\lambda^n)) \|u\|_{L_0} \|\bar{\pi}\bar{\pi}_n v\|_{L-\lambda^n M}.\end{aligned} \tag{38}$$

Further,

$$(\tilde{L}_n v, v) \ge ((L - \lambda^n M)\bar{\pi}_n v, \bar{\pi}_n v) \ge \|\bar{\pi}\bar{\pi}_n v\|^2_{L-\lambda^n M} - (\lambda^n - \lambda_0)\|\pi\bar{\pi}_n v\|^2_M$$

and (cf. the proof of Lemma 4.1)

$$\|\pi\bar{\pi}_n v\|^2_M \le \frac{\delta(\lambda^n)}{1-\delta(\lambda^n)} \|\bar{\pi}\bar{\pi}_n v\|^2_M = \tilde{\delta}(\lambda^n) \|\bar{\pi}\bar{\pi}_n v\|^2_M \le \frac{\tilde{\delta}(\lambda^n)}{\lambda_1 - \lambda^n} \|\bar{\pi}\bar{\pi}_n v\|^2_{L-\lambda^n M}.$$

Therefore,

$$(\tilde{L}_n v, v) \ge (1 - \tilde{\delta}(\lambda^n)^2) \|\bar{\pi}\bar{\pi}_n v\|^2_{L-\lambda^n M}$$

and we obtain from (38) the estimate

$$(B_n v, v) \le \frac{1+\tilde{\delta}(\lambda^n)}{\sqrt{1-\tilde{\delta}(\lambda^n)^2}} \|u\|_{L_0} \|v\|_{\tilde{L}_n} \le \frac{1+\tilde{\delta}(\lambda^n)}{\sqrt{1-\tilde{\delta}(\lambda^n)^2}} \sqrt{b^0} \|u\|_{L_0} \|v\|_{B_n},$$

which leads to the estimate

$$\|u\|^2_{L_0} \ge \frac{1}{b^0} \frac{1-\tilde{\delta}(\lambda^n)^2}{(1+\tilde{\delta}(\lambda^n))^2} \|v\|^2_{B_n} = \frac{1}{b^0} \frac{1-\tilde{\delta}(\lambda^n)}{1+\tilde{\delta}(\lambda^n)} \|v\|^2_{B_n} = \frac{1}{b^0_n} \|v\|^2_{B_n}, \tag{39}$$

where

$$b^0_n = b^0 \frac{1+\tilde{\delta}(\lambda^n)}{1-\tilde{\delta}(\lambda^n)} = \frac{b^0}{1-2\delta(\lambda^n)}.$$

Thus,

$$a^0_n \|u\|^2_{L_0} \le \|v\|^2_{B_n} \le b^0_n \|u\|^2_{L_0}.$$

Recalling that $v = B_n^{-1} \bar{\pi}_n^T L_0 u = K_n \bar{\pi}_n^T L_0 u$ we can rewrite the above inequalities as

$$a^0_n \|u\|^2_{L_0} \le (\tilde{K}_n L_0 u, L_0 u) \le b^0_n \|u\|^2_{L_0}.$$

To complete the proof, it remains to apply Lemma A.1.

A.4. Proof of Theorem 5.2. Since

$$(\alpha \pi^T M \pi + \bar{\pi}^T L_0 \bar{\pi})^{-1} = \alpha^{-1} \pi M^{-1} \pi^T + \bar{\pi} L_0^{-1} \bar{\pi}^T$$

from (28) we have for any $u \in \mathcal{E}$

$$a_0 \|\bar{\pi}^T u\|^2_{L_0^{-1}} \le (K_n u, u) \le b_0 \left(\alpha^{-1} \|\pi^T u\|^2_{M^{-1}} + \|\bar{\pi}^T u\|^2_{L_0^{-1}} \right). \tag{40}$$

In order to apply Lemma A.1 let us take $u = \bar{\pi}_n^T L_0 v$. We have

$$\pi^T u = \pi^T \bar{\pi}_n^T L_0 v = -\pi^T \pi_n^T L_0 v = -(L_0 v, u^n) M \pi u^n,$$

and hence

$$\begin{aligned} \alpha^{-1} \|\pi^T u\|^2_{M^{-1}} = \alpha^{-1} (L_0 v, u^n)^2 \|\pi u^n\|^2_M &\le \alpha^{-1} \|v\|^2_{L_0} \|u^n\|^2_{L_0} \\ &= \sigma \frac{\lambda^n - \lambda_0}{\lambda_1 - \lambda_0} \|v\|^2_{L_0}. \end{aligned}$$

Further,

$$\bar{\pi}^T u = \bar{\pi}^T \bar{\pi}_n^T L_0 v = \bar{\pi}^T L_0 v - \bar{\pi}^T \pi_n^T L_0 v = L_0 v - (L_0 v, u^n) \bar{\pi}^T M u^n.$$

For $w = (L_0 v, u^n) \bar{\pi}^T M u^n$ we have

$$\begin{aligned} \|w\|^2_{L_0^{-1}} &= (L_0 v, u^n)^2 \|M \bar{\pi} u^n\|^2_{L_0^{-1}} = (L_0 v, u^n)^2 \|L_0^{-1} M \bar{\pi} u^n\|^2_{L_0} \\ &\le \|v\|^2_{L_0} \|u^n\|^2_{L_0} \frac{\|u^n\|^2_{L_0}}{(\lambda_1 - \lambda_0)^2} = \left(\frac{\lambda^n - \lambda_0}{\lambda_1 - \lambda_0} \right)^2 \|v\|^2_{L_0} = \delta(\lambda^n)^2 \|v\|^2_{L_0}. \end{aligned}$$

Hence,

$$\|\bar{\pi}^T u\|_{L_0^{-1}} \le \|L_0 v\|_{L_0^{-1}} + \|w\|_{L_0^{-1}} = \|v\|_{L_0} + \|w\|_{L_0^{-1}} \le (1 + \delta(\lambda^n)) \|v\|_{L_0}$$

and

$$\|\bar{\pi}^T u\|_{L_0^{-1}} \ge \|v\|_{L_0} - \|w\|_{L_0^{-1}} \ge (1 - \delta(\lambda^n)) \|v\|_{L_0}.$$

Substituting the above estimates for $\pi^T u$ and $\bar{\pi}^T u$ into (40), we obtain

$$a_{0n} (L_0 v, v) \le (L_0 \bar{\pi}_n K_n \bar{\pi}_n^T L_0 v, v) \le b_{0n} (L_0 v, v),$$

where

$$a_{0n} = a_0 (1 - \delta(\lambda^n))^2, \quad b_{0n} = b_0 ((1 + \delta(\lambda^n))^2 + \sigma \delta(\lambda^n)). \tag{41}$$

To complete the proof it remains to apply Lemma A.1.

REFERENCES

[1] Z. Bai, J. Demmel, J. Dongarra, A. Ruhe, and H. van der Vorst, eds., *Templates for the Solution of Algebraic Eigenvalue Problems. A Practical Guide*, Software Environ. Tools 11, SIAM, Philadelphia, 2000.

[2] J. H. Bramble, J. E. Pasciak, and A. V. Knyazev, *A subspace preconditioning algorithm for eigenvector/eigenvalue computation*, Adv. Comput. Math., 6 (1996), pp. 159–189.

[3] M. Crouzeix, B. Philippe, and M. Sadkane, *The Davidson method*, SIAM J. Sci. Comput., 15 (1994), pp. 62–76.

[4] E. R. Davidson, *The iterative calculation of a few of the lowest eigenvalues and corresponding eigenvectors of large real symmetric matrices*, J. Computational Phys., 17 (1975), pp. 87–94.

[5] E. G. D'yakonov, *Iteration methods in eigenvalue problems*, Math. Notes, 34 (1983), pp. 945–953.

[6] E. G. D'yakonov, *Optimization in Solving Elliptic Problems*, CRC Press, 1996.

[7] E. G. D'yakonov and M. Yu. Orekhov, *Minimization of the computational labor in determining the first eigenvalues of differential operators*, Math. Notes, 27 (1980), pp. 382–391.

[8] J. van den Eshof, *The convergence of Jacobi-Davidson iterations for Hermitian eigenproblems*, Numer. Linear Algebra Appl., 9 (2002), 163–179.

[9] D. R. Fokkema, G. L. G. Sleijpen, and H. A. van der Vorst, *Jacobi–Davidson style QR and QZ algorithms for the reduction of matrix pencils*, SIAM J. Sci. Comput., 20 (1998), pp. 94–125.

[10] G. H. Golub and R. S. Varga, *Chebyshev semi-iterative methods, successive overrelaxation iterative methods and second order Richardson iterative methods*, Numer. Math., 3 (1961), pp. 147–168.

[11] A. V. Knyazev, *Convergence rate estimates for iterative methods for mesh symmetric eigenvalue problem*, Soviet J. Numer. Anal. Math. Modelling, 2 (1987), pp. 371–396.

[12] A. V. Knyazev, *A Preconditioned Conjugate Gradient Method for Eigenvalue Problems and Its Implementation in a Subspace*, Internat. Ser. Numer. Math. 96, Birkhäuser, Basel, 1991, pp. 143–154.

[13] A. V. Knyazev, *Preconditioned eigensolvers—an oxymoron?*, Electron. Trans. Numer. Anal., 7 (1998), pp. 104–123.

[14] A. V. Knyazev, *Toward the optimal preconditioned eigensolver: Locally optimal block preconditioned conjugate gradient method*, SIAM J. Sci. Comput., 23 (2001), pp. 517–541.

[15] A. V. Knyazev and K. Neymeyr, *A geometric theory for preconditioned inverse iteration.* III: *A short and sharp convergence estimate for generalized eigenvalue problems*, Linear Algebra Appl., 358 (2003), pp. 95–114.

[16] A. V. Knyazev and K. Neymeyr, *Efficient solution of symmetric eigenvalue problems using multigrid preconditioners in the locally optimal block conjugate gradient method*, Electron. Trans. Numer. Anal., to appear.

[17] K. Neymeyr, *A geometric theory for preconditioned inverse iteration.* I: *Extrema of the Rayleigh quotient*, Linear Algebra Appl., 322 (2001), pp. 61–85.

[18] K. Neymeyr, *A geometric theory for preconditioned inverse iteration.* II: *Convergence estimates*, Linear Algebra Appl., 322 (2001), pp. 87–104.

[19] K. Neymeyr, *Why Preconditioning Gradient Type Eigensolvers?* Sonderforschungsbereich 382, Report 146, revised version, Universität Tübingen, Tübingen, Germany, 2000.

[20] Y. Notay, *Combination of Jacobi–Davidson and conjugate gradients for the partial symmetric eigenproblem*, Numer. Linear Algebra Appl., 9 (2002), pp. 21–44.

[21] Y. Notay, *Convergence Analysis of Inexact Rayleigh Quotient Iterations*, Technical report, Université Libre de Bruxelles, Bruxelles, Belgium, 2001.

[22] S. Oliveira, *On the convergence rate of a preconditioned subspace eigensolver*, Computing, 63 (1999), pp. 219–231.

[23] E. Ovtchinnikov, *Convergence estimates for the generalized Davidson method for symmetric eigenvalue problems* II: *The subspace acceleration*, SIAM J. Numer. Anal., 41 (2003), pp. 272–286.

[24] B. N. Parlett, *The Rayleigh quotient iteration and some generalizations*, Math. Comp., 28 (1974), pp. 679–693.

[25] Y. Saad, *Numerical Methods for Large Eigenvalue Problems*, Algorithms Archit. Adv. Sci. Comput., Manchester University Press, Manchester, UK, Halsted Press, New York, 1992.

[26] B. Samokish, *The steepest descent method for an eigenvalue problem with semi-bounded operators*, Izv. Vyssh. Uchebn. Zaved. Mat., 5 (1958), pp. 105–114.

[27] G. L. G. Sleijpen, G. L. Booten, D. R. Fokkema, and H. A. van der Vorst, *Jacobi–Davidson type methods for generalized eigenproblems and polynomial eigenproblems*, BIT, 36 (1996), pp. 595–633.

[28] G. L. G. Sleijpen, H. A. van der Vorst, and E. Meijerink, *Efficient expansion of subspaces in the Jacobi–Davidson method for standard and generalized eigenproblems*, Electron. Trans. Numer. Anal., 7 (1998), pp. 75–89.

[29] G. L. G. Sleijpen and H. A. van der Vorst, *A Jacobi–Davidson iteration method for linear eigenvalue problems*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 401–425.

[30] P. Smith and M. H. C. Paardekooper, *The effects of inexact solvers in algorithms for symmetric eigenvalue problems*, Linear Algebra Appl., 287 (1999), pp. 337–357.

SIAM J. NUMER. ANAL.
Vol. 41, No. 1, pp. 272–286

CONVERGENCE ESTIMATES FOR THE GENERALIZED DAVIDSON METHOD FOR SYMMETRIC EIGENVALUE PROBLEMS II: THE SUBSPACE ACCELERATION*

E. OVTCHINNIKOV†

Abstract. The generalized Davidson (GD) method can be viewed as a generalization of the preconditioned steepest descent (PSD) method for solving symmetric eigenvalue problems. In the GD method, the new approximation is sought in the subspace that spans all the previous approximate eigenvectors, in addition to the current one and the preconditioned residual thereof used in PSD. In this respect, the relation between the GD and PSD methods is similar to that between the standard steepest descent method for linear systems and methods in Krylov subspaces. This paper presents convergence estimates for the (restarted) GD method that demonstrate convergence acceleration compared to the PSD method, similar to that achieved by methods in Krylov subspaces compared to the standard steepest descent.

Key words. iterative methods for eigenvalue problems, preconditioning, Krylov subspaces, generalized Davidson method, convergence estimates

AMS subject classifications. 65F15, 65N12, 65N22, 65N30

PII. S0036142902411768

1. Introduction.

This paper presents convergence estimates for the generalized Davidson (GD) method [1, 17, 11] for finding the smallest eigenvalue of the generalized eigenvalue problem

$$Lu = \lambda M u, \tag{1}$$

where L and M are, respectively, a symmetric and a symmetric positive definite linear operator in a Euclidean space $\mathcal{E}$. The GD method can be formulated as follows: given an arbitrary nontrivial vector u^0 we calculate a sequence λ^n of approximations to the smallest eigenvalue λ_0 of (1) using the following recurrent formula:

$$\begin{aligned} u^n &= \arg \min_{u \in \mathcal{D}^n,\ (Mu,u)=1} \lambda(u), \quad \lambda^n = \lambda(u^n), \\ \mathcal{D}^{n+1} &= \operatorname{span}\{u^0, \dots, u^n, K_n(L - \lambda^n M)u^n\}, \end{aligned} \tag{2}$$

where K_n are some linear operators in $\mathcal{E}$, $(\cdot,\cdot)$ is the scalar product in $\mathcal{E}$, and $\lambda(u)$ is the Rayleigh quotient, i.e.,

$$\lambda(u) \equiv \frac{(Lu,u)}{(Mu,u)}. \tag{3}$$

Two familiar instances of the GD method are the Davidson method (see, e.g., [3, 2]) and the Jacobi–Davidson method (see, e.g., [19] and the relevant chapters in [1]). In the former, L is a matrix, $M = I$ (the unit matrix), and $K_n = (D_L - \lambda^n I)^{-1}$, where D_L is the diagonal of L (i.e., a diagonal matrix of the same size and with the same

*Received by the editors January 11, 2002; accepted for publication September 16, 2002; published electronically March 26, 2003.
http://www.siam.org/journals/sinum/41-1/41176.html

†Harrow School of Computer Science, University of Westminster, Watford Road, London HA1 3TP, United Kingdom (eeo@wmin.ac.uk).

diagonal entries as in L). In the Jacobi–Davidson method $v = K_n r$ is calculated by approximately solving the linear system

$$(1-\pi_n)^T(L-\lambda^n M)(1-\pi_n)v = (1-\pi_n)^T r, \quad \pi_n v = 0, \tag{4}$$

where $\pi_n w = (Mw, u^n)u^n$.

It is easy to see that the number of arithmetic operations needed to perform the nth iteration (2) is asymptotically proportional to n^2. To avoid excessive computational expenses, in practical calculations the iterations (2) are restarted using the last approximation u^n as the initial one.[1] If the restart takes place every m iterations, then we arrive at the following iterative scheme:

$$\begin{aligned} u^{lm+k} &= \arg \min_{u \in \mathcal{D}^{lm,k},(Mu,u)=1} \lambda(u), \quad \lambda^{lm+k} = \lambda(u^{lm+k}), \\ \mathcal{D}^{l,k+1} &= \operatorname{span}\{u^{lm}, \dots, u^{lm+k}, K_{lm+k}(L-\lambda^{lm+k}M)u^{lm+k}\}. \end{aligned} \tag{5}$$

An alternative approach to restarting the GD method is to choose some $k > 0$ and to remove vectors $u^0,\dots,u^{n-k}$ from the definition of the subspace $\mathcal{D}^{n+1}$ for $n \geq k$, thereby bounding its dimension by $k+1$. For $k = 2$ this yields the so-called locally optimal preconditioned conjugate gradient (LOPCG) method [6]. Quite interestingly, there is numerical evidence (see [14]; cf. also [8, 9]) that the LOPCG method converges at the same rate as the GD method, despite using a smaller subspace. This remarkable phenomenon still awaits explanation (for some insights into the convergence behavior of LOPCG, see [6, 8]).

The GD method can be viewed as a generalization of the preconditioned steepest descent (PSD) iterations

$$u^{n+1} = u^n - \tau_n g^n, \quad g^n = K(L-\lambda(u^n)M)u^n, \tag{6}$$

where

$$\tau_n = \arg\min_{\tau} \lambda(u^n - \tau g^n), \tag{7}$$

and K is a symmetric positive definite operator usually referred to as *preconditioner*. Preconditioned eigensolvers based on (6) are among the oldest and best-studied to date (see, e.g., [7] and the references therein; for the comparative discussion of the latest results, see also [13, 15]). Owing to the minimax principle, any convergence estimate for (6) in terms of eigenvalues apply to (2) and (5) as well.[2] However, such estimates can only be used as preliminary ones because the GD method should obviously converge faster than the PSD method, by virtue of using larger subspaces for minimizing $\lambda(u)$. Indeed, if we take $K = L^{-1}$, then, assuming $\lambda^n < \lambda_1$, for (6) we have [5]

$$\frac{\mu_0 - \mu^n}{\mu^n - \mu_1} \leq \left(\frac{1-\xi}{1+\xi}\right)^2 \frac{\mu_0 - \mu^{n-1}}{\mu^{n-1} - \mu_1}, \tag{8}$$

where $\mu_0 = 1/\lambda_0$, $\mu^n = 1/\lambda^n$, and ξ is the relative distance between λ_0 and the second smallest distinct eigenvalue λ_1, i.e., $\xi = 1 - \lambda_0/\lambda_1$. Accordingly, if we take $K_n = L^{-1}$,

[1] A somewhat different restart strategy is suggested in [8].

[2] In fact, the only convergence estimates for the GD method that have been available so far are those for the PSD method (see [13] and the references therein).

then the subspaces $\mathcal{D}^n$ become Krylov subspaces, and (2) becomes equivalent to the Lanczos method (for finding the largest eigenvalue μ_0 of $L^{-1}M$), for which we have [5]

$$\frac{\mu_0 - \mu^n}{\mu^n - \mu_1} \le \left(T_n\left(\frac{1+\xi}{1-\xi}\right)\right)^{-2} \frac{\mu_0 - \mu^0}{\mu^0 - \mu_1}, \tag{9}$$

where $T_n(x) = \cos(n \arccos x)$ are the Chebyshev polynomials and we assume that $\mu^0 > \mu_1$. Since $T_n((1+\xi)/(1-\xi)) \ge 0.5((1+\sqrt{\xi})/(1-\sqrt{\xi}))^n$, the estimate (9) leads to the asymptotic convergence factor $((1-\sqrt{\xi})/(1+\sqrt{\xi}))^2$, which implies that (2) converges about twice as fast as (6) for a small ξ.

In the general case neither $\mathcal{D}^n$ nor $\mathcal{D}^{l,k}$ is a Krylov subspace. However, there are several Krylov subspaces that $\mathcal{D}^{l,k}$ approaches as l increases. Indeed, assuming for simplicity that $K_n = K$, we can define $\mathcal{D}^{l,k}$ equivalently as

$$\mathcal{D}^{l,k+1} = \operatorname{span}\{u^{lm}, K(L - \lambda^{lm}M)u^{lm}, \ldots, K(L - \lambda^{lm+k}M)u^{lm+k}\}$$

from which we see immediately that, as l increases, $\mathcal{D}^{l,k}$ approaches the Krylov subspace

$$\mathcal{K}_{k+1}(u^{lm}, K(L - \lambda^{lm}M)), \tag{10}$$

where

$$\mathcal{K}_n(u, A) = \operatorname{span}\{u, Au, \ldots, A^n u\}.$$

Some other Krylov subspaces that are "close" to $\mathcal{D}^{l,k}$ are discussed in section 4. We note that the subspace (10) appears in various preconditioned eigensolvers. For example, denoting by $\nu(t)$ the smallest eigenvalue of $A_t = K(L - tM)$, we can rewrite the problem (1) equivalently as $\nu(\lambda) = 0$. Applying the Newton method to solve this equation together with a method in Krylov subspaces for A_t to approximately compute $\nu(t)$, we obtain a two-level method that uses subspaces similar to (10) at the inner iteration level (see, e.g., [5]; cf. also [1, Algorithms 11.7 and 11.8]).

The present paper uses the above property of the Davidson subspaces $\mathcal{D}^{l,k}$ and the standard analysis in Krylov subspaces (cf. Lemma 4.1) to obtain convergence estimates for (5) that are similar to those for two-level methods using Krylov subspaces at the inner iteration level (cf. [5]). It should be noted that similar estimates for $m \le 2$ can be found in [6]. For $m > 2$ the estimates of the present paper are, to the best of the author's knowledge, the first estimates for the (restarted) GD method that reflect the subspace acceleration aspect of this method.

2. Notation.

In this paper we use the standard notation $A > 0$ (resp., $A \ge 0$) to declare that a symmetric linear operator A is positive definite (resp., semidefinite). Accordingly, $A \ge B$ stands for $A - B \ge 0$, etc. For $A \ge 0$ we denote (Au, v) by $(u, v)_A$, and $\|u\|_A$ stands for $\sqrt{(u,u)_A}$. By considering an auxiliary eigenvalue problem $Au = \mu Bu$ it is easy to show that $0 < A \le B$ implies $B^{-1} \le A^{-1}$, which, together with the fact that $A \le B$ implies $C^T AC \le C^T BC$ for any C, leads to the following elementary result:

$$\begin{gathered} 0 < A \le cB^{-1} \;\Rightarrow\; \|Au\|_B^2 \le c\|u\|_A^2 \le c^2\|u\|_{B^{-1}}^2, \\ \|BAu\|_A \le c\|u\|_A, \quad \|BAu\|_{B^{-1}} \le c\|u\|_{B^{-1}}. \end{gathered} \tag{11}$$

The minimal eigenvalue of (1) is denoted λ_0, and λ_1 denotes the second smallest *distinct* eigenvalue. The invariant subspace of $M^{-1}L$ corresponding to λ_0 is denoted by $\mathcal{I}_0$, and the $(\cdot,\cdot)_M$-orthogonal projection onto $\mathcal{I}_0$ is denoted by π. For any projection π' we denote $\bar{\pi}' \equiv I - \pi'$, where I is the identity operator. The projection π has the following simple properties:

$$L\pi = \pi^T L = \lambda_0 M\pi, \quad M\pi = \pi^T M, \quad L_0 \equiv L - \lambda_0 M = L_0\bar{\pi} = \bar{\pi}^T L_0,$$

and for any vector u

$$\|\bar{\pi}u\|_M^2 \le \delta(\lambda(u))\|u\|_M^2, \quad \|\bar{\pi}u\|_{L_0}^2 = (\lambda(u) - \lambda_0)\|u\|_M^2,$$

where

$$\delta(\lambda) = \frac{\lambda - \lambda_0}{\lambda_1 - \lambda_0}.$$

Finally, $r(u)$ denotes the residual $Lu - \lambda(u)Mu$, and $u_i(\mathcal{H})$ are the Ritz eigenvectors of the problem (1) in a subspace $\mathcal{H} \subset \mathcal{E}$, i.e.,

$$u_i(\mathcal{H}) \in \mathcal{H} \;:\; (r(u_i(\mathcal{H})), v) = 0 \quad \forall v \in \mathcal{H}. \tag{12}$$

We assume that $\|u_i(\mathcal{H})\|_M = 1$ and we enumerate the Ritz eigenpairs $\{\lambda_i(\mathcal{H}), u_i(\mathcal{H})\}$ in the ascending order of $\lambda_i(\mathcal{H}) = \lambda(u_i(\mathcal{H}))$.

3. Assumptions on the preconditioners. Preconditioning in the GD method has been discussed in detail in Part I of this paper [13]. Here we summarize that discussion in order to make proper assumptions on the preconditioners K_n.

The role of preconditioning in iterative methods is to improve the convergence. Accordingly, a usual assumption on a preconditioner K is

$$aK_* \le K \le bK_*, \tag{13}$$

where a and b are positive constants and K_* is some "ideal" preconditioner, i.e., the one that would lead to a very fast convergence if used. For example, in the case of a linear system $Lu = f$ with $L > 0$, taking $K = K_* = L^{-1}$ as a preconditioner in the PSD method makes it converge after just one iteration. Furthermore, convergence estimates for preconditioned algorithms for solving linear systems show that in the general case (13) (with $K_* = L^{-1}$) the ratio $\kappa = b/a$ can be used as a measure of the quality of preconditioning: the smaller, the better, the "ideal" case being $\kappa = 1$.

Many papers on preconditioned eigensolvers feature exactly the same choice for K_* despite the fact that it is no longer "ideal" in the case of eigenvalue problems, since it fails to deliver a superlinear convergence achievable with some other preconditioners. Part I of this paper considers two alternative choices $K_* = \tilde{L}^{-1}$ and $K_* = \tilde{L}_n^{-1}$, where

$$\tilde{L} = \alpha\,\pi^T M\pi + \bar{\pi}^T L_0\bar{\pi}$$

and

$$\tilde{L}_n = \omega\pi_n^T M\pi_n + \bar{\pi}_n^T(L - \lambda(u^n)M)\bar{\pi}_n$$

and α and ω are positive constants. These two choices are, in a sense, asymptotically equivalent, since $\tilde{L}_n$ with $\omega = \alpha$ converges to $\tilde{L}$ [13]. It should be noted that the first choice is not entirely new: the asymptotic estimates in [18] and [11] essentially use $K_* = \tilde{L}$ and assumption (13) in the subspace orthogonal to $\mathcal{I}_0$. The second choice is closely related to the Jacobi–Davidson method: if (4) is solved approximately using a linear iterative algorithm, then the respective Jacobi–Davidson method can be interpreted as the GD method with preconditioners satisfying (13) with $K_* = \tilde{L}_n$, $a = 1-\delta$, and $b = 1+\delta$, where δ is the accuracy in the $\tilde{L}_n$-norm to which (4) is solved [13].

The convergence estimates in [18, 11, 13] show that the suggested alternative assumptions on K are indeed more suitable for preconditioned eigensolvers than the "standard" one based on $K_* = L^{-1}$ because in the "ideal" case $a = b$ both lead to cubic convergence. In view of this, below we assume that either

$$a_0\tilde{L}^{-1} \le K_n \le b_0\tilde{L}^{-1} \tag{14}$$

or

$$a^0\tilde{L}_n^{-1} \le K_n \le b^0\tilde{L}_n^{-1}, \tag{15}$$

where a_0, b_0, a^0, and b^0 are positive constants. To simplify the proofs, we assume further that λ_0 is simple and that $K_{lm+i} = K_{lm}$ for $i = 1, \ldots, m-1$ (see [12] for a more general case).

4. Rayleigh–Ritz approximation in auxiliary Krylov subspaces. As mentioned in the introduction, the idea behind the estimates of this paper is to use the asymptotic closeness of the Davidson subspaces $\mathcal{D}^{l,k}$ to some Krylov subspaces. One such subspace is that in (10). Another one is obviously

$$\operatorname{span}\{u^{lm}, K(L-\lambda_0 M)u^{lm}, \ldots, (K(L-\lambda_0 M))^k u^{lm}\}. \tag{16}$$

In this section we introduce yet another Krylov subspace that appears to be asymptotically closer to $\mathcal{D}^{l,k}$ than (16) and provides better approximation for u_0 than (10).

Denoting $r^n \equiv r(u^n)$ and $g^n = K_n r^n$, we have

$$\mathcal{D}^{l,k+1} = \mathcal{D}^{l,k} + \operatorname{span}\{g^{lm+k}\}$$

and, therefore,

$$\begin{aligned}\mathcal{D}^{l,k+1} &= \operatorname{span}\{u^{lm}, g^{lm}, \ldots, g^{lm+k}\}\\ &= \operatorname{span}\{u^{lm}, \bar{\pi}_{lm}g^{lm}, \ldots, \bar{\pi}_{lm}g^{lm+k}\}.\end{aligned}$$

Since $(r^n, u^{lm}) = 0$ for $lm \le n \le (l+1)m$, we have $\bar{\pi}_{lm}^T r^n = r^n$ and $g^n = K_n\bar{\pi}_{lm}^T r^n = K_{lm}\bar{\pi}_{lm}^T r^n$. Therefore

$$\mathcal{D}^{l,k+1} = \operatorname{span}\{v_i^0\}_{i=-1,k},$$

where

$$v_{-1}^0 = u^{lm}, \quad v_i^0 = \bar{\pi}_{lm}K_{lm}\bar{\pi}_{lm}^T r^{lm+i} \equiv \hat{K}_{lm}r^{lm+i}, \quad i \ge 0. \tag{17}$$

Consider now the following Krylov subspaces:

$$\mathcal{K}^{l,k} = \mathcal{K}_{k+1}(u^{lm}, \hat{K}_{lm}L_0). \tag{18}$$

Using the standard technique for estimating the accuracy of the Rayleigh–Ritz approximation in Krylov subspaces (see, e.g., [5]), one easily obtains the following estimate for the smallest Ritz eigenvalue $\lambda^{l,k} = \lambda(u_0(\mathcal{K}^{l,k}))$ in $\mathcal{K}^{l,k}$.

LEMMA 4.1. *Assume that $\lambda^{l_0 m} < \lambda_1$ for some l_0 and that K_n satisfies* (14) *for $n \geq l_0 m$. Then for $l \geq l_0$ the following estimate holds for the smallest Ritz eigenvalue $\lambda^{l,k}$ in the subspace $\mathcal{K}^{l,k}$:*

$$0 \leq \lambda^{l,k} - \lambda_0 \leq \frac{\lambda^{lm} - \lambda_0}{T_k(\eta_l)^2}, \tag{19}$$

where $\eta_l = q(\kappa_0, \sigma, \delta(\lambda^{lm}))^{-1}$ and

$$\kappa_0 = \frac{b_0}{a_0}, \quad \sigma = \frac{\lambda_1 - \lambda_0}{\alpha},$$

$$q(u, v, w) = \frac{u\rho(v, w) - 1}{u\rho(v, w) + 1}, \quad \rho(v, w) = \frac{(1 + w)^2 + vw}{(1 - w)^2}.$$

Proof. Let us denote for brevity $n = lm$. We have

$$\lambda_0 \leq \lambda^{l,k} = \min_{u \in \mathcal{K}^{l,k}} \lambda(u) \leq \lambda(P_k(\hat{K}_n L_0) u^n),$$

where $P_k(x)$ is any polynomial of degree k. Let us take $P_k(x) \equiv \tilde{T}_k(x)$, where

$$\tilde{T}_k(x) = c_k T_k\left(2\frac{x - a_{0n}}{b_{0n} - a_{0n}} - 1\right), \quad c_k = T_k\left(-\frac{b_{0n} + a_{0n}}{b_{0n} - a_{0n}}\right)^{-1} = \frac{(-1)^k}{T_k(\eta_l)},$$

a_{0n} and b_{0n} are given in Lemma A.4, and $T_k(x) = \cos(k \arccos x)$ is the Chebyshev polynomial of degree k. Since $\tilde{T}_k(0) = 1$ we have

$$\tilde{T}_k(x) = 1 - \sum_{i=1}^{k} \tau_i x^i$$

and, thus,

$$v^k \equiv \tilde{T}_k(\hat{K}_n L_0) u^n = \left(1 - \sum_{i=1}^{k} \tau_i (\bar{\pi}_n K_n \bar{\pi}_n^T L_0)^i\right) u^n.$$

Since $v = v^k - u^n = \bar{\pi}_n v$ we have $(Mv, u^n) = 0$, and hence $\|v^k\|_M \geq 1$, and

$$\|v^k\|_{L_0}^2 = (\lambda(v^k) - \lambda_0)\|v^k\|_M^2 \geq \lambda(v^k) - \lambda_0 \geq \lambda^{l,k} - \lambda_0.$$

Further, since $L_0 = \bar{\pi}^T L_0 \bar{\pi}$ we have $\bar{\pi}(\hat{K}_n L_0)^i = (\bar{\pi}\hat{K}_n \bar{\pi}^T L_0)^i$, and, thus,

$$\|v^k\|_{L_0} = \|\bar{\pi}\tilde{T}_k(\hat{K}_n L_0) u^n\|_{L_0} = \|\tilde{T}_k(\bar{\pi}\hat{K}_n \bar{\pi}^T L_0) u^n\|_{L_0}.$$

The operator $\bar{\pi}\hat{K}_n \bar{\pi}^T L_0$ is symmetric in the semiscalar product $(\cdot,\cdot)_{L_0}$, and so is $\tilde{T}_k(\bar{\pi}\hat{K}_n \bar{\pi}^T L_0)$. From Lemma A.4 it follows that nonzero eigenvalues of $\bar{\pi}\hat{K}_n \bar{\pi}^T L_0$ lie in the interval $[a_{0n}, b_{0n}]$. Since $|\tilde{T}_k(x)| \le |c_k|$ for $a_{0n} \le x \le b_{0n}$, this implies that $\|v^k\|_{L_0} \le |c_k| \|u^n\|_{L_0}$, and, thus,

$$\lambda^{l,k} - \lambda_0 \le c_k^2 \|u^n\|_{L_0}^2 = c_k^2(\lambda^n - \lambda_0) = \frac{\lambda^{lm} - \lambda_0}{T_k(\eta_l)^2}. \qquad \square$$

Since $\bar{\pi}_{lm}^T M u^{lm} = 0$, we have $\hat{K}_{lm} r^{lm} = \hat{K}_{lm} L_0 u^{lm}$ and, thus, $\mathcal{D}^{l,k} = \mathcal{K}^{l,k}$ for $k = 0$ and $k = 1$. For $k > 1$, $\hat{K}_{lm} r^{lm+i} \to \hat{K}_{lm} L_0 u^{lm+i}$ as $l \to \infty$ and, thus, the subspace $\mathcal{D}^{l,k}$ approaches $\mathcal{K}^{l,k}$ as $l \to \infty$. The proof of Theorem 5.3 (see Appendix B) shows that for $0 \le i < k$ the distance between v_i^0 and $\mathcal{K}^{l,k}$ is of the order $\mathcal{O}\left(\delta(\lambda^{lm})^{3/2}\right)$ (cf. (37), (41), and (34)), whereas for (16) the author could only estimate the same distance as[3] $\mathcal{O}\left(\delta(\lambda^{lm})\right)$. Further, from Lemma 4.1 we observe that $\lambda^{l,k}$ converges to λ_0 as k increases, whereas the same is not true for the minimal Ritz eigenvalue in the Krylov subspace (10). Hence, this paper uses the subspaces $\mathcal{K}^{l,k}$ for the convergence analysis of (5), instead of those in (10) or (16).

5. Convergence estimates.

In Part I of this paper the following convergence results for the PSD method (6) were obtained.

THEOREM 5.1. *Assuming that $\lambda^{n_0} < \lambda_1$ for some n_0 and that the condition* (14) *is satisfied for $n \ge n_0$, the convergence of the iterations* (6) *for $n \ge n_0$ is described by the following estimate:*

$$0 \le \lambda^{n+1} - \lambda_0 \le q(\kappa_0, \sigma, \delta(\lambda^n))^2 (\lambda^n - \lambda_0), \tag{20}$$

where

$$\kappa_0 = \frac{b_0}{a_0}, \quad \sigma = \frac{\lambda_1 - \lambda_0}{\alpha},$$

$$q(u, v, w) = \frac{u\rho(v, w) - 1}{u\rho(v, w) + 1}, \quad \rho(v, w) = \frac{(1 + w)^2 + vw}{(1 - w)^2}.$$

THEOREM 5.2. *Assume that λ_0 is a simple eigenvalue of* (1). *If $\lambda^{n_0} < \frac{\lambda_0 + \lambda_1}{2}$ for some n_0 and the preconditioner K_n satisfies* (15) *for $n \ge n_0$, then the convergence of the iterations* (6) *for $n \ge n_0$ is described by the following estimate:*

$$0 \le \lambda^{n+1} - \lambda_0 \le \tilde{q}(\kappa^0, \nu, \delta(\lambda^n))^2 (\lambda^n - \lambda_0), \tag{21}$$

where

$$\kappa^0 = \frac{b^0}{a^0}, \quad \nu = \frac{\omega}{\lambda_1 - \lambda_0},$$

$$\tilde{q}(u, v, w) = \frac{u\tilde{\rho}(v, w) - 1}{u\tilde{\rho}(v, w) + 1}, \quad \tilde{\rho}(v, w) = \frac{1}{1 - 2w} \frac{(1 + w)^2 + vw}{(1 - w)^2}.$$

Due to the minimax principle, the above results apply to the restarted GD method as well. In this section we present respective improved estimates that reflect the acceleration of the convergence due to the increasing size of subspaces $\mathcal{D}^{l,k}$.

[3] This is still enough to obtain estimates similar to those in this paper but with somewhat larger asymptotically insignificant terms—cf. [12].

THEOREM 5.3. *Assume that λ_0 is simple, that $\lambda^{l_0 m} < \lambda_1$ for some l_0, and that K_n satisfies* (14) *for $n \geq l_0 m$. Then the convergence of the restarted generalized Davidson method* (5) *for $l \geq l_0$ is described by the following estimate:*

$$0 \leq \lambda^{lm+k} - \lambda_0 \leq q^2_{lm+k}(\lambda^{lm} - \lambda_0), \tag{22}$$

where

$$q_{lm+k} \leq \prod_{n=lm}^{lm+k-1} q(\kappa_0, \sigma, \delta(\lambda^n)) < 1 \tag{23}$$

and κ_0, σ, and $q(u,v,w)$ are given in Theorem 5.1. *Furthermore,*

$$q_{lm+k} = T_k\left(\frac{\kappa_0+1}{\kappa_0-1}\right)^{-1} + \mathcal{O}\left(\sqrt{\delta(\lambda^{lm})}\right), \quad k = 0, \dots, m-1. \tag{24}$$

Proof. The estimate (23) follows immediately from that in Theorem 5.3. The proof of the estimate (24) is rather long and technical, despite the simplicity of its main idea—the closeness between the Davidson subspaces $\mathcal{D}^{l,k}$ and the Krylov subspaces $\mathcal{K}^{l,k}$ discussed in the previous section—and it is therefore placed, together with other technicalities, in the appendix (see Appendix B). □

Remark 5.1. If K_n satisfies "standard" condition $aL^{-1} \leq K_n \leq bL^{-1}$, then $\kappa_0 = \frac{a}{b}\frac{\lambda_1}{\lambda_1-\lambda_0}$ and (24) becomes the asymptotic relationship proved earlier for $m \leq 2$ by Knyazev (see [6]).

Now, it is easy to see that

$$q(\kappa_0, \sigma, \delta(\lambda^n)) = \frac{\kappa_0-1}{\kappa_0+1} + \mathcal{O}(\delta(\lambda^n)). \tag{25}$$

Comparing the above asymptotics with (24) and recalling (8) and (9), we observe that the acceleration of convergence in Davidson subspaces is similar to that in Krylov subspaces.

Finally, using Theorem 4.2 and Lemma 5.1 from [13], we obtain from Theorem 5.3 the following convergence result for the restarted Jacobi–Davidson method.

THEOREM 5.4. *Assume that λ_0 is simple, that $\lambda^{l_0 m} < \frac{\lambda_0+\lambda_1}{2}$ for some l_0, and that K_n satisfies* (15) *for $n \geq l_0 m$. Then the convergence of the restarted generalized Davidson method* (5) *for $l \geq l_0$ is described by* (22), *where*

$$q_{lm+i} \leq \prod_{n=lm}^{lm+i-1} \tilde{q}(\kappa^0, \nu, \delta(\lambda^n)) < 1, \tag{26}$$

$$q_{lm+i} = T_i\left(\frac{\kappa^0+1}{\kappa^0-1}\right)^{-1} + \mathcal{O}\left(\sqrt{\delta(\lambda^{lm})}\right), \quad i = 0, \dots, m-1, \tag{27}$$

and κ^0, ν, and $\tilde{q}(u,v,w)$ are given in Theorem 5.2.

The above result allows one to make the following observation. Assume that $\lambda^0 < (\lambda_0+\lambda_1)/2$ and that preconditioners Z_n are available that satisfy the assumption $a\tilde{L}_n^{-1} \leq Z_n \leq b\tilde{L}_n^{-1}$. One has a choice: either to use the preconditioners Z_n in the Jacobi–Davidson method, i.e., in the iterative solution of (4), or to use them directly in the GD method. It was shown in [13] that if (4) is solved by a linear iterative method that reduces the $\tilde{L}_n$-norm of the error by a factor of $\delta < 1$, then the resulting two-level iterative method can be viewed as the GD method with preconditioners K_n satisfying (15) with $a^0 = 1 - \delta$ and $b^0 = 1 + \delta$. Assuming that j iterations of the Chebyshev semi-iterative method [4] are applied we have $\delta = T_j((b+a)/(b-a))^{-1}$, and hence by Theorem 5.4 we have

$$\begin{aligned} \lim_{l\to\infty} q_{lm+i} &= T_i(\delta^{-1})^{-1} = T_i(T_j((b+a)/(b-a)))^{-1} \\ &= T_{ij}((b+a)/(b-a))^{-1}. \end{aligned} \tag{28}$$

We observe that the above asymptotic relationship is precisely the same as that for q_{lm+ij} in the restarted GD method with $K_n = Z_n$. Thus, we may conclude that in the case at hand the use of two-level rather than one-level iterations does not accelerate the convergence of the GD method. At the same time, (28) suggests that the two-level option allows one to achieve the same asymptotic convergence rate using j times smaller Davidson subspaces, which obviously improves the performance of the GD method (cf. the introduction). The same, however, cannot be said about the LOPCG versus (the considered implementation of) the Jacobi–Davidson method. Assuming, on the basis of numerical evidence, that the estimate of Theorem 5.3 is valid for the LOPCG method (cf. [8, 14]), we observe from (28) that one should expect the latter method to have an asymptotic convergence rate similar to that of the Jacobi–Davidson method if the same preconditioners Z_n are used in both and, due to the use of a smaller subspace (of dimension three—cf. the introduction), to have better overall performance. We note that this theoretical conclusion is in agreement with the numerical comparisons in [9] between the LOPCG and two practical implementations of the Jacobi–Davidson method: JDCG by Notay [10] and JDRQ by Sleijpen (see, e.g., [1]).

Appendix A. Auxiliary results. In the convergence analysis below the following auxiliary scalar product and the associated norm in $\mathcal{E}$ are used:

$$\langle u, v\rangle = (M\pi_{lm}u, \pi_{lm}u) + (\hat{K}_{lm}^{-1}\bar{\pi}_{lm}u, \bar{\pi}_{lm}v), \quad \langle u\rangle = \sqrt{\langle u, u\rangle}, \tag{29}$$

where (and in the rest of the paper) $\hat{K}_n^{-1}$ is the inverse of the restriction of $\hat{K}_n \equiv \bar{\pi}_n K_n \bar{\pi}_n^T$ onto $\bar{\pi}_n^T\mathcal{E}$.

LEMMA A.1. *If K_n satisfies* (14)*, then*

$$\langle u\rangle^2 \le \|\pi_{lm}u\|_M^2 + \frac{1}{a_0}\|\bar{\pi}_{lm}u\|_{\tilde{L}}^2 \le \left(1 + \frac{\alpha}{a_0}\right)\|u\|_M^2 + \frac{1}{a_0}\|\bar{\pi}_{lm}u\|_{L_0}^2, \tag{30}$$

where the norm $\langle\cdot\rangle$ is given by (29).

Proof. Let us denote for brevity $n = lm$. Consider the system

$$\hat{K}_n v = \bar{\pi}_n u, \quad \pi_n^T v = 0.$$

Multiplying the first equation by $\bar{\pi}_n^T v$ and using (14) we obtain

$$a_0\|\bar{\pi}_n^T v\|_{\tilde{L}^{-1}}^2 \le (K_n\bar{\pi}_n^T v, \bar{\pi}_n^T v) = (\bar{\pi}_n u, \bar{\pi}_n^T v) \le \|\bar{\pi}_n u\|_{\tilde{L}}\|\bar{\pi}_n^T v\|_{\tilde{L}^{-1}},$$

that is, $\|\bar{\pi}_n^T v\|_{\tilde{L}^{-1}} \le a_0^{-1}\|\bar{\pi}_n u\|_{\tilde{L}}$. Therefore,

$$\begin{aligned} (\hat{K}_n^{-1}\bar{\pi}_n u, \bar{\pi}_n u) &= (v, \bar{\pi}_n u) = (\bar{\pi}_n u, \bar{\pi}_n^T v) \le \frac{1}{a_0}\|\bar{\pi}_n u\|_{\tilde{L}}^2 \\ &= \frac{\alpha}{a_0}\|\pi\bar{\pi}_n u\|_M^2 + \frac{1}{a_0}\|\bar{\pi}\bar{\pi}_n u\|_{L_0}^2 \le \frac{\alpha}{a_0}\|u\|_M^2 + \frac{1}{a_0}\|\bar{\pi}_n u\|_{L_0}^2, \end{aligned}$$

which leads to (30). □

Lemma A.2. *The vectors v_i^0 given by* (17) *are orthogonal in the scalar product $\langle\cdot,\cdot\rangle$ given by* (29).

Proof. From (29) we see that $\langle v_{-1}^0, v_i^0\rangle = 0$ for $i \geq 0$. For $0 \leq i \leq j$ we have

$$\begin{aligned} \langle v_i^0, v_j^0\rangle = \langle \hat{K}_{lm} r^{lm+i}, \hat{K}_{lm} r^{lm+j}\rangle &= (r^{lm+i}, \hat{K}_{lm} r^{lm+j}) \\ &= (r^{lm+i}, K_{lm} r^{lm+j}). \end{aligned} \tag{31}$$

Since u^{lm+i} is a Ritz vector in $\mathcal{D}^{l,i}$, we have (cf. (12))

$$(r^{lm+i}, v) = 0 \quad \forall v \in \mathcal{D}^{l,i}$$

and, thus, for $0 \leq j < i$

$$\langle v_i^0, v_j^0\rangle = (r^{lm+i}, K_{lm} r^{lm+j}) = 0$$

because $K_{lm} r^{lm+j} \in \mathcal{D}^{l,j+1} \subset \mathcal{D}^{l,i}$. □

Lemma A.3. *Let $\lambda(u^{l_0 m}) < \lambda_1$ for some l_0. Let x_i be the coordinates of u^{lm+k} in the basis $\{v_i^0\}_{i=-1,k-1}$, i.e.,*

$$u^{lm+k} = x_{-1} u^{lm} + \sum_{i=0}^{k-1} x_i v_i^0. \tag{32}$$

If K_n satisfies (14), *then $|x_i| \leq (a_0(1-\delta(\lambda^{l_0 m}))^2)^{-1}$ for $l \geq l_0$ and $i \geq 0$.*

Proof. For $i = j \geq 0$ we have from (31)

$$\begin{aligned} \langle v_i^0\rangle^2 = \|r^{lm+i}\|^2_{K_{lm}} &\geq a_0 \|\bar{\pi}^T r^{lm+i}\|^2_{L_0^{-1}} = \|(L - \lambda^{lm+i} M)\bar{\pi} u^{lm+i}\|^2_{L_0^{-1}} \\ &\geq a_0 \left(\frac{\lambda_1 - \lambda^{lm+i}}{\lambda_1 - \lambda_0}\right)^2 \|\bar{\pi} u^{lm+i}\|^2_{L_0} = a_0(1 - \delta(\lambda^{lm+i}))^2 (\lambda^{lm+i} - \lambda_0) \\ &\geq c_0^2 (\lambda^{lm+i} - \lambda_0), \end{aligned} \tag{33}$$

where $c_0 = \sqrt{a_0}(1 - \delta(\lambda^{l_0 m}))$. Multiplying (32) by v_i^0 in the scalar product $\langle\cdot,\cdot\rangle$ we obtain

$$x_i \langle v_i^0\rangle^2 = \langle u^{lm+k}, v_i^0\rangle.$$

We have

$$\begin{aligned} |\langle u^{lm+k}, v_i^0\rangle| &= |(u^{lm+k}, r^{lm+i})| = |(u^{lm+k}, (L - \lambda^{lm+i} M) u^{lm+i})| \\ &= |(\lambda^{lm+k} - \lambda^{lm+i})(M u^{lm+k}, u^{lm+i})| \leq \lambda^{lm+i} - \lambda_0 \end{aligned}$$

and, thus, $|x_i| \leq c_0^{-2} = (a_0(1 - \delta(\lambda^{l_0 m}))^2)^{-1}$. □

Lemma A.4.

$$a_{0,n} L_0 \leq L_0 \hat{K}_n L_0 \leq b_{0,n} L_0,$$

where $a_{0,n} = a_0(1 - \delta(\lambda^n))^2$, $b_{0,n} = b_0((1 + \delta(\lambda^n))^2 + \sigma\delta(\lambda^n))$, and σ is defined in Theorem 5.1.

Proof. See the proof of Theorem 5.2 of [13]. □

Appendix B. Proof of Theorem 5.3. Nonasymptotic result (23) follows immediately from Theorem 5.1. It remains to obtain the asymptotic result (24).

First, we note that if $\delta(\lambda^{lm+k_0}) \le \delta(\lambda^{lm})^3$ for some $k_0 \le m$, then $\delta(\lambda^{lm+k}) \le \delta(\lambda^{lm})^3$ for $k_0 \le k \le m$ and therefore we can take in (22)

$$q_{lm+k} = T_k\left(\frac{\kappa_0+1}{\kappa_0-1}\right)^{-1} + \delta(\lambda^{lm})^2, \quad k_0 \le k \le m.$$

Hence, in the rest of the proof we consider only the case $k < k_0$ (in other words, we assume that $\delta(\lambda^{lm+k}) > \delta(\lambda^{lm})^3$).

Let us show that the distance between the vector u^{lm+i} and the subspace $\mathcal{K}^{l,i}$ in $\|\cdot\|_{L_0}$ is $\mathcal{O}(\delta(\lambda^{lm})^{\frac{3}{2}})$. In other words, let us show that for any $0 \le i \le m$ there exists $\tilde{u}^{lm+i} \in \mathcal{K}^{l,i}$ such that

$$\|u^{lm+i} - \tilde{u}^{lm+i}\|_{L_0}^2 \le \alpha_i \delta(\lambda^{lm})^3, \tag{34}$$

where α_i are positive constants independent of l and $\delta(\lambda^{lm})$. Obviously, (34) is valid for $i = 0$ and $i = 1$. Therefore, it is enough to prove that if it is valid for $i < k$, then it is valid for $i = k$.

Let

$$\tilde{v}_{-1}^0 = u^{lm}, \quad \tilde{v}_i^0 = \hat{K}_{lm} L_0 \tilde{u}^{lm+i}, \ i \ge 0, \quad \tilde{v}_i = \langle v_i^0 \rangle^{-1} \tilde{v}_i^0,$$

$$\tilde{u}^{lm+i} = \sum_{j=-1}^{i-1} x_j^i \tilde{v}_j^0, \quad \delta u^i = \tilde{u}^{lm+i} - u^{lm+i},$$

where x_j^i are the coordinates of u^{lm+i} in the basis $\{v_j^0\}_{j=-1,i-1}$. Let us start with the estimates for $\tilde{v}_i^0 - v_i^0$ in M- and L_0-norms. We have

$$\begin{aligned}\tilde{v}_i^0 - v_i^0 &= \hat{K}_{lm}((L-\lambda_0 M)\tilde{u}^{lm+i} - (L - \lambda^{lm+i}M)u^{lm+i}) \\ &= \hat{K}_{lm} L_0 \delta u^i + (\lambda^{lm+i} - \lambda_0)\hat{K}_{lm} M u^{lm+i} \equiv \delta_1 + \delta_2.\end{aligned} \tag{35}$$

From (14) we observe that

$$K_n \le b_0 \max\left\{\frac{1}{\alpha}, \frac{1}{\lambda_1 - \lambda_0}\right\} M^{-1} \equiv b^M M^{-1}. \tag{36}$$

Hence,

$$\begin{aligned}\|\hat{K}_{lm} L_0 \delta u^i\|_M^2 &= \|\bar{\pi}_{lm} K_{lm} \bar{\pi}_{lm}^T L_0 \delta u^i\|_M^2 \le \|K_{lm}\bar{\pi}_{lm}^T L_0 \delta u^i\|_M^2 \\ &\le b^M \|\bar{\pi}_{lm}^T L_0 \delta u^i\|_{K_{lm}}^2 = b^M \|L_0 \delta u^i\|_{\hat{K}_{lm}}^2 \le b^M b_{0,lm} \|\delta u^i\|_{L_0}^2,\end{aligned}$$

where $b_{0,n}$ is given in Lemma A.4. Further,

$$\begin{aligned}\|\hat{K}_{lm} M u^{lm+i}\|_M &= \|\bar{\pi}_{lm} K_{lm} \bar{\pi}_{lm}^T M u^{lm+i}\|_M \le \|K_{lm}\bar{\pi}_{lm}^T M u^{lm+i}\|_M \\ &= \|K_{lm} M \bar{\pi}_{lm} u^{lm+i}\|_M \le b^M \|\bar{\pi}_{lm} u^{lm+i}\|_M.\end{aligned}$$

Since λ_0 is simple, we have

$$\begin{aligned}\|\bar{\pi}_{lm} u^{lm+i}\|_M &= \|(\pi_{lm+i} - \pi_{lm}) u^{lm+i}\|_M \le \|(\pi_{lm+i} - \pi) u^{lm+i}\|_M \\ &\quad + \|(\pi_{lm} - \pi) u^{lm+i}\|_M \le 2\sqrt{\delta(\lambda^{lm})},\end{aligned}$$

and hence $\|\hat{K}_{lm} M u^{lm+i}\|_{_M} < 2b^M \sqrt{\delta(\lambda^{lm})}$. Thus,

$$
\begin{aligned}
\|\tilde{v}_i^0 - v_i^0\|_{_M} &\le \sqrt{b^M b_{0,lm}} \|\delta u^i\|_{_{L_0}} + 2b^M (\lambda_1 - \lambda_0) \delta(\lambda^{lm+i})^{\frac{3}{2}} \\
&\le \sqrt{b^M (4+\sigma) b_0} \|\delta u^i\|_{_{L_0}} + 2b^M (\lambda_1 - \lambda_0) \delta(\lambda^{lm+i})^{\frac{3}{2}} \\
&\equiv b_1 \|\delta u^i\|_{_{L_0}} + b_2 \delta(\lambda^{lm+i})^{\frac{3}{2}} = (b_1 \sqrt{\alpha_i} + b_2) \delta(\lambda^{lm+i})^{\frac{3}{2}}. \qquad (37)
\end{aligned}
$$

In L_0-norm we have

$$
\|\delta_1\|_{_{L_0}} \le b_{0,lm} \|\delta u^i\|_{_{L_0}} \le (4+\sigma) b_0 \|\delta u^i\|_{_{L_0}} \equiv b_3 \|\delta u^i\|_{_{L_0}}. \qquad (38)
$$

Further, since (14) implies that $L_0 \le b_0 K_n^{-1}$, we have

$$
\begin{aligned}
\|\hat{K}_{lm} M u^{lm+i}\|_{_{L_0}} &= \|(1 - \pi_{lm}) K_{lm} M \bar{\pi}_{lm} u^{lm+i}\|_{_{L_0}} \\
&\le \|K_{lm} M \bar{\pi}_{lm} u^{lm+i}\|_{_{L_0}} + \|K_{lm} M \bar{\pi}_{lm} u^{lm+i}\|_{_M} \|u^{lm}\|_{_{L_0}} \\
&\le \left(\sqrt{b_0} + \sqrt{b^M} \sqrt{\lambda^{lm} - \lambda_0} \right) \|M \bar{\pi}_{lm} u^{lm+i}\|_{_{K_{lm}}} \\
&\le \left(\sqrt{b_0 b^M} + b^M \sqrt{\lambda^{l_0 m}} \right) \|\bar{\pi}_{lm} u^{lm+i}\|_{_M} \\
&\equiv b_4 \|\bar{\pi}_{lm} u^{lm+i}\|_{_M} = 2b_4 \sqrt{\delta(\lambda^{lm})}. \qquad (39)
\end{aligned}
$$

Thus,

$$
\|\delta_2\|_{_{L_0}} \le 2b_4 (\lambda_1 - \lambda_0) \delta(\lambda^{lm+i})^{\frac{3}{2}} \equiv b_5 \delta(\lambda^{lm+i})^{\frac{3}{2}} \qquad (40)
$$

and

$$
\|\tilde{v}_i^0 - v_i^0\|_{_{L_0}} \le b_3 \|\delta u^i\|_{_{L_0}} + b_5 \delta(\lambda^{lm+i})^{\frac{3}{2}} = (b_3 \sqrt{\alpha_i} + b_5) \delta(\lambda^{lm+i})^{\frac{3}{2}}. \qquad (41)
$$

Using Lemma A.3 together with (37) and (41) we obtain (34).

The outline of rest of the proof is as follows. First, we show that the vectors $\tilde{v}_i$ are linearly independent, by exploiting the fact that $\tilde{v}_i$ are close to v_i in the norm $\langle \cdot \rangle$, and v_i are orthogonal in the scalar product $\langle \cdot, \cdot \rangle$. Then, using $\tilde{v}_i$, $i = -1, \dots, k-1$, as a basis in $\mathcal{K}^{l,k}$ we obtain an estimate for the coordinates of the Ritz vector $v^{l,k} = u_0(\mathcal{K}^{l,k})$ in this basis. Finally, we use this estimate together with the above estimates for $\tilde{v}_i^0 - v_i^0$ in L_0-norm and Lemma 4.1 to obtain the asymptotic estimate (24).

Obviously, $\langle \tilde{v}_{-1}, \tilde{v}_i \rangle = 0$, $i \ge 0$. Let $\tilde{V}_{l,k-1}$ be the Gram matrix for $\tilde{v}_i$, $i = 0, \dots, k-1$, associated with the scalar product $\langle \cdot, \cdot \rangle$, i.e., the matrix with the entries $\langle \tilde{v}_i, \tilde{v}_j \rangle$. From (34) we have

$$
\|\delta u^i\|_{_{L_0}}^2 \le \gamma_0 \delta(\lambda^{lm})^3, \quad \gamma_0 = \max_{0 \le i < k_0} \alpha_i,
$$

and hence

$$
\langle \delta_1 \rangle^2 = (\hat{K}_{lm} L_0 \delta u^i, L_0 \delta u^i) \le b_{0,lm} \|\delta u^i\|_{_{L_0}}^2 \le b_3 \gamma_0 \delta(\lambda^{lm})^3.
$$

Further,

$$
\langle \delta_2 \rangle^2 = (\lambda^{lm+i} - \lambda_0)^2 \|M u^{lm+i}\|_{_{\hat{K}_{lm}}}^2 \le 4b^M (\lambda_1 - \lambda_0)^2 \delta(\lambda^{lm+i})^3,
$$

where δ_1 and δ_2 are from (35). Thus,

$$\langle \tilde{v}_i^0 - v_i^0 \rangle^2 \le \gamma_1^2 \delta(\lambda^{lm})^3,$$

where $\gamma_1 = \sqrt{b_3\gamma_0} + 2(\lambda_1 - \lambda_0)\sqrt{b^M}$. Denoting $v_i = \langle v_i^0 \rangle^{-1} v_i^0$ we have

$$\text{(42)} \qquad \langle \tilde{v}_i - v_i \rangle^2 = \langle v_i^0 \rangle^{-2} \langle \tilde{v}_i^0 - v_i^0 \rangle^2 \le \frac{\gamma_1^2}{c_0^2(\lambda_1 - \lambda_0)} \frac{\delta(\lambda^{lm})^3}{\delta(\lambda^{lm+i})} \equiv \gamma_2^2 \frac{\delta(\lambda^{lm})^3}{\delta(\lambda^{lm+i})},$$

where c_0 is given in the proof of Lemma A.2. From the above estimate, together with the trivial relationship

$$\langle \tilde{v}_i, \tilde{v}_j \rangle = \langle v_i, v_j \rangle + \langle \tilde{v}_i - v_i, v_j \rangle + \langle v_i, \tilde{v}_j - v_j \rangle + \langle \tilde{v}_i - v_i, \tilde{v}_j - v_j \rangle,$$

we obtain for $i > j \ge 0$

$$|\langle \tilde{v}_j \rangle^2 - 1| \le \epsilon_{l,j}, \quad |\langle \tilde{v}_i, \tilde{v}_j \rangle| \le \epsilon_{l,i},$$

where

$$\epsilon_{l,i} = \gamma_2 \left(2 + \gamma_2 \frac{\delta(\lambda^{lm})^{\frac{3}{2}}}{\sqrt{\delta(\lambda^{lm+i})}} \right) \frac{\delta(\lambda^{lm})^{\frac{3}{2}}}{\sqrt{\delta(\lambda^{lm+i})}}.$$

By the same logic as above, we may assume that $\delta(\lambda^{lm+k}) \ge \delta(\lambda^{lm})^2$ and, thus, $\epsilon_{l,i} = \mathcal{O}(\delta(\lambda^{lm})^{\frac{1}{2}})$, and the matrix $\tilde{V}_{l,k-1}$ can be represented as $\tilde{V}_{l,k-1} = 1 - \tilde{W}_{l,k-1}$ with $\|\tilde{W}_{l,k-1}\| = \mathcal{O}(\delta(\lambda^{lm})^{\frac{1}{2}})$. Thus, for $l \ge l_0$ we have $\|\tilde{W}_{l,k-1}\| < 1$, and hence the matrix $\tilde{V}_{l,k-1}$ is nondegenerate, which implies that the vectors $\tilde{v}_i$ are linearly independent.

Using $\tilde{v}_i$, $i = -1, \dots, k-1$, as the basis in $\mathcal{K}^{l,k}$ we have

$$\text{(43)} \qquad v^{l,k} = u_0(\mathcal{K}^{l,k}) = x_{-1}u^{lm} + \sum_{i=0}^{k-1} x_j \tilde{v}_j.$$

The last technically difficult step in the proof is to obtain for the coordinates x_i, $i \ge 0$, the estimate $x_i = \mathcal{O}(\delta(\lambda^{lm})^{\frac{1}{2}})$. Multiplying (43) by $\tilde{v}_i$, $i = 0, \dots, k-1$, in the scalar product $\langle \cdot, \cdot \rangle$ we obtain the linear system

$$\sum_{j=0}^{k-1} x_j \langle \tilde{v}_i, \tilde{v}_j \rangle = \langle \tilde{v}_i, v^{l,k} \rangle \equiv y_i, \quad i = 0, \dots, k-1,$$

or, in matrix form,

$$\tilde{V}_{l,k-1} x = y,$$

which is equivalent to

$$\tilde{V}_{l,k-1} z = \tilde{W}_{l,k-1} y,$$

where $z = x - y$. Using the above estimates for the norm of $\tilde{W}_{l,k-1}$, we have

$$\|z\| \le \|\tilde{W}_{l,k-1}\| \, \|y\| = \mathcal{O}(\delta(\lambda^{lm})^{\frac{1}{2}}) \|y\|,$$

which shows that $\|x\| = \mathcal{O}(\|y\|)$. To estimate $\|y\|$ we observe that

$$\begin{aligned}\langle v^{l,k}, v_i^0\rangle &= (v^{l,k}, r^i) = (v^{l,k}, (L - \lambda^{lm+i}M)u^{lm+i})\\ &= (v^{l,k}, L_0 u^{lm+i}) - (\lambda^{lm+i} - \lambda_0)(Mv^{l,k}, u^{lm+i}),\end{aligned}$$

and hence

$$\begin{aligned}|\langle v^{l,k}, v_i^0\rangle| &\le \|v^{l,k}\|_{L_0}\|u^{lm+i}\|_{L_0} + |\lambda^{lm+i} - \lambda_0||(Mv^{l,k}, u^{lm+i})|\\ &\le \sqrt{\lambda^{l,k} - \lambda_0}\sqrt{\lambda^{lm+i} - \lambda_0} + \lambda^{lm+i} - \lambda_0.\end{aligned}$$

Now, if $\lambda^{lm+i} - \lambda_0 \le \lambda^{l,k} - \lambda_0$, then $\lambda^{lm+k} - \lambda_0 \le \lambda^{l,k} - \lambda_0$, and, by Lemma 4.1, we can take $q_{lm+k} = T_k(\eta_l)^{-1}$. Otherwise, we have

$$|\langle v^{l,k}, v_i^0\rangle| \le 2(\lambda^{lm+i} - \lambda_0), \quad |\langle v^{l,k}, v_i\rangle| \le \frac{2}{c_0}\sqrt{\lambda^{lm+i} - \lambda_0}.$$

Further,

$$\begin{aligned}\|\bar{\pi}_{lm} v^{l,k}\|_{L_0} &\le \|v^{l,k}\|_{L_0} + |(Mv^{l,k}, u^{lm})|\|u^{lm}\|_{L_0}\\ &\le \sqrt{\lambda^{l,k} - \lambda_0} + \sqrt{\lambda^{lm} - \lambda_0} \le 2\sqrt{\lambda^{lm} - \lambda_0},\end{aligned}$$

and hence, using Lemma A.1, we have

$$\langle v^{l,k}\rangle \le 1 + \frac{\alpha}{a_0} + 4(\lambda^{l_0 m} - \lambda_0) \equiv \gamma_3.$$

Thus,

$$\begin{aligned}|\langle v^{l,k}, \tilde{v}_i\rangle| &\le |\langle v^{l,k}, \hat{v}_i\rangle| + |\langle v^{l,k}, (\tilde{v}_i - v_i)\rangle|\\ &\le \frac{2}{c_0}\sqrt{\lambda^{lm+i} - \lambda_0} + \gamma_2\gamma_3\frac{\delta(\lambda^{lm})^{\frac{3}{2}}}{\sqrt{\delta(\lambda^{lm+i})}} = \mathcal{O}(\delta(\lambda^{lm})^{\frac{1}{2}})\end{aligned}$$

and

$$\|x\| = \mathcal{O}(\|y\|) = \mathcal{O}(\delta(\lambda^{lm})^{\frac{1}{2}}). \tag{44}$$

The rest of the proof is fairly simple. Let us denote

$$u^{l,k} = x_{-1}u^{lm} + \sum_{i=0}^{k-1} x_j v_j.$$

From (37) and (41) we have

$$\|\tilde{v}_i - v_i\|_M = \mathcal{O}(\delta(\lambda^{lm})^{\frac{1}{2}}), \quad \|\tilde{v}_i - v_i\|_{L_0} = \mathcal{O}(\delta(\lambda^{lm})^{\frac{1}{2}}).$$

Using the above estimates and (44), we have for $\delta v^{l,k} = v^{l,k} - u^{l,k}$

$$\|\delta v^{l,k}\|_M = \mathcal{O}(\delta(\lambda^{lm})), \quad \|\delta v^{l,k}\|_{L_0} = \mathcal{O}(\delta(\lambda^{lm})).$$

Using Lemma 4.1 we obtain

$$\begin{aligned}\|u^{l,k}\|_{L_0} \le \|v^{l,k}\|_{L_0} + \|\delta v^{l,k}\|_{L_0} &= \sqrt{\lambda^{l,k} - \lambda_0} + \mathcal{O}(\delta(\lambda^{lm}))\\ &\le T_k(\eta)^{-1}\sqrt{\lambda^{lm} - \lambda_0} + \mathcal{O}(\delta(\lambda^{lm}))\\ &= (T_k(\eta_l)^{-1} + \mathcal{O}(\delta(\lambda^{lm})^{\frac{1}{2}}))\sqrt{\lambda^{lm} - \lambda_0}.\end{aligned}$$

Since $\|u^{l,k}\|_M \geq \|v^{l,k}\|_M - \|\delta v^{l,k}\|_M = 1 - \|\delta v^{l,k}\|_M = \mathcal{O}(1)$, we finally obtain

$$\begin{aligned}\lambda^{lm+k} - \lambda_0 \leq \lambda(u^{l,k}) - \lambda_0 &\leq (\lambda(u^{l,k}) - \lambda_0)(\|u^{l,k}\|_M + \|\delta v^{l,k}\|_M)^2 \\ &= \|u^{l,k}\|_{L_0}^2 \left(1 + \frac{\|\delta v^{l,k}\|_M}{\|u^{l,k}\|_M}\right)^2 \leq q_{lm+k}^2(\lambda^{lm} - \lambda_0),\end{aligned}$$

where

$$\begin{aligned}q_{lm+k} &= (T_k(\eta_l))^{-1} + \mathcal{O}(\delta(\lambda^{lm})^{\frac{1}{2}}))(1 + \mathcal{O}(\delta(\lambda^{lm})) \\ &= T_k(\eta_l)^{-1} + \mathcal{O}(\delta(\lambda^{lm})^{\frac{1}{2}}).\end{aligned}$$

REFERENCES

[1] Z. Bai, J. Demmel, J. Dongarra, A. Ruhe, and H. van der Vorst, eds., *Templates for the Solution of Algebraic Eigenvalue Problems: A Practical Guide*, Software Environ. Tools 11, SIAM, Philadelphia, 2000.

[2] M. Crouzeix, B. Philippe, and M. Sadkane, *The Davidson method*, SIAM J. Sci. Comput., 15 (1994), pp. 62–76.

[3] E. R. Davidson, *The iterative calculation of a few of the lowest eigenvalues and corresponding eigenvectors of large real symmetric matrices*, J. Computational Phys., 17 (1975), pp. 87–94.

[4] G. H. Golub and R. S. Varga, *Chebyshev semi-iterative methods, successive overrelaxation iterative methods and second order Richardson iterative methods*, Numer. Math., 3 (1961), pp. 147–168.

[5] A. V. Knyazev, *Convergence rate estimates for iterative methods for a mesh symmetric eigenvalue problem*, Soviet J. Numer. Anal. Math. Modelling, 2 (1987), pp. 371–396.

[6] A. V. Knyazev, *A Preconditioned Conjugate Gradient Method for Eigenvalue Problems and Its Implementation in a Subspace*, Internat. Ser. Numer. Math. 96, Birkhäuser, Basel, 1991, pp. 143–154.

[7] A. V. Knyazev, *Preconditioned eigensolvers—an oxymoron?*, Electron. Trans. Numer. Anal., 7 (1998), pp. 104–123.

[8] A. V. Knyazev, *Toward the optimal preconditioned eigensolver: Locally optimal block preconditioned conjugate gradient method*, SIAM J. Sci. Comput., 23 (2001), pp. 517–541.

[9] A. V. Knyazev and K. Neymeyr, *A geometric theory for preconditioned inverse iteration.* III: *A short and sharp convergence estimate for generalized eigenvalue problems*, Linear Algebra Appl., 358 (2003), pp. 95–114.

[10] Y. Notay, *Combination of Jacobi–Davidson and conjugate gradients for the partial symmetric eigenproblem*, Numer. Linear Algebra Appl., 9 (2002), pp. 21–44.

[11] S. Oliveira, *On the convergence rate of a preconditioned subspace eigensolver*, Computing, 63 (1999), pp. 219–231.

[12] E. Ovtchinnikov, *Convergence Estimates for the Generalized Davidson Method for Symmetric Eigenvalue Problems*, Technical report, University of Westminster, London, UK, 2001.

[13] E. Ovtchinnikov, *Convergence estimates for the generalized Davidson method for symmetric eigenvalue problems* I: *The preconditioning aspect*, SIAM J. Numer. Anal., 41 (2003), pp. 258–271.

[14] E. Ovtchinnikov, *Generalized Davidson Versus Conjugate Gradient Methods: A Numerical Study*, manuscript, 2002.

[15] E. Ovtchinnikov, *Convergence Estimates for Preconditioned Gradient Subspace Iteration Eigensolvers*, Technical report 1244, University of Utrecht, Utrecht, the Netherlands, June 2002.

[16] B. N. Parlett, *The Rayleigh quotient iteration and some generalizations*, Math. Comp., 28 (1974), pp. 679–693.

[17] Y. Saad, *Numerical Methods for Large Eigenvalue Problems*, Algorithms Archit. Adv. Sci. Comput., Manchester University Press, Manchester, UK, Halsted Press, New York, 1992.

[18] B. Samokish, *The steepest descent method for an eigenvalue problem with semi-bounded operators*, Izv. Vyssh. Uchebn. Zaved. Mat., 5 (1958), pp. 105–114.

[19] G. L. G. Sleijpen and H. A. van der Vorst, *A Jacobi–Davidson iteration method for linear eigenvalue problems*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 401–425.

SIAM J. NUMER. ANAL.
Vol. 41, No. 1, pp. 287–305

ON OPTIMAL FINITE-DIFFERENCE APPROXIMATION OF PML*

SERGEY ASVADUROV[†], VLADIMIR DRUSKIN[†], MURTHY N. GUDDATI[‡], AND LEONID KNIZHNERMAN[§]

Abstract. A technique derived from two related methods suggested earlier by some of the authors for optimization of finite-difference grids and absorbing boundary conditions is applied to discretization of perfectly matched layer (PML) absorbing boundary conditions for wave equations in Cartesian coordinates. We formulate simple sufficient conditions for optimality and implement them. It is found that the minimal error can be achieved using pure imaginary coordinate stretching. As such, the PML discretization is algebraically equivalent to the rational approximation of the square root on $[0, 1]$ conventionally used for approximate absorbing boundary conditions. We present optimal solutions for two cost functions, with exponential (and exponential of the square root) rates of convergence with respect to the number of the discrete PML layers using a second order finite-difference scheme with optimal grids. Results of numerical calculations are presented.

Key words. absorbing boundary conditions, exponential convergence, finite differences, hyperbolic problems, perfectly matched layers, wave propagation, optimal rational approximations

AMS subject classifications. 65N06, 73C02

PII. S0036142901391451

1. Introduction.

This paper is a sequel to a number of papers on so-called optimal finite-difference grids or finite-difference Gaussian rules [11, 12, 3, 4, 18, 13], where exponential superconvergence of standard second order finite-difference schemes at a priori given points was obtained due to a special grid optimization procedure. This approach was successfully applied to the approximation of many nontrivial practically important problems, including elliptic PDEs for both bounded and unbounded domains. For the latter, in [18] the optimal finite-difference grid was obtained, which can be considered as the boundary condition requiring minimal arithmetic work for a given spectral interval. For hyperbolic problems, however, optimal grids were introduced only for the approximation in the interior part of the domain. Here we consider exterior hyperbolic problems. For this kind of problem a closely related method of continued fraction boundary conditions was suggested in [16], where absorbing boundary conditions were reduced to a three-term equations resembling finite-difference relations. Combining the approaches of [18] and [16] we obtain frequency independent finite-difference discretization of Berenger's perfectly matched layer (PML) absorbing boundary conditions (ABCs) [7], which produces the minimal possible impedance error for a given number of discrete layers. Similarly to the optimal grid for the Laplace equation with a solution from a Sobolev space considered in [18], the obtained discretizations show the exponential of the square root rates of convergence, though they use only the three-point stencil for second derivatives. Our solution exhibits much smaller reflection coefficients compared to examples of optimized PMLs (for the same

*Received by the editors June 26, 2001; accepted for publication (in revised form) September 12, 2002; published electronically March 26, 2003.

http://www.siam.org/journals/sinum/41-1/39145.html

[†] Schlumberger-Doll Research, Old Quarry Road, Ridgefield, CT 06877-4108 (asvadurov@slb.com, druskin@slb.com).

[‡] Department of Civil Engineering, North Carolina State University, Raleigh, NC 27695-7908 (mnguddat@eos.ncsu.edu).

[§] Central Geophysical Expedition, Narodnogo Opolcheniya St., 40-3, Moscow 123298, Russia (mmd@cge.ru).

numbers of discrete layers) known from the literature [10]. The optimal solution represents a limiting case of PML with pure imaginary coordinate stretching that results in a different type of the equation in time domain. The drawback of the proposed method is that, unlike a conventional PML discretization, it cannot be implemented in time domain using the same time-stepping realization in the interior and the exterior parts. Nevertheless, the arithmetical cost per grid node of the new method is close to that of the conventional PML.

Let us consider a model problem for the scalar wave equation on $\mathbb{R}^2 \times [0, +\infty)$ in Cartesian coordinates

$$u_{xx} + u_{yy} - u_{tt} = 0. \tag{1.1}$$

After the Fourier transform with respect to y and t, (1.1) becomes

$$u_{xx} - \left(l^2 - \omega^2\right) u = 0, \tag{1.2}$$

where l and ω are real spatial and temporal frequencies, respectively. We here abuse notation slightly by using the same name for a function and its Fourier transform; the use of the time or frequency domain will always be clear from the context. The ratio $\sigma = l/\omega$ is the sine of the incidence angle of the wave on the plane $x = 0$; this angle is labelled θ.

We assume that only so-called propagative modes with $|\sigma| \leq 1$ are present in the spectrum of the solution and that for positive x the solution contains only waves moving to the right; i.e., we are considering solutions of the form

$$u(x) = ce^{-i\omega\sqrt{\lambda}x} \text{ if } x > 0, \tag{1.3}$$

where $\lambda = 1 - \sigma^2 = \cos^2\theta$ is positive.

The solutions of this form do not vanish at infinity, which is the origin of the notorious problem of domain truncation in the numerical solution of wave problems on unbounded domains. Functions given by (1.3) satisfy the impedance boundary condition

$$u_x\big|_{x=0} = -i\omega\sqrt{\lambda}u\big|_{x=0}. \tag{1.4}$$

With the help of this condition, the subdomain $x > 0$ can be truncated. Many approximate ABCs are based on rational approximation of $\sqrt{\lambda}$, e.g., in [14, 16]. The authors of these investigations used the fact that the approximant after the inverse Fourier transform to (y, t) coordinates becomes a solution of a PDE in (y, t) plane; i.e., it can be computed within the finite-difference time-stepping framework of the solution of (1.1).

Instead of the direct implementation of the approximate condition (1.4), one can modify (1.2) for $x > 0$ in such a way that it would be easier to solve and the new solution would approximate (1.4) well. The methods of this sort that recently received wide attention in the literature, e.g., in [7, 9, 10, 19], are Berenger's PML or sponge layer methods. These methods generate special artificial media layers that add exponential decay or attenuation to the propagative modes so that the new solution satisfies the same boundary condition at $x = 0$. Thus, the domain for the new equation can be truncated, which will produce only a small reflection for $x < 0$. For brevity we will call PML all methods from this group.

It is known that Berenger's PML can be obtained using complex coordinate stretching [9]:

$$dx = \left(\alpha + \frac{\beta}{i\omega}\right) d\bar{x} \quad \text{if } x > 0, \qquad x = \bar{x} \quad \text{otherwise,} \tag{1.5}$$

where α and β are some real nonnegative functions. The ω^{-1} dependence of the imaginary part is introduced to simplify the time-domain formulation and weaken the dependence on ω of the PML error. The new function $\bar{u}(\bar{x}) = u(x)$ is defined, and the equation is modified in the following way: the new variable $\bar{x}$ is taken to be real. This transformation does not change the solution for negative $\bar{x}$, and for positive $\bar{x}$ it transforms (1.3) to

$$\bar{u}(\bar{x}) = c \exp\left[-\sqrt{\lambda}\int_0^{\bar{x}} \beta(\xi) d\xi\right] \exp\left[-i\omega\sqrt{\lambda}\int_0^{\bar{x}} \alpha(\xi) d\xi\right];$$

i.e., the exponential attenuation is added to the resulting function. Because of this attenuation of the solution in the new coordinate, the subdomain $[0, +\infty)$ can be truncated to a finite length L, with the logarithm of the absolute value of the impedance error (or reflection coefficient) approximately proportional to

$$R = -\sqrt{\lambda}\int_0^L \beta(\xi) d\xi.$$

From this estimate one might conclude that just choosing large enough β alone would make error negligibly small, regardless of the temporal frequency. Unfortunately, this consideration cannot be applied to *discretized* PML because of the numerical dispersion, which is frequency dependent and increases with the increase of β. Conventionally, ad hoc rules or general nonlinear optimization algorithms have been used for the choice of the discrete PML parameters, e.g., in [10]. It is obvious that proper use of the analytical structure of the cost function can greatly improve the efficiency of such optimization; this is the main motivation of our investigation.

Let u^k denote the solution of (1.2) after a three-point second order finite-difference discretization of PML with k primary and k dual nodes. In this case, α and β become finite $2k$-dimensional vectors that determine the discrete transformation at the primary and dual grid nodes. We consider the following *PML optimization problem*:

- Find α, β minimizing the error functional

$$\delta_k^{\alpha,\beta} = \sup_{0\le\lambda\le1,\ 0<\omega\le\omega_{\max}} s(\lambda)\left|\sqrt{\lambda} - \frac{u_x^k}{i\omega u^k}\Big|_{x=0}\right|, \tag{1.6}$$

where u_x^k is the finite-difference derivative of the discrete solution at the boundary. Here we assume that the temporal spectrum of the solution is uniformly bounded at $(0, \omega_{\max}]$ by a positive constant, where $\omega_{\max}$ is a cutoff frequency (as we shall see, the value of $\omega_{\max}$ is unimportant for the construction of our optimal solution), and $s(\lambda)$ is a nonnegative weight, chosen depending on the distribution of the incident waves. We will consider two cases:

$$s(\lambda) = \frac{1}{\sqrt{\lambda}} \text{ if } \lambda \in [\lambda_{\min}, 1], \lambda_{\min} > 0, \qquad s(\lambda) = 0 \text{ otherwise,} \tag{1.7}$$

and

$$s(\lambda) = 1, \quad \lambda \in [0, 1]. \tag{1.8}$$

Here the first case corresponds to all possible waves with the incidence angles not exceeding $\arccos(\sqrt{\lambda_{\min}}) < \frac{\pi}{2}$; the second case assumes that the wave amplitudes are uniformly bounded for all incidence angles on $[0, \pi/2]$. The first case will be used when a priori information limiting the range of incidence angles is available; the second case will generally be used when no such information can be obtained from the geometry of the model.

Traditionally, only equidistant grids were used for the PML discretization, and the optimization was performed by adjusting the distribution of α or β. Such discretization can be equivalently presented as the one on a nonuniform grid with *complex* grid steps but with $\alpha = 1$, $\beta = 0$. So, we can apply the approach of [11, 18], which reduces the problem of the finite-difference grid optimization to rational approximation and allows us to make the following conclusions, which will be proven in the following sections:

- The minimum of $\delta_k^{\alpha,\beta}$ can be achieved with pure imaginary stretching:
$$\min_{\alpha,\beta} \delta_k^{\alpha,\beta} = \min_{\beta} \delta_k^{0,\beta},$$
though it is not clear if the condition $\alpha \equiv 0$ is necessary for optimality.
- With pure imaginary stretching the PML discretization becomes *independent* of ω and depends only on the incidence angle. In other words, the PML discretization becomes a well-studied problem of rational approximation of the $\sqrt{\lambda}$ on a nonnegative interval of real axis.

Choosing appropriate grid steps (or β) for PML with pure imaginary stretching, one can make it algebraically equivalent to known approximate ABCs of, e.g., [14, 16]. Instead, using known results of approximation theory, we give optimal solutions for the two weight functions considered here, which are efficient for wide bandwidth of the incidence angles. For a problem similar to (1.7) the optimal PML discretization is based on a closed form solution, obtained in 1877 by Zolotarjov [20]. If $0 < \lambda_{\min} \ll 1$ its error decays with the increase of k approximately as

$$O\left\{\exp\left[\frac{\pi^2 k}{\log\left(\sqrt{\lambda_{\min}}/4\right)}\right]\right\}. \tag{1.9}$$

For the case (1.8) the optimal grid was computed in the manner of [24], where optimal rational approximants of a slightly different type were computed to very high precision. For the latter case the optimal error asymptotically decays as

$$8e^{-\pi\sqrt{2k}}. \tag{1.10}$$

Both of these estimates again highlight the phenomenon of exponential superconvergence of three-point second order finite-different approximations, which has been earlier used for efficient approximation of elliptic and hyperbolic problems in [11, 3, 18, 12].

The condition $\alpha \equiv 0$ does not allow standard split time-domain PML realization; however, it makes possible a simple nonsplit realization, which is, in fact, similar to the one of [16].

2. S-fraction representation of discrete PML.

We want the propagating solution $u = ce^{-i\sqrt{\omega^2 - l^2}x}$ to become nonoscillating evanescent for $x > 0$. We consider a limiting case of transformation (1.5) with $\alpha = 0$ and $\beta = 1$:

$$\bar{x} = i\omega x \text{ if } x > 0, \qquad \bar{x} = x \text{ otherwise.}$$

Define a new function $\bar{u}(\bar{x}) = u(x)$ and again take $\bar{x}$ to be real. The equation that is satisfied by this new function can be written in divergence form as

$$\frac{d}{d\bar{x}}\left[\gamma(\bar{x})\frac{d\bar{u}(\bar{x})}{d\bar{x}}\right] - \lambda\rho(\bar{x})\bar{u}(\bar{x}) = 0, \tag{2.1}$$

where $\lambda = 1 - l^2/\omega^2 = \cos^2\theta$ and

$$\gamma = i\omega,\ \rho = i\omega \text{ if } \bar{x} > 0, \qquad \gamma = 1,\ \rho = -\omega^2 \text{ otherwise.} \tag{2.2}$$

Equation (2.1) is a standard divergence equation with discontinuous coefficients; $\bar{u}(\bar{x})$ and $\gamma(\bar{x})\frac{d\bar{u}(\bar{x})}{d\bar{x}}$ are continuous across the interface at $\bar{x} = 0$.

The transformed Helmholtz equation (2.1) becomes diffusive (absorbing) for $\bar{x} > 0$ and remains the same as the original oscillating (1.2) otherwise. Its solution vanishing at $+\infty$ can be written as

$$\bar{u} = ce^{-\sqrt{\lambda}\bar{x}} \text{ if } \bar{x} > 0, \qquad \bar{u} = ce^{-i\omega\sqrt{\lambda}\bar{x}} = u \text{ otherwise.}$$

From now on we will approximate the above solution for $\bar{x} \in \mathbb{R}$, so we will drop the bars over all the symbols. The impedance condition (1.4) is then written as

$$u_x\big|_{x=0+} = -\sqrt{\lambda}u\big|_{x=0+}, \tag{2.3}$$

and the error functional (1.6) for the discrete absorbing condition is

$$\delta_k = \max_{0\le\lambda\le1} s(\lambda)\left|\sqrt{\lambda} - \frac{u_x^k}{u^k}|_{x=0}\right|, \tag{2.4}$$

where u^k is the discrete solution of (2.1), and u_x^k stands for the finite-difference analogue of the derivative of u.

Now, using a second order finite-difference approximation of the diffusive part of (2.1) as it was done for diffusive problems in [11, 18], we obtain a formulation equivalent to the continued fraction of [16].

We want to approximate the ABC (2.3) and to get an explicit expression for the error functional (2.4) in terms of the steps of the finite-difference discretization. For this purpose we first consider (2.1) for $x \in [0, +\infty)$ with the following boundary conditions:

$$-\gamma u'(0) = 1, \qquad u(+\infty) = 0.$$

Let us approximate the solution of this problem by a staggered three-point finite-difference scheme. In a staggered scheme, the numerical solution is defined at primary nodes

$$x_j,\quad j = 1, \ldots, k+1, \qquad \text{with}\quad x_1 = 0 \quad\text{and}\quad x_{j+1} > x_j\ (1 \le j \le k),$$

and the finite-difference derivatives are defined at dual nodes

$$\hat{x}_j,\quad j = 0, \ldots, k, \qquad \text{with}\quad \hat{x}_0 = 0 \quad\text{and}\quad \hat{x}_{j+1} > \hat{x}_j\ (1 \le j \le k-1).$$

We denote the step sizes by

$$h_j = x_{j+1} - x_j \quad\text{and}\quad \hat{h}_j = \hat{x}_j - \hat{x}_{j-1}$$

and solve the following finite-difference problem:

$$\frac{1}{\hat{h}_j}\left(\hat{\gamma}_j\frac{u_{j+1}-u_j}{h_j}-\hat{\gamma}_{j-1}\frac{u_j-u_{j-1}}{h_{j-1}}\right)-\lambda\rho_j u_j=0, \qquad j=2,\dots,k, \tag{2.5}$$

with the boundary conditions

$$\frac{\hat{\gamma}_1}{\hat{h}_1}\left(\frac{u_2-u_1}{h_1}\right)-\lambda\rho_1 u_1=-\frac{1}{\hat{h}_1} \tag{2.6}$$

and

$$u_{k+1}=0. \tag{2.7}$$

Note that the first boundary condition (2.6) is consistent with the differential equation since it is the same as creating a dummy node u_0, allowing $j=1$ in (2.5) and setting

$$-\hat{\gamma}_0\frac{u_1-u_0}{h_0}=1.$$

Discrete analogues of γ and ρ equal $\hat{\gamma}_j=\rho_j=i\omega$ in accordance with the definition (2.2) of their continuous counterparts on $[0,+\infty)$.

As follows from [11, 18], $u_1=-\frac{f_k}{\gamma}$, where the so-called discrete impedance function f_k is the Stieltjes fraction or S-fraction

$$f_k(\lambda)=\cfrac{1}{\hat{h}_1\lambda+\cfrac{1}{h_1+\cfrac{1}{\hat{h}_2\lambda+\dots\cfrac{1}{h_{k-1}+\cfrac{1}{\hat{h}_k\lambda+\cfrac{1}{h_k}}}}}}. \tag{2.8}$$

It is easy to see that this S-fraction formulation is algebraically equivalent to the continued fraction formulation of [16] with a proper choice of the steps.

Overall, for the error functional we get

$$\delta_k=\max_{\lambda\in[0,1]} s(\lambda)\left|\sqrt{\lambda}-\frac{1}{\gamma u_1}\right|=\max_{\lambda\in[0,1]} s(\lambda)\left|\sqrt{\lambda}-\frac{1}{f_k(\lambda)}\right|. \tag{2.9}$$

We are now prepared to show that the proper choice of steps for the finite-difference discretization makes δ_k attain the minimum value for all $\delta_k^{\alpha,\beta}$.

3. Optimality.

Let us return to the finite-difference PML for the general case given by (1.5) and show that the problem of its optimal finite-difference approximation can be reduced (possibly not uniquely) to the problem considered in the previous section, i.e.,

$$\min_{\alpha,\beta}\delta_k^{\alpha,\beta}=\min_{\beta}\delta_k^{0,\beta}\equiv\min_{h,\hat{h}}\delta_k.$$

Fortunately, the latter is a well-studied problem of Chebyshev rational approximation.

In terms of the new variable $\bar{x}$ defined by (1.5) we obtain the same problem as (2.1) but with the following new coefficients:

$$\gamma = \frac{i\omega}{i\omega\alpha + \beta}, \quad \rho = -\omega^2 \left(\alpha + \frac{\beta}{i\omega}\right) \text{ if } \bar{x} > 0.$$

We use a finite-difference discretization similar to (2.5)–(2.7). Let us denote the discrete counterparts of α and β, respectively, by $\alpha_j, \hat{\alpha}_j$ and $\beta_j, \hat{\beta}_j$. Formally we assume that α_j, β_j and $\hat{\alpha}_j, \hat{\beta}_j$ reside at points x_j and $\hat{x}_j$, respectively. Then the PML finite-difference solution satisfies (2.5)–(2.7) but with $\hat{\gamma}_j = \rho_j = 1$ and the new grid "steps"

$$a_j = h_j \left(\hat{\alpha}_j + \frac{\hat{\beta}_j}{i\omega}\right), \qquad \hat{a}_j = -\omega^2 \hat{h}_j \left(\alpha_j + \frac{\beta_j}{i\omega}\right), \qquad j = 1, \ldots, k. \tag{3.1}$$

The general finite-difference PML solution can be defined similarly to the one from the previous section if we use the new steps (3.1) in (2.5)–(2.7). We denote by $f_k^{\alpha,\beta} = f_k^{\alpha,\beta}(\lambda, \omega)$ the discrete impedance function presented by formula (2.8) with the new steps (3.1). This function $f_k^{\alpha,\beta}$ is still a regular continued fraction but since its coefficients can be complex now, it is generally not an S-fraction anymore. We can equivalently rewrite the PML error (1.6) as

$$\delta_k^{\alpha,\beta} = \max_{\lambda \in [0,1],\ \omega \in [0, \omega_{\max}]} s(\lambda) \left| \sqrt{\lambda} - \frac{1}{i\omega f_k^{\alpha,\beta}(\lambda, \omega)} \right|. \tag{3.2}$$

It is easy to check that when $\alpha = 0$ the dependence on ω of function $f_k^{0,\beta}(\lambda, \omega)$ is such that $i\omega f_k^{0,\beta}(\lambda, \omega) = f_k^{\beta}(\lambda)$. Obviously, for every set β_j, $\hat{\beta}_j$ we can find real steps h_j, $\hat{h}_j$, so that $f_k^{\beta}(\lambda) = f_k(\lambda)$, and hence the error functional (3.2) with $\alpha = 0$ coincides with (2.9), i.e.,

$$\delta_k = \delta_k^{0,\beta}.$$

The following proposition shows that the general complex transformations cannot produce better PML approximations than their purely imaginary counterparts.

PROPOSITION 3.1. *For any α and any $\beta \neq 0$ we have $\delta_k^{\alpha,\beta} \geq \delta_k$.*

Proof. From the definitions it follows that

$$\left[i\omega f_k^{\alpha,\beta}(\lambda, \omega) \right] \Big|_{\omega=0} = i\omega f_k^{0,\beta}(\lambda, \omega) = f_k^{\beta}(\lambda);$$

therefore,

$$\delta_k^{\alpha,\beta} = \max_{\lambda \in [0,1],\ \omega \in [0, \omega_{\max}]} s(\lambda) \left| \sqrt{\lambda} - \frac{1}{i\omega f_k^{\alpha,\beta}(\lambda, \omega)} \right| \geq \max_{\lambda \in [0,1],\ \omega = 0} s(\lambda) \left| \sqrt{\lambda} - \frac{1}{i\omega f_k^{\alpha,\beta}(\lambda, \omega)} \right|$$

$$= \max_{\lambda \in [0,1]} s(\lambda) \left| \sqrt{\lambda} - \frac{1}{i\omega f_k^{0,\beta}(\lambda, \omega)} \right| = \max_{\lambda \in [0,1]} s(\lambda) \left| \sqrt{\lambda} - \frac{1}{f_k(\lambda)} \right| = \delta_k^{0,\beta} = \delta_k. \qquad \square$$

The proposition allows us to reduce minimization of $\delta_k^{\alpha,\beta}$ to minimization of δ_k, which is a well-studied problem of Chebyshev rational optimization. Using the fact that $\sqrt{\lambda}$ is hypernormal on the support $[a, b]$ of s [23], we conclude that the existence and uniqueness theorems for the Chebyshev (optimal) rational approximation, which can be found in [2, Chapter II, Theorems 33 and 34], are applicable. Since the function f_k is an S-fraction, it is a $[(k-1)/k]$ real rational function (though the opposite statement is not true in general). However, we will be looking for the optimal approximation in the form of the $[(k-1)/k]$ real rational function. Such an approximant is irreducible and has exactly $2k+1$ alternating points (set of ordered noncoinciding points λ_j, where the weighted error is equal to $(-1)^j\delta_k$) on $[a, b]$. Hence, the $[(k-1)/k]$ optimal rational approximant must have $2k$ interpolation points on (a, b), so it must be a Markov–Stieltjes function [15]. But any Markov–Stieltjes function can be presented as an S-fraction [6, Theorem 5.1.2, Corollary 2], i.e., as (2.8) with all h_j and $\hat{h}_j$ being positive. The above consideration can be summarized in the following proposition.

Proposition 3.2. *Let the support of s be a segment $[a, b] \subseteq [0, 1]$, and let s be continuous and positive on $[a, b]$. Then there exists the unique minimum of δ_k with the optimal approximant being an S-fraction* (2.8)*.*

The optimal approximant can be uniquely characterized as the one that has exactly $2k+1$ alternation points, all of which are located on $[a, b]$.

From the above proposition we obtain the following corollary.

Corollary 3.3. *The choice of parameters $\alpha_j^{opt} = \hat{\alpha}_j^{opt} = 0$, $\beta_j^{opt} = h_j$, $\hat{\beta}_j^{opt} = \hat{h}_j$, which minimize δ_k, defines a solution that minimizes the functional $\delta_k^{\alpha,\beta}$.*

Remark 3.1. Generally, there is no uniqueness in optimal approximation by complex rational functions even in approximating real functions on real intervals [25, Chapter 5]; that is why we cannot say if our optimal solution is unique. Though, obviously, β^{opt} is unique, so the remaining question is whether there exists an optimal pair $(\alpha^{\text{opt}}, \beta^{\text{opt}})$ with $\alpha^{\text{opt}} \neq 0$.

Remark 3.2. There is an important reason why we chose the L^∞ norm in our "cost" functional, instead of the L_2 norm traditionally used in the literature on PML, e.g., in [10], in addition to the fact that the former provides more reliable bounds. Even real rational approximation problems in L^p, $1 \leq p < \infty$, may have more than one optimal solution [20, section 2.3], so using the L^∞ norm at least provides us the uniqueness for β^{opt}.

Remark 3.3. An interesting question is if a high order or spectral Galerkin PML discretization can perform better than the optimal finite-difference scheme, and the answer is that it cannot. The reason is that a Galerkin (and, generally, any Galerkin–Petrov) process on any k-dimensional subspace also generates a $[(k-1)/k]$ rational impedance [13], so it cannot do better than the optimal $[(k-1)/k]$ approximation.

4. Rational approximation and optimal grids.

Perhaps the simplest way to construct a rational approximation is to compute a Padé approximant satisfying the conditions

$$\frac{d^i}{d\lambda^i}\left[f_k(\lambda) - \sqrt{\lambda}\right]\Big|_{\lambda=1} = 0, \quad i = 0, \ldots, 2k-1,$$

in which case the PML becomes algebraically equivalent to the ABC by Engquist and Majda [14]. However, such an approximation is not efficient for small λ. Better results can be achieved using a more general continued fraction ABC based on multipoint Padé approximants [16], but the double root interpolation used there does not produce the alternation of the error and thus cannot arrive at the optimal approximation.

Real optimal rational approximations in rare cases can be obtained in a closed form, and in most cases they are obtained numerically, but in all cases the algorithms are based on the alternation property which was stated in Proposition 3.2.

4.1. Zolotarjov's approximation. Consider first the case $\lambda_{\min} > 0$. The interval $[\lambda_{\min}, 1]$ can be linearly shifted onto $[1, 1/\lambda_{\min}] = [1, 1/\kappa'^2]$ with $\kappa' = \sqrt{\lambda_{\min}}$. Let $\kappa = \sqrt{1-\kappa'^2}$. Zolotarjov found a $[(k-1)/k]$ rational function $\tilde{r}$ such that

$$\left\| 1 - \sqrt{\lambda}\tilde{r}(\lambda) \right\|_{C[1,1/\kappa'^2]} = \inf_r \left\{ \left\| 1 - \sqrt{\lambda} r(\lambda) \right\|_{C[1,1/\kappa'^2]} \right\}.$$

When the error is small, this optimization problem is close to (1.6)–(1.7), in the sense that if the error of one of these approximation is equal to ε, then the error of the other is $\varepsilon + O(\varepsilon^2)$.

THEOREM 4.1 (see Zolotarjov, 1887 [20]). *The best approximant is given by*

$$\tilde{r}(\lambda) = D \frac{\prod_{l=1}^{k-1}(\lambda + c_{2l})}{\prod_{l=1}^{k}(\lambda + c_{2l-1})}, \tag{4.1}$$

where

$$c_l = \frac{\mathrm{sn}^2\,(lK/(2k);\kappa)}{\mathrm{cn}^2\,(lK/(2k);\kappa)}, \qquad l = 1, \ldots, 2k-1,$$

$K = K(\kappa)$ *is the complete elliptic integral, and the number* D *is uniquely determined by the condition*

$$\max_{C[1,1/\kappa'^2]} \left[1 - \sqrt{\lambda}\tilde{r}(\lambda) \right] = - \min_{C[1,1/\kappa'^2]} \left[1 - \sqrt{\lambda}\tilde{r}(\lambda) \right].$$

Remark 4.1. If $\lambda_{\min}$ is small, then κ is close to 1. Standard subroutines for computing elliptic functions fail at the very beginning, because they accept κ and compute $\kappa' = \sqrt{1-\kappa^2}$, losing significant digits. We recommend computing the elliptic functions by the arithmetic-geometric mean method [1, Chapters 16 and 17] in terms of $\kappa' = \sqrt{\lambda_{\min}}$.

Recall that the asymptotic convergence factor is given by formula (1.9).

In Figure 1 one can see the Chebyshev alternation of the error via λ on the optimal interval for a Zolotarjov's grid. The L^∞ norm of the error for $\lambda_{\min} = 10^{-3}$ is given in Figure 2; it is in good agreement with the estimate (1.9).

4.2. Approximation on [0, 1]. In the case $\lambda_{\min} = 0$ we numerically solved the optimization problem (1.6), (1.8). Here we followed the solution of this problem given in [24] with some minor modifications.

For small values of k we used the unconditionally converging differential correction method [20, section 2.5], the convergence of which, however, deteriorates rapidly with the increase of k. To solve the associated convex nonsmooth optimization problem, we exploited the Fortran 90 package SOLVOPT by A. Kuntsevich and F. Kappel (see http://bedvgm.kfunigraz.ac.at:8001/alex/solvopt).

For larger k the Remez method [20, section 2.5] was exploited. This algorithm converges quadratically, provided a good initial guess is given, but otherwise it may diverge. Therefore, we implemented an extrapolation procedure in the spirit of [24] to obtain good initial iterants. The Fortran 90 multiple precision package [5] was incorporated into our Fortran program realizing the Remez method.

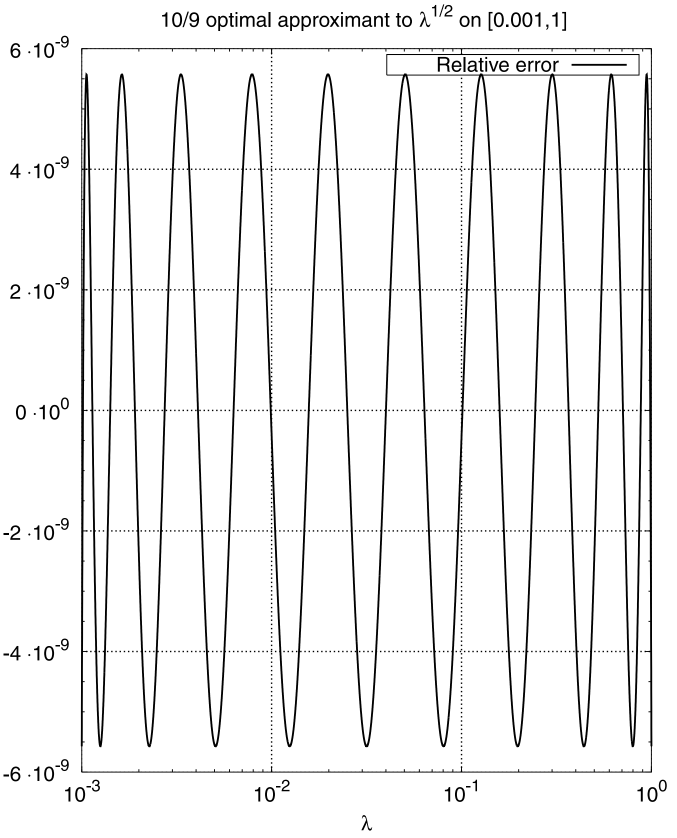


FIG. 1. *Impedance error* $\sqrt{\lambda}\tilde{r}(\lambda) - 1$ *of Zolotarjov's grid for* $\lambda_{\min} = 10^{-3}$ *and* $k = 10$ *as a function of* λ.

An error estimate is presented by formula (1.10); it follows from [21]. The actual distribution of the error for $k = 10$ is plotted in Figure 3; similarly to the Zolotarjov's error it exhibits Chebyshev alternating properties according to Proposition 3.2. The error graph in Figure 4 is in very good agreement with the error estimate (1.10).

4.3. Computation of grid steps. The optimal rational approximations obtained above can be represented in terms of poles and residues as

$$f_k(\lambda) = \sum_{i=1}^{k} \frac{y_i}{\lambda - \theta_i},$$

but to construct the finite-difference scheme we need them in the form of an S-fraction (2.8). A recursive algorithm of computing h_i and $\hat{h}_i$ from y_i and θ_i based on the Lanczos method is given in [11, subsection 3.1]. In Figure 5 we show grids for Zolotarjov approximation on $[0.001, 1]$- and $[0, 1]$-optimal approximations. Qualitatively, both the grids behave similarly to other optimal or Gaussian grids described in [11, 12, 3, 18, 13]; i.e., they exhibit gradual refinement towards the origin and alternation of primary and dual nodes. Specifically, both grids are close to geometric progressions, which corresponds well to the asymptotic property of optimal grids on spectral intervals with large condition numbers, discovered in [18]. Since the $[0, 1]$-

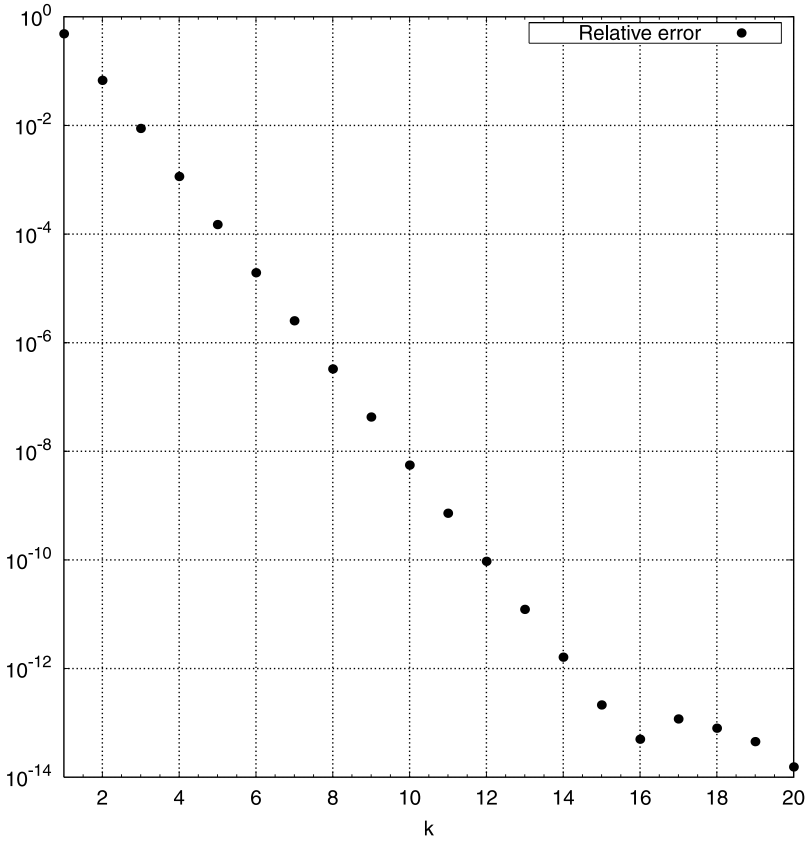


FIG. 2. *The computed impedance error of Zolotarjov's grids* $\max_{[10^{-3},1]}|\sqrt{\lambda}\tilde{r}(\lambda)-1|$.

grid is designed to absorb at all incidence angles, it requires thicker PML than the Zolotarjov grid.

4.4. Comparison with standard PML. The first advantage of the new method of implementation of PML over the standard implementation is that the errors of the newly computed meshes depend only on the incidence angle $\theta = \arccos(\sqrt{\lambda})$ and not on the frequency. The practical importance of this fact is that the user can fix the mesh only once, depending on the requirements of the accuracy and the spectral range in question. Of course, the reflection coefficient of such a mesh will also depend on the discretization of the interior domain, and thus the total error of the ABC will not be better than the error of the interior discretization.

The second important property is that these meshes give an optimal or near-optimal error for a chosen spectral interval, and no ad hoc or numerical optimization of parameters is necessary after the mesh is defined.

Previously, in the literature devoted to PML absorbing conditions, near-optimal solutions for the parameters were obtained by various numerical optimization schemes [10, 8]. Comparing the magnitudes of these errors to the errors of the new optimal PML, we conclude that the new scheme produces a much smaller error. From Figure 2 we see that a 5 point Zolotarjov's mesh gives an error of approximately 10^{-4} for all $\lambda \in [10^{-3}, 1]$, which corresponds to the incident angles $\theta \in [0, 88°]$, and a 10 point

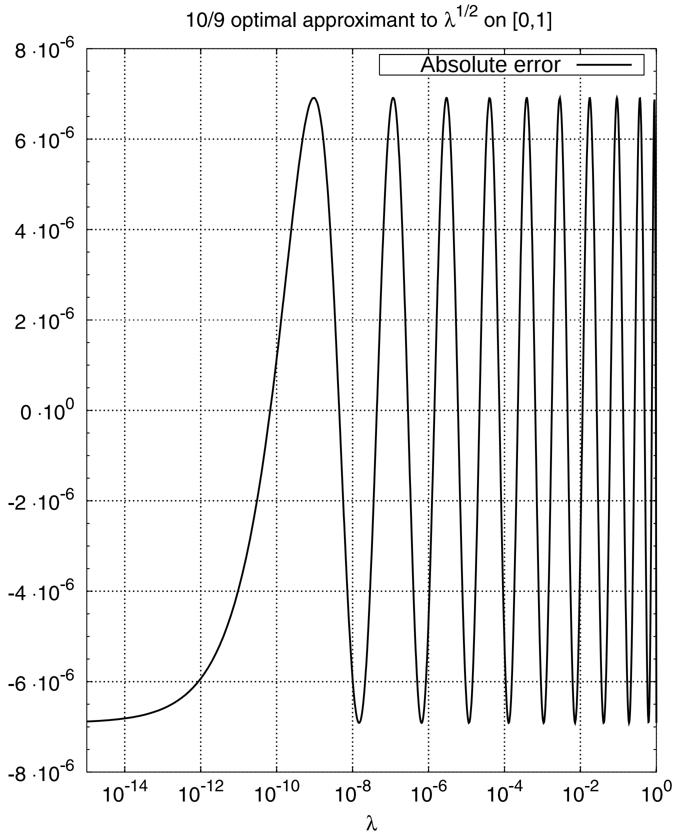


FIG. 3. *The impedance error* $\sqrt{\lambda} - \tilde{r}(\lambda)$ *of optimal grid for the segment* $[0,1]$ *and* $k = 10$ *as a function of* λ*.*

mesh gives an error of 10^{-8} for the same angles, all independent of the frequency. For comparison, the PML discretizations in [10] for approximately the same range of the angles give an error that varies (depending on the frequency) between 10^{-2} and 10^{-1} (5 point mesh), and between 10^{-3} and 10^{-2} (10 point mesh).

Among the disadvantages of the proposed method we see the nontrivial discretization of the time-domain problem.

5. Time-domain realization. The S-fraction representation of the impedance function f_k provides stability of the time-domain solution. The poles of S-fractions are real negative and their residues are real positive, which is sufficient for the absence of exponentially growing modes [22].

Unfortunately, our modification of the equation in the absorbing region makes impossible the simple variable-splitting time-domain realization that is commonly used in the PML methods. Moreover, the equation in the absorbing region becomes noncausal, and hence the discretized system has to be implicit.

We will first consider the discretization of our two-dimensional wave equation with the absorbing region in the half-space $x > 0$; the case with absorbing layers on the four edges of a square differs only by the treatment of the corners; i.e., the subdomains in which the imaginary stretching is applied to both coordinates. The treatment of the two-dimensional domains will be briefly discussed in the end of the

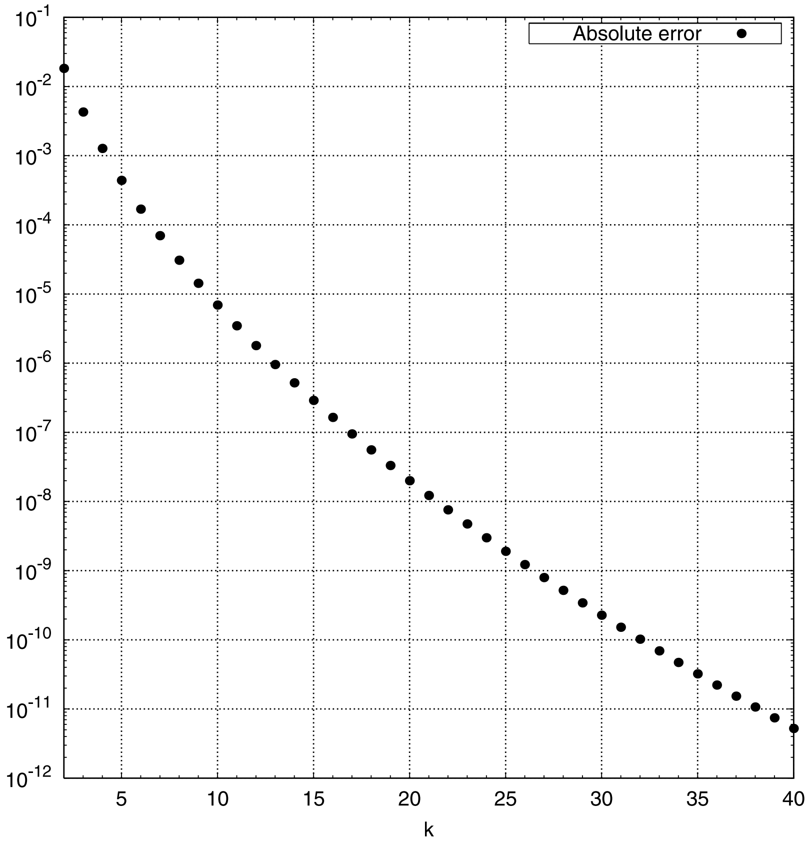


FIG. 4. *The impedance error* $\max_{[0,1]} |\sqrt{\lambda} - \tilde{r}(\lambda)|$ *for different* k.

section.

Let us first Fourier transform (2.1)–(2.2) back to the two-dimensional wave problem while again keeping the same name for the function:

$$u_{xx} + u_{yy} - u_{tt} = 0 \quad \text{if } x < 0, \qquad u_{xxtt} + u_{yy} - u_{tt} = 0 \quad \text{otherwise,}$$

with the interface conjugation conditions

$$u(0-, y, t) = u(0+, y, t), \qquad u(0-, y, t)_x = u(0+, y, t)_{xt}$$

and the boundary condition $u(+\infty, y, t) = 0$. We will demonstrate the discretization in time and in the x-direction, while leaving the problem continuous in the y-direction.

For $x < 0$ we write the discretization of the equations in the standard way:

$$\frac{d}{dt} u_j^- = \frac{V_j^- - V_{j-1}^-}{\hat{h}_j^-} + W_{j,y}^-, \qquad j = 1, \ldots, k, \tag{5.1}$$

$$\frac{d}{dt} V_j^- = \frac{u_{j+1}^- - u_j^-}{h_j^-}, \qquad u_{k+1}^- = 0, \qquad j = 1, \ldots, k, \tag{5.2}$$

$$\frac{d}{dt} W_j^- = u_{j,y}^-, \qquad j = 1, \ldots, k. \tag{5.3}$$

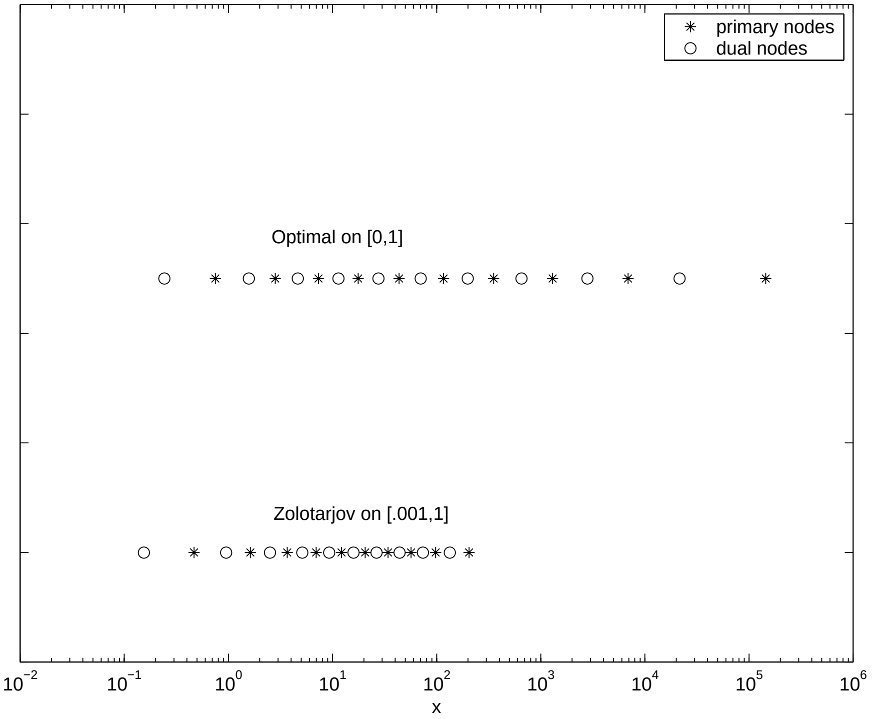


FIG. 5. *Zolotarjov and* $[0, 1]$*-optimal grids,* $k = 10$.

Here the unknowns are numbered from right to left with u_1^- and V_0^- being placed at the boundary of the PML region. We note that the right-to-left numbering of the unknowns implies that the steps h_j^-, $\hat{h}_j^-$ are negative.

The only change in the PML region, i.e., for $x > 0$, is that the steps h_j^+ and $\hat{h}_j^+$ get divided by $i\omega$, which leads to the following system in time domain:

$$\frac{d}{dt}u_j^+ = \frac{d}{dt}\frac{V_j^+ - V_{j-1}^+}{\hat{h}_j^+} + W_{j,y}^+, \qquad j = 1, \ldots, k, \tag{5.4}$$

$$\frac{d}{dt}V_j^+ = \frac{d}{dt}\frac{u_{j+1}^+ - u_j^+}{h_j^+}, \qquad u_{k+1}^+ = 0, \qquad j = 1, \ldots, k, \tag{5.5}$$

$$\frac{d}{dt}W_j^+ = u_{j,y}^+, \qquad j = 1, \ldots, k. \tag{5.6}$$

Here the numbering of the unknowns is left to right; hence the steps h_j^+, $\hat{h}_j^+$ are positive.

It is clear that to get the total number of equations equal to the total number of unknowns we need two extra equations; these are

$$u_1^- = u_1^+, \qquad V_0^- = V_0^+. \tag{5.7}$$

The time discretization follows naturally: we place variables u_j^- at time levels $n\Delta t$ and variables V_j^- and W_j^- at time levels $(n + 1/2)\Delta t$, where $\Delta t > 0$ is a time step. The system (5.1)–(5.3) is thus solved in the standard explicit "leap-frog" fashion.

For $x > 0$, however, we have to place variables u_j^+ and V_j^+ at the same time levels $n\Delta t$, and only the variables W_j^+ at time levels $(n + 1/2)\Delta t$; thus (5.4), (5.5), and the second equation in (5.7) will compose the system of equations that needs to be solved at every time step. With the use of the notation

$$[D_t f]^{m+1/2} = \frac{f^{m+1} - f^m}{\Delta t}, \qquad [A_t f]^{m+1/2} = \frac{f^{m+1} + f^m}{2},$$

where one can put $m = n$ or $m = n + 1/2$, we rewrite the scheme (5.1)–(5.7) as

$$\begin{aligned}
&[D_t u_1]^{n+1/2} = \frac{V_1^{-,n+1/2} - [A_t V_0^+]^{n+1/2}}{\hat{h}_1^-} + W_{1,y}^{-,n+1/2}, && \\
&[D_t u_j^-]^{n+1/2} = \frac{V_j^{-,n+1/2} - V_{j-1}^{-,n+1/2}}{\hat{h}_j^-} + W_{j,y}^{-,n+1/2}, && j = 2, \dots, k, \\
&[D_t V_j^-]^n = \frac{u_{j+1}^{-,n} - u_j^{-,n}}{h_j^-}, && j = 1, \dots, k, \\
&[D_t W_j^-]^n = u_{j,y}^{-,n}, && j = 1, \dots, k,
\end{aligned}$$

for $x \le 0$, and

$$\begin{aligned}
&[D_t u_j^+]^{n+1/2} = \frac{[D_t V_j^+]^{n+1/2} - [D_t V_{j-1}^+]^{n+1/2}}{\hat{h}_j^+} + W_{j,y}^{+,n+1/2}, && j = 1, \dots, k, \\
&[D_t V_j^+]^{n+1/2} = \frac{[D_t u_{j+1}^+]^{n+1/2} - [D_t u_j^+]^{n+1/2}}{h_j^-}, && j = 1, \dots, k, \\
&[D_t W_j^+]^n = u_{j,y}^{+,n}, && j = 1, \dots, k,
\end{aligned}$$

for $x > 0$. We note that variables V_j^+ with $j > 0$ can be excluded from the scheme analytically by substituting the second of the three equations above into the first. The time marching follows in an obvious fashion.

The discretization of the two-dimensional model is performed as a tensor product of the two one-dimensional discretizations. It can be seen that in the corners of a rectangular domain in this case the equation is independent of time. In fact, it turns out that the Helmholtz equation $u - u_{xx} - u_{yy} = 0$ is solved there, with the time dependent boundary conditions. For such an equation on a given mesh one can compute the Neumann-to-Dirichlet map on the boundary and include only the variables defined on the boundary in the total scheme.

It becomes clear that the computational cost of the new boundary condition amounts to solving a linear system with a tridiagonal matrix of dimension k for each point on the interface at every time step. This cost will be linear in k with the constant close to the one for the implementation of standard Berenger's PML with variable splitting. The computation of the corner regions for a two-dimensional problem requires computing at each time step a partial Neumann-to-Dirichlet map on a $k \times k$ square, which again is an operation proportional to the total number of nodes in the absorbing region.

6. Numerical experiments. We performed a series of experiments for the following elongated model: the source and the receiver, positioned on the same vertical

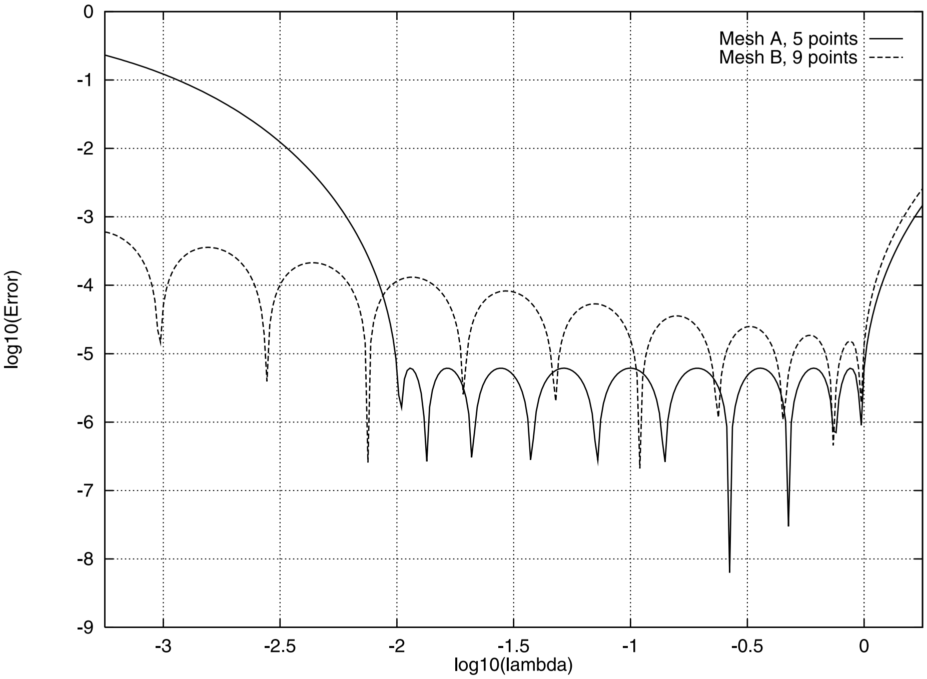


FIG. 6. *The relative error for meshes* A *and* B *of Experiment* 1.

line, are separated by 10 wavelengths with the new PML boundary conditions positioned at two wavelengths on both sides of that line, as well as below the source and above the receiver. For such an elongation the average incidence angle is $\theta = \arctan 0.2$ and the maximum incidence angle is $\theta = \arctan 0.1$; hence the minimum value of λ is approximately 10^{-2}.

We take the source wavelet as the second derivative of a Blackman–Harris pulse [17] with the maximum frequency $F_{\max} = 1$. The signal recorded at the receiver was compared with the signal recorded in the "infinite" discrete experiment, i.e., the one in which the discrete computational domain is big enough so that any reflections from the boundary do not get recorded within the assumed time frame. The error of the two signals was calculated as $L^\infty(t)$, relative to the signal of this larger computation. As this error obviously depends on the discretization of the propagative part of the computational domain, we performed the experiments using 8, 16, and 32 points per wavelength in the interior part.

The experiment was carried out using two different absorbing meshes: mesh A was taken as Zolotarjov's optimized for the interval $\lambda \in [10^{-2}, 1]$ with the total of five points, and mesh B was taken to be optimal on $\lambda \in [0, 1]$ with the total of nine points. The first mesh was chosen to take advantage of the a priori knowledge of the range of λ; the second was taken to be of approximately the same error but assuming no such knowledge. As we see, a priori knowledge of the range of incidence angles lets one decrease the amount of work required for absorption almost twice for the given tolerance.

The relative errors (i.e., with $s(\lambda)$ defined as in (1.7)—notice that mesh B was designed to provide minimum error in a different norm!) for these meshes are presented in Figure 6, and we see that these two meshes have approximately the same order of

Table 1
The $L^\infty(t)$ relative error in percent for Experiment 1.

	Mesh A, $k = 5$			Mesh B, $k = 9$		
Frequency	8 ppw.	16 ppw.	32 ppw.	8 ppw.	16 ppw.	32 ppw.
$F_{\max} = 1$	1.286	0.712	0.710	1.618	0.947	0.906
$F_{\max} = 2$	1.191	0.641	0.387	1.258	0.605	0.460
$F_{\max} = 4$	1.129	0.591	0.335	1.142	0.563	0.333

approximation on the spectral interval of interest. It is thus not surprising that the $L^\infty(t)$ errors that we obtain are similar too, as seen in the line $F_{\max} = 1$ of Table 1.

It is clear that the error is partly due to the reflection of the evanescent modes from the boundary between the interior propagative part of the domain and the absorbing part. To consider the importance of evanescent reflections, we increase the frequency of the source, thus keeping the incidence angles of the experiment unchanged, while effectively increasing the distance from the absorbing boundary to the receiver in terms of the wavelengths, and decreasing the effect of the evanescent reflections. The errors resulting from this change (with frequencies $F_{\max} = 2$ and $F_{\max} = 4$) are shown on the corresponding lines of Table 1. We note that with this change the boundary conditions become more effective for all values of the discretization of the interior for both meshes, but the effect of the decrease of the evanescent modes is especially notable for fine discretization of 32 points per wavelength, at which the boundary conditions approach the saturation error.

For the ABCs to be most effective, they need to be able to absorb not only the propagative modes of the solution but also the evanescent modes. Our proposed PML ABCs are optimal in terms of the absorption of the propagative part; however, the evanescent modes were left untreated. The obvious solution to this problem is to place this boundary condition at a distance to the closest receiver; this will ensure that the evanescent modes of the solution get absorbed. However, in the problems in which the receiver stays closer to the boundary conditions than any scatterer, the necessity of placing the boundary conditions away from the receiver clearly decreases the effectiveness of such conditions. This problem can be cured by using the regular nonabsorbing optimal meshes [3, 11, 12, etc.] in the layers between the receiver and the boundary conditions. This approach allows one to solve the problem of absorbing both the evanescent and propagative modes in an optimal way.

It can be seen from the results of these experiments that in the cases in which the spectral interval is known a priori (as above) it is reasonable to use Zolotarjov's meshes, specifically designed for such an interval. However, when such a mesh is not available, or in the case where the spectral interval is not known in advance (such as in models with multiple reflections), we expect that it will be advantageous to use the universal optimal mesh of the same type as "B" above, which satisfies the accuracy requirements.

7. Conclusions. We have shown that the optimal choice of the attenuation parameters for Berenger PML ABCs can be achieved in the limiting case of the purely imaginary coordinate stretching approach. Indeed, it was questioned by one of the reviewers if the proposed approach can and should be included in the class of PML boundary conditions, because, unlike those conditions, it leads to a diffusion-type equation in the absorbing region. However, the situation when a minimum is attained on the boundary of the closure of a set is often encountered in optimization theory, and thus our new ABC is a part of the closure of the classical PML set.

For this limiting case, the problem becomes frequency independent and can be viewed as a classical Chebyshev rational approximation of the square root function, and the parameters that need to be chosen are the steps of the grid of the resulting finite-difference scheme. We consider two different error functions and show how to obtain the optimal grid steps for them. The resulting scheme exhibits attenuative properties superior to those of the classical PML known to the authors; the goal attenuation of 1% can be achieved using as few as five points in the absorbing region. The drawback of the resulting scheme is that, unlike the classic PML, it needs to be implemented in a fashion that is different in the absorbing and the propagating regions. The details of the discretization of a two-dimensional scalar wave problem are discussed and the results of numerical experiments are presented.

Acknowledgments. The authors are grateful to R. Varga for useful discussions and to D. Bailey (NASA Ames Research Center) for the publication of package [5] on the Internet.

REFERENCES

[1] M. Abramowitz and J. Stegan, *Handbook of Mathematical Functions*, Applied Math. 55, National Bureau of Standards, Washington, D.C., 1964.

[2] N. Akhiezer, *Theory of Approximation*, Dover, New York, 1992.

[3] S. Asvadurov, V. Druskin, and L. Knizhnerman, *Application of the difference Gaussian rules to solution of hyperbolic problems*, J. Comput. Phys., 158 (2000), pp. 116–135.

[4] S. Asvadurov, V. Druskin, and L. Knizhnerman, *Applications of the difference Gaussian rules to solutions of hyperbolic problems:* II. *Global expansion*, J. Comput. Phys., 175 (2002), pp. 24–49.

[5] D. Bailey, *A Fortran*-90 *Based Multiprecision System*, Technical report RNR-94-013, RNR, NASA Ames Research Center, Moffett Field, CA, 1994.

[6] G. Baker and P. Graves-Morris, *Padé Approximants*, Addison-Wesley, London, UK, 1996.

[7] J.-P. Berenger, *Perfectly matched layer for the absorption of electromagnetic waves*, J. Comput. Phys., 114 (1994), pp. 185–200.

[8] W. Chew, J. Jin, and E. H. Michielssen, *Complex Coordinate Stretching as a Generalized Absorbing Boundary Condition*, Technical report CCEM-20-96, Center for Computational Electromagnetics, University of Illinois at Urbana-Champaign, Urbana, IL, 1996.

[9] W. Chew and B. Weedon, *A* 3*D perfectly matched medium from modified Maxwell's equations with stretched coordinates*, Microwave Opt. Technol. Lett., 7 (1994), pp. 599–604.

[10] F. Collino and P. Monk, *Optimizing the perfectly matched layer*, Comput. Methods Appl. Mech. Engrg., 164 (1998), pp. 157–171.

[11] V. Druskin and L. Knizhnerman, *Gaussian spectral rules for the three-point second differences:* I. *A two-point positive definite problem in a semi-infinite domain*, SIAM J. Numer. Anal., 37 (1999), pp. 403–422.

[12] V. Druskin and L. Knizhnerman, *Gaussian spectral rules for second order finite-differences*, Numer. Algorithms, 25 (2000), pp. 139–159.

[13] V. Druskin and S. Moskow, *Three-point finite difference schemes, Padé and the spectral Galerkin method:* I. *One-sided impedance approximation*, Math. Comp., 71 (2002), pp. 995–1019.

[14] B. Engquist and A. Majda, *Radiation boundary conditions for acoustic and elastic wave calculations*, Comm. Pure Appl. Math., 32 (1979), pp. 313–357.

[15] A. Gonchar and G. Lopez, *On Markov's theorem for multipoint Padé approximants*, Math. USSR-Sb., 34 (1978), pp. 449–459.

[16] M. Guddati and J. Tassoulas, *Continued-fraction absorbing boundary conditions for the wave equation*, J. Comput. Acoust., 8 (2000), pp. 139–156.

[17] F. Harris, *On the use of windows for harmonic analysis with the discrete Fourier transform*, Proc. IEEE, 66 (1978), pp. 51–83.

[18] D. Ingerman, V. Druskin, and L. Knizhnerman, *Optimal finite-difference grids and rational approximations of square root:* I. *Elliptic problems*, Comm. Pure Appl. Math., 53 (2000), pp. 1039–1066.

[19] P. G. Petropoulos, *Reflectionless sponge layers as absorbing boundary conditions for the*

numerical solution of Maxwell equations in rectangular, cylindrical, and spherical coordinates, SIAM J. Appl. Math, 60 (2000), pp. 1037–1058.

[20] P. Petrushev and V. Popov, *Rational Approximation of Real Functions*, Encyclopedia Math. Appl. 28, Cambridge University Press, Cambridge, UK, 1987.

[21] H. Stahl, *Best uniform rational approximation of x^α on* $[0, 1]$, Bull. Amer. Math. Soc. (N.S.), 28 (1993), pp. 116–122.

[22] L. Trefethen and L. Halpern, *Well-posedness of absorbing boundary conditions and one-way wave equations*, Math. Comp., 47 (1986), pp. 421–435.

[23] R. Varga, *private communication*, 2000.

[24] R. Varga, A. Ruttan, and A. Carpenter, *Numerical results on best uniform rational approximation of $|x|$ on* $[-1, +1]$, Mat. Sb., 182 (1991), pp. 271–290.

[25] R. S. Varga, *Scientific Computations on Mathematical Problems and Conjectures*, CBMS-NSF Regional Conf. Ser. in Appl. Math. 60, SIAM, Philadelphia, PA, 1990.

SIAM J. NUMER. ANAL.
Vol. 41, No. 1, pp. 306–324

POINCARÉ–FRIEDRICHS INEQUALITIES FOR PIECEWISE H^1 FUNCTIONS[*]

SUSANNE C. BRENNER[†]

Abstract. Poincaré–Friedrichs inequalities for piecewise H^1 functions are established. They can be applied to classical nonconforming finite element methods, mortar methods, and discontinuous Galerkin methods.

Key words. Poincaré–Friedrichs inequalities, piecewise H^1 functions, nonconforming finite elements, mortar methods, discontinuous Galerkin methods

AMS subject classifications. 65N30, 46E35

PII. S0036142902401311

1. Introduction.

Let Ω be a connected open polyhedral domain in $\mathbb{R}^d$ ($d = 2, 3$), and let $H^1(\Omega)$ be the Sobolev space of (real-valued) functions in $L_2(\Omega)$ whose first order distributional derivatives also belong to $L_2(\Omega)$. The Poincaré–Friedrichs inequalities (cf. [23, 29]) state that

$$\|\zeta\|_{L_2(\Omega)} \le C\left(|\zeta|_{H^1(\Omega)} + \left|\int_\Gamma \zeta\, ds\right|\right) \qquad \forall\, \zeta \in H^1(\Omega), \tag{1.1}$$

$$\|\zeta\|_{L_2(\Omega)} \le C\left(|\zeta|_{H^1(\Omega)} + \left|\int_\Omega \zeta\, dx\right|\right) \qquad \forall\, \zeta \in H^1(\Omega), \tag{1.2}$$

where

$$|\zeta|_{H^1(\Omega)} = \left(\int_\Omega |\nabla\zeta|^2\, dx\right)^{1/2},$$

Γ is a measurable subset of $\partial\Omega$ with a positive $(d-1)$-dimensional measure, and ds is the infinitesimal $(d-1)$-dimensional volume.

In this paper we establish analogues of (1.1) and (1.2) for piecewise H^1 functions with respect to a partition $\mathcal{P}$ of Ω by open polygons ($d = 2$) or polyhedra ($d = 3$), which is not necessarily a triangulation of Ω. In other words, we assume only that

$$D \cap D' = \emptyset \text{ if } D \text{ and } D' \text{ are distinct members of } \mathcal{P}, \text{ and } \overline{\Omega} = \bigcup_{D\in\mathcal{P}} \overline{D}.$$

Typical two- and three-dimensional examples of partitions are depicted in Figure 1.1, where the square is partitioned into seven subdomains and the cube is partitioned into five subdomains. The space $H^1(\Omega, \mathcal{P})$ of piecewise H^1 functions is defined by

$$H^1(\Omega, \mathcal{P}) = \{\zeta \in L_2(\Omega) : \zeta_D = \zeta\big|_D \in H^1(D) \quad \forall\, D \in \mathcal{P}\},$$

[*] Received by the editors January 22, 2002; accepted for publication (in revised form) August 29, 2002; published electronically March 26, 2003. This work was supported in part by the National Science Foundation under grant DMS-00-74246.

http://www.siam.org/journals/sinum/41-1/40131.html

[†] Department of Mathematics, University of South Carolina, Columbia, SC 29208 (brenner@math.sc.edu).

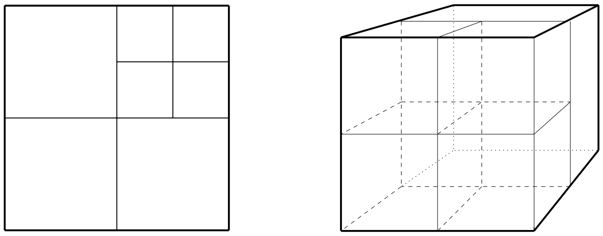


FIG. 1.1. *Examples of general partitions.*

and the seminorm $|\cdot|_{H^1(\Omega,\mathcal{P})}$ is defined by

$$|\zeta|_{H^1(\Omega,\mathcal{P})} = \left(\sum_{D\in\mathcal{P}} \int_D |\nabla\zeta|^2\,dx\right)^{1/2}.$$

We will denote by $S(\mathcal{P},\Omega)$ the set of all the (open) sides (i.e., edges ($d=2$) or faces ($d=3$)) common to two subdomains in $\mathcal{P}$. For example, there are ten such edges in the two-dimensional example in Figure 1.1 and eight such faces in the three-dimensional example. (The precise definition of $S(\mathcal{P},\Omega)$ will be given in sections 6 and 7.)

The following are analogues of the Poincaré–Friedrichs inequalities for $\zeta \in H^1(\Omega,\mathcal{P})$:

$$\|\zeta\|^2_{L_2(\Omega)} \le C\left[|\zeta|^2_{H^1(\Omega,\mathcal{P})} + \sum_{\sigma\in S(\mathcal{P},\Omega)} |\sigma|^{d/(1-d)}\left(\int_\sigma [\zeta]\,ds\right)^2 + \left(\int_\Gamma \zeta\,ds\right)^2\right], \tag{1.3}$$

$$\|\zeta\|^2_{L_2(\Omega)} \le C\left[|\zeta|^2_{H^1(\Omega,\mathcal{P})} + \sum_{\sigma\in S(\mathcal{P},\Omega)} |\sigma|^{d/(1-d)}\left(\int_\sigma [\zeta]\,ds\right)^2 + \left(\int_\Omega \zeta\,dx\right)^2\right], \tag{1.4}$$

where $|\sigma|$ is the $(d-1)$-dimensional volume of the side σ, $[\zeta]$ denotes the jump of the function ζ across a side, and the positive constant C depends only on the shape regularity of the partition $\mathcal{P}$. In particular, these inequalities are valid for partitions that are not quasi-uniform. (More details on the shape regularity assumptions are given in sections 6 and 7.)

Remark 1.1. The Poincaré–Friedrichs inequalities (1.3) and (1.4) can also be written in the following equivalent forms:

$$\|\zeta\|^2_{L_2(\Omega)} \le C\left[|\zeta|^2_{H^1(\Omega,\mathcal{P})} + \sum_{\sigma\in S(\mathcal{P},\Omega)} (\operatorname{diam}\sigma)^{-1}\|\pi_{0,\sigma}[\zeta]\|^2_{L_2(\sigma)} + \left(\int_\Gamma \zeta\,ds\right)^2\right],$$

$$\|\zeta\|^2_{L_2(\Omega)} \le C\left[|\zeta|^2_{H^1(\Omega,\mathcal{P})} + \sum_{\sigma\in S(\mathcal{P},\Omega)} (\operatorname{diam}\sigma)^{-1}\|\pi_{0,\sigma}[\zeta]\|^2_{L_2(\sigma)} + \left(\int_\Omega \zeta\,dx\right)^2\right],$$

where $\pi_{0,\sigma}$ is the orthogonal projection operator from $L_2(\sigma)$ onto $P_0(\sigma)$, the space of constant functions on σ.

The inequalities (1.3) and (1.4) imply that

$$\|\zeta\|_{L_2(\Omega)} \le C\left(|\zeta|_{H^1(\Omega,\mathcal{P})} + \left|\int_\Gamma \zeta\, ds\right|\right), \tag{1.5}$$

$$\|\zeta\|_{L_2(\Omega)} \le C\left(|\zeta|_{H^1(\Omega,\mathcal{P})} + \left|\int_\Omega \zeta\, dx\right|\right), \tag{1.6}$$

provided

$$\int_\sigma [\zeta]\, ds = 0 \qquad \forall\, \sigma \in S(\mathcal{P},\Omega)\,. \tag{1.7}$$

Thus we immediately obtain (1.5) and (1.6) for ζ belonging to many classical nonconforming finite element spaces [15, 19, 21, 18, 25, 17, 12, 11] where $\mathcal{P}$ is a triangulation of Ω, and for ζ belonging to mortar element spaces [5, 6, 26, 4, 30, 20] for a general partition $\mathcal{P}$.

The estimates (1.3) and (1.4) also imply that

$$\|\zeta\|^2_{L_2(\Omega)} \le C\left(|\zeta|^2_{H^1(\Omega,\mathcal{P})} + \sum_{\sigma\in S(\mathcal{P},\Omega)} |\sigma|^{1/(1-d)} \int_\sigma [\zeta]^2\, ds + \int_\Gamma |\zeta|^2\, ds\right), \tag{1.8}$$

$$\|\zeta\|^2_{L_2(\Omega)} \le C\left(|\zeta|^2_{H^1(\Omega,\mathcal{P})} + \sum_{\sigma\in S(\mathcal{P},\Omega)} |\sigma|^{1/(1-d)} \int_\sigma [\zeta]^2\, ds + \left(\int_\Omega \zeta\, dx\right)^2\right) \tag{1.9}$$

for all $\zeta \in H^1(\Omega,\mathcal{P})$. Such inequalities are useful for the analysis of discontinuous Galerkin methods (cf. [14, 24, 3] and the references therein). It should be pointed out that for $d = 2$ and $\Gamma = \partial\Omega$ the inequality (1.8) is slightly stronger than an earlier one obtained in [2], which has the form

$$\|\zeta\|^2_{L_2(\Omega)} \le C\left(|\zeta|^2_{H^1(\Omega,\mathcal{P})} + \sum_{\sigma\in S(\mathcal{P},\Omega)} \frac{1}{|\sigma|} \int_\sigma [\zeta]^2\, ds + \sum_{\sigma\subseteq\partial\Omega} \frac{1}{|\sigma|} \int_\sigma |\zeta|^2\, ds\right), \tag{1.10}$$

where the integrals on the edges along $\partial\Omega$ are also penalized.

Note that the classical inequalities (1.1) and (1.2) are simple consequences of the compactness of the embedding of $H^1(\Omega)$ in $L_2(\Omega)$ (cf. [23, 29]), while the proof of (1.10) in [2] relies on the much deeper regularity theory for the Laplace operator on nonsmooth domains. In contrast, the approach in this paper is based on the classical Poincaré–Friedrichs inequalities and therefore the elementary nature of (1.3) and (1.4) (and hence of (1.8) and (1.9)) is restored.

The rest of the paper is organized as follows. We will first establish (1.3) and (1.4) for the simpler case where $\mathcal{P}$ is a simplicial triangulation of Ω. The key idea is to bridge the classical estimates and the discontinuous estimates through the construction of a nonconforming P_1 interpolant and through the connections between nonconforming P_1 finite elements (in two and three dimensions) and their conforming relatives. These are carried out in sections 2 and 3. Poincaré–Friedrichs inequalities for nonconforming P_1 finite element functions are then derived in section 4, followed by Poincaré–Friedrichs inequalities for piecewise H^1 functions with respect to a simplicial triangulation of Ω in section 5. Finally, Poincaré–Friedrichs inequalities for piecewise H^1 functions with respect to general partitions are given in sections 6 and 7 for two- and three-dimensional domains, and some concluding remarks are given in section 8.

2. A nonconforming P_1 interpolant. In this and the next three sections we restrict our attention to the case where the partition is actually a simplicial triangulation $\mathcal{T}$ of Ω consisting of triangles ($d = 2$) or tetrahedra ($d = 3$); i.e., the intersection of the closures of two members of $\mathcal{T}$ is either empty, a vertex, a closed edge, or a closed face. In this case $S(\mathcal{T}, \Omega)$ is just the set of interior open edges ($d = 2$) or open faces ($d = 3$). We will also denote the set of boundary edges or faces by $S(\mathcal{T}, \partial\Omega)$ and the minimum angle of the triangles or tetrahedra in $\mathcal{T}$ by $\theta_{\mathcal{T}}$.

To avoid the proliferation of constants, we henceforth use the notation $A \lesssim B$ to represent the statement $A \le \kappa(\theta_{\mathcal{T}})B$, where the (generic) function $\kappa : \mathbb{R}_+ \longrightarrow \mathbb{R}_+$ is continuous and independent of $\mathcal{T}$. The notation $A \approx B$ is equivalent to $A \lesssim B$ and $B \lesssim A$.

The nonconforming P_1 finite element space (cf. [15]) associated with the triangulation $\mathcal{T}$ is $V_{\mathcal{T}} = \{v \in L_2(\Omega) : v_{\scriptscriptstyle T} = v\big|_T \in P_1(T)$ for any $T \in \mathcal{T}$ and v is continuous at the center of the common side of any two neighboring triangles ($d = 2$) or tetrahedra ($d = 3$)$\}$. A function in $V_{\mathcal{T}}$ is completely determined by its nodal values at the centers of the sides of the triangles or tetrahedra in $\mathcal{T}$ (cf. Figure 2.1).

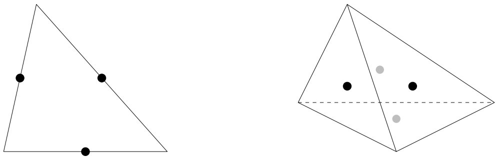


FIG. 2.1. *P_1 nonconforming finite elements.*

The interpolation operator $\mathcal{I} : H^1(\Omega, \mathcal{T}) \longrightarrow V_{\mathcal{T}}$ is defined by

$$(\mathcal{I}\zeta)(c_\sigma) = \frac{1}{|\sigma|}\int_\sigma \{\zeta\}\, ds \qquad \forall\, \sigma \in S(\mathcal{T}, \Omega) \cup S(\mathcal{T}, \partial\Omega), \tag{2.1}$$

where c_σ is the center of the side σ and $\{\zeta\}$ is the average of the traces from the two sides of σ. For $\sigma \subset \partial\Omega$, we take $\{\zeta\}$ to be ζ.

Let $\Pi_T : H^1(T) \longrightarrow P_1(T)$ be the local interpolation operator defined by

$$(\Pi_T\zeta)(c_\sigma) = \frac{1}{|\sigma|}\int_\sigma \zeta\, ds \qquad \forall\, \sigma \subset \partial T\,. \tag{2.2}$$

From (2.1) and (2.2) we see that the difference of the two interpolants on $T \in \mathcal{T}$ is given by

$$(\mathcal{I}\zeta - \Pi_T\zeta)(c_\sigma) = \begin{cases} \dfrac{1}{2|\sigma|}\displaystyle\int_\sigma [\zeta]\, ds & \text{if} \quad \sigma \subset \partial T \setminus \partial\Omega, \\ 0 & \text{if} \quad \sigma \subset \partial T \cap \partial\Omega, \end{cases} \tag{2.3}$$

where the jump $[\zeta]$ is measured by subtracting the interior trace from the exterior trace.

Using (2.3) and standard finite element estimates (cf. [13, 10]), we find

$$\begin{aligned}|\mathcal{I}\zeta-\Pi_T\zeta|^2_{H^1(T)} &\lesssim |T|^{1-(2/d)}\sum_{\sigma\subset\partial T}\big[\big(\mathcal{I}\zeta-\Pi_T\zeta\big)(c_\sigma)\big]^2\\ &\lesssim |T|^{1-(2/d)}\sum_{\sigma\subset\partial T\setminus\partial\Omega}\frac{1}{|\sigma|^2}\left(\int_\sigma[\zeta]\,ds\right)^2 \qquad (2.4)\\ &\lesssim \sum_{\sigma\subset\partial T\setminus\partial\Omega}|\sigma|^{d/(1-d)}\left(\int_\sigma[\zeta]\,ds\right)^2,\end{aligned}$$

$$\begin{aligned}\|\mathcal{I}\zeta-\Pi_T\zeta\|^2_{L_2(T)} &\lesssim |T|\sum_{\sigma\subset\partial T}\big[\big(\mathcal{I}\zeta-\Pi_T\zeta\big)(c_\sigma)\big]^2 \qquad (2.5)\\ &\lesssim \sum_{\sigma\subset\partial T\setminus\partial\Omega}|\sigma|^{(2-d)/(d-1)}\left(\int_\sigma[\zeta]\,ds\right)^2,\end{aligned}$$

where $|T|$ is the d-dimensional volume of T. Note that

$$|T|\approx|\sigma|^{d/(d-1)} \qquad \text{for} \quad \sigma\subset\partial T\,. \tag{2.6}$$

On the other hand, we also have the following well-known estimates (cf. [15]) for the local interpolation operator:

$$\|\zeta-\Pi_T\zeta\|^2_{L_2(T)}+|T|^{2/d}\,|\Pi_T\zeta|^2_{H^1(T)}\lesssim|T|^{2/d}\,|\zeta|^2_{H^1(T)}\,. \tag{2.7}$$

Combining the estimates (2.4)–(2.7) and summing over all $T\in\mathcal{T}$ we find

$$|\mathcal{I}\zeta|^2_{H^1(\Omega,\mathcal{T})}\lesssim|\zeta|^2_{H^1(\Omega,\mathcal{T})}+\sum_{\sigma\in S(\mathcal{T},\Omega)}|\sigma|^{d/(1-d)}\left(\int_\sigma[\zeta]\,ds\right)^2, \tag{2.8}$$

$$\|\zeta-\mathcal{I}\zeta\|^2_{L_2(\Omega)}\lesssim\sum_{T\in\mathcal{T}}|T|^{2/d}\,|\zeta|^2_{H^1(T)}+\sum_{\sigma\in S(\mathcal{T},\Omega)}|\sigma|^{(2-d)/(d-1)}\left(\int_\sigma[\zeta]\,ds\right)^2. \tag{2.9}$$

3. The connections between nonconforming P_1 finite elements and their conforming relatives. Let $W_{\mathcal{T}}\subset H^1(\Omega)$ be the P_2 Lagrange finite element space associated with $\mathcal{T}$ for $d=2$ and the P_3 Lagrange finite element space associated with $\mathcal{T}$ for $d=3$. The nodal variables (degrees of freedom) of these elements are depicted in Figure 3.1. We will denote the set of the centers of the sides of T by $\mathcal{C}(T)$, denote the set of the other nodes by $\mathcal{N}(T)$, and define $\mathcal{C}(\mathcal{T})=\bigcup_{T\in\mathcal{T}}\mathcal{C}(T)$ and $\mathcal{N}(\mathcal{T})=\bigcup_{T\in\mathcal{T}}\mathcal{N}(T)$.

Remark 3.1. The nonconforming P_1 finite element and the conforming P_2 (respectively, P_3) finite element in two (respectively, three) dimensions are *relatives* in

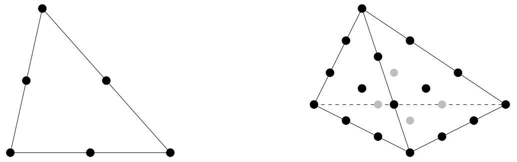


FIG. 3.1. *Two-dimensional P_2 and three-dimensional P_3 Lagrange elements.*

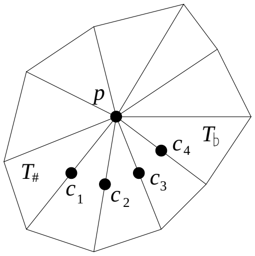


FIG. 3.2. *A sequence of c_j's connecting $T_\sharp$ and $T_\flat$.*

the sense that the shape functions of the first are also shape functions of the latter and the nodal variables of the first are also nodal variables of the latter (cf. Figures 2.1 and 3.1).

The finite element spaces $V_{\mathcal{T}}$ and $W_{\mathcal{T}}$ are connected by the operators $E : V_{\mathcal{T}} \longrightarrow W_{\mathcal{T}}$ and $F : W_{\mathcal{T}} \longrightarrow V_{\mathcal{T}}$ defined by

$$\big(Ev\big)(p) = \frac{1}{|\Xi_p|} \sum_{T \in \Xi_p} v_{_T}(p) \qquad \forall\, p \in \mathcal{N}(\mathcal{T}) \cup \mathcal{C}(\mathcal{T}), \tag{3.1}$$

$$\big(Fw\big)(p) = w(p) \qquad \forall\, p \in \mathcal{C}(\mathcal{T}), \tag{3.2}$$

where $\Xi_p = \{T \in \mathcal{T} : p \in \partial T\}$ is the set of the simplexes sharing p as a vertex and $|\Xi_p|$ is the number of elements in Ξ_p. Note that $\big(Ev\big)(p) = v(p)$ for $p \in \mathcal{C}(\mathcal{T})$ since v is continuous at the centers of the sides. For $d = 2$, these connection operators were introduced in [8, 9] and used in the analysis of domain decomposition methods and multigrid methods for nonconforming finite elements.

The following lemma contains the basic estimates for E and F.

LEMMA 3.2. *It holds that*

$$\|Ev - v\|^2_{L_2(\Omega)} \approx \sum_{T \in \mathcal{T}} |T|^{2/d} |v|^2_{H^1(T)} \qquad \forall\, v \in V_{\mathcal{T}}, \tag{3.3}$$

$$\|Fw - w\|^2_{L_2(\Omega)} \approx \sum_{T \in \mathcal{T}} |T|^{2/d} |w|^2_{H^1(T)} \qquad \forall\, w \in W_{\mathcal{T}}\,. \tag{3.4}$$

Proof. Let $p \in \mathcal{N}(\mathcal{T})$ and $T_\sharp, T_\flat \in \Xi_p$. We can find a sequence $c_1, \ldots, c_m$ in $\mathcal{C}(\mathcal{T})$ so that $c_1 \in \partial T_\sharp$, $c_m \in \partial T_\flat$, and c_j, c_{j+1} belong to the boundary of $T_j \in \Xi_p$ for $j = 1, \ldots, m-1$ (cf. Figure 3.2 for a two-dimensional example with $m = 4$). Note that $|\Xi_p|$ and hence m are bounded by a constant depending continuously on the minimum angle $\theta_{\mathcal{T}}$. Hence it follows from the Cauchy–Schwarz inequality and the mean value theorem that

$$\begin{aligned}
[v_{_{T_\sharp}}(p) - v_{_{T_\flat}}(p)]^2 &\lesssim [v_{_{T_\sharp}}(p) - v_{_{T_\sharp}}(c_1)]^2 + \sum_{j=1}^{m-1} [v_{_{T_j}}(c_j) - v_{_{T_j}}(c_{j+1})]^2 \\
&\qquad + [v_{_{T_\flat}}(c_m) - v_{_{T_\flat}}(p)]^2 \\
&\lesssim \sum_{T' \in \Xi_p} |T'|^{(2/d)-1} |v|^2_{H^1(T')}\,.
\end{aligned} \tag{3.5}$$

Using (3.1), (3.5), and the Cauchy–Schwarz inequality we find for any $T \in \Xi_p$

$$\big[\big(Ev - v_{_T}\big)(p)\big]^2 \lesssim \sum_{T' \in \Xi_p} |T'|^{(2/d)-1} |v|^2_{H^1(T')} \qquad \forall\, v \in V_{\mathcal{T}}\,. \tag{3.6}$$

Let $T \in \mathcal{T}$. We have, by (3.6),

$$\begin{aligned} \|Ev - v\|^2_{L_2(T)} &\approx |T| \sum_{p \in \mathcal{N}(T) \cup \mathcal{C}(T)} \big[\big(Ev - v_{_T}\big)(p)\big]^2 \\ &= |T| \sum_{p \in \mathcal{N}(T)} \big[\big(Ev - v_{_T}\big)(p)\big]^2 \\ &\lesssim \sum_{p \in \mathcal{N}(T)} \sum_{T' \in \Xi_p} |T'|^{2/d} |v|^2_{H^1(T')}, \end{aligned} \tag{3.7}$$

where we have also used the fact that

$$|T| \approx |T'| \qquad \text{for} \quad T' \in \Xi_p \ \text{ and } \ p \in \mathcal{N}(T)\,. \tag{3.8}$$

The estimate (3.3) then follows by summing (3.7) over all $T \in \mathcal{T}$.

Observe that, on each $T \in \mathcal{T}$, Fw is just the linear nodal interpolant of w with the nodes placed at the centers of the sides of T. It follows from standard interpolation and inverse estimates (cf. [13, 10]) that

$$\|Fw - w\|^2_{L_2(T)} \lesssim |T|^{4/d} |w|^2_{H^2(T)} \lesssim |T|^{2/d} |w|^2_{H^1(T)} \qquad \forall\, w \in W_{\mathcal{T}}\,. \tag{3.9}$$

The estimate (3.4) follows by summing (3.9) over all $T \in \mathcal{T}$. □

COROLLARY 3.3. *It holds that*

$$\|Ev\|_{L_2(\Omega)} \approx \|v\|_{L_2(\Omega)} \qquad \forall\, v \in V_{\mathcal{T}}, \tag{3.10}$$

$$|Ev|_{H^1(\Omega)} \approx |v|_{H^1(\Omega,\mathcal{T})} \qquad \forall\, v \in V_{\mathcal{T}}\,. \tag{3.11}$$

Proof. It follows from (3.3) and a standard inverse estimate (cf. [13, 10]) that

$$\begin{aligned} \|Ev\|_{L_2(\Omega)} &\le \|Ev - v\|_{L_2(\Omega)} + \|v\|_{L_2(\Omega)} \\ &\lesssim \Big(\sum_{T \in \mathcal{T}} |T|^{2/d} |v|^2_{H^1(T)}\Big)^{1/2} + \|v\|_{L_2(\Omega)} \\ &\lesssim \|v\|_{L_2(\Omega)} \qquad \forall\, v \in V_{\mathcal{T}}\,. \end{aligned} \tag{3.12}$$

Similarly from (3.4) we have

$$\|Fw\|_{L_2(\Omega)} \lesssim \|w\|_{L_2(\Omega)} \qquad \forall\, w \in W_{\mathcal{T}}\,. \tag{3.13}$$

It is clear from definitions (3.1) and (3.2) that

$$F(Ev) = v \qquad \forall\, v \in V_{\mathcal{T}}, \tag{3.14}$$

and hence, by (3.13),

$$\|v\|_{L_2(\Omega)} = \|F(Ev)\|_{L_2(\Omega)} \lesssim \|Ev\|_{L_2(\Omega)} \qquad \forall\, v \in V_{\mathcal{T}}\,. \tag{3.15}$$

The estimates (3.12) and (3.15) together yield (3.10).

By (3.3) and a standard inverse estimate, we also find

$$
\begin{aligned}
|Ev|_{H^1(\Omega)} &\le |Ev - v|_{H^1(\Omega,\mathcal{T})} + |v|_{H^1(\Omega,\mathcal{T})} \\
&\lesssim \left(\sum_{T\in\mathcal{T}} |T|^{-(2/d)} \|Ev - v\|^2_{L_2(T)} \right)^{1/2} + |v|_{H^1(\Omega,\mathcal{T})} \qquad (3.16)\\
&\lesssim |v|_{H^1(\Omega,\mathcal{T})} \qquad \forall\, v \in V_{\mathcal{T}}\,.
\end{aligned}
$$

Similarly, we derive from (3.4) that

$$
|Fw|_{H^1(\Omega,\mathcal{T})} \lesssim |w|_{H^1(\Omega)} \qquad \forall\, w \in W_{\mathcal{T}}\,. \tag{3.17}
$$

Again, the estimate (3.11) follows from (3.14), (3.16), and (3.17). □

4. Poincaré–Friedrichs inequalities for nonconforming P_1 finite element functions.

Functions in the nonconforming P_1 finite element space $V_{\mathcal{T}}$ are known to satisfy Poincaré–Friedrichs inequalities (cf. [28, 27, 16, 22]). Here we give a simple derivation of such inequalities using the connection between $V_{\mathcal{T}}$ and its conforming relative $W_{\mathcal{T}}$ developed in section 3.

Let Φ be a seminorm on $H^1(\Omega)$ with the following properties:

$$
\Phi(\phi) \le C\|\phi\|_{H^1(\Omega)} \qquad \forall\, \phi \in H^1(\Omega), \tag{4.1}
$$

where C is a positive constant, and

(4.2) for a constant function c, $\Phi(c) = 0$ if and only if $c = 0$.

Then we have the generalized Poincaré–Friedrichs inequality (cf. [23]) for $H^1(\Omega)$:

$$
\|\phi\|_{L_2(\Omega)} \le C\big[|\phi|_{H^1(\Omega)} + \Phi(\phi)\big] \qquad \forall\, \phi \in H^1(\Omega), \tag{4.3}
$$

which follows from the compactness of the embedding of $H^1(\Omega)$ in $L_2(\Omega)$.

We can derive from (4.3) a generalized Poincaré–Friedrichs inequality for nonconforming P_1 finite element functions.

THEOREM 4.1. *Let Φ be a seminorm on $H^1(\Omega,\mathcal{T})$ that satisfies* (4.1), (4.2), *and the additional condition that*

$$
\Phi(Ev - v) \lesssim |v|_{H^1(\Omega,\mathcal{T})} \qquad \forall\, v \in V_{\mathcal{T}}, \tag{4.4}
$$

where $E : V_{\mathcal{T}} \longrightarrow W_{\mathcal{T}}$ is defined by (3.1). *Then there exists a continuous function $\kappa : \mathbb{R}_+ \longrightarrow \mathbb{R}_+$, independent of $\mathcal{T}$ such that*

$$
\|v\|_{L_2(\Omega)} \le \kappa(\theta_{\mathcal{T}})\big(|v|_{H^1(\Omega,\mathcal{T})} + \Phi(v)\big) \qquad \forall\, v \in V_{\mathcal{T}}\,. \tag{4.5}
$$

Proof. Combining (3.10), (3.11), (4.3), and (4.4), we find

$$
\begin{aligned}
\|v\|_{L_2(\Omega)} \approx \|Ev\|_{L_2(\Omega)} &\lesssim |Ev|_{H^1(\Omega)} + \Phi(Ev) \\
&\lesssim |v|_{H^1(\Omega,\mathcal{T})} + \Phi(Ev - v) + \Phi(v) \\
&\lesssim |v|_{H^1(\Omega,\mathcal{T})} + \Phi(v) \qquad \forall\, v \in V_{\mathcal{T}}\,.
\end{aligned}
$$

□

The following are examples of seminorms that satisfy (4.1), (4.2), and (4.4).

Example 4.2. Let ψ be a square integrable function on $\partial\Omega$ such that

$$\int_{\partial\Omega} \psi \, ds \neq 0,$$

and let $\Phi_1 : H^1(\Omega, \mathcal{T}) \longrightarrow \mathbb{R}$ be defined by

$$\Phi_1(\zeta) = \left| \int_{\partial\Omega} \psi\zeta \, ds \right| . \tag{4.6}$$

Then (4.2) is obvious and (4.1) follows from the trace theorem. Condition (4.4) can be established as follows:

$$\begin{aligned}
[\Phi_1(Ev - v)]^2 &\leq \|\psi\|^2_{L_2(\partial\Omega)} \|Ev - v\|^2_{L_2(\partial\Omega)} \\
&\lesssim \sum_{\sigma \in S(\mathcal{T}, \partial\Omega)} \|Ev - v\|^2_{L_2(\sigma)} \\
&\lesssim \sum_{\sigma \in S(\mathcal{T}, \partial\Omega)} |\sigma| \sum_{p \in \mathcal{N}(\sigma)} \left[(Ev - v_{T_\sigma})(p) \right]^2,
\end{aligned}$$

where $\mathcal{N}(\sigma)$ denotes the set of the nodes on $\bar{\sigma}$ excluding the center of σ, and T_σ is the member of $\mathcal{T}$ whose boundary contains σ. Combining this last estimate with (2.6), (3.6), and (3.8), we find

$$[\Phi_1(Ev - v)]^2 \lesssim \sum_{\substack{T \in \mathcal{T} \\ |\partial T \cap \partial\Omega| > 0}} |T|^{1/d} |v|^2_{H^1(T)} \lesssim |v|^2_{H^1(\Omega, \mathcal{T})} .$$

Example 4.3. Let ψ be a square integrable function on Ω such that

$$\int_{\Omega} \psi \, dx \neq 0,$$

and let $\Phi_2 : H^1(\Omega, \mathcal{T}) \longrightarrow \mathbb{R}$ be defined by

$$\Phi_2(\zeta) = \left| \int_{\Omega} \psi\zeta \, dx \right| . \tag{4.7}$$

Then (4.1) and (4.2) are trivial, and (4.4) follows from (3.3).

Remark 4.4. The Poincaré–Friedrichs inequalities (1.5) and (1.6) for P_1 nonconforming finite element functions follow immediately from Theorem 4.1 if we take ψ to be the characteristic function of Γ in (4.6) and the characteristic function of Ω in (4.7).

5. Poincaré–Friedrichs inequalities for $H^1(\Omega, \mathcal{T})$. Let $\mathcal{T}$ be a simplicial triangulation of Ω. The last ingredient for establishing a generalized Poincaré–Friedrichs inequality for $H^1(\Omega, \mathcal{T})$ is the following condition on Φ:

$$[\Phi(\mathcal{I}\zeta - \zeta)]^2 \lesssim |\zeta|^2_{H^1(\Omega, \mathcal{T})} + \sum_{\sigma \in S(\mathcal{T}, \Omega)} |\sigma|^{d/(1-d)} \left(\int_{\sigma} [\zeta] \, ds \right)^2 \tag{5.1}$$

for all $\zeta \in H^1(\Omega, \mathcal{T})$.

THEOREM 5.1. *Let Φ be a seminorm on $H^1(\Omega,\mathcal{T})$ that satisfies the conditions* (4.1), (4.2), (4.4), *and* (5.1). *Then there exists a continuous function $\kappa : \mathbb{R}_+ \longrightarrow \mathbb{R}_+$, independent of $\mathcal{T}$, such that*

$$\|\zeta\|^2_{L_2(\Omega)} \le \kappa(\theta_{\mathcal{T}}) \left[|\zeta|^2_{H^1(\Omega,\mathcal{T})} + \sum_{\sigma\in S(\mathcal{T},\Omega)} |\sigma|^{d/(1-d)} \left(\int_\sigma [\zeta]\, ds \right)^2 + [\Phi(\zeta)]^2 \right] \tag{5.2}$$

for all $\zeta \in H^1(\Omega,\mathcal{T})$.

Proof. Combining (2.8), (2.9), (4.5), and (5.1) we have

$$\begin{aligned} \|\zeta\|^2_{L_2(\Omega)} &\lesssim \|\zeta - \mathcal{I}\zeta\|^2_{L_2(\Omega)} + \|\mathcal{I}\zeta\|^2_{L_2(\Omega)} \\ &\lesssim \|\zeta - \mathcal{I}\zeta\|^2_{L_2(\Omega)} + |\mathcal{I}\zeta|^2_{H^1(\Omega,\mathcal{T})} + [\Phi(\mathcal{I}\zeta)]^2 \\ &\lesssim \|\zeta - \mathcal{I}\zeta\|^2_{L_2(\Omega)} + |\mathcal{I}\zeta|^2_{H^1(\Omega,\mathcal{T})} + [\Phi(\mathcal{I}\zeta - \zeta)]^2 + [\Phi(\zeta)]^2 \\ &\lesssim |\zeta|^2_{H^1(\Omega,\mathcal{T})} + \sum_{\sigma\in S(\mathcal{T},\Omega)} |\sigma|^{d/(1-d)} \left(\int_\sigma [\zeta]\, ds \right)^2 + [\Phi(\zeta)]^2 . \qquad \square \end{aligned}$$

Note that condition (5.1) is satisfied by the seminorms Φ_1 (Example 4.2) and Φ_2 (Example 4.3). The case of Φ_2 follows immediately from (2.9). The case of Φ_1 can be established as follows.

First of all we have, by the Cauchy–Schwarz inequality,

$$\begin{aligned} [\Phi_1(\mathcal{I}\zeta - \zeta)]^2 &\lesssim \sum_{\substack{T\in\mathcal{T} \\ |\partial T\cap\partial\Omega|>0}} \|\mathcal{I}\zeta - \zeta\|^2_{L_2(\partial T\cap\partial\Omega)} \\ &\lesssim \sum_{T\in\mathcal{T}} \|\mathcal{I}\zeta - \Pi_T\zeta\|^2_{L_2(\partial T)} + \sum_{T\in\mathcal{T}} \|\Pi_T\zeta - \zeta\|^2_{L_2(\partial T)} . \end{aligned} \tag{5.3}$$

Using the equivalence of norms on finite-dimensional spaces and a scaling argument, we have

$$\|\mathcal{I}\zeta - \Pi_T\zeta\|^2_{L_2(\partial T)} \lesssim |T|^{-(1/d)} \|\mathcal{I}\zeta - \Pi_T\zeta\|^2_{L_2(T)} . \tag{5.4}$$

On the other hand, it follows from the trace theorem, the Bramble–Hilbert lemma (cf. [7]), and scaling that

$$\|\Pi_T\zeta - \zeta\|^2_{L_2(\partial T)} \lesssim |T|^{1/d} |\zeta|^2_{H^1(T)} \lesssim |\zeta|^2_{H^1(T)} . \tag{5.5}$$

From (2.5), (2.6), and (5.3)–(5.5) we then obtain the estimate (5.1):

$$\begin{aligned} [\Phi_1(\mathcal{I}\zeta - \zeta)]^2 &\lesssim |\zeta|^2_{H^1(\Omega,\mathcal{T})} + \sum_{T\in\mathcal{T}} \sum_{\sigma\subset\partial T\setminus\partial\Omega} |\sigma|^{-1/(d-1)} |\sigma|^{(2-d)/(d-1)} \left(\int_\sigma [\zeta]\, ds \right)^2 \\ &\lesssim |\zeta|^2_{H^1(\Omega,\mathcal{T})} + \sum_{T\in\mathcal{T}} \sum_{\sigma\subset\partial T\setminus\partial\Omega} |\sigma|^{-1} \left(\int_\sigma [\zeta]\, ds \right)^2 \\ &\lesssim |\zeta|^2_{H^1(\Omega,\mathcal{T})} + \sum_{\sigma\in S(\mathcal{T},\Omega)} |\sigma|^{d/(1-d)} \left(\int_\sigma [\zeta]\, ds \right)^2 . \end{aligned}$$

Remark 5.2. Let $\mathfrak{T}_\Omega$ be the set of all simplicial triangulations of Ω. From here on we assume that $\Phi : \bigcup_{\mathcal{T}\in\mathfrak{T}_\Omega} H^1(\Omega,\mathcal{T}) \longrightarrow \mathbb{R}$ is a seminorm for every $\mathcal{T} \in \mathfrak{T}_\Omega$ and that it satisfies the conditions (4.1), (4.2), (4.4), and (5.1) for every $\mathcal{T} \in \mathfrak{T}_\Omega$.

Remark 5.3. The Poincaré–Friedrichs inequalities (1.3) and (1.4) for $\zeta \in H^1(\Omega,\mathcal{T})$ follow immediately from Theorem 5.1 if the functions ψ in (4.6) and (4.7) are chosen as in Remark 4.4. Similar remarks also apply in the next two sections.

6. Poincaré–Friedrichs inequalities for $H^1(\Omega,\mathcal{P})$ on two-dimensional domains. Let us first give the precise definition of interior edges for a general partition $\mathcal{P}$. We define a vertex of the partition $\mathcal{P}$ to be a vertex of any of the subdomains in $\mathcal{P}$. (For example, the partition of the square in Figure 1.1 has 14 vertices.) We then define an open edge of $\mathcal{P}$ to be an open line segment on the boundary of a subdomain in $\mathcal{P}$ bounded between two of the vertices of $\mathcal{P}$. The set $S(\mathcal{P},\Omega)$ contains the open edges of $\mathcal{P}$ that are common to the boundaries of two members of $\mathcal{P}$.

Remark 6.1. The concept of an edge of a polygon $D \in \mathcal{P}$ and the concept of an edge of $\mathcal{P}$ on ∂D should be distinguished. For example, a square always has four edges, while there are five edges of the two-dimensional partition in Figure 1.1 on the boundary of the square at the lower right corner.

In order to apply Theorem 5.1 we define the set

$$\begin{aligned}(6.1)\qquad \mathfrak{T}_\mathcal{P} = \{&\mathcal{T} : \mathcal{T} \text{ is a triangulation of } \Omega \text{ by triangles and each member of}\\ &S(\mathcal{P},\Omega) \text{ is also an edge of } \mathcal{T}\}.\end{aligned}$$

(A triangulation belonging to $\mathfrak{T}_\mathcal{P}$ for the two-dimensional example in Figure 1.1 is depicted in Figure 6.1.) By definition (6.1), $H^1(\Omega,\mathcal{P})$ is a subspace of $H^1(\Omega,\mathcal{T})$ for any $\mathcal{T} \in \mathfrak{T}_\mathcal{P}$. Observing that for a function $\zeta \in H^1(\Omega,\mathcal{P})$ the jump $[\zeta]$ is zero across the edges of $\mathcal{T}$ that are not in $S(\mathcal{P},\Omega)$, we immediately deduce the following result from Theorem 5.1.

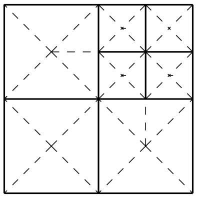


Fig. 6.1. *A triangulation in $\mathfrak{T}_\mathcal{P}$ for a partition of a square.*

Theorem 6.2. *Let Φ be as in Remark* 5.2. *Then we have*

$$(6.2)\quad \|\zeta\|^2_{L_2(\Omega)} \le \left[\inf_{\mathcal{T}\in\mathfrak{T}_\mathcal{P}} \kappa(\theta_\mathcal{T})\right]\left[|\zeta|^2_{H^1(\Omega,\mathcal{P})} + \sum_{\sigma\in S(\mathcal{P},\Omega)} |\sigma|^{-2}\left(\int_\sigma [\zeta]\,ds\right)^2 + [\Phi(\zeta)]^2\right]$$

for all $\zeta \in H^1(\Omega,\mathcal{P})$, where $\kappa : \mathbb{R}_+ \longrightarrow \mathbb{R}_+$ is a continuous function independent of $\mathcal{P}$.

In an abstract sense the set $\{\theta_\mathcal{T} : \mathcal{T} \in \mathfrak{T}_\mathcal{P}\}$ provides a measure of the shape regularity of the partition $\mathcal{P}$ and $\inf_{\mathcal{T}\in\mathfrak{T}_\mathcal{P}} \kappa(\theta_\mathcal{T})$ is a constant depending on the shape regularity of $\mathcal{P}$.

On the other hand, we can also describe the shape regularity of $\mathcal{P}$ more concretely in terms of the shape regularity of individual subdomains and the relative positions of subdomains sharing a common edge of $\mathcal{P}$.

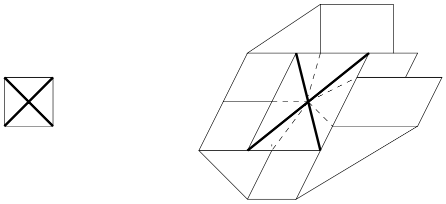


FIG. 6.2. *An example of the construction of* $\mathfrak{T}_{\mathcal{P}_i}$.

The shape regularity of a polygonal (or polyhedral) domain D can be measured by using an affine homeomorphism between D and a reference domain, and by using the aspect ratio of D defined by the quotient (diameter of D)/(diameter of the largest disc (or ball) in $\bar{D}$).

We will use the quantity

$$\rho(\mathcal{P}) = \max\{|\partial D|/|\sigma| : \sigma \in S(\mathcal{P},\Omega),\ D \in \mathcal{P} \text{ and } \sigma \subset \partial D\} \tag{6.3}$$

to measure the relative positions between subdomains sharing a common edge of $\mathcal{P}$.

We can now formulate and prove the following corollary, which gives an application of Theorem 6.2 to a fairly general class of two-dimensional partitions.

COROLLARY 6.3. *Let* Φ *be as in Remark* 5.2, *and let* $\{\mathcal{P}_i : i \in I\}$ *be a family of partitions of* Ω. *Suppose that the polygons appearing in all the partitions* $\mathcal{P}_i$ *are affine homeomorphic to a fixed finite set of reference polygons and the aspect ratios of the polygons in all the* $\mathcal{P}_i$*'s are uniformly bounded. Assume also that the set* $\{\rho(\mathcal{P}_i) : i \in I\}$ *is bounded. Then there exists a positive constant* C, *independent of* $i \in I$, *such that*

$$\|\zeta\|^2_{L_2(\Omega)} \le C\left[|\zeta|^2_{H^1(\Omega,\mathcal{P}_i)} + \sum_{\sigma\in S(\mathcal{P}_i,\Omega)} |\sigma|^{-2}\left(\int_\sigma [\zeta]\,ds\right)^2 + [\Phi(\zeta)]^2\right] \tag{6.4}$$

for any $\zeta \in H^1(\Omega,\mathcal{P}_i)$ *and* $i \in I$.

Proof. First we impose on each reference polygon a triangulation by triangles such that each edge of the polygon is also the edge of a triangle in the triangulation and each triangle of the triangulation has at most one edge on the boundary of the polygon.

Let $D \in \mathcal{P}_i$. We can induce a triangulation $\mathcal{T}_D$ on D using the triangulation on a reference polygon and the corresponding affine homeomorphism. Let $p \in \partial D$ be a vertex of $\mathcal{P}$ but not a vertex of D. Then p belongs to an edge of D which is a side of a triangle $T \in \mathcal{T}_D$, and we connect p to the vertex of T not on ∂D by a straight line (cf. Figure 6.2 where the construction is carried out for a square reference domain and a subdomain D which is a parallelogram). In this way we have created a triangulation $\mathcal{T}_i \in \mathfrak{T}_{\mathcal{P}_i}$.

Let $\hat{D}$ be the reference polygon affine homeomorphic to D, and let $\hat{x} \mapsto \alpha(\hat{x}) = B\hat{x}+b$ be the corresponding affine map from $\hat{D}$ to D. The uniform boundedness of the aspect ratios implies (cf. Theorem 3.1.3 in [13]) the existence of a positive constant C_*, independent of $i \in I$, such that

$$\|B\| \le C_*(\operatorname{diam} D) \text{ and } \|B^{-1}\| \le C_*(\operatorname{diam} D)^{-1}, \tag{6.5}$$

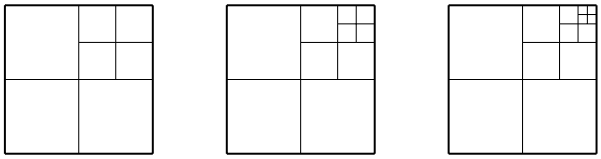


FIG. 6.3. *A family of partitions of a square.*

where $\|\cdot\|$ is the matrix 2-norm induced by the Euclidean vector norm. Hence we have

$$C_*^{-2}\frac{|\hat{x}_1-\hat{x}_2|}{|\hat{x}_3-\hat{x}_4|}\le\frac{|x_1-x_2|}{|x_3-x_4|}\le C_*^2\frac{|\hat{x}_1-\hat{x}_2|}{|\hat{x}_3-\hat{x}_4|}, \tag{6.6}$$

where $x_j=\alpha(\hat{x}_j)$, and $\hat{x}_1,\hat{x}_2,\hat{x}_3,\hat{x}_3$ are any four points such that $\hat{x}_1\neq\hat{x}_2$ and $\hat{x}_3\neq\hat{x}_4$.

We conclude from (6.6) and the boundedness of the set $\{\rho(\mathcal{P}_i): i\in I\}$ that $\theta_{\mathcal{T}_i}$ is bounded away from zero. The estimate (6.4) then follows from (6.2) if we take C to be an upper bound of the bounded set $\{\kappa(\theta_{\mathcal{T}_i}): i\in I\}$. □

Remark 6.4. If the family of partitions $\{\mathcal{P}_i: i\in I\}$ in Corollary 6.3 is actually a family of triangulations, then the condition on the boundedness of $\{\rho(\mathcal{P}_i): i\in I\}$ is redundant.

An example of a family of partitions satisfying the assumptions of Corollary 6.3 is depicted in Figure 6.3, where the partition of the square in Figure 1.1 is being refined successively towards the upper right corner.

7. Poincaré–Friedrichs inequalities for $H^1(\Omega,\mathcal{P})$ on three-dimensional domains. We first give the precise definition of interior faces for a partition $\mathcal{P}$. We define an edge of $\mathcal{P}$ to be an edge of any of the subdomains in $\mathcal{P}$. We then define an open face of $\mathcal{P}$ to be an open subset of the boundary of a member of $\mathcal{P}$ enclosed by edges of $\mathcal{P}$. The set $S(\mathcal{P},\Omega)$ contains the open faces of $\mathcal{P}$ that are common to the boundaries of two members of $\mathcal{P}$.

Remark 7.1. Again the concept of a face of a polyhedron $D\in\mathcal{P}$ and the concept of a face of $\mathcal{P}$ on ∂D should be distinguished. For example, there are always six faces on a parallelepiped, while there are nine faces of the three-dimensional partition in Figure 1.1 on the boundary of the subdomain in the back.

The situation here is more complicated than the two-dimensional case discussed in section 6, since the faces in $S(\mathcal{P},\Omega)$ may not be triangles. Accordingly, we introduce the set

$$\begin{aligned}\mathfrak{T}_{\mathcal{P}}=\{\mathcal{T}:\ &\mathcal{T}\text{ is a triangulation of }\Omega\text{ by tetrahedra such that each face in}\\ &S(\mathcal{P},\Omega)\text{ is triangulated by the (triangular) faces in }S(\mathcal{T},\Omega)\}.\end{aligned} \tag{7.1}$$

Since a face in $S(\mathcal{P},\Omega)$ may not be a face in $S(\mathcal{T},\Omega)$, we cannot immediately obtain an analogue of Theorem 6.2. In order to apply Theorem 5.1 to derive a generalized Poincaré–Friedrichs inequality for $H^1(\Omega,\mathcal{P})$, we need to introduce two more parameters related to the shape regularity of $\mathcal{P}$ besides the parameter $\rho(\mathcal{P})$ defined by (6.3).

Let $\mathcal{T}\in\mathfrak{T}_{\mathcal{P}}$. For $\sigma\in S(\mathcal{P},\Omega)$ we will denote by $\mathcal{T}_\sigma$ the triangulation of σ by faces of $S(\mathcal{T},\Omega)$, i.e.,

$$\mathcal{T}_\sigma=\{\tilde{\sigma}\in S(\mathcal{T},\Omega):\tilde{\sigma}\subseteq\sigma\},$$

and define the parameter

$$\rho(\mathcal{P},\mathcal{T}) = \max\{|\sigma|/|\tilde{\sigma}| : \sigma \in S(\mathcal{P},\Omega) \text{ and } \tilde{\sigma} \in \mathcal{T}_\sigma\}\,. \tag{7.2}$$

Note that we have the following obvious bound for $|\mathcal{T}_\sigma|$ (the number of elements in $\mathcal{T}_\sigma$):

$$|\mathcal{T}_\sigma| \le \rho(\mathcal{P},\mathcal{T}) \qquad \forall\, \sigma \in S(\mathcal{P},\Omega)\,. \tag{7.3}$$

Moreover, (6.3) and (7.2) imply that

$$\frac{|\partial D|}{|\tilde{\sigma}|} \le \rho(\mathcal{P})\rho(\mathcal{P},\mathcal{T}) \qquad \text{for} \quad D \in \mathcal{P},\ \tilde{\sigma} \in \mathcal{T}_\sigma \ \text{ and } \ \sigma \subset \partial D\,. \tag{7.4}$$

The other parameter is the smallest number $\lambda(\mathcal{P}) \ge 1$ with the property that

$$\left(\frac{1}{|\mathcal{F}|}\int_{\mathcal{F}} |\zeta - \bar{\zeta}|\, ds\right)^2 \le \frac{\lambda(\mathcal{P})}{\operatorname{diam} D} |\zeta|^2_{H^1(D)}, \tag{7.5}$$

where D is any member of $\mathcal{P}$, $\mathcal{F}$ is any face of D, ζ is any function in $H^1(D)$, and $\bar{\zeta} = |D|^{-1}\int_D \zeta\, dx$ is the mean of ζ over D. The existence of $\lambda(\mathcal{P})$ is a consequence of the trace theorem, the classical Poincaré–Friedrichs inequalities, and scaling.

Note that (7.5) implies

$$\left(\frac{1}{|G|}\int_{G} |\zeta - \bar{\zeta}|\, ds\right)^2 \le \left(\frac{1}{|G|}\int_{\mathcal{F}} |\zeta - \bar{\zeta}|\, ds\right)^2 \le \left(\frac{|\mathcal{F}|}{|G|}\right)^2 \frac{\lambda(\mathcal{P})}{\operatorname{diam} D} |\zeta|^2_{H^1(D)}, \tag{7.6}$$

where G is any measurable subset of $\mathcal{F}$ with a positive $(d-1)$-dimensional measure.

Theorem 7.2. *Let Φ be as in Remark 5.2. Then we have*

$$\begin{aligned} \|\zeta\|^2_{L_2(\Omega)} \le &\left[\inf_{\mathcal{T}\in\mathfrak{T}_\mathcal{P}} K\big(\rho(\mathcal{P}),\lambda(\mathcal{P}),\rho(\mathcal{P},\mathcal{T}),\theta_\mathcal{T}\big)\right] \\ &\times \left[|\zeta|^2_{H^1(\Omega,\mathcal{P})} + \sum_{\sigma\in S(\mathcal{P},\Omega)} |\sigma|^{-\frac{3}{2}} \left(\int_\sigma [\zeta]\, ds\right)^2 + [\Phi(\zeta)]^2\right] \end{aligned} \tag{7.7}$$

for all $\zeta \in H^1(\Omega,\mathcal{P})$, where $K : \mathbb{R}^4_+ \longrightarrow \mathbb{R}_+$ is a continuous function independent of $\mathcal{P}$.

Proof. Since definition (7.1) implies that $H^1(\Omega,\mathcal{P})$ is a subspace of $H^1(\Omega,\mathcal{T})$ for any $\mathcal{T} \in \mathfrak{T}_\mathcal{P}$, we deduce from Theorem 5.1 the estimate

$$\begin{aligned} \|\zeta\|^2_{L_2(\Omega)} \le \kappa(\theta_\mathcal{T}) \Bigg[|\zeta|^2_{H^1(\Omega,\mathcal{P})} &+ \sum_{\sigma\in S(\mathcal{P},\Omega)} \sum_{\tilde{\sigma}\in\mathcal{T}_\sigma} |\tilde{\sigma}|^{1/2} \left(\frac{1}{|\tilde{\sigma}|}\int_{\tilde{\sigma}} [\zeta]\, ds\right)^2 \\ &+ [\Phi(\zeta)]^2 \Bigg] \end{aligned} \tag{7.8}$$

for any $\zeta \in H^1(\Omega,\mathcal{P})$ and $\mathcal{T} \in \mathfrak{T}_\mathcal{P}$, where $\kappa : \mathbb{R}_+ \longrightarrow \mathbb{R}_+$ is a continuous function.

Let $\sigma \in S(\mathcal{P},\Omega)$. Since $\mathcal{T}_\sigma$ is a triangulation of σ, it follows from the Cauchy–Schwarz inequality and (7.3) that

$$\sum_{\tilde{\sigma}\in\mathcal{T}_\sigma} |\tilde{\sigma}|^{1/2} \le |\mathcal{T}_\sigma|^{1/2} \left(\sum_{\tilde{\sigma}\in\mathcal{T}_\sigma} |\tilde{\sigma}|\right)^{1/2} \le [\rho(\mathcal{P},\mathcal{T})]^{1/2} |\sigma|^{1/2}\,. \tag{7.9}$$

Let $\mathcal{P}_\sigma$ be the set of the two polyhedra in $\mathcal{P}$ which share σ as a common face. Then we have, by the Cauchy–Schwarz inequality,

$$\left(\frac{1}{|\tilde\sigma|}\int_{\tilde\sigma}[\zeta]\,ds\right)^2 \le 3\left(\frac{1}{|\sigma|}\int_{\sigma}[\zeta]\,ds\right)^2 + 3\sum_{D\in\mathcal{P}_\sigma}\left(\frac{1}{|\tilde\sigma|}\int_{\tilde\sigma}\zeta_{_D}\,ds - \frac{1}{|\sigma|}\int_{\sigma}\zeta_{_D}\,ds\right)^2 \tag{7.10}$$

for any $\zeta\in H^1(\Omega,\mathcal{P})$.

From (6.3), (7.4), and (7.6) we have

$$\begin{aligned}&\left(\frac{1}{|\tilde\sigma|}\int_{\tilde\sigma}\zeta_{_D}\,ds - \frac{1}{|\sigma|}\int_{\sigma}\zeta_{_D}\,ds\right)^2\\ &\quad= \left(\frac{1}{|\tilde\sigma|}\int_{\tilde\sigma}(\zeta_{_D}-\bar\zeta_{_D})\,ds - \frac{1}{|\sigma|}\int_{\sigma}(\zeta_{_D}-\bar\zeta_{_D})\,ds\right)^2\\ &\quad\le 2\left[\frac{1}{|\tilde\sigma|}\int_{\tilde\sigma}|\zeta_{_D}-\bar\zeta_{_D}|\,ds\right]^2 + 2\left[\frac{1}{|\sigma|}\int_{\sigma}|\zeta_{_D}-\bar\zeta_{_D}|\,ds\right]^2\\ &\quad\le 2[\rho(\mathcal{P})]^2\big([\rho(\mathcal{P},\mathcal{T})]^2+1\big)\frac{\lambda(\mathcal{P})}{\operatorname{diam} D}|\zeta|^2_{H^1(D)}\end{aligned} \tag{7.11}$$

for $D\in\mathcal{P}_\sigma$, where $\bar\zeta_{_D} = |D|^{-1}\int_D \zeta\,dx$ is the mean of ζ over D.

Combining (7.9)–(7.11), we obtain

$$\begin{aligned}\sum_{\tilde\sigma\in\mathcal{T}_\sigma}|\tilde\sigma|^{1/2}\left(\frac{1}{|\tilde\sigma|}\int_{\tilde\sigma}[\zeta]\,ds\right)^2 &\le 3[\rho(\mathcal{P},\mathcal{T})]^{1/2}|\sigma|^{-(3/2)}\left(\int_\sigma[\zeta]\,ds\right)^2\\ &\quad+ 6[\rho(\mathcal{P},\mathcal{T})]^{1/2}[\rho(\mathcal{P})]^2\big([\rho(\mathcal{P},\mathcal{T})]^2+1\big)\lambda(\mathcal{P})\sum_{D\in\mathcal{P}_\sigma}|\zeta|^2_{H^1(D)},\end{aligned} \tag{7.12}$$

where we have used the fact that $|\sigma|^{1/2} < \operatorname{diam} D$.

Finally, we observe that the number of faces of $S(\mathcal{P},\Omega)$ on the boundary of any subdomain in $\mathcal{P}$ is less than or equal to $\rho(\mathcal{P})$, and hence

$$\sum_{\sigma\in S(\mathcal{P},\Omega)}\sum_{D\in\mathcal{P}_\sigma}|\zeta|^2_{H^1(D)} \le \rho(\mathcal{P})|\zeta|^2_{H^1(\Omega,\mathcal{P})}\,. \tag{7.13}$$

The generalized Poincaré–Friedrichs inequality (7.7) then follows from (7.8), (7.12), and (7.13), with the function K given by, for example,

$$K\big(\rho(\mathcal{P}),\lambda(\mathcal{P}),\rho(\mathcal{P},\mathcal{T}),\theta_{\mathcal{T}}\big) = 13[\rho(\mathcal{P})]^3\lambda(\mathcal{P})[\rho(\mathcal{P},\mathcal{T})]^{5/2}\kappa(\theta_{\mathcal{T}})\,. \qquad \square$$

Again in an abstract sense the set $\big\{\big(\rho(\mathcal{P}),\lambda(\mathcal{P}),\rho(\mathcal{P},\mathcal{T}),\theta_{\mathcal{T}}\big) : \mathcal{T}\in\mathfrak{T}_{\mathcal{P}}\big\}$ provides a measure of the shape regularity of the partition $\mathcal{P}$, and the number

$$\inf_{\mathcal{T}\in\mathfrak{T}_{\mathcal{P}}} K\big(\rho(\mathcal{P}),\lambda(\mathcal{P}),\rho(\mathcal{P},\mathcal{T}),\theta_{\mathcal{T}}\big)$$

is a constant depending on the shape regularity of $\mathcal{P}$. In applications one can use Theorem 7.2 to derive Poincaré–Friedrichs inequalities with a uniform constant for a family of partitions under appropriate concrete shape regularity assumptions. Since the geometry of three-dimensional partitions can be much more varied than that of two-dimensional ones, here we are content with giving only an analogue of Corollary 6.3 for partitions by *convex* polyhedra.

Note that a face of a partition of Ω by convex polyhedra is a convex polygon, and therefore it can be triangulated by connecting the center to the vertices by straight lines. Such a triangulation will be referred to as the *canonical triangulation* of the convex polygon.

COROLLARY 7.3. *Let Φ be as in Remark* 5.2, *and let $\{\mathcal{P}_i : i \in I\}$ be a family of partitions of Ω. Assume that the following conditions are satisfied:*

(i) *The polyhedra appearing in all the partitions $\mathcal{P}_i$ are affine homeomorphic to a fixed finite set of convex reference polyhedra and the aspect ratios of the polyhedra in all the $\mathcal{P}_i$'s are uniformly bounded.*

(ii) *The set $\{\rho(\mathcal{P}_i) : i \in I\}$ is bounded.*

(iii) *The angles of the triangles in the canonical triangulations of the faces of all the partitions $\mathcal{P}_i$ are bounded below by a positive constant.*

Then there exists a positive constant C, independent of $i \in I$, such that

$$\|\zeta\|^2_{L_2(\Omega)} \le C\left[|\zeta|^2_{H^1(\Omega,\mathcal{P}_i)} + \sum_{\sigma\in S(\mathcal{P}_i,\Omega)} |\sigma|^{-\frac{3}{2}}\left(\int_\sigma [\zeta]\,ds\right)^2 + [\Phi(\zeta)]^2\right] \tag{7.14}$$

for any $\zeta \in H^1(\Omega,\mathcal{P}_i)$ and $i \in I$.

Proof. Let $D \in \mathcal{P}_i$, $\hat{D}$ be a reference polyhedron affine homeomorphic to D, and let $\alpha(\hat{x}) = B\hat{x} + b$ be the corresponding affine map from $\hat{D}$ to D. Then the estimates (6.5) and (6.6) again follow from condition (i).

Let $\hat{\mathcal{F}}$ be a face of $\hat{D}$ corresponding to a face $\mathcal{F}$ of D. From the Poincaré–Friedrichs inequalities for $\hat{D}$ and the trace theorem, we have

$$\left(\frac{1}{|\hat{\mathcal{F}}|}\int_{\hat{\mathcal{F}}} |\hat{\zeta} - \bar{\hat{\zeta}}|\,d\hat{s}\right)^2 \le \lambda(\hat{D})|\hat{\zeta}|^2_{H^1(\hat{D})} \qquad \forall\,\hat{\zeta} \in H^1(\hat{D}), \tag{7.15}$$

where $\bar{\hat{\zeta}}$ is the mean of $\hat{\zeta}$ over $\hat{D}$ and $\lambda(\hat{D})$ is a positive constant depending only on $\hat{D}$. Combining (6.5) and (7.15) we find

$$\left(\frac{1}{|\mathcal{F}|}\int_{\mathcal{F}} |\zeta - \bar{\zeta}|\,ds\right)^2 = \left(\frac{1}{|\hat{\mathcal{F}}|}\int_{\hat{\mathcal{F}}} |\hat{\zeta} - \bar{\hat{\zeta}}|\,d\hat{s}\right)^2 \le C_\dagger \frac{\lambda(\hat{D})}{\operatorname{diam} D}|\zeta|^2_{H^1(D)} \tag{7.16}$$

for all $\zeta \in H^1(D)$, where $\bar{\zeta}$ is the mean of ζ over D, $\hat{\zeta} = \zeta \circ \alpha$, and $C_\dagger$ is a positive constant independent of $i \in I$. It follows from (7.16) and the finiteness of the number of reference polyhedra that the set

$$\{\lambda(\mathcal{P}_i) : i \in I\} \text{ is bounded.} \tag{7.17}$$

Let $i \in I$. We construct a triangulation $\mathcal{T}_i \in \mathfrak{T}_{\mathcal{P}_i}$ by first imposing the canonical triangulation on each face of $\mathcal{P}_i$ and then triangulating each member of $\mathcal{P}_i$ using its center and the triangles on its faces.

Since the number of edges on each face of $\mathcal{P}_i$ is limited by the number of edges appearing on the faces of the reference polyhedra, condition (iii) implies that the set

$$\{\rho(\mathcal{P}_i, \mathcal{T}_i) : i \in I\} \text{ is bounded.} \tag{7.18}$$

Moreover, it follows from conditions (ii) and (iii) that the triangulation induced by $\mathcal{T}_i$ on the boundary of any subdomain in $\mathcal{P}_i$ is quasi-uniform, which together with (6.6) implies

$$\inf\{\theta_{\mathcal{T}_i} : i \in I\} > 0\,. \tag{7.19}$$

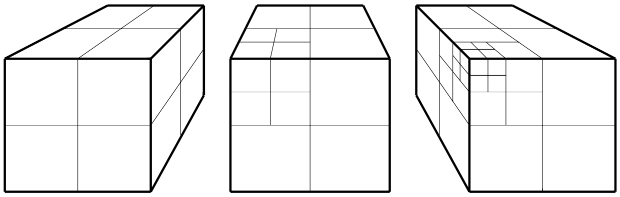


FIG. 7.1. *A family of partitions of a cube.*

Combining condition (ii) and (7.17)–(7.19), we see that the set

$$\{(\rho(\mathcal{P}_i), \lambda(\mathcal{P}_i), \rho(\mathcal{P}_i, \mathcal{T}_i), \theta_{\mathcal{T}_i}) : i \in I\}$$

is a precompact subset of $\mathbb{R}^4_+$. The estimate (7.14) then follows from (7.7) if we take C to be an upper bound of the bounded set

$$\{K(\rho(\mathcal{P}_i), \lambda(\mathcal{P}_i), \rho(\mathcal{P}_i, \mathcal{T}_i), \theta_{\mathcal{T}_i}) : i \in I\}. \qquad \square$$

Remark 7.4. If the family of partitions in Corollary 7.3 is actually a family of triangulations, then conditions (ii) and (iii) are redundant.

An example of a family of partitions satisfying the assumptions of Corollary 7.3 is depicted in Figure 7.1, where a cube is being refined successively towards the upper left corner in the front.

8. Concluding remarks. The approach to Poincaré–Friedrichs inequalities in this paper depends only on the classical Poincaré–Friedrichs inequalities which in turn depend only on the compactness of the embedding of $H^1(\Omega)$ in $L_2(\Omega)$. Since this is valid for any Ω satisfying the cone condition (cf. [1]), the results of this paper are also valid for domains with cracks.

We can, of course, also treat the one-dimensional case where Ω is an interval. In this case we can take the interpolant $\mathcal{I}\zeta \in H^1(\Omega)$ to be the piecewise linear function which takes the average value of ζ at the internal nodes of $\mathcal{P}$ and agrees with ζ at the endpoints. The resulting Poincaré–Friedrichs inequalities for $\zeta \in H^1(\Omega, \mathcal{P})$ are

$$(8.1) \qquad \|\zeta\|^2_{L_2(\Omega)} \le C \left(|\zeta|^2_{H^1(\Omega,\mathcal{P})} + \sum_{p\in\mathcal{P}_0} \sum_{e\in\Xi_p} |e|^{-1}[\zeta(p)]^2 + |\zeta(q)|^2 \right),$$

$$(8.2) \qquad \|\zeta\|^2_{L_2(\Omega)} \le C \left(|\zeta|^2_{H^1(\Omega,\mathcal{P})} + \sum_{p\in\mathcal{P}_0} \sum_{e\in\Xi_p} |e|^{-1}[\zeta(p)]^2 + \left|\int_\Omega \zeta\, dx\right|^2 \right),$$

where $\mathcal{P}_0$ is the set of the internal nodes of the partition $\mathcal{P}$, Ξ_p is the set of the two subintervals sharing p as an endpoint, $[\zeta(p)]$ is the jump of ζ across the point p, q is an endpoint of Ω, and C is a universal positive constant.

Let $\Omega = (0, 1)$, $\mathcal{P}_n$ be the uniform partition of Ω by n subintervals, and let ζ_n be the piecewise constant function defined by

$$(8.3) \qquad \zeta_n(x) = i \qquad \text{for} \quad \frac{i}{n} < x < \frac{i+1}{n} \ \text{ and } \ 0 \le i \le n-1.$$

Then $\zeta_n \in H^1(\Omega, \mathcal{P}_n)$ and $\zeta_n(0) = |\zeta_n|_{H^1(\Omega,\mathcal{P}_n)} = 0$. A straightforward calculation shows that both sides of (8.1) (with $q = 0$) grow at the rate of n^2 for the function ζ_n, and hence the weight $|e|^{-1}$ in (8.1) cannot be improved. If we define

$$\zeta_n(x) = -\frac{n-1}{2} + i \qquad \text{for} \quad \frac{i}{n} < x < \frac{i+1}{n} \ \text{ and } \ 0 \le i \le n-1, \tag{8.4}$$

then $\int_\Omega \zeta_n dx = |\zeta_n|_{H^1(\Omega,\mathcal{P}_n)} = 0$. Again both sides of (8.2) grow at the rate of n^2 for the function ζ_n, which shows that the weight $|e|^{-1}$ in (8.2) also cannot be improved.

Similarly, we can show that the weight $|\sigma|^{d/(1-d)}$ in (1.3) and (1.4) is sharp. Indeed, let Ω be the unit square $(0,1)^2$ and $\mathcal{P}_n$ be the uniform partition of Ω by n^2 squares. Consider the piecewise constant function (an analogue of the function in (8.3)) defined by

$$\zeta_n(x_1,x_2) = i \qquad \text{for} \quad \frac{i}{n} < x_1 < \frac{i+1}{n},\ \frac{j}{n} < x_2 < \frac{j+1}{n}, \ \text{ and } \ 0 \le i,j \le n-1.$$

We have $\zeta_n \in H^1(\Omega,\mathcal{P}_n)$ and $|\zeta_n|_{H^1(\Omega,\mathcal{P}_n)} = \zeta_n\big|_\Gamma = 0$, where Γ is the side $\{(0,t) : 0 < t < 1\}$. Again both sides of (1.3) grow at the rate of n^2 for ζ_n, and hence the weight $|\sigma|^{-2}$ cannot be improved. The sharpness of $|\sigma|^{-2}$ in (1.4) is established by constructing piecewise constant functions analogous to the ones defined by (8.4), and the sharpness of the weight $|\sigma|^{-3/2}$ for three-dimensional domains can be handled by similar constructions.

Finally, we remark that the techniques of this paper can also be used to derive Poincaré–Friedrichs inequalities for piecewise W_p^1 functions and $p \neq 2$.

REFERENCES

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

[2] D. N. Arnold, *An interior penalty finite element method with discontinuous elements*, SIAM J. Numer. Anal., 19 (1982), pp. 742–760.

[3] D. N. Arnold, F. Brezzi, B. Cockburn, and L. D. Marini, *Unified analysis of discontinuous Galerkin methods for elliptic problems*, SIAM J. Numer. Anal., 39 (2002), pp. 1749–1779.

[4] F. Ben Belgacem, *The mortar finite element method with Lagrange multipliers*, Numer. Math., 84 (1999), pp. 173–197.

[5] C. Bernardi, Y. Maday, and A. Patera, *A new nonconforming approach to domain decomposition: the mortar element method*, in Collége de France Seminar, H. Brezis and J.-L. Lions, eds., Pitman, Essex, 1990, pp. 13–51.

[6] C. Bernardi, Y. Maday, and A. Patera, *Domain decomposition by the mortar element method*, in Asymptotic and Numerical Methods for Partial Differential Equations with Critical Parameters, H. Kaper, M. Garby, and G. Pieper, eds., Kluwer Academic Publishers, Dordrecht, The Netherlands, 1993, pp. 269–286.

[7] J. H. Bramble and S. J. Hilbert, *Estimation of linear functionals on Sobolev spaces with applications to Fourier transforms and spline interpolation*, SIAM J. Numer. Anal., 7 (1970), pp. 112–124.

[8] S. C. Brenner, *Two-level additive Schwarz preconditioners for nonconforming finite element methods*, Math. Comp., 65 (1996), pp. 897–921.

[9] S. C. Brenner, *Convergence of nonconforming multigrid methods without full elliptic regularity*, Math. Comp., 68 (1999), pp. 25–53.

[10] S. C. Brenner and L. R. Scott, *The Mathematical Theory of Finite Element Methods*, 2nd ed., Springer-Verlag, New York, Berlin, Heidelberg, 2002.

[11] Z. Cai, J. Douglas, Jr., J. Santos, D. Sheen, and X. Ye, *Nonconforming quadrilateral finite elements: A correction*, Calcolo, 37 (2000), pp. 253–254.

[12] Z. Cai, J. Douglas, Jr., and X. Ye, *A stable nonconforming quadrilateral finite element method for the stationary Stokes and Navier-Stokes equations*, Calcolo, 36 (1999), pp. 215–232.

[13] P. Ciarlet, *The Finite Element Method for Elliptic Problems*, North-Holland, Amsterdam, 1978.

[14] B. Cockburn, G. Karniadakis, and C.-W. Shu, eds., *Discontinuous Galerkin Methods*, Springer-Verlag, Berlin, Heidelberg, 2000.

[15] M. Crouzeix and P.-A. Raviart, *Conforming and nonconforming finite element methods for solving the stationary Stokes equations* I, Rev. Française Automat. Informat. Recherche Opérationnelle Ser. Anal. Numér., 7 (1973), pp. 33–75.

[16] V. Dolejší, M. Feistauer, and J. Felcman, *On the discrete Friedrichs inequality for nonconforming finite elements*, Numer. Funct. Anal. Optim., 20 (1999), pp. 437–447.

[17] J. Douglas, Jr., J. Santos, D. Sheen, and X. Ye, *Nonconforming Galerkin methods based on quadrilateral elements for second order elliptic problems*, M2AN Math. Model. Numer. Anal., 33 (1999), pp. 747–770.

[18] M. Fortin, *A three-dimensional quadratic nonconforming element*, Numer. Math., 46 (1985), pp. 269–279.

[19] M. Fortin and M. Soulie, *A non-conforming piecewise quadratic finite element on triangles*, Internat. J. Numer. Methods Engrg., 19 (1983), pp. 505–520.

[20] J. Gopalakrishnan, *Mortar estimates independent of number of subdomains*, East-West J. Numer. Math., 8 (2000), pp. 111–125.

[21] H. Han, *A finite element approximation of Navier-Stokes equations using nonconforming elements*, J. Comput. Math., 2 (1984), pp. 77–88.

[22] P. Knobloch, *Uniform validity of discrete Friedrichs' inequality for general nonconforming finite element spaces*, Numer. Funct. Anal. Optim., 22 (2001), pp. 107–126.

[23] J. Nečas, *Les Méthodes Directes en Théorie des Équations Elliptiques*, Masson, Paris, 1967.

[24] S. Prudhomme, F. Pascal, J. Oden, and A. Romkes, *Review of A Priori Error Estimation for Discontinuous Galerkin Methods*, Tech. report 00-27, TICAM, Austin, TX, 2000.

[25] R. Rannacher and S. Turek, *Simple nonconforming quadrilateral Stokes element*, Numer. Methods Partial Differential Equations, 8 (1992), pp. 97–111.

[26] D. Stefanica, *Poincaré and Friedrichs Inequalities for Mortar Finite Element Methods*, Tech. report 1998-774, Courant Institute of Mathematical Sciences, New York, 1998.

[27] R. Temam, *Navier-Stokes Equations*, North-Holland, Amsterdam, 1977.

[28] J.-M. Thomas, *Sur l'analyse numérique des méthodes d'eléments finis hybrides et mixtes*, Ph.D. thesis, Université Pierre et Marie Curie, Paris, 1977.

[29] J. Wloka, *Partial Differential Equations*, Cambridge University Press, Cambridge, UK, 1987.

[30] B. I. Wohlmuth, *A mortar finite element method using dual spaces for the Lagrange multiplier*, SIAM J. Numer. Anal., 38 (2000), pp. 989–1012.

SIAM J. Numer. Anal.
Vol. 41, No. 1, pp. 325–363

ORDERED UPWIND METHODS FOR STATIC HAMILTON–JACOBI EQUATIONS: THEORY AND ALGORITHMS*

JAMES A. SETHIAN† and ALEXANDER VLADIMIRSKY‡

Abstract. We develop a family of fast methods for approximating the solutions to a wide class of static Hamilton–Jacobi PDEs; these fast methods include both semi-Lagrangian and fully Eulerian versions. Numerical solutions to these problems are typically obtained by solving large systems of coupled nonlinear discretized equations. Our techniques, which we refer to as "Ordered Upwind Methods" (OUMs), use partial information about the characteristic directions to decouple these nonlinear systems, greatly reducing the computational labor. Our techniques are considered in the context of control-theoretic and front-propagation problems.

We begin by discussing existing OUMs, focusing on those designed for isotropic problems. We then introduce a new class of OUMs which decouple systems for general (anisotropic) problems. We prove convergence of one such scheme to the viscosity solution of the corresponding Hamilton–Jacobi PDE. Next, we introduce a set of finite-differences methods based on an analysis of the role played by anisotropy in the context of front propagation and optimal trajectory problems.

The performance of the methods is analyzed, and computational experiments are performed using test problems from computational geometry and seismology.

Key words. ordered upwind methods, fast marching methods, Dijkstra-like methods, anisotropic optimal control, dynamic programming, viscosity solution, anisotropic front propagation

AMS subject classifications. 65N12, 65N06, 49L20, 49L25, 49N90, 35F30, 35B37

PII. S0036142901392742

1. Introduction.

In this paper we present a family of noniterative methods applicable to the boundary value problem for static Hamilton–Jacobi equations of the form[1]

$$(1)\qquad \begin{array}{lcll} H(\nabla u, \boldsymbol{x}) & = & 1, & \boldsymbol{x} \in \Omega \subset R^2, \\ u(\boldsymbol{x}) & = & q(\boldsymbol{x}), & \boldsymbol{x} \in \partial\Omega, \end{array}$$

where the Hamiltonian H is assumed to be Lipschitz-continuous, convex, and homogeneous of degree 1 in the first argument:

$$(2)\qquad H(\nabla u, \boldsymbol{x}) = \|\nabla u\| F\left(\boldsymbol{x}, \frac{\nabla u}{\|\nabla u\|}\right)$$

for some function F. We will further assume that the function q is also Lipschitz-continuous, and that

$$0 < F_1 \le F(\boldsymbol{x}, \boldsymbol{p}) \le F_2,$$
$$q_1 \le q(\boldsymbol{x}) \le q_2$$

*Received by the editors July 23, 2001; accepted for publication (in revised form) August 30, 2002; published electronically April 9, 2003.

http://www.siam.org/journals/sinum/41-1/39274.html

†Department of Mathematics, University of California at Berkeley, Berkeley, CA 94720 (sethian@math.berkeley.edu). The research of this author was supported in part by the Applied Mathematical Sciences, Office of Energy Research, U.S. Department of Energy, grant DE-AC03-76SF00098, and by the Division of Mathematical Sciences, National Science Foundation.

‡Department of Mathematics, Cornell University, Ithaca, NY 14853 (vlad@math.cornell.edu). The research of this author was supported in part by a National Science Foundation Postdoctoral Research Fellowship.

[1] For the sake of notational clarity we restrict our discussion to R^2; all results can be restated for R^n and for meshes on manifolds.

for all $\boldsymbol{x} \in \Omega$ and for all the vectors of unit length $\boldsymbol{p}$.

Even for arbitrarily smooth H, q, and $\partial\Omega$, a smooth solution on Ω need not exist. In general, there are infinitely many weak Lipschitz-continuous solutions, but the unique *viscosity solution* can be defined using additional conditions on the smooth test functions (see [10, 9]).

To obtain a numerical solution, one often starts with a mesh X covering the domain Ω. Let $U_i = U(\boldsymbol{x_i})$ be the numerical solution at the mesh point $\boldsymbol{x_i} \in X$. Denote the set of mesh points adjacent to $\boldsymbol{x_i}$ as $N(\boldsymbol{x_i})$ and the set of values adjacent to U_i as $NU(\boldsymbol{x_i}) = \{U_j | \boldsymbol{x_j} \in N(\boldsymbol{x_i})\}$. Let $\overline{H}$ be a consistent discretization of H such that one can write

$$\overline{H}(U_i, NU(\boldsymbol{x_i}), \boldsymbol{x_i}) = 1. \tag{3}$$

If M is the total number of mesh points, then one needs to solve M coupled nonlinear equations simultaneously. One typical approach is to solve this nonlinear system iteratively.

Our ultimate goal is to introduce a set of "single-pass" numerical methods. By this, we mean that each U_i is recalculated at most r times, where r depends only upon the PDE (1) and the mesh structure and not upon the number of mesh points.

To construct single-pass algorithms with efficient update orderings, one can utilize the fact that the value of $u(\boldsymbol{x})$ for the first-order PDE depends only on the value of u along the characteristic(s) passing through the point $\boldsymbol{x}$. If $\boldsymbol{x_{i_1}}, \boldsymbol{x_{i_2}} \in N(\boldsymbol{x_i})$ are such that the characteristic for the mesh point $\boldsymbol{x_i}$ lies in the simplex $\boldsymbol{x_i x_{i_1} x_{i_2}}$, then it is useful to consider an *upwind* discretization of the PDE:

$$\overline{H}(U_i, U_{i_1}, U_{i_2}, \boldsymbol{x_i}) = 1. \tag{4}$$

This reduces the coupling in the system: U_i *depends* only upon U_{i_1} and U_{i_2} and not on all of the $NU(\boldsymbol{x_i})$. A recursive construction allows one to build the entire *dependency graph* for $\boldsymbol{x_i}$.

If two or more characteristics collide at the point $\boldsymbol{x}$, the solution loses smoothness. The entropy condition does not allow characteristics to be created at these collision points; hence, if $\boldsymbol{x_i}$ is far enough from these collision points, then, for a suitably chosen discretization, its dependency graph is actually a tree. If the characteristic directions of the PDE were known in advance, then the dependency-ordering of the grid points would be known as well, leading to a fully decoupled system. Formally, this construction would lead to an $O(M)$ method.

In general, characteristic directions are not known in advance due to the nonlinearity of (1). Nonetheless, single-pass methods can be devised to determine the mesh point ordering (and the characteristic directions) in the process of decoupling the system. We refer to such methods as "Ordered Upwind Methods" (OUMs) and show that they have computational complexity of $O(M \log M)$.

Since (1) can be interpreted as a description of a continuous control problem, we start in section 2 by viewing the discrete control problem and by considering Dijkstra's method as a prototype for the OUMs to be built for the continuous case.

Next, in section 3, we view the Hamilton–Jacobi PDE (1) as an anisotropic min-time optimal trajectory problem. In this control-theoretic setting, the speed of a vehicle's motion depends not only on its position, but also on the direction. The corresponding value function u is the viscosity solution of the static Hamilton–Jacobi–

Bellman equation

$$\begin{aligned} &\max_{\boldsymbol{a}\in S_1}\{(\nabla u(\boldsymbol{x})\cdot(-\boldsymbol{a}))f(\boldsymbol{x},\boldsymbol{a})\}=1, \quad && \boldsymbol{x}\in\Omega,\\ &u(\boldsymbol{x})=q(\boldsymbol{x}), && \boldsymbol{x}\in\partial\Omega. \end{aligned} \tag{5}$$

Here, $\boldsymbol{a}$ is the unit vector determining the direction of motion, $f(\boldsymbol{x},\boldsymbol{a})$ is the speed of motion in the direction $\boldsymbol{a}$ starting from the point $\boldsymbol{x}\in\Omega$, and $q(\boldsymbol{x})$ is the time-penalty for exiting the domain at the point $\boldsymbol{x}\in\partial\Omega$. The maximizer $\boldsymbol{a}$ corresponds to the characteristic direction for the point $\boldsymbol{x}$.

If the speed functions F and f depend only upon their first argument, both forms of the Hamilton–Jacobi PDE reduce to the Eikonal equation

$$\|\nabla u(\boldsymbol{x})\| = K(\boldsymbol{x}), \tag{6}$$

where $K(\boldsymbol{x})=\frac{1}{F(\boldsymbol{x})}=\frac{1}{f(\boldsymbol{x})}$. In this case, the characteristics of the PDE coincide with the gradient lines of its viscosity solution u. This property is the foundation for two single-pass methods for the Eikonal equation: Tsitsiklis' algorithm (1994) and Sethian's Fast Marching Method (1996), representing Dijkstra-like decoupling of a semi-Lagrangian and a fully Eulerian discretization of (1), respectively (section 4).

We then proceed to the central part of this paper. First, we show (in section 5) that neither of these single-pass Eikonal solvers can be directly applied in the general anisotropic case. Nonetheless, the underlying ideas can be used to build the OUMs which are applicable for much more general equations. Such methods hinge on two properties of the unique viscosity solution:

1. The viscosity solution $u(\boldsymbol{x})$ is strictly increasing along the characteristics of the PDE (1).
2. We can derive a precise upper bound on the maximum angle between the characteristic and the gradient of u.

In section 6, we introduce the first general OUM with computational complexity of $O(\frac{F_2}{F_1}M\log M)$ based on a semi-Lagrangian discretization; an announcement of this algorithm without details or proof was first made in [39]. The method's convergence to the viscosity solution is proven in section 7.

Next, in section 8, we reinterpret the Hamilton–Jacobi PDE (1), this time as describing an anisotropic front expansion (contraction) problem. In this context, $F(\boldsymbol{x},\boldsymbol{n})$ is interpreted as the speed of the front in the normal direction $\boldsymbol{n}$, and $\partial\Omega$ as the initial position of the front.

$$\begin{aligned} &\|\nabla u\|F\left(\boldsymbol{x},\tfrac{\nabla u}{\|\nabla u\|}\right)=1, \quad && \boldsymbol{x}\in\Omega,\\ &u(\boldsymbol{x})=0, && \boldsymbol{x}\in\partial\Omega. \end{aligned}$$

The anisotropy is the result of the dependence of F on $\boldsymbol{n}$. The level sets of the viscosity solution u correspond to the positions of the front at different times. We consider only those front expansion (contraction) problems in which the speed F is such that the Hamiltonian H is convex.

The fully Eulerian OUMs, introduced in section 8, use the finite-difference approximations developed as a generalization of the Fast Marching Method and are based on the analysis of the role played by anisotropy in the front propagation and optimal trajectory problems. These single-pass methods also have the same computational complexity of $O(\frac{F_2}{F_1}M\log M)$. The appendix examines the relationship between the first-order semi-Lagrangian and Eulerian OUMs.

Finally, in section 9, we analyze the efficiency of the new methods and consider several anisotropic test problems from computational geometry and seismology.

2. Dijkstra's method and discrete optimal trajectories. We begin by considering the problem of computing the shortest path on a network. (See, for example, [3] for a catalogue of available algorithms.) Computing the shortest path can be viewed as a discrete dynamic programming problem. Here, it serves as a simpler analogue for the continuous optimal trajectory problem considered in the next section.

For the case in which the network is sparsely connected and all arc-costs are positive, the heap-sort version of Dijkstra's method [12] is one of the most widely used algorithms. We will now reinterpret Dijkstra's method as a single-pass OUM since it serves as a model for building the OUMs for the continuous front propagation and optimal trajectory problems.

2.1. Shortest paths and value function. Consider a discrete network of nodes $X = \{\boldsymbol{x_1}, \dots, \boldsymbol{x_M}\}$. A vehicle starts somewhere in the network and travels from node to node until it reaches one of the *exit nodes* $\boldsymbol{x} \in Q \subset X$. A vehicle's trajectory is a sequence of nodes $(\boldsymbol{y_1}, \dots, \boldsymbol{y_r})$ such that $\boldsymbol{y_r} \in Q$ and $\boldsymbol{y_k} \notin Q$ for $k < r$. There is a *time-penalty* $K(\boldsymbol{x_i}, \boldsymbol{x_j}) = K_{ij} > 0$ for passing from $\boldsymbol{x_i}$ to $\boldsymbol{x_j}$. ($K_{ij} = +\infty$ if there is no link from $\boldsymbol{x_i}$ to $\boldsymbol{x_j}$.) For all $\boldsymbol{x} \in Q$ there is an *exit time-penalty* $q(\boldsymbol{x}) < \infty$. Thus, the total time needed for a trajectory $(\boldsymbol{y_1}, \dots, \boldsymbol{y_r})$ is

$$\text{TotalTime}(\boldsymbol{y_1}, \dots, \boldsymbol{y_r}) = \sum_{j=1}^{r} K(\boldsymbol{y_j}, \boldsymbol{y_{j+1}}) + q(\boldsymbol{y_r}). \tag{7}$$

The goal is to find the optimal trajectory for each node $\boldsymbol{x} \in X \backslash Q$.

The key idea of *dynamic programming* [5, 6] is to solve for all of the nodes at once. Instead of searching for a particular optimal trajectory, one derives an equation for the *value function* $U(\boldsymbol{x})$, defined as the minimum time to exit the network if one starts at $\boldsymbol{x}$:

$$\begin{cases} U(\boldsymbol{x}) = \min\limits_{\substack{\text{all the paths} \\ \text{starting at } \boldsymbol{x}}} \text{TotalTime}(\boldsymbol{x}, \dots), & \boldsymbol{x} \in X \backslash Q, \\ U(\boldsymbol{x}) = q(\boldsymbol{x}), & \boldsymbol{x} \in Q. \end{cases} \tag{8}$$

Bellman's optimality principle [5] shows the relationship between $U(\boldsymbol{x})$ and the values of U on the set of adjacent nodes $N(\boldsymbol{x}) = \{\boldsymbol{y} \in X \mid K(\boldsymbol{x}, \boldsymbol{y}) < \infty\}$, namely,

$$U(\boldsymbol{x}) = \min_{\boldsymbol{y} \in N(\boldsymbol{x})} \{K(\boldsymbol{x}, \boldsymbol{y}) + U(\boldsymbol{y})\} \quad \text{for all } \boldsymbol{x} \in X \backslash Q. \tag{9}$$

Equation (9) is nonlinear, and it has to hold for each node in $X \backslash Q$. Thus, if there are M such nodes, we have to solve a coupled system of M nonlinear equations.

2.2. Dijkstra's method. Dijkstra's method [12] provides a way of decoupling system (9) and is based on the following monotonicity observations.

OBSERVATION 2.1. *If* $(\boldsymbol{y_1}, \dots, \boldsymbol{y_r})$ *is an optimal trajectory for* $\boldsymbol{y_1}$, *then we have* $U(\boldsymbol{y_1}) > \dots > U(\boldsymbol{y_r})$.

OBSERVATION 2.2. *If* $N_-(\boldsymbol{x}) = \{\boldsymbol{y} \in N(\boldsymbol{x}) \mid U(\boldsymbol{y}) < U(\boldsymbol{x})\}$, *then Bellman's equation* (9) *can be rewritten as*

$$U(\boldsymbol{x}) = \min_{\boldsymbol{y} \in N_-(\boldsymbol{x})} \{K(\boldsymbol{x}, \boldsymbol{y}) + U(\boldsymbol{y})\} \quad \textit{for all } \boldsymbol{x} \in X \backslash Q. \tag{10}$$

If the nodes were somehow sorted by the value of U, one could solve equations (10) one by one, yielding a method with an overall complexity of $O(M)$. Even though this ordering on X is not known in advance, Dijkstra's method reconstructs it (one node at a time) as follows.

All the nodes are divided into three classes: *Far* (no information about the correct value of U is known), *Accepted* (the correct value of U has been computed), and *Considered* (adjacent to *Accepted*). For every *Considered* $\boldsymbol{x}$ we define the set $\mathrm{NF}(\boldsymbol{x}) = \{\boldsymbol{y} \in N(\boldsymbol{x}) \,|\boldsymbol{y} \text{ is } \textit{Accepted}\}$.

1. Start with all the nodes in *Far*.
2. Move the exit nodes ($\boldsymbol{y} \in Q$) to *Accepted* ($U(\boldsymbol{y}) = q(\boldsymbol{y})$).
3. Move all the nodes $\boldsymbol{x}$ adjacent to the boundary into *Considered* and evaluate the tentative values

$$V(\boldsymbol{x}) := \min_{\boldsymbol{y} \in \mathrm{NF}(\boldsymbol{x})} \{K(\boldsymbol{x},\boldsymbol{y}) + U(\boldsymbol{y})\}\,. \tag{11}$$

4. Find the node $\bar{\boldsymbol{x}}$ with the smallest value of V among all the *Considered*.
5. Move $\bar{\boldsymbol{x}}$ to *Accepted* ($U(\bar{\boldsymbol{x}}) = V(\bar{\boldsymbol{x}})$).
6. Move the *Far* nodes adjacent to $\bar{\boldsymbol{x}}$ into *Considered*.
7. Reevaluate V for all the *Considered* $\boldsymbol{x}$ adjacent to $\bar{\boldsymbol{x}}$

$$V(\boldsymbol{x}) := \min\{V(\boldsymbol{x}), K(\boldsymbol{x},\bar{\boldsymbol{x}}) + U(\bar{\boldsymbol{x}})\}\,. \tag{12}$$

8. If *Considered* is not empty, then go to 4.

The described algorithm has the computational complexity of $O(M\log(M))$; the factor of $\log(M)$ reflects the necessity to maintain a sorted list of the *Considered* values $V(\boldsymbol{x_i})$ to determine the next *Accepted* node.[2]

On a grid-like network, we can reinterpret Dijkstra's method as an upwind finite difference scheme. Consider a uniform Cartesian grid of grid size h, where the time-penalty $K_{ij} > 0$ is given for passing through each grid point $x_{ij} = (ih, jh)$. The minimal total time-to-exit U_{ij} starting from the node $\boldsymbol{x}_{ij}$ can be written in terms of the minimal total time-to-exit starting at its neighbors:

$$U_{ij} = \min\left(U_{i-1,j}, U_{i+1,j}, U_{i,j-1}, U_{i,j+1}\right) + K_{ij}. \tag{13}$$

As pointed out by Sethian in [36], the U_{ij} obtained through Dijkstra's method is formally a first-order approximation to the solution $u(x,y)$ of the differential equation

$$H(\nabla u(\boldsymbol{x}), u(\boldsymbol{x}), \boldsymbol{x}) = \max(|u_x|, |u_y|) = K(\boldsymbol{x}), \tag{14}$$

provided that the time-penalties are $K_{ij} = hK(\boldsymbol{x})$.

3. Continuous optimal trajectory problems and semi-Lagrangian discretization.

3.1. Statement of problem.

Consider an optimal trajectory problem for a vehicle moving inside the domain Ω, with the speed f depending upon the direction of motion and the current position of the vehicle inside the domain. The dynamics of the vehicle is defined by

$$\begin{aligned} \boldsymbol{y}'(t) &= f(\boldsymbol{y}(t), \boldsymbol{a}(t))\boldsymbol{a}(t), \\ \boldsymbol{y}(0) &= \boldsymbol{x} \ \in \Omega, \end{aligned} \tag{15}$$

[2] This variant of Dijkstra's method is often referred to as a *heap-sort Dijkstra's method* since its implementation requires the use of a binary heap, d-heap, or Fibonacci heap to maintain the ordering of the *Considered* nodes efficiently [3].

The complexity estimate for the densely connected network would be $O(M^2\log(M))$, but for our case, when X is a grid or a mesh, the precise complexity estimate is $O(rM\log(M))$, where r is the maximum number of nodes connected to a single node in X.

where $\boldsymbol{y}(t)$ is the position of the vehicle at time t, $S_1 = \{\boldsymbol{a} \in R^2 \mid \|\boldsymbol{a}\| = 1\}$ is the set of *admissible control values*, and $\mathcal{A} = \{\boldsymbol{a} : R_{+,0} \mapsto S_1 \mid \boldsymbol{a}(\cdot) \text{ is measurable}\}$ is the set of *admissible controls*. We are interested in studying $\boldsymbol{y}(t)$ only while the vehicle remains inside Ω, i.e., until the exit time

$$T(\boldsymbol{x}, \boldsymbol{a}(\cdot)) = \inf\{t \in R_{+,0} | \boldsymbol{y}(t) \in \partial\Omega\}.$$

If the function $q(\boldsymbol{x}) \geq 0$ is the *time-penalty* for exiting the domain at the point $\boldsymbol{x} \in \partial\Omega$, then a *min-time* optimal trajectory problem is the task of finding an optimal control $\boldsymbol{a}(\cdot)$ which minimizes the total time:[3]

$$\text{TotalTime}\left(\boldsymbol{x}, \boldsymbol{a}(\cdot)\right) = T\left(\boldsymbol{x}, \boldsymbol{a}(\cdot)\right) + q\left(\boldsymbol{y}\left(T(\boldsymbol{x}, \boldsymbol{a}(\cdot))\right)\right).$$

We will alternatively refer to the above quantity as a *total cost* of using the control: $\text{Cost}\left(\boldsymbol{x}, \boldsymbol{a}(\cdot)\right) = \text{TotalTime}\left(\boldsymbol{x}, \boldsymbol{a}(\cdot)\right)$.

Unless otherwise explicitly specified, we will assume that both f and q are Lipschitz-continuous and that there exist constants f_1, f_2, q_1, q_2 such that

$$\begin{aligned} 0 < f_1 \leq f(\boldsymbol{x}, \boldsymbol{a}) \leq f_2 < \infty \quad & \text{for all } \boldsymbol{x} \in \Omega \text{ and } \boldsymbol{a} \in S_1, \\ 0 < q_1 \leq q(\boldsymbol{x}) \leq q_2 < \infty \quad & \text{for all } \boldsymbol{x} \in \partial\Omega. \end{aligned} \tag{16}$$

For notational convenience, we will also define the *anisotropy coefficient* $\Upsilon = \frac{f_2}{f_1}$. Strictly speaking, since f_1 and f_2 are global bounds, the coefficient Υ reflects the measure of anisotropy only in the homogeneous domain (i.e., when $f(\boldsymbol{x}, \boldsymbol{a}) = f(\boldsymbol{a})$). We will use Υ in deriving the worst-case-scenario computational complexity of the algorithms. In section 9.2, the more accurate *local anisotropy coefficient* $\Upsilon(\boldsymbol{x})$ will be defined and used for a more detailed computational complexity analysis.

As in the discrete case, the key idea of dynamic programming [5] is to define the value function $u(\boldsymbol{x})$ such that

$$\begin{cases} u(\boldsymbol{x}) = \inf_{\boldsymbol{a}(\cdot)} \text{Cost}(\boldsymbol{x}, \boldsymbol{a}(\cdot)), & \boldsymbol{x} \in \Omega \backslash \partial\Omega, \\ u(\boldsymbol{x}) = q(\boldsymbol{x}), & \boldsymbol{x} \in \partial\Omega. \end{cases} \tag{17}$$

In general, an optimal control $\boldsymbol{a}(\cdot)$ does not have to exist; therefore, when proving properties of the value function u, one uses *ϵ-suboptimal controls* $\boldsymbol{a}(\cdot)$ such that $\text{Cost}(\boldsymbol{x}, \boldsymbol{a}(\cdot)) < u(\boldsymbol{x}) + \epsilon$. To simplify the presentation, we will somewhat loosely refer to the optimal controls and trajectories. If such optimal controls do not exist, the corresponding properties can be formulated and proven for the ϵ-suboptimal controls and trajectories.

3.2. Properties of the value function. The following lemmas enumerate several well-known properties of the value function (see proofs in [46], for example). In section 7 we will prove the similar properties of the numerical approximation constructed by the OUM.

LEMMA 3.1 (Fixed-time optimality principle). *Let $d(\boldsymbol{x})$ be the minimum distance to the boundary $\partial\Omega$. Then for every point $\boldsymbol{x} \in \Omega \backslash \partial\Omega$ and for any $\tau < d(\boldsymbol{x})$*

$$u(\boldsymbol{x}) = \tau + \inf_{\boldsymbol{a}(\cdot)} \left\{ u(\boldsymbol{y}(\tau)) \right\}, \tag{18}$$

[3] A different optimal trajectory problem can be formulated for minimizing the total cost of moving the vehicle with a unit speed, when the running cost depends upon both the vehicle's position and the direction of motion; see [45, 15], for example. It is not hard to show the equivalence of this *min-cost* problem to the *min-time* optimal trajectory problem considered here [46].

where $\boldsymbol{y}(\cdot)$ is a trajectory corresponding to a chosen control $\boldsymbol{a}(\cdot)$.

LEMMA 3.2 (Fixed-boundary optimality principle). *Consider a simple closed curve $\Gamma \subset \Omega \backslash \partial\Omega$ and an arbitrary point $\boldsymbol{x}$ inside Γ. For every control $\boldsymbol{a}(\cdot)$, we define $T_\Gamma(\boldsymbol{x}, \boldsymbol{a}(\cdot))$ to be the earliest time at which the corresponding trajectory $\boldsymbol{y}(\cdot)$ reaches the curve Γ. Then*

$$u(\boldsymbol{x}) = \inf_{\boldsymbol{a}(\cdot)} \left\{ T_\Gamma\left(\boldsymbol{x}, \boldsymbol{a}(\cdot)\right) + u\left(\boldsymbol{y}\left(T_\Gamma(\boldsymbol{x}, \boldsymbol{a}(\cdot))\right)\right) \right\}. \tag{19}$$

LEMMA 3.3. *The value function $u(\boldsymbol{x})$ is Lipschitz-continuous*[4] *and bounded on $\Omega \backslash \partial\Omega$. If $\boldsymbol{y}'(t) = \boldsymbol{a}(t)$ defines an optimal trajectory for a point $\boldsymbol{x}$ (i.e., $\boldsymbol{y}(0) = \boldsymbol{x}$ and $u(\boldsymbol{x}) = \mathrm{Cost}(\boldsymbol{x}, \boldsymbol{a}(\cdot))$), then the function $u(\boldsymbol{y}(t))$ is strictly decreasing for $t \in [0, T(\boldsymbol{x}, \boldsymbol{a}(\cdot))]$.*

The following two lemmas utilize the *fixed-time optimality principle* and provide the key motivation for constructing OUMs for this problem (sections 6 and 8).

LEMMA 3.4. *Consider a point $\bar{\boldsymbol{x}} \in \Omega \backslash \partial\Omega$. Then, for any constant C such that $q_2 \leq C \leq u(\bar{\boldsymbol{x}})$, the optimal trajectory for $\bar{\boldsymbol{x}}$ will intersect the level set $u(\boldsymbol{x}) = C$ at some point $\tilde{\boldsymbol{x}}$. If $\bar{\boldsymbol{x}}$ is distance d_1 away from that level set, then*

$$\|\tilde{\boldsymbol{x}} - \bar{\boldsymbol{x}}\| \leq d_1 \Upsilon. \tag{20}$$

Proof. Let $\boldsymbol{a}(\cdot)$ be an optimal control for $\bar{\boldsymbol{x}}$. The intersection point $\tilde{\boldsymbol{x}} = \boldsymbol{y}(\tau)$ exists because of the continuity of the value function and of the optimal trajectory: $u(\bar{\boldsymbol{x}}) \geq C \geq q_2 \geq u(\boldsymbol{y}(T(\bar{\boldsymbol{x}}, \boldsymbol{a}(\cdot))))$.

Therefore,

$$u(\bar{\boldsymbol{x}}) = \tau \, + \, u(\tilde{\boldsymbol{x}}) \geq \frac{\|\tilde{\boldsymbol{x}} - \bar{\boldsymbol{x}}\|}{f_2} + C.$$

There also exists some point $\hat{\boldsymbol{x}}$ on the level set such that $\|\bar{\boldsymbol{x}} - \hat{\boldsymbol{x}}\| = d_1$. Consider a control $\boldsymbol{a_1}(t) = \frac{\hat{\boldsymbol{x}} - \bar{\boldsymbol{x}}}{d_1}$, and suppose it takes time τ_1 to reach $\hat{\boldsymbol{x}}$ along the corresponding straight-line trajectory. By the optimality principle,

$$u(\bar{\boldsymbol{x}}) \leq \tau_1 \, + \, u(\hat{\boldsymbol{x}}) \leq \frac{d_1}{f_1} + C.$$

Thus, $\|\tilde{\boldsymbol{x}} - \bar{\boldsymbol{x}}\| \leq d_1 \Upsilon$. □

LEMMA 3.5. *Consider an unstructured (triangulated) mesh X of diameter h on Ω. Consider a simple closed curve $\Gamma \subset \Omega \backslash \partial\Omega$ with the property that for any point $\boldsymbol{x}$ on Γ there exists a mesh point $\hat{\boldsymbol{x}}$ inside Γ such that $\|\boldsymbol{x} - \hat{\boldsymbol{x}}\| < h$. Suppose a mesh point $\bar{\boldsymbol{x}}$ is such that $u(\bar{\boldsymbol{x}}) \leq u(\boldsymbol{x_i})$ for all the mesh points $\boldsymbol{x_i} \in X$ inside Γ. The optimal trajectory for $\bar{\boldsymbol{x}}$ will intersect Γ at some point $\tilde{\boldsymbol{x}}$ such that*

$$\|\tilde{\boldsymbol{x}} - \bar{\boldsymbol{x}}\| \leq h \Upsilon. \tag{21}$$

Proof. Let $\boldsymbol{a}(\cdot)$ be an optimal control for $\bar{\boldsymbol{x}}$. The intersection point $\tilde{\boldsymbol{x}} = \boldsymbol{y}(\tau)$ exists because of the continuity of Γ and of the optimal trajectory. Therefore,

$$u(\bar{\boldsymbol{x}}) = \tau \, + \, u(\tilde{\boldsymbol{x}}) \geq \frac{\|\tilde{\boldsymbol{x}} - \bar{\boldsymbol{x}}\|}{f_2} + \, u(\tilde{\boldsymbol{x}}).$$

[4] This holds in the interior of Ω even in the presence of *state constraints*: the assumption $f_1 > 0$ is sufficient for the *local controllability* near $\partial\Omega$ (as defined in [4], for example).

Let $\hat{\boldsymbol{x}}$ be a mesh point inside Γ such that $\|\tilde{\boldsymbol{x}} - \hat{\boldsymbol{x}}\| \leq h$. Consider a control $\boldsymbol{a_1}(t) = \frac{\tilde{\boldsymbol{x}}-\hat{\boldsymbol{x}}}{\|\tilde{\boldsymbol{x}}-\hat{\boldsymbol{x}}\|}$, and let τ_1 be the time required to reach $\tilde{\boldsymbol{x}}$ from $\hat{\boldsymbol{x}}$ using $\boldsymbol{a_1}(\cdot)$. Then, by the optimality principle,

$$u(\hat{\boldsymbol{x}}) \leq \tau_1 \,+\, u(\tilde{\boldsymbol{x}}) \leq \frac{h}{f_1} \,+\, u(\tilde{\boldsymbol{x}}).$$

To complete the proof, we recall that $u(\bar{\boldsymbol{x}}) \leq u(\hat{\boldsymbol{x}})$. □

3.3. Hamilton–Jacobi–Bellman PDE. As in the discrete case, Bellman's optimality principle can be used to formally derive the local equation for $u(\boldsymbol{x})$ if the value function is smooth around $\boldsymbol{x}$:

$$\begin{aligned} \min_{\boldsymbol{a}\in S_1}\{(\nabla u(\boldsymbol{x}) \cdot \boldsymbol{a}) f(\boldsymbol{x},\boldsymbol{a})\} \,+\, 1 = 0, \quad \boldsymbol{x} \in \Omega, \\ u(\boldsymbol{x}) = q(\boldsymbol{x}), \quad \boldsymbol{x} \in \partial\Omega. \end{aligned} \tag{22}$$

The above *Hamilton–Jacobi–Bellman PDE* can be rewritten in the form $H(\nabla u, \boldsymbol{x}) = 1$, where the Hamiltonian $H = -\min_{\boldsymbol{a}\in S_1}\{(\boldsymbol{p}\cdot\boldsymbol{a}) f(\boldsymbol{x},\boldsymbol{a})\} = \max_{\boldsymbol{a}\in S_1}\{(\boldsymbol{p}\cdot(-\boldsymbol{a})) f(\boldsymbol{x},\boldsymbol{a})\}$. Moreover, this Hamiltonian is convex and homogeneous of degree one in the first argument; thus, this PDE belongs to the class of problems described in section 1. We also note that the characteristics of this PDE can be formally shown to be the optimal trajectories for the corresponding min-time control problem.

In an important case of isotropic optimal speed function ($f(\boldsymbol{x},\boldsymbol{a}) = f(\boldsymbol{x})$), equation (22) reduces to the *Eikonal equation* $\|\nabla u(\boldsymbol{x})\| = \frac{1}{f(\boldsymbol{x})}$. We particularly emphasize one property of the Eikonal equations: if ∇u is defined at the point, then the minimizer is $\boldsymbol{a} = \frac{-\nabla u}{\|\nabla u\|}$. Thus, the gradient lines of $u(\boldsymbol{x})$ coincide with the characteristics of the Eikonal PDE (i.e., the optimal trajectories for the isotropic min-time control problem). This is the main reason for the following *causality property*, a foundation for the noniterative Eikonal solvers.

PROPERTY 3.6 (Causality for the Eikonal equation). *If $\nabla u(\boldsymbol{x})$ is defined and $\boldsymbol{x}\boldsymbol{x_1}\boldsymbol{x_2}$ is a sufficiently small acute simplex, which contains the characteristic for $\boldsymbol{x}$, then $u(\boldsymbol{x}) \geq \max\{u(\boldsymbol{x_1}), u(\boldsymbol{x_2})\}$.*

Unfortunately, a smooth solution to (22) might not exist even for smooth f, q, and $\partial\Omega$. Generally, this equation has infinitely many weak Lipschitz-continuous solutions, but the unique *viscosity solution* [10] can be defined using the conditions on smooth test functions [9] as follows.

A bounded, uniformly continuous function u is the *viscosity solution* of (22) if the following holds for each smooth test function[5] $\phi \in C_c^\infty(\Omega)$:

(i) if $u - \phi$ has a local minimum at $\boldsymbol{x_0} \in \Omega$, then

$$\min_{\boldsymbol{a}\in S_1}\{(\nabla\phi(\boldsymbol{x_0}) \cdot \boldsymbol{a}) f(\boldsymbol{x_0},\boldsymbol{a})\} + 1 \leq 0; \tag{23}$$

(ii) if $u - \phi$ has a local maximum at $\boldsymbol{x_0} \in \Omega$, then

$$\min_{\boldsymbol{a}\in S_1}\{(\nabla\phi(\boldsymbol{x_0}) \cdot \boldsymbol{a}) f(\boldsymbol{x_0},\boldsymbol{a})\} + 1 \geq 0. \tag{24}$$

Moreover, the optimality principle (Lemma 3.1) can be used to demonstrate that the value function of the min-time optimal trajectory problem satisfies the inequalities

[5] The standard definition of the viscosity solution (see [9, 8], for example) uses the test functions $\phi \in C^1(\Omega)$. However, as shown in [9], the definition using the test functions $\phi \in C_c^\infty(\Omega)$ is equivalent. This second formulation enables us to use the upper bounds on the second derivatives of ϕ in the convergence proof in section 7.

(24) and (23) and thus is the viscosity solution of the Hamilton–Jacobi–Bellman PDE (see [8, 7] or [16], for example[6]).

3.4. Modified definition of the viscosity solution. Define $S_1^{\phi,\boldsymbol{x}} = \{\boldsymbol{a} \in S_1 \mid \boldsymbol{a}\cdot\nabla\phi(\boldsymbol{x}) \leq -\|\nabla\phi(\boldsymbol{x})\|\Upsilon^{-1}\}$. In [46] we demonstrate that using $S_1^{\phi,\boldsymbol{x_0}}$ instead of S_1 in the inequalities (24) and (23) yields an equivalent definition of the viscosity solution for (22).

Proof. We first observe that, since $f > f_1 > 0$, if the minimum is attained for some $\boldsymbol{a} = \boldsymbol{a_1}$, then $(\boldsymbol{a_1} \cdot \nabla\phi(\boldsymbol{x_0})) < 0$. Let $\boldsymbol{b} = \frac{-\nabla\phi(\boldsymbol{x_0})}{\|\nabla\phi(\boldsymbol{x_0})\|}$. Since $\boldsymbol{a_1}$ is the minimizer, we have

$$(\nabla\phi(\boldsymbol{x_0}) \cdot \boldsymbol{a_1}) f(\boldsymbol{x_0}, \boldsymbol{a_1}) \leq (\nabla\phi(\boldsymbol{x_0}) \cdot \boldsymbol{b}) f(\boldsymbol{x_0}, \boldsymbol{b}) \leq -\|\nabla\phi(\boldsymbol{x_0})\| f_1.$$

Therefore,

$$\boldsymbol{a_1} \cdot \nabla\phi(\boldsymbol{x_0}) \leq -\|\nabla\phi(\boldsymbol{x_0})\| \frac{f_1}{f(\boldsymbol{x_0}, \boldsymbol{a_1})} \leq -\|\nabla\phi(\boldsymbol{x_0})\|\Upsilon^{-1}. \qquad \square$$

Remark 3.7. We have just established a bound on the angle between the characteristic of the PDE (22) and the gradient line of its viscosity solution. If the gradient $\nabla u(\boldsymbol{x_0})$ exists, then $\nabla u(\boldsymbol{x_0}) = \nabla\phi(\boldsymbol{x_0})$. Therefore, $\boldsymbol{a_1}\cdot\nabla u(\boldsymbol{x_0}) \leq -\|\nabla u(\boldsymbol{x_0})\|\Upsilon^{-1}$. If γ is the angle between $\nabla u(\boldsymbol{x_0})$ and $(-\boldsymbol{a_1})$, then $\cos(\gamma) \geq \frac{1}{\Upsilon}$. (If the level sets of $u(\boldsymbol{x})$ were straight lines, the last inequality would trivially follow from Lemma 3.4.) We note that the above argument heavily uses the existence of a positive lower bound f_1 and, therefore, does not directly apply to more general control problems.

3.5. A semi-Lagrangian discretization for the Hamilton–Jacobi–Bellman PDE. Assume that a triangulated mesh X of diameter h is defined on Ω. For every mesh point $\boldsymbol{x} \in X$, define $S(\boldsymbol{x})$ to be a set of all the simplexes in the mesh adjacent to $\boldsymbol{x}$ (i.e., the simplexes that have $\boldsymbol{x}$ as one of their vertices). If $s \in S(\boldsymbol{x})$, we will use the notation $\boldsymbol{x_{s,1}}$ and $\boldsymbol{x_{s,2}}$ for the other vertices of the simplex s.

A simple control-theoretic discretization of (19) follows from the assumption that, as the vehicle starts to move from a mesh point $\boldsymbol{x}$ inside a simplex $s \in S(x)$, its direction of motion $\boldsymbol{a}$ does not change until the vehicle reaches the edge $\boldsymbol{x_{s,1}x_{s,2}}$ (see Figure 1). The value $u(\tilde{\boldsymbol{x}})$ at the point of intersection can be approximated[7] using the values $u(\boldsymbol{x_{s,1}})$ and $u(\boldsymbol{x_{s,2}})$.

Defining $\tau(\zeta) = \|\tilde{\boldsymbol{x}} - \boldsymbol{x}\| = \|(\zeta\boldsymbol{x_{s,1}} + (1-\zeta)\boldsymbol{x_{s,2}}) - \boldsymbol{x}\|$ and $\boldsymbol{a}_\zeta = \frac{\tilde{\boldsymbol{x}}-\boldsymbol{x}}{\tau(\zeta)}$, we can now write the equation for the numerical approximation U:

$$\begin{aligned} U(\boldsymbol{x}) &= \min_{s\in S(\boldsymbol{x})} V_s(\boldsymbol{x}), \\ V_s(\boldsymbol{x}) &= \min_{\zeta\in[0,1]} \left\{ \frac{\tau(\zeta)}{f(\boldsymbol{x}, \boldsymbol{a}_\zeta)} + \zeta U(\boldsymbol{x_{s,1}}) + (1-\zeta)U(\boldsymbol{x_{s,2}}) \right\}. \end{aligned} \tag{25}$$

The above "naive" derivation is based on a direct application of Bellman's optimality principle rather than on discretization of the corresponding Hamilton–Jacobi–Bellman

[6] The control-theoretic problems discussed in these papers are slightly different. They consider infinite horizon or exit time problems with time-discounted running costs, e.g., $\text{Cost}(\boldsymbol{x}, \boldsymbol{a}(\cdot)) = \int_0^\infty e^{-\lambda s} K(\boldsymbol{y}(s), \boldsymbol{a}(s))ds$. Thus, the resulting PDE is also slightly different, but Kruzhkov's transform [24] can be used to obtain the mapping from one to another; see [8], [4], for example. In addition, the iterative methods in these papers are also applicable for a more general case of $f_1 = 0$.

[7] Since the interpolation has to be used to approximate the value at a non–mesh point $\tilde{\boldsymbol{x}}$, we refer to this and similar discretizations as *semi-Lagrangian*.

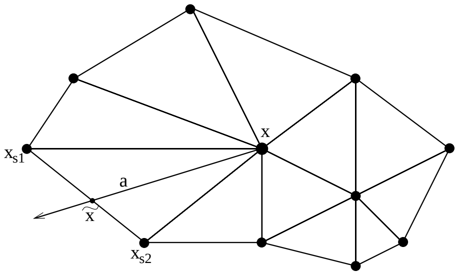


Fig. 1. *A control-theoretic discretization: each trajectory is approximated by a straight line within a simplex.*

PDE; a number of related methods, treatment of more general control problems (including the case $f_1 = 0$), and the proof of convergence can be found in [25, 26], and [18]. Similar higher-order control-theoretic numerical methods can be found in [17].

The discretized equation (25) has to be satisfied at every mesh point in X; this results in a coupled system of M nonlinear equations, which usually have to be solved simultaneously through the iterations. Due to the structure of the system, each iteration involves solving a local minimization problem for each mesh point, and even in the simplest problems the number of iterations will be proportional to the diameter of the mesh-graph. The number of iterations can be reduced using Gauss–Seidel relaxation (as in [18]), but we know of no theoretical guarantees of the rate of convergence.

4. OUMs for the isotropic case: Dijkstra-like Eikonal solvers.

Until recently, the Eikonal equation, corresponding to the isotropic optimal trajectory and front propagation problems, was the only case for which single-pass methods were available. Several fast algorithms have been introduced to solve the corresponding discretized system as efficiently as Dijkstra's method solves the shortest path problems on discrete networks. These methods are based on an observation that a particular upwind discretization possesses a *causality* property similar to that of the Eikonal equation (Property 3.6).

Property 4.1 (Causality). *If s is an acute simplex in $S(\boldsymbol{x})$ and $V_s(\boldsymbol{x})$ is a value of $U(\boldsymbol{x})$ computed under the assumption that the characteristic for $\boldsymbol{x}$ lies in s, then $V_s(\boldsymbol{x}) \geq \max\{U(\boldsymbol{x_{s,1}}), U(\boldsymbol{x_{s,2}})\}$.*

Any upwind discretization possessing this property leads to equations which can be decoupled by computing the value function at the mesh points in the increasing order. Since the ordering is not known in advance, we can structure these Dijkstra-like solvers in the spirit of section 2.2 as follows.

All the mesh points are divided into three classes: *Far* (no information about the correct value of U is known), *Accepted* (the correct value of U has been computed), and *Considered* (adjacent to *Accepted*), for which V has already been computed, but it is still unclear if $V = U$. For every *Considered* $\boldsymbol{x}$, we define the set $\mathrm{NS}(\boldsymbol{x}) = \{s \in S(\boldsymbol{x}) \mid \boldsymbol{x_{s,1}}$ and $\boldsymbol{x_{s,2}}$ are *Accepted*$\}$. We will also use a set $S(\boldsymbol{x_1}, \boldsymbol{x_2})$ to denote the simplexes adjacent to both of these mesh points.

1. Start with all the mesh points in *Far*.
2. Move the mesh points on the boundary ($\boldsymbol{y} \in \partial\Omega$) to *Accepted* ($U(\boldsymbol{y}) = q(\boldsymbol{y})$).
3. Move all the mesh points $\boldsymbol{x}$ adjacent to the boundary into *Considered* and

evaluate the tentative values

$$V(\boldsymbol{x}) := \min_{s \in NS(\boldsymbol{x})} V_s(\boldsymbol{x}). \tag{26}$$

4. Find the mesh point $\bar{\boldsymbol{x}}$ with the smallest value of V among all the *Considered*.
5. Move $\bar{\boldsymbol{x}}$ to *Accepted* ($U(\bar{\boldsymbol{x}}) = V(\bar{\boldsymbol{x}})$).
6. Move the *Far* mesh points adjacent to $\bar{\boldsymbol{x}}$ into *Considered*.
7. Reevaluate V for all the *Considered* $\boldsymbol{x}$ adjacent to $\bar{\boldsymbol{x}}$

$$V(\boldsymbol{x}) := \min \left\{ V(\boldsymbol{x}), \min_{s \in (S(\boldsymbol{x}, \bar{\boldsymbol{x}}) \bigcap \mathrm{NS}(\boldsymbol{x}))} V_s(\boldsymbol{x}) \right\}. \tag{27}$$

8. If *Considered* is not empty, then go to 4.

Such Dijkstra-like methods use heap-sort data structures to achieve Dijkstra-like efficiency of $O(M \log M)$ and compute the numerical solutions converging to the viscosity solution (due to the upwinding structure of discretization).

The first Dijkstra-like method, introduced by Tsitsiklis in 1994, evolved from studying isotropic min-cost optimal trajectory problems and was based on a direct approximation of the characteristic directions at each mesh point [44, 45]. Tsitsiklis proved that Property 4.1 holds for the particular first-order semi-Lagrangian discretization (i.e., formula (25)) of the Eikonal equation, when used on a uniform Cartesian grid in R^n. The algorithm requires solving a local minimization problem to update the solution at each mesh point; however, as shown in [45], the Kuhn–Tucker optimality conditions can be used to find a quadratic equation satisfied by the minimum value instead. In the appendix we provide a more general proof that the same causality property is possessed by the discretization (25) on an arbitrary unstructured mesh and derive the corresponding quadratic equation for the minimum value.

The family of Fast Marching Methods, introduced by Sethian in [33] and extended by Sethian and several co-authors in [35, 21, 38], evolved from studying isotropic front propagation problems (see section 8.1 for the recasting of Eikonal PDE in this context). Those discretizations were based on upwinding approximations of the gradient and were all obtained in a fully Eulerian frame of reference. Sethian proved that the causality Property 4.1 holds for a wide class of upwind finite-difference discretizations. Following that approach, upwind finite-difference operators were then used to obtain higher-order Cartesian versions [35], extensions to triangulated meshes [21], and general higher-order versions for the unstructured meshes in R^n [38]. In addition, the "lifting-to-surface" technique introduced in [38] allowed the Fast Marching Method to be used to solve a limited class of non-Eikonal (elliptically anisotropic) problems. We note that these extensions are all OUMs, relying on an upwinding criterion that establishes a monotonicity-preserving update procedure. Early applications of the Fast Marching Methods included the narrow band level set method [1], photolithography [34], a comparison of a similar algorithm with volume-of-fluid techniques [19], and a fast algorithm for image segmentation [27]. More recent applications include problems in robotic navigation [22], extension velocity computation [2], visibility evaluation [35], geophysics [32, 37], and computational geometry [23]. To produce an update for each mesh point, these methods require solving a quadratic equation, which will depend on the particular upwind finite-difference operator used. The original Fast Marching Method, as defined in [33], was based on the first-order Godunov-type discretization on a uniform Cartesian grid, and the corresponding quadratic update equation coincides with the equation derived from the Kuhn–Tucker conditions in [45].

(See the appendix for a discussion of the correspondence between the first-order semi-Lagrangian and fully Eulerian discretizations on unstructured meshes.)

Several different higher-order versions of the Fast Marching Method are available for structured and unstructured meshes, while there are currently no higher-order Dijkstra-like methods based on a semi-Lagrangian approach. The reason is the difficulty of finding the higher-order semi-Lagrangian discretization that would provably possess the causality property. Finally, when such a discretization is found, it will generally require performing a local minimization at every mesh point, since it is not obvious whether the Kuhn–Tucker conditions can be used to produce a quadratic equation in this more general case.

5. Characteristics versus gradients. In this section, we show why the Dijkstra-like methods cannot be directly applied to handle general (non-Eikonal) Hamilton–Jacobi equations. As a test problem, consider the "distance from the origin" Eikonal equation ($\|\nabla d\| = 1$, $d(0,0) = 0$), in a plane $z = c_1 x + c_2 y$ for some vector $\boldsymbol{c} = \left[\begin{smallmatrix} c_1 \\ c_2 \end{smallmatrix}\right]$. The level sets of d will be just the circles around the origin in that plane. Projecting those circles orthogonally onto the $x-y$ plane, we will see a set of concentric ellipses. As expected, the function $u(\boldsymbol{x})$, whose level sets coincide with these ellipses, may be obtained in two ways:

- *As an optimal-trajectory problem.* The function d can be considered as a value function for a vehicle moving with a unit speed in the plane $z = c_1 x + c_2 y$. As shown in [46], if one considers another vehicle which moves as a shadow of the first one in the $x-y$ plane, its value function will be the viscosity solution $u(\boldsymbol{x})$ of the Hamilton–Jacobi–Bellman equation

$$\min_{\boldsymbol{a} \in S_1} \{(\nabla u(\boldsymbol{x}) \cdot \boldsymbol{a}) f(\boldsymbol{x}, \boldsymbol{a})\} + 1 = 0, \qquad \boldsymbol{x} \in \Omega, \tag{28}$$

  and the vehicle's speed function in the direction $\boldsymbol{a} = \left[\begin{smallmatrix} a_1 \\ a_2 \end{smallmatrix}\right]$ will be given by

$$f(\boldsymbol{a}, x, y) = \left(1 + (c_1 a_1 + c_2 a_2)^2\right)^{-\frac{1}{2}}.$$

- *As a front propagation problem.* As shown in [36], this same problem can be viewed in the front propagation framework, using the speed function

$$F(\boldsymbol{x}, \boldsymbol{n}) = \sqrt{\frac{(1 + c_2^2) n_1^2 + (1 + c_1^2) n_2^2 - 2 c_2 c_1 n_1 n_2}{1 + c_1^2 + c_2^2}}.$$

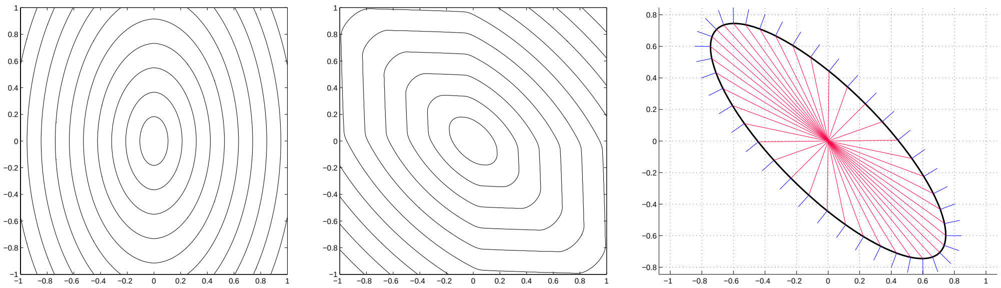
  The Hamilton–Jacobi PDE corresponding to this speed function F is

$$\sqrt{\frac{(1 + c_2^2) u_x^2(\boldsymbol{x}) + (1 + c_1^2) u_y^2(\boldsymbol{x}) - 2 c_2 c_1 u_x(\boldsymbol{x}) u_y(\boldsymbol{x})}{1 + c_1^2 + c_2^2}} = 1. \tag{29}$$

It would appear that Tsitsiklis' algorithm (defined for the isotropic case in section 4) can be applied to this anisotropic problem without any changes at all, except that the dependence of the speed f upon the direction $\boldsymbol{a}$ will now be present in the update-from-a-single-simplex formula:

$$V_s(\boldsymbol{x}) = \min_{\zeta \in [0,1]} \left\{ \frac{\tau(\zeta)}{f(\boldsymbol{x}, \boldsymbol{a}_\zeta)} + \zeta U(\boldsymbol{x}_{\boldsymbol{s},\boldsymbol{1}}) + (1 - \zeta) U(\boldsymbol{x}_{\boldsymbol{s},\boldsymbol{2}}) \right\}. \tag{30}$$

What happens when this algorithm is used to compute the expansion of the ellipse (as defined by (28))? In Figure 2 we show the level sets of the numerical solution



(a) (b) (c)

Fig. 2. (a) *and* (b) *Ellipse expansion computed by Tsitsiklis' algorithm. Both computations were performed on a* 129×129 *uniform Cartesian grid.* (c) *The characteristics and the gradient directions for the second expanding ellipse.*

U obtained by this method for two different expanding ellipses. The first contour plot corresponds to the vector $\boldsymbol{c} = \left[\begin{smallmatrix}\sqrt{2}\\0\end{smallmatrix}\right]$. The numerical solution converges to the value function $u(\boldsymbol{x})$ and is first-order accurate as the grid size tends to 0. (We will return to this example later when we discuss fast methods relying on a particular grid orientation in section 5.) The second contour plot corresponds to the vector $\boldsymbol{c} = \left[\begin{smallmatrix}1\\1\end{smallmatrix}\right]$. In this case, it is obvious that $U(\boldsymbol{x})$ does not approximate the viscosity solution very well. Nor does it improve under a grid refinement.

In order to understand what is different in the second example, we recall that all Dijkstra-like methods are fundamentally dependent on the *causality property* (3.6) of the Eikonal equation. Each of these single-pass methods is based on the observation that a certain discretization also possesses a similar *causality property*. This *causality* results from the fact that the characteristics of the Eikonal equation coincide with the gradient lines of its viscosity solution u. However, for the anisotropic problems this property does not hold. When the characteristic and gradient directions are different, the simplex $\boldsymbol{x x_j x_k}$ may contain the characteristic for the point $\boldsymbol{x}$, even if the gradient $\nabla u(\boldsymbol{x})$ is not pointing from that simplex. Thus, no matter how small that simplex is, it is still possible that $u(\boldsymbol{x}) < u(\boldsymbol{x_j})$. This is an intrinsic problem with Dijkstra-like methods in the anisotropic case: to produce the numerical solution efficiently, these methods attempt to compute $U(\boldsymbol{x})$ in the ascending order (i.e., from the simplex containing $(-\nabla u)$), whereas, in order to maintain the upwinding, $U(\boldsymbol{x})$ has to be computed from the simplex containing the characteristic. That phenomenon is also quite obvious from comparing Figures 2(b) and 2(c): the Dijkstra-like method fails exactly at those points where the gradient line and the characteristic do not lie in the same coordinate quadrant (or, more generally, in the same simplex—the quadrants are used because the numerical solution in Figure 2 is computed on a Cartesian grid).

However, it is still possible that, for a given PDE and for a chosen discretization scheme, Dijkstra-like decoupling will produce a convergent numerical solution, provided that it is used on *a specially oriented grid* (e.g., Figure 2(a)). A criterion based on this observation was introduced by Sethian in [36] as follows.

Criterion 5.1 (Applicability of the Fast Marching Method). *For a static Hamilton–Jacobi equation* $H(\nabla u, \boldsymbol{x}) = 0$*, if the convex Hamiltonian* H *is approximated on a Cartesian grid by a consistent difference operator*

$$H_{ij}(U_{i,j}, U_{i-1,j}, U_{i+1,j}, U_{i,j-1}, U_{i,j+1}, \boldsymbol{x_{i,j}}) = 0,$$

and if it is known that $U_{i,j}$ depends only on the smaller *values of U at the neighboring points, then the Fast Marching Method can be used to compute $U_{i,j}$'s efficiently.*

Remark 5.2. In the context of upwinding discretizations on unstructured meshes, the above criterion is equivalent to requiring that the characteristics and the (numerically approximated) vector $(-\nabla u)$ should always lie in the same simplex. Several sufficient conditions for a class of numerical Hamiltonians to satisfy the above criterion on Cartesian grids were presented by Osher and Fedkiw in [29]. For instance, the causality property was proven in [29] for the Godunov-type upwinding discretization H^G_{ij}, provided that the original Hamiltonian $H(\nabla u, \boldsymbol{x})$ has a special form $H(\nabla u, \boldsymbol{x}) = G(u_x^2, u_y^2)$ for some function G. We note that, even for a relatively simple elliptical front propagation (29), this condition is satisfied only in the case when c_1 or c_2 is equal to zero, i.e., only when the axes of the ellipse are exactly aligned with the grid coordinate directions. This is precisely the situation illustrated by Figure 2(a).

In general, finding discretizations which satisfy Criterion 5.1 is a difficult task. We note the following problems associated with this approach:

- *Whether or not the criterion is satisfied depends upon a particular grid/mesh-orientation.* Indeed, the two test problems in Figure 2 are actually the same (modulo a rotation by 45°), yet only one of them satisfies the criterion.
- *For any anisotropic problem, there are infinitely many grid orientations such that the criterion is not satisfied.* If an angle between the characteristic and the gradient line is not zero, then any grid line lying inside that angle will violate the criterion. Correspondingly, the bigger the *anisotropy coefficient* Υ is, the harder it is to find the grid orientation satisfying the criterion.
- *The criterion is infinitely sensitive to grid perturbations.*
- *If the criterion is not satisfied, the numerical solution does not lose stability under grid refinement.* In other words, when it does not work, it is not immediately obvious.
- *If the criterion is not satisfied even at a single grid point, the numerical solution need not converge to the viscosity solution.* Criterion 5.1 is the basis for determining the order for computing the values of U. Computing even one of them from a wrong quadrant can greatly affect the ordering of the remaining computations.
- *For many anisotropic problems, the criterion cannot be satisfied for any choice of the grid directions.* Indeed, if the angle between gradient lines and the characteristics is sufficiently wide, and if the medium is substantially inhomogeneous (i.e., if the speed $f(\boldsymbol{x}, \boldsymbol{a})$ varies significantly in different parts of Ω), then any Cartesian grid might violate Criterion 5.1 for some grid point $\boldsymbol{x} \in X$.

As a result, we have chosen to concentrate on a family of robust single-pass methods, which are independent of the grid choice[8] and applicable to a wider class of control problems. Nevertheless, for the limited class of problems in which Criterion 5.1 can be analytically demonstrated for a certain choice of grid, the original Dijkstra-like solvers will perform better than the new OUMs introduced in the next section.

[8] Of course, it is just the fact of *convergence* that is independent of the grid choice for our methods; the *speed of convergence* is certainly influenced by the choice of the grid and its alignment with the shock lines.

In fact, if the computational mesh is not fixed for some application-specific reasons, the convergence of our single-pass methods can be further improved by using the computed characteristic information to dynamically add the mesh points inside the AF, wherever the shock is suspected.

5.1. Causality in the Hamilton–Jacobi–Bellman PDE. A different (weaker) *causality property* for the more general Hamilton–Jacobi–Bellman equation results from Bellman's optimality principle (see section 3). Since the characteristics of that PDE are, in fact, the optimal trajectories of the corresponding control problem, we know that the value function u is strictly increasing along the characteristics. Our OUMs for the general anisotropic problems will be based on the fixed-boundary optimality principle (Lemma 3.2).

Let $u(\boldsymbol{x})$ be the value function for the anisotropic min-time optimal trajectory problem defined on Ω in section 3. We will use the notation $T_{\hat{\boldsymbol{x}}}(\boldsymbol{x})$ for the minimum time required to reach the point $\hat{\boldsymbol{x}}$ starting from the point $\boldsymbol{x}$. If Γ is a simple closed curve in $\Omega\backslash\partial\Omega$ and a point $\boldsymbol{x}$ is inside Γ, then Lemma 3.2 shows that

$$u(\boldsymbol{x}) = \inf_{\hat{\boldsymbol{x}} \in \Gamma} \left\{ T_{\hat{\boldsymbol{x}}}(\boldsymbol{x}) + u(\hat{\boldsymbol{x}}) \right\}.$$

Because of the properties of the speed function f and by the continuity of Γ, that infimum is actually a minimum achieved at some point $\tilde{\boldsymbol{x}}$, i.e., $u(\boldsymbol{x}) = T_{\tilde{\boldsymbol{x}}}(\boldsymbol{x}) + u(\tilde{\boldsymbol{x}})$. The point $\tilde{\boldsymbol{x}}$ can be interpreted as an intersection of the optimal trajectory for $\boldsymbol{x}$ with the curve Γ. Thus, knowing u on Γ is sufficient for evaluating u at any point inside Γ. Moreover, if Γ is a level set of $u(\boldsymbol{x})$, then, by Lemma 3.4, we know that $\|\tilde{\boldsymbol{x}} - \boldsymbol{x}\| \leq d_1\Upsilon$, where $\Upsilon = \frac{f_2}{f_1}$ and d_1 is the distance from $\boldsymbol{x}$ to Γ (see Figure 3). The last observation[9] necessary for constructing a computational algorithm is that, if d_1 is small relative to the size of Γ, then the optimal time $T_{\tilde{\boldsymbol{x}}}(\boldsymbol{x})$ cannot be much smaller than the time required to traverse the straight line trajectory from $\boldsymbol{x}$ to $\tilde{\boldsymbol{x}}$.

6. Control-theoretic OUM. We now describe our control-theoretic OUM, which was first discussed without convergence proof in [39]. Consider an unstructured triangulated mesh X of diameter h (i.e., if the mesh points $\boldsymbol{x_j}$ and $\boldsymbol{x_k}$ are adjacent, then $\|\boldsymbol{x_j} - \boldsymbol{x_k}\| \leq h$).

Let $\boldsymbol{x_j}$ and $\boldsymbol{x_k}$ be two adjacent mesh points. Define the upwinding approximation for $U(\boldsymbol{x})$ from a "virtual simplex" $\boldsymbol{x_j x x_k}$:

$$V_{\boldsymbol{x_j},\boldsymbol{x_k}}(\boldsymbol{x}) = \min_{\zeta\in[0,1]} \left\{ \frac{\tau(\zeta)}{f(\boldsymbol{x}, \boldsymbol{a}_\zeta)} + \zeta U(\boldsymbol{x_j}) + (1-\zeta)U(\boldsymbol{x_k}) \right\}, \tag{31}$$

where $\tau(\zeta) = \|(\zeta\boldsymbol{x_j} + (1-\zeta)\boldsymbol{x_k}) - \boldsymbol{x}\|$ and $\boldsymbol{a}_\zeta = \frac{(\zeta\boldsymbol{x_j}+(1-\zeta)\boldsymbol{x_k})-\boldsymbol{x}}{\tau(\zeta)}$.

Remark 6.1. The above update formula is basically the same as the upwind formula for simplex s given by (25). The difference is that $V_{\boldsymbol{x_j},\boldsymbol{x_k}}(\boldsymbol{x})$ is defined even when $\boldsymbol{x_j}$ and $\boldsymbol{x_k}$ are not adjacent to $\boldsymbol{x}$.

Control-theoretic OUM for anisotropic problems. As before, mesh points are divided into three classes (*Far*, *Considered*, *Accepted*). The *AcceptedFront* is defined as a set of *Accepted* mesh points, which are adjacent to some not-yet-accepted (i.e., *Considered*) mesh points. Define the set AF of the line segments $\boldsymbol{x_j x_k}$, where $\boldsymbol{x_j}$ and $\boldsymbol{x_k}$ are adjacent mesh points on the *AcceptedFront*, such that there exists a *Considered* mesh point $\boldsymbol{x_i}$ adjacent to both $\boldsymbol{x_j}$ and $\boldsymbol{x_k}$. For each *Considered* mesh point $\boldsymbol{x}$ we define the "near front" as the part of AF "relevant to $\boldsymbol{x}$":

$$\mathrm{NF}(\boldsymbol{x}) = \{\boldsymbol{x_j x_k} \in AF \ \mid \exists \tilde{\boldsymbol{x}} \text{ on } \boldsymbol{x_j x_k} \text{ s.t. } \|\tilde{\boldsymbol{x}} - \boldsymbol{x}\| \leq h\Upsilon\}.$$

[9] Since Γ generally is not a level set of u, the logic of the method is more subtle and cannot really be based on Lemma 3.4. Instead, it relies on Lemma 3.5, which provides a weaker version of this inequality, but for any Γ "well-resolved" by an underlying mesh X.

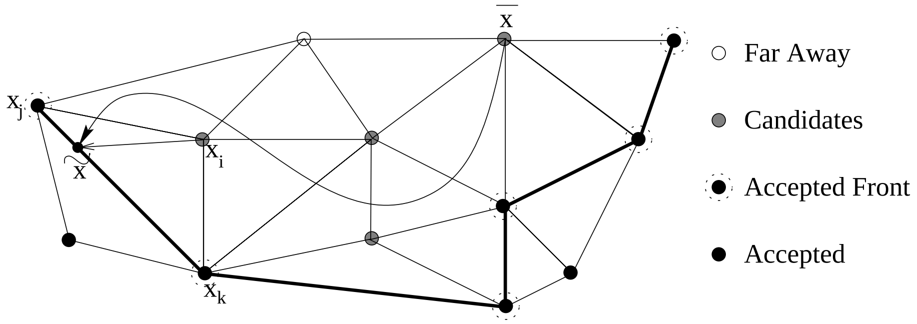


FIG. 3. *The AcceptedFront and the Considered mesh points. Segments of AF are shown in bold. The optimal trajectory for $\bar{\boldsymbol{x}}$ cannot intersect AF too far away from $\bar{\boldsymbol{x}}$, for if $\|\tilde{\boldsymbol{x}} - \bar{\boldsymbol{x}}\| > h\Upsilon$, then $u(\boldsymbol{x_i}) < u(\bar{\boldsymbol{x}})$.*

1. Start with all the mesh points in *Far*.

2. Move the mesh points on the boundary ($\boldsymbol{y} \in \partial\Omega$) to *Accepted* ($U(\boldsymbol{y}) = q(\boldsymbol{y})$).

3. Move all the mesh points $\boldsymbol{x}$ adjacent to the boundary into *Considered* and evaluate the tentative values

$$V(\boldsymbol{x}) := \min_{\boldsymbol{x_j x_k} \in \mathrm{NF}(\boldsymbol{x})} V_{\boldsymbol{x_j},\boldsymbol{x_k}}(\boldsymbol{x}). \tag{32}$$

4. Find the mesh point $\bar{\boldsymbol{x}}$ with the smallest value of V among all the *Considered*.

5. Move $\bar{\boldsymbol{x}}$ to *Accepted* ($U(\bar{\boldsymbol{x}}) = V(\bar{\boldsymbol{x}})$) and update the *AcceptedFront*.

6. Move the *Far* mesh points adjacent to $\bar{\boldsymbol{x}}$ into *Considered* and compute their tentative values by (32).

7. Recompute the value for all the other *Considered* $\boldsymbol{x}$ such that $\bar{\boldsymbol{x}} \in \mathrm{NF}(\boldsymbol{x})$

$$V(\boldsymbol{x}) := \min\left\{V(\boldsymbol{x}), \min_{\bar{\boldsymbol{x}}\boldsymbol{x_i} \in \mathrm{NF}(\boldsymbol{x})} V_{\bar{\boldsymbol{x}},\boldsymbol{x_i}}(\boldsymbol{x})\right\}. \tag{33}$$

8. If *Considered* is not empty, then go to 4.

We note that the resulting algorithm

- is "single-pass," since it produces the numerical solution U in $O(\Upsilon^2 M \log(M))$ steps. This is because there are a total of M points to *Accept*, every time a mesh point is accepted there are at most Υ^2 *Considered* points to reevaluate, and we must maintain an ordering of *Considered* based on V, which contributes a factor of $\log(M)$.
- produces the numerical solution U that converges to u as the diameter of the mesh tends to zero (see the proof in section 7);
- is at most first-order accurate;
- works equally well on acute and nonacute triangulated meshes;
- is applicable for a general anisotropic optimal trajectory problem described in section 3.

An extension of this method to R^n and manifolds is straightforward, since the update formula (31) can easily be generalized for these cases. The only part of the program which needs to be modified to handle a manifold-approximating mesh is the algorithm for sorting and searching the *AcceptedFront*.

Remark 6.2 (Comments on computational complexity).

1. In the above complexity analysis, the calculation of an upwind-update-from-a-single-simplex value $V_{\boldsymbol{x_j},\boldsymbol{x_k}}(\boldsymbol{x})$ was counted as a single operation. We note that

the optimization problem solved to compute $V_{\boldsymbol{x_j},\boldsymbol{x_k}}(\boldsymbol{x})$ is *local* (i.e., $V_{\boldsymbol{x_j},\boldsymbol{x_k}}(\boldsymbol{x})$ can be computed independently from any other $V_{\boldsymbol{x_i},\boldsymbol{x_m}}(\boldsymbol{x_l})$) and, thus, should not be confused with the iterations necessary to solve the coupled system of nonlinear equations (25) simultaneously. More details on algorithmic efficiency can be found in section 9.

2. As we will show in section 7, AF can be considered as an approximation for a level set of U. Thus, if the mesh diameter h is sufficiently small, then the number of $Considered$ points which have to be updated after each acceptance becomes closer to Υ, since the $Considered$ points are immediately adjacent to AF. Thus, as h decreases, the computational complexity of the method tends to $O\left(\Upsilon M \log(M)\right)$.

3. If the problem is formulated in R^n, the complexity of the corresponding algorithm is $O(\Upsilon^{n-1} M \log(M))$, where M remains the total number of mesh points.

Remark 6.3 (A comment on the rate of convergence). Our proof of convergence in section 7 does not provide an estimate for the rate of convergence. We believe that this method is first-order; this belief is based on the first-order accuracy of the approximations behind the semi-Lagrangian discretization (used to calculate $V_{\boldsymbol{x_j},\boldsymbol{x_k}}(\boldsymbol{x})$) and is also confirmed by the numerical evidence (see section 9.4 and [38, 46]). Based on numerical experiments, we note that for sufficiently small h, $\|U-u\|_\infty$ is at worst $O(\Upsilon h)$. This is not surprising, since Υh is the largest distance over which the first-order accurate approximation might be performed when $V_{\boldsymbol{x_j},\boldsymbol{x_k}}(\boldsymbol{x})$ is computed.

Remark 6.4 (A comment on mesh degeneracy). The fast Eikonal solvers described in section 4 rely on the *causality property*, which holds only for the acute simplexes. An additional "splitting section" construction is required to handle the nonacute case [21, 38]. Not surprisingly, the acuteness of simplexes in X is not required for the OUM introduced here. After all, the algorithm uses the mesh connectivity only to determine what becomes $Considered$, when a new mesh point is *Accepted*. All of the upwind-update-from-a-single-simplex values $V_{\boldsymbol{x_j},\boldsymbol{x_k}}(\boldsymbol{x})$ are computed from the simplexes defined by the position of $AcceptedFront$ rather than from the simplexes present in X.

Nevertheless, in order to prove the convergence of $U(\boldsymbol{x})$ to the viscosity solution, we will have to assume that the mesh X cannot be arbitrarily degenerate. Namely, we will assume that if h is the diameter of X and h_{min} is the smallest triangle height in X, then the ratio $\eta = h/h_{min}$ is bounded for all sufficiently small h. See the proof of Lemma 7.5 for details.

Remark 6.5 (A comment on the order of *Acceptance*). Unlike in Sethian's Fast Marching Methods or in Tsitsiklis' algorithm, in the above method the mesh points *are not Accepted* in the order of increasing U. As was pointed out in section 5, for the anisotropic optimal trajectory problems the fact that the characteristic for $\boldsymbol{x}$ lies inside the simplex $\boldsymbol{x}\boldsymbol{x_j}\boldsymbol{x_k}$ does not mean that the gradient is pointing from that simplex. Thus, it is entirely possible that $U(\boldsymbol{x}) \leq V_{\boldsymbol{x_j},\boldsymbol{x_k}}(\boldsymbol{x}) < U(\boldsymbol{x_j})$. Nevertheless, we will show in Lemma 7.3 that the order of *Acceptance* is monotone, albeit in a much weaker sense than for the single-pass Eikonal solvers.

Remark 6.6 (Decoupling of the "extended semi-Lagrangian scheme"). Define the extended set of neighbors

$$N_K(\boldsymbol{x}) = \{\boldsymbol{x_1}\boldsymbol{x_2} \in X \mid \boldsymbol{x_1} \text{ and } \boldsymbol{x_2} \text{ are adjacent and } \exists \tilde{\boldsymbol{x}} \text{ on } \boldsymbol{x_1}\boldsymbol{x_2} \text{ s.t. } \|\tilde{\boldsymbol{x}} - \boldsymbol{x}\| \leq h\Upsilon\}.$$

Note that if we replace NF$(\boldsymbol{x})$ by $N_K(\boldsymbol{x})$, the formula (32) becomes

$$U(\boldsymbol{x}) = \min_{\boldsymbol{x_1}\boldsymbol{x_2} \in N_K(\boldsymbol{x})} V_{\boldsymbol{x_1}\boldsymbol{x_2}}(\boldsymbol{x}). \tag{34}$$

This is an "extended" version of the semi-Lagrangian scheme (25), and it is easy to show that its solution U converges to the viscosity solution u. Equation (34) can be solved by successive approximation techniques described in [18], for example. However, a single-pass algorithm cannot be used to find U since we need to consider all possible directions of motion for the vehicle starting at the point $\boldsymbol{x}$ (i.e., $U(\boldsymbol{x})$ might potentially depend upon $U(\boldsymbol{y})$ for all $\boldsymbol{y} \in N_K(\boldsymbol{x})$, including the values $U(\boldsymbol{y}) > U(\boldsymbol{x})$). Therefore, the formula (32) can be interpreted as an upwinding analogue of (34).

The above comparison is just an analogue—not an equivalence. The numerical values produced by executing the above OUM will be different from those obtained by solving the coupled system (34); thus, the convergence of the OUM has to be proven separately.

7. Proof of convergence.

In this section we prove the convergence of the above algorithm to the viscosity solution.[10]

We will assume that the numerical solution $U(\boldsymbol{x})$ is computed for each $x \in X$, using the OUM described in section 6. For the points $\boldsymbol{x} \in \Omega \backslash X$, $U(\boldsymbol{x})$ is defined by linear interpolation as follows.

If $\boldsymbol{x}$ is inside Ω but is not a mesh point, then it lies in some simplex $\boldsymbol{x_1 x_2 x_3}$. In that case, there exist $\zeta_1, \zeta_2, \zeta_3 \geq 0$ such that

$$\zeta_1 + \zeta_2 + \zeta_3 = 1, \qquad \zeta_1 \boldsymbol{x_1} + \zeta_2 \boldsymbol{x_2} + \zeta_3 \boldsymbol{x_3} = \boldsymbol{x}. \tag{35}$$

The value at $\boldsymbol{x}$ is defined to be $U(\boldsymbol{x}) = \zeta_1 U(\boldsymbol{x_1}) + \zeta_2 U(\boldsymbol{x_2}) + \zeta_3 U(\boldsymbol{x_3})$.

Suppose that h_{min} is the smallest triangle height in the mesh X. We will use the constant $\eta = \frac{h}{h_{min}}$ to characterize the degree of "degeneracy" of the mesh X.

7.1. Properties of the numerical solution.

The following lemmas demonstrate several properties of the numerical solution U, which are necessary to prove the convergence and also mirror the properties of the value function for the optimal trajectory problem (section 3.2).

7.1.1. Is NF(x) big enough?

Suppose that the mesh point $\bar{\boldsymbol{x}}$ is about to be *Accepted* (hence, $V(\bar{\boldsymbol{x}}) = \min_{\boldsymbol{x} \in Considered} V(\boldsymbol{x})$).

LEMMA 7.1. *For every Considered mesh point $\boldsymbol{x}$ define*

$$W(\boldsymbol{x}) = \min_{\boldsymbol{x_1 x_2} \in AF} \; \min_{\zeta \in [0,1]} \left\{ \frac{\tau(\zeta)}{f(\boldsymbol{x}, \boldsymbol{a})} + (\zeta U(\boldsymbol{x_1}) + (1-\zeta) U(\boldsymbol{x_2})) \right\}, \tag{36}$$

where $\tau(\zeta) = \|(\zeta \boldsymbol{x_1} + (1-\zeta)\boldsymbol{x_2}) - \boldsymbol{x}\|$ and $\boldsymbol{a} = \frac{\zeta(\boldsymbol{x_1} - \boldsymbol{x}) + (1-\zeta)(\boldsymbol{x_2} - \boldsymbol{x})}{\tau(\zeta)}$. If $\bar{\boldsymbol{x}}$ is about to be Accepted, *then $U(\bar{\boldsymbol{x}}) = V(\bar{\boldsymbol{x}}) = W(\bar{\boldsymbol{x}})$.*

Proof. First, $U(\bar{\boldsymbol{x}}) = V(\bar{\boldsymbol{x}})$ simply because $\bar{\boldsymbol{x}}$ is about to be *Accepted.*

Recall that $V(\boldsymbol{x})$ for every *Considered* mesh point $\boldsymbol{x}$ is computed by formula (32) as follows:

$$V(\boldsymbol{x}) = \min_{\boldsymbol{x_1 x_2} \in \mathrm{NF}(\boldsymbol{x})} \; \min_{\zeta \in [0,1]} \left\{ \frac{\tau(\zeta)}{f(\boldsymbol{x}, \boldsymbol{a})} + (\zeta U(\boldsymbol{x_1}) + (1-\zeta) U(\boldsymbol{x_2})) \right\},$$

[10] As of now, we do not know of any natural discretized version of the Hamilton–Jacobi–Bellman PDE that would be exactly satisfied by the numerical solution $U(\boldsymbol{x})$ produced by the OUM in section 6. Since $U(\boldsymbol{x})$ is defined constructively (i.e., by an algorithm to compute it), we cannot rely on properties of a discretized equation for the proof of convergence; thus, the proof in this section has to rely on the properties of the algorithm itself.

where $\mathrm{NF}(\boldsymbol{x})$ is the part of the *AcceptedFront* "relevant to $\boldsymbol{x}$": $\mathrm{NF}(\boldsymbol{x}) = \{\boldsymbol{x_1 x_2} \in AF \mid \exists \tilde{\boldsymbol{x}}$ on the line segment $\boldsymbol{x_1 x_2}$ s.t. $\|\tilde{\boldsymbol{x}} - \boldsymbol{x}\| \leq h\Upsilon\}$. Since $\mathrm{NF}(\boldsymbol{x}) \subset AF$, we immediately see that for any *Considered* mesh point $\boldsymbol{x}$

$$V(\boldsymbol{x}) \geq W(\boldsymbol{x}). \tag{37}$$

Let $\boldsymbol{x_1 x_2} \in AF$ and $\zeta \in [0,1]$ be such that the minimum in formula (36) is attained; i.e., if $\hat{\boldsymbol{x}} = (\zeta \boldsymbol{x_1} + (1-\zeta)\boldsymbol{x_2})$, then

$$W(\bar{\boldsymbol{x}}) = \frac{\|\hat{\boldsymbol{x}} - \bar{\boldsymbol{x}}\|}{f(\bar{\boldsymbol{x}}, \frac{\hat{\boldsymbol{x}}-\bar{\boldsymbol{x}}}{\|\hat{\boldsymbol{x}}-\bar{\boldsymbol{x}}\|})} + (\zeta U(\boldsymbol{x_1}) + (1-\zeta)U(\boldsymbol{x_2}))\,. \tag{38}$$

Let $\boldsymbol{x_3}$ be the *Considered* mesh point adjacent to both $\boldsymbol{x_1}$ and $\boldsymbol{x_2}$. Then $U(\bar{\boldsymbol{x}}) = V(\bar{\boldsymbol{x}}) \leq V(\boldsymbol{x_3})$ since $\bar{\boldsymbol{x}}$ is about to be accepted. $V(\boldsymbol{x_3})$ is also computed by formula (32); thus,

$$V(\boldsymbol{x_3}) \leq \frac{\|\hat{\boldsymbol{x}} - \boldsymbol{x_3}\|}{f(\boldsymbol{x_3}, \frac{\hat{\boldsymbol{x}}-\boldsymbol{x_3}}{\|\hat{\boldsymbol{x}}-\boldsymbol{x_3}\|})} + (\zeta U(\boldsymbol{x_1}) + (1-\zeta)U(\boldsymbol{x_2})) \leq \frac{h}{f_1} + (\zeta U(\boldsymbol{x_1}) + (1-\zeta)U(\boldsymbol{x_2}))\,.$$

Combining this with the inequalities (37) and(38), we obtain

$$\begin{aligned} \frac{\|\hat{\boldsymbol{x}} - \bar{\boldsymbol{x}}\|}{f_2} + (\zeta U(\boldsymbol{x_1}) + (1-\zeta)U(\boldsymbol{x_2})) \leq W(\bar{\boldsymbol{x}}) \leq V(\bar{\boldsymbol{x}}) &\leq V(\boldsymbol{x_3}) \\ &\leq \frac{h}{f_1} + (\zeta U(\boldsymbol{x_1}) + (1-\zeta)U(\boldsymbol{x_2}))\,, \end{aligned}$$

which implies $\|\hat{\boldsymbol{x}} - \bar{\boldsymbol{x}}\| \leq h\Upsilon$. Therefore, $\boldsymbol{x_1 x_2} \in \mathrm{NF}(\bar{\boldsymbol{x}})$ and $W(\bar{\boldsymbol{x}}) = V(\bar{\boldsymbol{x}})$. □

7.1.2. Uniform upper bound.

LEMMA 7.2. *If Ω is convex and $d(\boldsymbol{x})$ is the distance from $\boldsymbol{x} \in \Omega$ to the boundary $\partial\Omega$, then*

$$U(\boldsymbol{x}) \leq \frac{d(\boldsymbol{x})}{f_1} + q_2. \tag{39}$$

Proof. If $\boldsymbol{x} \in \partial\Omega$, then the inequality holds trivially since $0 \leq q(\boldsymbol{x}) \leq q_2$.

If $\boldsymbol{x}$ is a mesh point inside Ω, we prove the lemma by induction: assume that the inequality (39) holds for all the mesh points that are on the *AcceptedFront* just before $\boldsymbol{x} = \bar{\boldsymbol{x}}$ is *Accepted*. Consider a (possibly nonunique) shortest path from $\bar{\boldsymbol{x}}$ to the boundary. By the properties of the distance function $d(\cdot)$, that shortest path is a straight line. Moreover, suppose that line intersects the segment of *AcceptedFront* $\boldsymbol{x_1 x_2} \in AF$ at the point $\hat{\boldsymbol{x}} = (\zeta \boldsymbol{x_1} + (1-\zeta)\boldsymbol{x_2})$. It is trivial to show that $d(\bar{\boldsymbol{x}}) = d(\hat{\boldsymbol{x}}) + \|\hat{\boldsymbol{x}} - \bar{\boldsymbol{x}}\|$. Using Lemma 7.1,

$$U(\bar{\boldsymbol{x}}) \;=\; W(\bar{\boldsymbol{x}}) \;\leq\; \frac{\|\hat{\boldsymbol{x}} - \bar{\boldsymbol{x}}\|}{f(\bar{\boldsymbol{x}}, \frac{\hat{\boldsymbol{x}}-\bar{\boldsymbol{x}}}{\|\hat{\boldsymbol{x}}-\bar{\boldsymbol{x}}\|})} + (\zeta U(\boldsymbol{x_1}) + (1-\zeta)U(\boldsymbol{x_2}))\,.$$

Based on the assumption of induction,

$$U(\bar{\boldsymbol{x}}) \;\leq\; \frac{\|\hat{\boldsymbol{x}} - \bar{\boldsymbol{x}}\|}{f_1} + \zeta\left(\frac{d(\boldsymbol{x_1})}{f_1} + q_2\right) + (1-\zeta)\left(\frac{d(\boldsymbol{x_2})}{f_1} + q_2\right).$$

By the convexity of Ω, the distance-to-boundary function $d(\boldsymbol{x})$ is concave and $d(\hat{\boldsymbol{x}}) \geq \zeta d(\boldsymbol{x_1}) + (1-\zeta) d(\boldsymbol{x_2})$. Therefore,

$$U(\bar{\boldsymbol{x}}) \;\leq\; \frac{1}{f_1}(\|\hat{\boldsymbol{x}} - \bar{\boldsymbol{x}}\| + d(\hat{\boldsymbol{x}})) + q_2 \;=\; \frac{d(\bar{\boldsymbol{x}})}{f_1} + q_2,$$

which completes the proof by induction. (The base of induction is obvious, since only the mesh points on the boundary $\partial\Omega$ are already *Accepted* when the algorithm starts.)

If $\boldsymbol{x}$ is inside Ω but is not a mesh point, then it lies in some simplex $\boldsymbol{x_1 x_2 x_3}$ and there exist $\zeta_1, \zeta_2, \zeta_3 \geq 0$ such that

(40)
$$\zeta_1 + \zeta_2 + \zeta_3 = 1, \quad \boldsymbol{x} = \zeta_1 \boldsymbol{x_1} + \zeta_2 \boldsymbol{x_2} + \zeta_3 \boldsymbol{x_3}, \quad U(\boldsymbol{x}) = \zeta_1 U(\boldsymbol{x_1}) + \zeta_2 U(\boldsymbol{x_2}) + \zeta_3 U(\boldsymbol{x_3}).$$

Once again, using the concavity of the distance function,

$$\begin{aligned} U(\boldsymbol{x}) &\leq \zeta_1 \left(\frac{d(\boldsymbol{x_1})}{f_1} + q_2 \right) + \zeta_2 \left(\frac{d(\boldsymbol{x_2})}{f_1} + q_2 \right) + \zeta_3 \left(\frac{d(\boldsymbol{x_3})}{f_1} + q_2 \right) \\ &\leq q_2 + \frac{1}{f_1} \left(\zeta_1 d(\boldsymbol{x_1}) + \zeta_2 d(\boldsymbol{x_2}) + \zeta_3 d(\boldsymbol{x_3}) \right) \;\leq\; \frac{d(\boldsymbol{x})}{f_1} + q_2. \qquad \square \end{aligned}$$

The obtained bound is "uniform" since it is independent of the diameter h of the mesh X. We also note that a uniform upper bound on U can be derived even for a nonconvex Ω, assuming that η remains bounded and the boundary $\partial\Omega$ is adequately represented by the mesh as h tends to zero.

7.1.3. Relaxed monotonicity of the *Accepted*. In contrast with Dijkstra-like Eikonal solvers, the OUM introduced in section 6 is not computing (and accepting) the values in a monotone fashion: "$\boldsymbol{x_i}$ is *Accepted* after $\boldsymbol{x_j}$" does not imply "$U(\boldsymbol{x_i}) \geq U(\boldsymbol{x_j})$" (see Remark 6.5). However, a weaker monotonicity property can still be formulated, based on the evolution of AF during the computation. Recall that AF is defined as the set of the line segments $\boldsymbol{x_j x_k}$, where $\boldsymbol{x_j}$ and $\boldsymbol{x_k}$ are adjacent mesh points on the $AcceptedFront$ such that there exists a $Considered$ mesh point $\boldsymbol{x_i}$ adjacent to both $\boldsymbol{x_j}$ and $\boldsymbol{x_k}$. Define U_{min}^{AF} (and U_{max}^{AF}) as the min (max) value of U on the set AF. Note that, since U is defined by the linear interpolation, both U_{min}^{AF} and U_{max}^{AF} are attained at the mesh points.

The following definitions are useful for discussing the evolution of $AcceptedFront$:

— $AF_{\bar{\boldsymbol{x}}}$ is the state of AF *immediately before* $\bar{\boldsymbol{x}}$ *is Accepted.*

— $U_{min}^{AF_{\bar{\boldsymbol{x}}}}$ and $U_{max}^{AF_{\bar{\boldsymbol{x}}}}$ are the minimum and maximum values of U on $AF_{\bar{\boldsymbol{x}}}$.

— $AF^{\bar{\boldsymbol{x}}}$ is the state of AF *immediately after* $\bar{\boldsymbol{x}}$ *is Accepted.*

— $U_{min}^{AF^{\bar{\boldsymbol{x}}}}$ and $U_{max}^{AF^{\bar{\boldsymbol{x}}}}$ are the minimum and maximum values of U on $AF^{\bar{\boldsymbol{x}}}$.

LEMMA 7.3 (Monotonicity of AF's evolution). *Suppose that h_{min} is the smallest triangle height in the triangulated mesh X on Ω. Then the following weak monotonicity results hold for the numerical solution U:*

(i)

$$U_{min}^{AF_{\bar{\boldsymbol{x}}}} + \frac{h_{min}}{f_2} \;\leq\; U(\bar{\boldsymbol{x}}) \;\leq\; U_{max}^{AF_{\bar{\boldsymbol{x}}}} + \frac{h}{f_1}. \tag{41}$$

(ii)

$$U_{min}^{AF_{\bar{\boldsymbol{x}}}} \;\leq\; U_{min}^{AF^{\bar{\boldsymbol{x}}}}. \tag{42}$$

(iii) *If* $\boldsymbol{x_i}$ *is Accepted before* $\boldsymbol{x_j}$, *then* $U_{min}^{AF_{\boldsymbol{x_i}}} \le U_{min}^{AF_{\boldsymbol{x_j}}}$.

(iv) *If* $U_{max}^{AF_{\bar{\boldsymbol{x}}}} \le U_{min}^{AF_{\bar{\boldsymbol{x}}}} + \frac{h}{f_1}$, *then* $U_{max}^{AF^{\bar{\boldsymbol{x}}}} \le U_{min}^{AF^{\bar{\boldsymbol{x}}}} + \frac{h}{f_1}$.

Proof. (i) Let $\boldsymbol{x_1}$ be a mesh point on $AF_{\bar{\boldsymbol{x}}}$ such that $U(\boldsymbol{x_1}) = U_{min}^{AF_{\bar{\boldsymbol{x}}}}$. Since it is on *AcceptedFront* immediately before $\bar{\boldsymbol{x}}$ is *Accepted*, there exists at that time a *Considered* mesh point $\boldsymbol{x_2}$ adjacent to $\boldsymbol{x_1}$. Thus,

$$V(\boldsymbol{x_2}) \le U(\boldsymbol{x_1}) + \frac{\|\boldsymbol{x_2} - \boldsymbol{x_1}\|}{f(\boldsymbol{x_2}, \frac{\boldsymbol{x_2}-\boldsymbol{x_1}}{\|\boldsymbol{x_2}-\boldsymbol{x_1}\|})}.$$

Since $\bar{\boldsymbol{x}}$ is about to be *Accepted*, $U(\bar{\boldsymbol{x}}) = V(\bar{\boldsymbol{x}}) \le V(\boldsymbol{x_2}) \le U_{min}^{AF_{\bar{\boldsymbol{x}}}} + \frac{h}{f_1}$. On the other hand,

$$U(\bar{\boldsymbol{x}}) = V(\bar{\boldsymbol{x}}) = U(\tilde{\boldsymbol{x}}) + \frac{\|\tilde{\boldsymbol{x}} - \bar{\boldsymbol{x}}\|}{f(\bar{\boldsymbol{x}}, \frac{\tilde{\boldsymbol{x}}-\bar{\boldsymbol{x}}}{\|\tilde{\boldsymbol{x}}-\bar{\boldsymbol{x}}\|})}$$

for some $\tilde{\boldsymbol{x}} \in AF_{\bar{\boldsymbol{x}}}$. Thus, $U(\bar{\boldsymbol{x}}) \ge U_{min}^{AF_{\bar{\boldsymbol{x}}}} + \frac{h_{min}}{f_2}$.

(ii) As $\bar{\boldsymbol{x}}$ is *Accepted*, several mesh points might be removed from the *AcceptedFront*, but the only point possibly added to the *AcceptedFront* is $\bar{\boldsymbol{x}}$ itself. ($\bar{\boldsymbol{x}}$ will be added if there still is a not-yet-*Accepted* mesh point adjacent to it.) Since $U_{min}^{AF_{\bar{\boldsymbol{x}}}} \le U(\bar{\boldsymbol{x}})$, it follows that $U_{min}^{AF_{\bar{\boldsymbol{x}}}} \le U_{min}^{AF^{\bar{\boldsymbol{x}}}}$.

(iii) This point follows trivially by induction from the inequality (42).

(iv) Since $\bar{\boldsymbol{x}}$ is the only point possibly added to the *AcceptedFront*,

$$U_{max}^{AF^{\bar{\boldsymbol{x}}}} \le \max\left(U_{max}^{AF_{\bar{\boldsymbol{x}}}}, U(\bar{\boldsymbol{x}})\right) \le U_{min}^{AF_{\bar{\boldsymbol{x}}}} + \frac{h}{f_1} \le U_{min}^{AF^{\bar{\boldsymbol{x}}}} + \frac{h}{f_1}. \qquad \square$$

Remark 7.4. It immediately follows from the above Lemma that if $q_2 \le q_1 + h/f_1$, then $U_{max}^{AF} \le U_{min}^{AF} + h/f_1$ at all times. Thus, if the exit time-penalty q is approximately constant on $\partial\Omega$, then the AF will be approximately a level set of U throughout the computation. Moreover, even if q is not approximately constant, the AF will still approximate a level set of U as soon as U_{min}^{AF} becomes bigger than $(q_2 - h/f_1)$.

7.1.4. Uniform Lipschitz-continuity.

LEMMA 7.5. (i) *Let* $L_1 = \eta/f_1$. *If* $\boldsymbol{x_1}$ *and* $\boldsymbol{x_2}$ *are two adjacent mesh points inside* Ω, *then*

$$|U(\boldsymbol{x_1}) - U(\boldsymbol{x_2})| \le L_1\|\boldsymbol{x_1} - \boldsymbol{x_2}\|. \tag{43}$$

(ii) *Let* $L_2 = \eta L_1$. *If* $\nabla U(\boldsymbol{x})$ *is defined for some* $\boldsymbol{x} \in \Omega\backslash\partial\Omega$ *such that* $d(\boldsymbol{x}) > h$ *(i.e.,* $\boldsymbol{x}$ *is not in a simplex immediately adjacent to* $\partial\Omega$*), then*

$$\|\nabla U(\boldsymbol{x})\| \le L_2. \tag{44}$$

(iii) *Finally, for arbitrary points* $\boldsymbol{x_1}, \boldsymbol{x_2} \in \Omega$,

$$|U(\boldsymbol{x_1}) - U(\boldsymbol{x_2})| \le L_2\|\boldsymbol{x_1} - \boldsymbol{x_2}\|. \tag{45}$$

Proof. (i) Suppose that $\boldsymbol{x_1}, \boldsymbol{x_2} \in \Omega\backslash\partial\Omega$ are two adjacent mesh points. Without loss of generality, assume that $\boldsymbol{x_1}$ was *Accepted* before $\boldsymbol{x_2}$. Thus, immediately before $\boldsymbol{x_2}$ is *Accepted*, $\boldsymbol{x_1}$ will still be on the *AcceptedFront* and

$$U(\boldsymbol{x_2}) \le \frac{\|\boldsymbol{x_1} - \boldsymbol{x_2}\|}{f(\boldsymbol{x_2}, \frac{\boldsymbol{x_1}-\boldsymbol{x_2}}{\|\boldsymbol{x_1}-\boldsymbol{x_2}\|})} + U(\boldsymbol{x_1}) \le \frac{\|\boldsymbol{x_1} - \boldsymbol{x_2}\|}{f_1} + U(\boldsymbol{x_1}) \le L_1\|\boldsymbol{x_1} - \boldsymbol{x_2}\| + U(\boldsymbol{x_1}).$$

Since U's are not necessarily *Accepted* in the ascending order, it is not generally true that $U(\boldsymbol{x_2}) \geq U(\boldsymbol{x_1})$, but from Lemma 7.3,

$$\begin{aligned} U(\boldsymbol{x_1}) &\leq U_{min}^{AF\boldsymbol{x_1}} + \frac{h}{f_1} \leq U_{min}^{AF\boldsymbol{x_2}} + \frac{h}{f_1} \leq U(\boldsymbol{x_2}) + \frac{h}{f_1} \\ &= U(\boldsymbol{x_2}) + \eta\frac{h_{min}}{f_1} \leq U(\boldsymbol{x_2}) + L_1\|\boldsymbol{x_1} - \boldsymbol{x_2}\|, \end{aligned}$$

which concludes the proof of inequality (43).

(ii) Let $\boldsymbol{x_1}, \boldsymbol{x_2}, \boldsymbol{x_3}$ be the vertices of the simplex in the mesh X which contains $\boldsymbol{x}$. Since $d(\boldsymbol{x}) > h$, we know that $\boldsymbol{x_1}$, $\boldsymbol{x_2}$, and $\boldsymbol{x_3}$ are also inside Ω, i.e., not on the boundary. Inside each simplex, U is defined by the linear interpolation, and ∇U is a constant. Whatever the direction of ∇U, a straight line parallel to it passes through one of the vertices and intersects the opposite side of the triangle. Without loss of generality, assume that that line passes through $\boldsymbol{x_1}$ and intersects the side $\boldsymbol{x_2 x_3}$ at the point $\boldsymbol{x_4}$. Since $\boldsymbol{x_4}$ lies on $\boldsymbol{x_2 x_3}$, either $(\boldsymbol{x_2} - \boldsymbol{x_1}) \cdot (\boldsymbol{x_4} - \boldsymbol{x_1}) \geq \|\boldsymbol{x_4} - \boldsymbol{x_1}\|^2$ or $(\boldsymbol{x_3} - \boldsymbol{x_1}) \cdot (\boldsymbol{x_4} - \boldsymbol{x_1}) \geq \|\boldsymbol{x_4} - \boldsymbol{x_1}\|^2$. Without loss of generality, assume the latter. Since $\|\boldsymbol{x_4} - \boldsymbol{x_1}\| \geq h_{min}$,

$$\|\nabla U\| h_{min} \leq \|\nabla U\| \|\boldsymbol{x_4} - \boldsymbol{x_1}\| = |U(\boldsymbol{x_4}) - U(\boldsymbol{x_1})| \leq |U(\boldsymbol{x_3}) - U(\boldsymbol{x_1})| \leq L_1 h.$$

Thus, $\|\nabla U\| \leq L_1 \frac{h}{h_{min}} = L_2$.

(iii) This point is obvious, since U is piecewise linear, with the slope bounded by L_2 in every simplex. □

Remark 7.6. Better estimates of L_1 and L_2 can be derived if f and q are smooth and h is sufficiently small. However, to prove the uniform convergence of U to the value function u, it is just necessary to show that some such L_2 independent of h does indeed exist. The dependence of L_2 upon η is not dangerous: if the triangulated mesh X_r does not become more and more "degenerate" as $h_r \to 0$, then η_r will be bounded.

7.2. Convergence to a viscosity solution.

THEOREM 7.7. *Consider a sequence of meshes* $\{X_r\}$ *such that* $h_r \to 0$ *but* $\eta^r = \frac{h_r}{h_{r_{min}}} < \eta$ *as* $r \to \infty$. *Let* U^r *be the approximate solution obtained on the mesh* X_r *by the algorithm described in section* 6. *As* $h_r \to 0$, U^r *uniformly converges to the viscosity solution of* (22) *(defined by the inequalities* (23) *and* (24) *in section* 3.3*).*

Proof. Since $\{U^r\}$ are bounded and uniformly Lipschitz-continuous, by the Arzela–Ascoli theorem, there exists a subsequence $\{X_p\}$ of the sequence $\{X_r\}$ such that $h_p \to 0$ as $p \to \infty$, and a function u such that $U^p \to u$ uniformly as $p \to \infty$. Boundedness and uniform continuity of u immediately follow from the properties of U^p.

(i) Consider any function $\phi \in C_c^\infty(\Omega)$, and suppose that $(u - \phi)$ has a *strict local minimum* at $\boldsymbol{x_0} \in \Omega$. Define B_δ to be the closed ball of radius δ around $\boldsymbol{x_0}$. Then there exists some $\delta > 0$ such that $B_\delta \subset \Omega$ and $\boldsymbol{x} \in B_\delta$ implies

$$(u - \phi)(\boldsymbol{x_0}) < (u - \phi)(\boldsymbol{x}). \tag{46}$$

If $D_2(\boldsymbol{x})$ is the matrix of second derivatives of $\phi(\boldsymbol{x})$, then there exists $\mu > 0$ such that $\|D_2(\boldsymbol{x})\|_2 \leq \mu$ for all $\boldsymbol{x} \in B_\delta$. Now let $\boldsymbol{x_0}^p$ be a minimum point for $(U^p - \phi)$ over B_δ; from (46) and from the uniform convergence of U^p's it follows that

$$\lim_{p \to \infty} \boldsymbol{x_0}^p = \boldsymbol{x_0}. \tag{47}$$

If $\boldsymbol{x_0}^p$ is not a mesh point of X_p and lies in the interior of some simplex s, we define $\boldsymbol{x_1}^p$ to be the vertex of s closest to it. If $\boldsymbol{x_0}^p$ lies on the edge of a simplex, we take $\boldsymbol{x_1}^p$ to be the closest endpoint of that edge; finally, we use $\boldsymbol{x_1}^p = \boldsymbol{x_0}^p$ if $\boldsymbol{x_0}^p$ is a mesh point. In any case, since ϕ is smooth and U^p is linear on every simplex, we know that $U^p(\boldsymbol{x_1}^p) - U^p(\boldsymbol{x_0}^p) = \nabla\phi(\boldsymbol{x_0}^p) \cdot (\boldsymbol{x_1}^p - \boldsymbol{x_0}^p)$ and

$$(U^p - \phi)(\boldsymbol{x_1}^p) \leq (U^p - \phi)(\boldsymbol{x_0}^p) + \frac{\mu {h_p}^2}{2}. \tag{48}$$

Moreover, using inequality (48), we obtain for every $\boldsymbol{x} \in B_\delta$

$$\begin{aligned} \phi(\boldsymbol{x}) - \phi(\boldsymbol{x_1}^p) &= (\phi(\boldsymbol{x}) - \phi(\boldsymbol{x_0}^p)) + (\phi(\boldsymbol{x_0}^p) - \phi(\boldsymbol{x_1}^p)) \\ &\leq (U^p(\boldsymbol{x}) - U^p(\boldsymbol{x_0}^p)) + \left(U^p(\boldsymbol{x_0}^p) - U^p(\boldsymbol{x_1}^p) + \frac{\mu {h_p}^2}{2} \right) = U^p(\boldsymbol{x}) - U^p(\boldsymbol{x_1}^p) + \frac{\mu {h_p}^2}{2}. \end{aligned} \tag{49}$$

Since $\|\boldsymbol{x_0}^p - \boldsymbol{x_1}^p\| \leq h_p$, it is also clear that $\lim_{p \to \infty} \boldsymbol{x_1}^p = \boldsymbol{x_0}$. So, for big enough p, $h_p \Upsilon \leq \delta$; thus, by the update formula (31), there exists $\tilde{\boldsymbol{x}}^{\boldsymbol{p}} \in AF_{\boldsymbol{x_1}^p} \bigcap B_\delta$ such that

$$U^p(\boldsymbol{x_1}^p) = \frac{\tau_p}{f(\boldsymbol{x_1}^p, \boldsymbol{a}^p)} + U^p(\tilde{\boldsymbol{x}}^{\boldsymbol{p}}), \tag{50}$$

where $\tau_p = \|\tilde{\boldsymbol{x}}^p - \boldsymbol{x_1}^p\|$ and $\boldsymbol{a}^p = \frac{\tilde{\boldsymbol{x}}^p - \boldsymbol{x_1}^p}{\|\tilde{\boldsymbol{x}}^p - \boldsymbol{x_1}^p\|}$.

Using the smoothness of ϕ, the inequality (49), and the equality (50), we obtain

$$\begin{aligned} \nabla\phi(\boldsymbol{x_1}^p) \cdot \boldsymbol{a}^p + \frac{1}{f(\boldsymbol{x_1}^p, \boldsymbol{a}^p)} &\leq \frac{\phi(\boldsymbol{x_1}^p + \tau_p \boldsymbol{a}^p) - \phi(\boldsymbol{x_1}^p)}{\tau_p} + \frac{1}{f(\boldsymbol{x_1}^p, \boldsymbol{a}^p)} + \tau_p \mu \\ &\leq \frac{U^p(\boldsymbol{x_1}^p + \tau_p \boldsymbol{a}^p) - U^p(\boldsymbol{x_1}^p) + (\mu {h_p}^2/2)}{\tau_p} + \frac{1}{f(\boldsymbol{x_1}^p, \boldsymbol{a}^p)} + \tau_p \mu \\ &= \frac{U^p(\tilde{\boldsymbol{x}}^{\boldsymbol{p}}) - U^p(\boldsymbol{x_1}^p) + \tau_p / f(\boldsymbol{x_1}^p, \boldsymbol{a}^p)}{\tau_p} + \frac{\mu {h_p}^2}{2\tau_p} + \tau_p \mu = \frac{\mu {h_p}^2}{2\tau_p} + \tau_p \mu. \end{aligned} \tag{51}$$

Since $\tilde{\boldsymbol{x}}$ lies on the $AF_{\boldsymbol{x_1}^p}$ and $\tau_p = \|\tilde{\boldsymbol{x}}^p - \boldsymbol{x_1}^p\|$, it is at least as big as the minimal triangle height in the mesh X_p, i.e., $\tau_p \geq \frac{h_p}{\eta}$. On the other hand, $\tau_p \leq h_p \Upsilon$ because $\tilde{\boldsymbol{x}}^p \in \mathrm{NF}(\boldsymbol{x_1}^p)$. Combining these bounds with inequality (51), we see that

$$\nabla\phi(\boldsymbol{x_1}^p) \cdot \boldsymbol{a}^p + \frac{1}{f(\boldsymbol{x_1}^p, \boldsymbol{a}^p)} \leq \mu \left(\frac{\eta}{2} + \Upsilon \right) h_p. \tag{52}$$

The sequence $\{\boldsymbol{a}^p\}$ has to have a subsequence converging to some vector $\boldsymbol{b} \in S_1$; we restrict our attention to that subsequence, but will still use the subscript p to avoid further cluttering of the notation. Now we can use the continuity of f, the smoothness of ϕ, and the uniformity of convergence of U^p to pass to a limit as $p \to \infty$ in the inequality (52):

$$\nabla\phi(\boldsymbol{x_0}) \cdot \boldsymbol{b} + \frac{1}{f(\boldsymbol{x_0}, \boldsymbol{b})} \leq 0 \implies (\nabla\phi(\boldsymbol{x_0}) \cdot \boldsymbol{b}) f(\boldsymbol{x_0}, \boldsymbol{b}) + 1 \leq 0,$$

which completes the first half of the proof, since

$$\min_{\boldsymbol{a} \in S_1} \{(\nabla\phi(\boldsymbol{x_0}) \cdot \boldsymbol{a}) f(\boldsymbol{x_0}, \boldsymbol{a})\} \leq (\nabla\phi(\boldsymbol{x_0}) \cdot \boldsymbol{b}) f(\boldsymbol{x_0}, \boldsymbol{b}).$$

(ii) Consider any function $\phi \in C_c^\infty(\Omega)$, and suppose that $(u-\phi)$ has a *strict local maximum* at $\boldsymbol{x_0} \in \Omega$. Define B_δ to be the closed ball of radius δ around $\boldsymbol{x_0}$. Then there exists some $\delta > 0$ such that $B_\delta \subset \Omega$ and $\boldsymbol{x} \in B_\delta$ implies

$$(u-\phi)(\boldsymbol{x_0}) > (u-\phi)(\boldsymbol{x}). \tag{53}$$

If $\nabla\phi(\boldsymbol{x_0}) = 0$, then the inequality (24) is trivially satisfied. Thus, we will further assume that $\|\nabla\phi(x)\| \geq \nu > 0$ for all $\boldsymbol{x} \in B_\delta$. If $D_2(\boldsymbol{x})$ is the matrix of second derivatives of $\phi(\boldsymbol{x})$, then there exists $\mu > 0$ such that $\|D_2(\boldsymbol{x})\|_2 \leq \mu$ for all $\boldsymbol{x} \in B_\delta$. Now let $\boldsymbol{x_0}^p$ be a maximum point for $(U^p - \phi)$ over B_δ; from (53) and from the uniform convergence of U^p's it follows that

$$\lim_{p\to\infty} \boldsymbol{x_0}^p = \boldsymbol{x_0}. \tag{54}$$

If $\boldsymbol{x_0}^p$ is not a mesh point of X_p and lies in the interior of some simplex s, we define $\boldsymbol{x_1}^p$ to be the vertex of s closest to it. If $\boldsymbol{x_0}^p$ lies on the edge of a simplex, we take $\boldsymbol{x_1}^p$ to be the closest endpoint of that edge; finally, we use $\boldsymbol{x_1}^p = \boldsymbol{x_0}^p$ if $\boldsymbol{x_0}^p$ is a mesh point. In any case, since ϕ is smooth and U^p is linear on every simplex, we know that $U^p(\boldsymbol{x_1}^p) - U^p(\boldsymbol{x_0}^p) = \nabla\phi(\boldsymbol{x_0}^p) \cdot (\boldsymbol{x_1}^p - \boldsymbol{x_0}^p)$ and

$$(U^p - \phi)(\boldsymbol{x_1}^p) \geq (U^p - \phi)(\boldsymbol{x_0}^p) - \frac{\mu {h_p}^2}{2}. \tag{55}$$

Moreover, using the inequality (55), we obtain for every $\boldsymbol{x} \in B_\delta$,

$$\begin{aligned}&\phi(\boldsymbol{x}) - \phi(\boldsymbol{x_1}^p) = (\phi(\boldsymbol{x}) - \phi(\boldsymbol{x_0}^p)) + (\phi(\boldsymbol{x_0}^p) - \phi(\boldsymbol{x_1}^p))\\ &\geq (U^p(\boldsymbol{x}) - U^p(\boldsymbol{x_0}^p)) + \left(U^p(\boldsymbol{x_0}^p) - U^p(\boldsymbol{x_1}^p) - \frac{\mu {h_p}^2}{2}\right) = U^p(\boldsymbol{x}) - U^p(\boldsymbol{x_1}^p) - \frac{\mu {h_p}^2}{2}.\end{aligned} \tag{56}$$

Since $\|\boldsymbol{x_0}^p - \boldsymbol{x_1}^p\| \leq h_p$, it is also clear that $\lim_{p\to\infty} \boldsymbol{x_1}^p = \boldsymbol{x_0}$. As proven in section 3.4,

$$\min_{\boldsymbol{a}\in S_1} \{(\nabla\phi(\boldsymbol{x_0}) \cdot \boldsymbol{a}) f(\boldsymbol{x_0}, \boldsymbol{a})\} = \min_{\boldsymbol{a}\in S_1^{\phi,\boldsymbol{x_0}}} \{(\nabla\phi(\boldsymbol{x_0}) \cdot \boldsymbol{a}) f(\boldsymbol{x_0}, \boldsymbol{a})\}. \tag{57}$$

Thus, to prove (24), we need to consider only $\boldsymbol{a} \in S_1^{\phi,\boldsymbol{x_0}}$, i.e., only $\boldsymbol{a}$ such that

$$\boldsymbol{a} \cdot \nabla\phi(\boldsymbol{x_0}) \leq -\Upsilon^{-1}\|\nabla\phi(\boldsymbol{x_0})\| \leq -\nu\Upsilon^{-1}. \tag{58}$$

Suppose that a particular $\boldsymbol{a} \in S_1^{\phi,\boldsymbol{x_0}}$ was chosen. We would like to show that, for sufficiently small h_p,

(*) *if we start at $\boldsymbol{x_1}^p$ and go some distance $\tau_p = O(h_p)$ in the direction $\boldsymbol{a}$, then we will have to intersect the $AF_{\boldsymbol{x_1}^p}$.*

If the local maximum were attained at the mesh point (i.e., the case $\boldsymbol{x_1}^p = \boldsymbol{x_0}^p$) and the test function ϕ were linear, then (*) would be almost obvious: ϕ would be linearly decreasing in the direction $\boldsymbol{a}$, and so would U^p, because of the local maximum condition, and, as we know from Lemma 7.3,

$$U^p(x) \geq U_{min}^{AF_{\boldsymbol{x_1}^p}} + \frac{h_p}{\eta f_2}$$

for every mesh point $x \in X_p$ *Accepted* after the point $\boldsymbol{x_1}^p$. Since ϕ is generally not linear and $\boldsymbol{x_1}^p \neq \boldsymbol{x_0}^p$, we will have to be more careful.

Suppose we start moving from $\boldsymbol{x_1}^p$ in the direction $\boldsymbol{a}$ for some time t_p. Using the inequality (56), the smoothness of ϕ, and the inequality (58), we obtain

$$\begin{aligned} U^p(\boldsymbol{x_1}^p + t_p\boldsymbol{a}) - U^p(\boldsymbol{x_1}^p) &\leq (\phi(\boldsymbol{x_1}^p + t_p\boldsymbol{a}) - \phi(\boldsymbol{x_1}^p)) + \frac{\mu {h_p}^2}{2} \\ &\leq \left(t_p(\nabla\phi(\boldsymbol{x_1}^p)\cdot\boldsymbol{a}) + \frac{\mu {t_p}^2}{2} \right) + \frac{\mu {h_p}^2}{2} \leq -\nu t_p \Upsilon^{-1} + \mu\frac{{h_p}^2 + {t_p}^2}{2}. \end{aligned} \tag{59}$$

In order to prove (*) we will need the following inequality to be satisfied:

$$-\nu t_p \frac{f_1}{f_2} + \mu\frac{{h_p}^2 + {t_p}^2}{2} \leq -\frac{h_p}{f_1}. \tag{60}$$

Let $t_p = Ah_p$. We will now show that the constant A can be chosen such that (60) holds for small enough h_p. Indeed, (60) can be rewritten as

$$h_p^2\left(\mu\frac{1+A^2}{2}\right) + h_p\left(\frac{1}{f_1} - \nu A\frac{f_1}{f_2}\right) \leq 0. \tag{61}$$

If A is such that $(\frac{1}{f_1} - \nu A\frac{f_1}{f_2}) > 0$, then we also have that (60) is satisfied for all $h_p \in [0, (\nu A\frac{f_1}{f_2} - f_2)\frac{2}{\mu(1+A^2)}]$. Thus, choosing any $A > (\frac{f_2^2}{\nu f_1})$, we ensure that (60) is satisfied for the sufficiently small h_p. Combining this with the inequality (59) and using the monotonicity result in Lemma 7.3, we see that

$$U^p(\boldsymbol{x_1}^p + t_p\boldsymbol{a}) \leq U^p(\boldsymbol{x_1}^p) - \frac{h_p}{f_1} \leq U_{min}^{AF_{\boldsymbol{x_1}^p}},$$

i.e., the point $(\boldsymbol{x_1}^p + t_p\boldsymbol{a})$ cannot be inside the $AF_{\boldsymbol{x_1}^p}$. Since $\boldsymbol{x_1}^p$ is inside $AF_{\boldsymbol{x_1}^p}$, that means that (*) holds: there exists some $\tau_p \in [0, t_p]$ such that

$$\tilde{\boldsymbol{x}}^{\boldsymbol{p}} = (\boldsymbol{x_1}^p + \tau_p\boldsymbol{a}) \in AF_{\boldsymbol{x_1}^p}.$$

By Lemma 7.1,

$$U^p(\boldsymbol{x_1}^p) = W^p(\boldsymbol{x_1}^p) \leq \frac{\tau_p}{f(\boldsymbol{x_1}^p, \boldsymbol{a})} + U^p(\tilde{\boldsymbol{x}}^{\boldsymbol{p}}). \tag{62}$$

The remainder of the proof is similar to what we have done to prove (i).

Using the smoothness of ϕ, the inequality (56), and the inequality (62), we obtain

$$\begin{aligned} \nabla\phi(\boldsymbol{x_1}^p)\cdot\boldsymbol{a} + \frac{1}{f(\boldsymbol{x_1}^p, \boldsymbol{a})} &\geq \frac{\phi(\boldsymbol{x_1}^p + \tau_p\boldsymbol{a}) - \phi(\boldsymbol{x_1}^p)}{\tau_p} + \frac{1}{f(\boldsymbol{x_1}^p, \boldsymbol{a})} - \tau_p\mu \\ &\geq \frac{U^p(\boldsymbol{x_1}^p + \tau_p\boldsymbol{a}) - U^p(\boldsymbol{x_1}^p) - (\mu {h_p}^2/2)}{\tau_p} + \frac{1}{f(\boldsymbol{x_1}^p, \boldsymbol{a})} - \tau_p\mu \\ &= \frac{U^p(\tilde{\boldsymbol{x}}^{\boldsymbol{p}}) - U^p(\boldsymbol{x_1}^p) + \tau_p/f(\boldsymbol{x_1}^p, \boldsymbol{a})}{\tau_p} - \frac{\mu {h_p}^2}{2\tau_p} - \tau_p\mu \geq -\frac{\mu {h_p}^2}{2\tau_p} - \tau_p\mu. \end{aligned} \tag{63}$$

Since $\tilde{\boldsymbol{x}}$ lies on the $AF_{\boldsymbol{x_1}^p}$ and $\tau_p = \|\tilde{\boldsymbol{x}}^p - \boldsymbol{x_1}^p\|$, it is at least as big as the minimal triangle height in the mesh X_p, i.e., $\tau_p \geq \frac{h_p}{\eta}$. On the other hand, $\tau_p \leq t_p = Ah_p$, where A is a constant chosen for this particular ϕ, but independent of h_p. Combining these bounds with inequality (63), we see that

$$\nabla\phi(\boldsymbol{x_1}^p)\cdot\boldsymbol{a}+\frac{1}{f(\boldsymbol{x_1}^p,\boldsymbol{a})}\geq-\mu\left(\frac{\eta}{2}+A\right)h_p. \tag{64}$$

We can now use the continuity of f, the smoothness of ϕ, and the uniformness of convergence of U^p to pass to a limit as $p\to\infty$ in the inequality (64):

$$\nabla\phi(\boldsymbol{x_0})\cdot\boldsymbol{a}+\frac{1}{f(\boldsymbol{x_0},\boldsymbol{a})}\geq 0 \implies (\nabla\phi(\boldsymbol{x_0})\cdot\boldsymbol{a})f(\boldsymbol{x_0},\boldsymbol{a})+1\geq 0,$$

which completes the proof of inequality (24), since $\boldsymbol{a}$ was chosen to be an arbitrary vector in $S_1^{\phi,\boldsymbol{x_0}}$.

In this proof we have several times passed to a subsequence. If some other subsequence of U^r converges to a different limit $\bar{u}$, the above argument could be repeated for that subsequence to prove that $\bar{u}$ also satisfies (23) and (24). The uniqueness of the viscosity solution (proven in [10, 9]) implies $u=\bar{u}$; thus, the entire sequence U^r converges to u uniformly as $r\to\infty$. □

7.3. Additional comments. First, the above proof of convergence, as well as the preceding lemmas, can be easily repeated for the corresponding method in higher dimensions. Moreover, if the update formula (31) in the description of the method is replaced by any other update-from-a-single-simplex formula such that

1. the update formula is consistent (converges to the PDE as $h\to 0$),
2. the update formula is upwinding (the update is computed/accepted only from the simplex, which contains the characteristic direction), and
3. the update formula is stable (there exists a uniform bound for U),

then the resulting numerical solutions should converge to the viscosity solution of the Hamilton–Jacobi–Bellman PDE. This conjecture is the basis for the methods based on finite-difference discretization discussed in section 8.2.

Second, the above proof uses the continuity of the speed function f, but not the Lipschitz-continuity. Thus, if the viscosity solution can be defined for $f\in C(\Omega)$, then the above proof will still be valid. The majority of the control-theoretic papers to which we are referring in this work require the speed of motion f (or, equivalently, the running cost K) to be Lipschitz-continuous. Nevertheless, the value function u can be defined for a much broader class of control problems, including those for which f is discontinuous. Some examples of our method applied to such problems can be found in section 9.4. Numerical evidence confirms that the method described above works correctly in that more general case. This is not surprising, since Bellman's optimality principle is valid even when the speed $f(\boldsymbol{x},\boldsymbol{a})$ is very ill-behaved, and our numerical methods merely mimic the logic of that principle.

8. Front propagation problems and OUMs. Anisotropic aspects of front propagation have been studied in several different contexts, including geometric optics, geophysics, tomography, and crystal growth; our primary emphasis is on an application-neutral analysis, concentrating on the properties of the particular class of static Hamilton–Jacobi PDEs. We begin with the correspondence between anisotropic optimal trajectory problems and a class of anisotropic front expansion (contraction) problems described in section 8.1. Our goal is to determine a set of the anisotropic front expansion (contraction) problems, which can be solved efficiently by our semi-Lagrangian OUM, and to construct a family of OUMs based on fully Eulerian discretizations.

8.1. Front propagation problems: Static Hamilton–Jacobi approach. Consider a simple curve Γ_t moving in R^2 in the direction normal to itself with some speed $F(\boldsymbol{x}, \boldsymbol{n})$, where $\boldsymbol{n}$ is an "outwards pointing" unit vector normal to the curve as it passes through the point $\boldsymbol{x}$. If the curve is not smooth, then $\boldsymbol{n}$ is not defined, but a geometric construction based on a variant of Huygens' principle can still be used to define the evolution of Γ_t. An important subclass of the front propagation problems consists of applications in which the speed function F never changes sign. If the function F is strictly positive (or negative), then the front always expands (or contracts). This implies that the front passes through each point only once. Thus, we can define $u(\boldsymbol{x})$ to be the arrival time: $u(\boldsymbol{x}) = t \iff \boldsymbol{x} \in \Gamma_t$. If F is always nonnegative, the outwards unit normal vector can be expressed as $\boldsymbol{n}(\boldsymbol{x}) = \frac{\nabla u(\boldsymbol{x})}{\|\nabla u(\boldsymbol{x})\|}$ and, assuming that $\boldsymbol{n}(\boldsymbol{x})$ is always defined, it is straightforward to show that the arrival time u satisfies the PDE

$$\|\nabla u(\boldsymbol{x})\| F\left(\boldsymbol{x}, \frac{\nabla u(\boldsymbol{x})}{\|\nabla u(\boldsymbol{x})\|}\right) = 1, \tag{65}$$

$$u = 0 \text{ on } \Gamma_0.$$

This is a static Hamilton–Jacobi PDE of the form $H(\nabla u, \boldsymbol{x}) = 1$, where the Hamiltonian H is homogeneous of degree one in the first argument. To interpret (65) where ∇u does not exist, one normally uses the unique viscosity solution, as defined in [9, 10]. As follows from the results in [14] and [41], the level sets of the viscosity solution u of (65) will correspond to the evolution of Γ_0 defined by Huygens' principle.

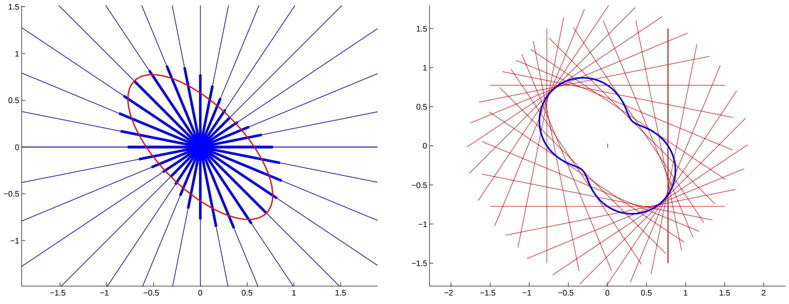
We note that, in general, the Hamiltonian $H(\nabla u, \boldsymbol{x}) = \|\nabla u\| F(\boldsymbol{x}, \frac{\nabla u}{\|\nabla u\|})$ is not convex. As shown in [14], such $H(\nabla u, \boldsymbol{x})$ can always be considered as a result of a differential game model. Several iterative numerical schemes are based on this approach (see [4] or [16], for example). However, if H is convex, it can be alternatively considered as a product of a dual min-time optimal control problem [41]. Using two interpretations of the Hamiltonian, we can show that the speeds F and f are related by a homogeneous Legendre transform:

$$F(\boldsymbol{x}, \boldsymbol{n}) = \max_{\boldsymbol{a} \in S_1} \{(\boldsymbol{n} \cdot (-\boldsymbol{a})) f(\boldsymbol{x}, \boldsymbol{a})\}, \tag{66}$$

$$f(\boldsymbol{x}, \boldsymbol{a}) = \min_{\boldsymbol{n} \in S_1,\, (\boldsymbol{n} \cdot \boldsymbol{a}) < 0} \left\{ \frac{F(\boldsymbol{x}, \boldsymbol{n})}{(-\boldsymbol{n} \cdot \boldsymbol{a})} \right\}. \tag{67}$$

Remark 8.1. We note that $F(\boldsymbol{x}, \boldsymbol{n})$ is the speed of the front's movement in the direction normal to itself (here, $\boldsymbol{n} = \frac{\nabla u(\boldsymbol{x})}{\|\nabla u(\boldsymbol{x})\|}$), whereas $f(\boldsymbol{x}, \boldsymbol{a})$ is the speed of the vehicle's motion in the direction $\boldsymbol{a}$. Correspondingly, the correct $\boldsymbol{n}$ is fully determined by the gradient direction of the function $u(\boldsymbol{x})$, while the optimal $\boldsymbol{a} \in S_1$ is determined by the direction of the characteristic passing through the point $\boldsymbol{x}$ and, therefore, is a function of the particular Hamilton–Jacobi–Bellman equation. In the isotropic case, however, there is no difference since (66) yields $f(\boldsymbol{x}) = F(\boldsymbol{x})$.

Define the vehicle's speed profile $S_f(\boldsymbol{x}) = \{\boldsymbol{a} f(\boldsymbol{x}, \boldsymbol{a}) \mid \boldsymbol{a} \in S_1\}$, its flipped (center symmetry applied) version $S_{-f}(\boldsymbol{x}) = \{-\boldsymbol{a} f(\boldsymbol{x}, \boldsymbol{a}) \mid \boldsymbol{a} \in S_1\}$, and the front propagation speed profile $S_F(\boldsymbol{x}) = \{\boldsymbol{n} F(\boldsymbol{x}, \boldsymbol{n}) \mid \boldsymbol{n} \in S_1\}$. The formulas (66) and (67) can now be



(a) (b)

FIG. 4. (a) *Using S_f to construct S_F;* (b) *using S_F to construct S_f.*

interpreted geometrically:[11]

— $F(\boldsymbol{x}, \boldsymbol{n})$ can be obtained by projecting $S_{-f}(\boldsymbol{x})$ onto a unit vector $\boldsymbol{n}$ and then by taking the maximum of this (signed) projection (see Figure 4(a));

— $S_{-f}(x)$ can be obtained as an envelope of lines perpendicular to $\boldsymbol{n}$ drawn at every point of $S_F(x)$ (see Figure 4(b)).

By the above construction, S_f is always convex. Thus, different optimal trajectory problems will yield the same Hamilton–Jacobi–Bellman equation, provided that the speed profiles have the same convex hull.[12] See [30] and the references therein for an additional discussion of Wulff shapes and mutual properties of functions related by the homogeneous Legendre transform.

Finally, we note that the correspondence between these two types of anisotropic problems can be used to build an alternative definition of Huygens' principle: using scaled speed profiles S_{-f} at each point of the wavefront instead of the circles of front-direction-dependent radius; see [14], [40], or [46].

8.2. OUMs for Eulerian discretizations. Given $0 < F_1 \leq F(\boldsymbol{x}, \boldsymbol{n}) < F_2$ for all $\boldsymbol{x}$ and $\boldsymbol{n}$, we can use the formula (67) to prove that $0 < F_1 = f_1 \leq f(\boldsymbol{x}, \boldsymbol{a}) < f_2 = F_2$ for all $\boldsymbol{x}$ and $\boldsymbol{a}$. Thus, the corresponding control problem can be treated by the OUMs with the semi-Lagrangian update formula described in section 6.

We now proceed to construct the OUMs for the fully Eulerian approximation of $\|\nabla u\| F(\boldsymbol{x}, \frac{\nabla u}{\|\nabla u\|}) = 1$. The key idea is that any consistent upwind finite difference discretization can be used to compute an update-from-a-single-simplex $V_{\boldsymbol{x_j}, \boldsymbol{x_k}}(\boldsymbol{x})$. Our derivation of such discretizations generalizes the approach used in defining the Fast Marching Method on unstructured meshes given in [38].

[11] In wave physics, $F(\boldsymbol{x}, \boldsymbol{n})$ corresponds to the *"phase velocity"* if $\boldsymbol{n}$ is the direction normal to the wavefront [11]. In crystalline variational problems, $F(\boldsymbol{x}, \boldsymbol{n})$ corresponds to the *"surface free energy"* if $\boldsymbol{n}$ is the direction normal to the surface [43]. Additionally, $f(\boldsymbol{x}, \boldsymbol{a})$ corresponds to the *"group velocity,"* i.e., the speed with which a blob of energy is moving in the direction $\boldsymbol{a}$ [11]. Finally, this speed profile is often referred to as a *"ray surface"* or *"impulse-response surface"* [11, 31]. The corresponding object in crystalline variational problems is the *"Wulff shape"*—the shape which minimizes the free surface energy for a fixed volume with no additional constraints [43].

[12] Similar geometric construction is common in tomography; the formulas (67) and (66) are related to the *inverse Radon transform* [20]. Analytic expressions for F in terms of f and for f in terms of F can be easily derived for R^2 (see [30, 46], for example). Similar formulas expressing the relationship between the *group speed* and the *phase speed* were known in wave physics [28] at least as early as 1837.

8.2.1. Upwind finite-difference discretization. Consider an unstructured triangulated mesh X of diameter h (i.e., if the mesh points $\boldsymbol{x_j}$ and $\boldsymbol{x_k}$ are adjacent, then $\|\boldsymbol{x_j} - \boldsymbol{x_k}\| \leq h$). Let $\boldsymbol{x_j}$ and $\boldsymbol{x_k}$ be two adjacent mesh points and choose some other mesh point $\boldsymbol{x} \in \Omega\backslash\partial\Omega$. Define the unit vectors $\boldsymbol{P_1} = \frac{\boldsymbol{x}-\boldsymbol{x_j}}{\|\boldsymbol{x}-\boldsymbol{x_j}\|}$ and $\boldsymbol{P_2} = \frac{\boldsymbol{x}-\boldsymbol{x_k}}{\|\boldsymbol{x}-\boldsymbol{x_k}\|}$. Assume that $\boldsymbol{P_1}$ and $\boldsymbol{P_2}$ are linearly independent, and consider the 2×2 nonsingular matrix P having $\boldsymbol{P_1}$ and $\boldsymbol{P_2}$ as its rows. Let $v_r(\boldsymbol{x})$ be the value of the directional derivative for the direction $\boldsymbol{P_r}$ evaluated at the point $\boldsymbol{x}$. Assuming that the function u is differentiable at $\boldsymbol{x}$, we have $P\nabla u(\boldsymbol{x}) = \boldsymbol{v}(\boldsymbol{x})$, where $\boldsymbol{v}(\boldsymbol{x}) = \left[\begin{smallmatrix} v_1(\boldsymbol{x}) \\ v_2(\boldsymbol{x}) \end{smallmatrix}\right]$. Recall that the front propagation equation (65) can be written as $\|\nabla u(\boldsymbol{x})\|^2 F^2(\boldsymbol{x}, \frac{\nabla u(\boldsymbol{x})}{\|\nabla u(\boldsymbol{x})\|}) = 1$, which can be restated in terms of $\boldsymbol{v}(\boldsymbol{x})$:

$$\boldsymbol{v}(\boldsymbol{x})^T (PP^T)^{-1} \boldsymbol{v}(\boldsymbol{x}) F^2 \left(\boldsymbol{x}, \frac{P^{-1}\boldsymbol{v}(\boldsymbol{x})}{\|P^{-1}\boldsymbol{v}(\boldsymbol{x})\|}\right) = 1. \tag{68}$$

To obtain the discretized equation, we now replace each v_r with a difference approximation: $v_r(\boldsymbol{x}) \approx w_r \equiv a_r U + b_r$, where the b_r's linearly depend on the values of U (and possibly of ∇U for higher-order schemes) at the mesh points $\boldsymbol{x_j}$ and $\boldsymbol{x_k}$.

Remark 8.2. In general, the choice of the difference approximation will depend upon the structure of the mesh and will also affect the rate of convergence of the method. The simplest first-order finite-difference approximation is obtained by choosing

$$a_1 = \frac{1}{\|\boldsymbol{x} - \boldsymbol{x_j}\|}, \quad b_1 = \frac{-U(\boldsymbol{x_j})}{\|\boldsymbol{x} - \boldsymbol{x_j}\|}, \quad a_2 = \frac{1}{\|\boldsymbol{x} - \boldsymbol{x_k}\|}, \quad b_2 = \frac{-U(\boldsymbol{x_k})}{\|\boldsymbol{x} - \boldsymbol{x_k}\|}.$$

Higher-order accurate operators can be built using the computed value for ∇U at the mesh points (see [38, 46] for further details).

For convenience, let $Q = (PP^T)^{-1}$ and use $\boldsymbol{v}(\boldsymbol{x}) \approx \boldsymbol{w}(\boldsymbol{x}) = V_{\boldsymbol{x_j},\boldsymbol{x_k}}(\boldsymbol{x})\boldsymbol{a} + \boldsymbol{b}$. Then the discretized version of (68) can be used as an equation for the upwind-update-from-a-single-simplex $V_{\boldsymbol{x_j},\boldsymbol{x_k}}(\boldsymbol{x})$:

$$\left((\boldsymbol{a}^T Q \boldsymbol{a})(V_{\boldsymbol{x_j},\boldsymbol{x_k}}(\boldsymbol{x}))^2 + (2\boldsymbol{a}^T Q \boldsymbol{b}) V_{\boldsymbol{x_j},\boldsymbol{x_k}}(\boldsymbol{x}) + (\boldsymbol{b}^T Q \boldsymbol{b})\right) F^2 \left(\boldsymbol{x}, \frac{P^{-1}\boldsymbol{w}}{\|P^{-1}\boldsymbol{w}\|}\right) = 1. \tag{69}$$

Remark 8.3. In the isotropic case, the analogous equation was just a quadratic (see the appendix and [38]). Equation (69) is a more complex nonlinear equation since $\boldsymbol{w}(\boldsymbol{x})$ also depends on $V_{\boldsymbol{x_j},\boldsymbol{x_k}}(\boldsymbol{x})$. In general, this equation will have to be solved approximately, and the overall efficiency of the method will also depend on the iterative numerical method used to solve (69). Since these iterations are generally unavoidable, we will consider solving this equation as a single operation in the further analysis of computational complexity. We note that the iterative zero-finding required to compute $V_{\boldsymbol{x_j},\boldsymbol{x_k}}(\boldsymbol{x})$ is *local* (i.e., $V_{\boldsymbol{x_j},\boldsymbol{x_k}}(\boldsymbol{x})$ can be computed independently from any other $V_{\boldsymbol{x_i},\boldsymbol{x_l}}(\boldsymbol{x_m})$) and thus should not be confused with the iterations necessary to solve a coupled system of nonlinear equations (such as (25)) simultaneously.

8.2.2. Upwinding criterion and combined update formula. We need to ensure that the value of $V_{\boldsymbol{x_j},\boldsymbol{x_k}}(\boldsymbol{x})$ computed from (69) is truly upwind, i.e., that the characteristic for the mesh point $\boldsymbol{x}$ lies inside the simplex $\boldsymbol{x}\boldsymbol{x_j}\boldsymbol{x_k}$. The approximate gradient $P^{-1}(V_{\boldsymbol{x_j},\boldsymbol{x_k}}(\boldsymbol{x})\boldsymbol{a} + \boldsymbol{b})$ can be used to compute an approximation to the characteristic direction $\boldsymbol{a}(\boldsymbol{x})$. For R^n, the requirement that the characteristic direction should point into the simplex $\boldsymbol{x}\boldsymbol{x_1}\ldots\boldsymbol{x_n}$ is equivalent to the condition that all the elements of the vector $(P^T)^{-1}\boldsymbol{a}(\boldsymbol{x})$ should be positive.

The unfortunate feature of this upwinding criterion is that it is based on the approximate rather than the exact characteristic direction. Due to the approximation error, it is possible that an upwinding criterion will not be satisfied even though the true characteristic for the mesh point $\boldsymbol{x}$ lies inside the simplex $\boldsymbol{x}\boldsymbol{x_j}\boldsymbol{x_k}$. If that simplex is small enough, this can happen only when one of the elements of the vector $(P^T)^{-1}\boldsymbol{a}(\boldsymbol{x})$ is close to zero, i.e., only when the characteristic direction almost coincides with $(-\boldsymbol{P_1})$ or $(-\boldsymbol{P_2})$. That corresponds to the situation in which $U(\boldsymbol{x})$ can be computed based on either $U(\boldsymbol{x_j})$ or $U(\boldsymbol{x_k})$. Thus, we define the "one-sided-update" formula in a manner consistent with the control-theoretic perspective:

$$V_{\boldsymbol{x_i}}(\boldsymbol{x}) = \frac{\|\boldsymbol{x_i}-\boldsymbol{x}\|}{f(\boldsymbol{x}, \frac{\boldsymbol{x_i}-\boldsymbol{x}}{\|\boldsymbol{x_i}-\boldsymbol{x}\|})} + U(\boldsymbol{x_i}). \tag{70}$$

Therefore, the final formula for the upwind-update-from-a-single-simplex becomes

(71)

$$V_{\boldsymbol{x_j},\boldsymbol{x_k}}(\boldsymbol{x}) = \begin{cases} \text{solution of (69)} & \text{if } \boldsymbol{P_1} \text{ and } \boldsymbol{P_2} \text{ are linearly independent} \\ & \text{and the upwinding criterion is satisfied,} \\ \min(V_{\boldsymbol{x_j}}(\boldsymbol{x}),\, V_{\boldsymbol{x_k}}(\boldsymbol{x})) & \text{otherwise.} \end{cases}$$

Using the finite-difference update formula (71) instead of the formula (31) in the algorithm described in section 6, we obtain a new OUM for solving the front expansion problem (Hamilton–Jacobi equation (65)). In fact, this defines a whole family of such methods, since different upwind finite-difference operators can be used to approximate $w_r(\boldsymbol{x})$ in (68).

We note that the resulting methods

- are single-pass and have the same computational complexity as the semi-Lagrangian OUM introduced in section 6,
- work equally well on acute and nonacute triangulated meshes,
- are applicable for a general anisotropic optimal trajectory problem described in section 3,
- can be easily extended to R^n and manifolds. (The generalizations of the mapping $\boldsymbol{n} \mapsto \boldsymbol{a}$, of (68), and of the upwinding criterion are obvious.)

Remark 8.4 (Convergence). In the appendix, we show the connection between a particular first-order Eulerian OUM (based on Remark 8.2) and the first-order semi-Lagrangian OUM (introduced in section 6); in this case, the convergence to the viscosity solution follows from section 7. However, as of right now, we do not have a proof of convergence for the general (higher-order) OUMs based on the Eulerian discretization. We rely on general convergence considerations (see the remarks following the proof of Theorem 7.7) and on the numerical evidence (section 9.4 and [39, 46]). In all of our numerical experiments the numerical solution U produced by these methods converges to the viscosity solution of the original PDE. The rate of convergence depends on the particular finite-difference operators used to approximate $w_r(\boldsymbol{x})$ in (68).

9. Implementation and numerical results.

9.1. Implementation notes. An efficient implementation of the described numerical methods for the anisotropic optimal-trajectory and front-propagation problems requires dealing with several algorithmic issues. Storing and sorting the current

AcceptedFront, for example, has to be implemented rather carefully to enable efficient search for the "*AcceptedFront* neighborhood" $\mathrm{NF}(\boldsymbol{x})$ for every *Considered* point $\boldsymbol{x}$. The inverse operation (searching for all *Considered* $\boldsymbol{x}$ such that $\bar{\boldsymbol{x}} \in \mathrm{NF}(\boldsymbol{x})$) is another major component of the implementation. Efficient use of data structures allows us to construct an algorithm with the computational complexity of $O(\Upsilon M \log M)$.

The connection between a particular class of anisotropic front-propagation and optimal-trajectory problems allows us to build both semi-Lagrangian and fully Eulerian single-pass methods. On a fixed mesh X, the computational complexity of these methods will be the same. However, the overall efficiency of each program will be affected by the chosen upwind-update-from-a-single-simplex formula. The optimal choice depends on the particular speed functions F and f and on the details of implementation: the semi-Lagrangian method requires performing a local minimization at each mesh point (using (31)), whereas the finite-differences upwind update formula requires finding the roots of the nonlinear equation (69). Generally, both the minimization and the root-finding have to be done approximately, and the overall efficiency depends on the particular numerical method used to compute the approximate update. Our implementations used the "golden section search" and the Newton–Raphson method to numerically resolve the control-theoretic and finite-difference update formulas, respectively; the implementation details of these and other numerical minimization and zero-finding techniques can be found in [42].

We note that the above complexity and efficiency discussion is limited to finding a numerical solution on a fixed grid. The speed of convergence (of the numerical approximation $U(\boldsymbol{x})$ to the viscosity solution $u(\boldsymbol{x})$ as the grid is refined) is a separate issue. Thus, the availability of the higher-order accurate upwind update formulas is a significant advantage of the Eulerian approach.

9.2. Using a local anisotropy coefficient. So far we have always used the global bounds on the speed function $0 < F_1 \leq F(\boldsymbol{x}, \boldsymbol{p}) \leq F_2$ for all $\boldsymbol{p}$ and $\boldsymbol{x}$. We now define the local bounds on F,

$$F_1(\boldsymbol{x}) = \min_{\boldsymbol{p} \in S_1} F(\boldsymbol{x}, \boldsymbol{p}), \qquad F_2(\boldsymbol{x}) = \max_{\boldsymbol{p} \in S_1} F(\boldsymbol{x}, \boldsymbol{p}),$$

and the local anisotropy coefficient $\Upsilon(\boldsymbol{x}) = \frac{F_2(\boldsymbol{x})}{F_1(\boldsymbol{x})}$.

We note that many of the lemmas stated in section 3 for $u(\boldsymbol{x})$ in terms of F_1 and F_2 can be restated in terms of $F_1(\boldsymbol{x})$ and $F_2(\boldsymbol{x})$. Most importantly, this is true for Remark 3.7, which establishes a bound on the angle between the characteristic and gradient directions. Thus, it is also possible to build the numerical method using $\Upsilon(\boldsymbol{x})$ instead of Υ in the definition of $\mathrm{NF}(\boldsymbol{x})$. Moreover, if F is smooth and the maximum/minimum in defining $F_1(\boldsymbol{x})$ and $F_2(\boldsymbol{x})$ are taken not just at the point but over some closed ball B centered at $\boldsymbol{x}$, then the resulting algorithm provably converges to the viscosity solution. (Indeed, for small enough h, $\mathrm{NF}(\boldsymbol{x}) \subset B$ even if $NF(\boldsymbol{x})$ were defined using the global anisotropy coefficient Υ.)

This observation leads to a substantially more efficient algorithm since the global anisotropy coefficient Υ can be much larger than $\sup_{\boldsymbol{x} \in \Omega} \Upsilon(\boldsymbol{x})$ for the front propagating in a strongly inhomogeneous medium.

9.3. Heuristic: Relaxing the update procedure. In the algorithm described in section 6, there are two different situations in which the tentative value $V(\boldsymbol{x})$ is recomputed for a *Considered* point $\boldsymbol{x}$:

- $V(\boldsymbol{x})$ is first computed using the entire $\mathrm{NF}(\boldsymbol{x})$ at the moment when $\boldsymbol{x}$ is added to *Considered*;

- $V(\boldsymbol{x})$ is then recalculated from at most two simplexes every time the newly *Accepted* mesh point $\bar{\boldsymbol{x}}$ belongs to $\mathrm{NF}(\boldsymbol{x})$.

If the boundary condition for the PDE is nearly constant (i.e., if $q_2 \leq q_1 + h/f_1$, where h is the diameter of the triangulated mesh), Lemma 7.3 shows that the AF will also approximate the level set throughout the execution of the algorithm. On the other hand, Lemma 3.4 shows that the optimal trajectory for $\boldsymbol{x}$ intersects a level set at some point $\tilde{\boldsymbol{x}}$ such that $\|\tilde{\boldsymbol{x}} - \boldsymbol{x}\| \leq d_1 \Upsilon$, where d_1 is the distance from $\boldsymbol{x}$ to that level set. This means that if AF were exactly the level set, the initial evaluation of $V(\boldsymbol{x})$ would capture all the necessary information about all the potential characteristic directions for $\boldsymbol{x}$; thus, the further reevaluations of $V(\boldsymbol{x})$ would not be necessary. Since AF is only approximating the level set, capturing all the necessary directions requires "widening" the set $\mathrm{NF}(\boldsymbol{x})$. Carefully combining Lemmas 7.3 and 3.4, and assuming $U \approx u$ on AF, we can show that all the characteristic directions are still covered if $\mathrm{NF}(\boldsymbol{x})$ is taken to be two times "wider":

$$\widehat{\mathrm{NF}}(\boldsymbol{x}) \;=\; \{\boldsymbol{x_j x_k} \in AF \;\mid \exists \tilde{\boldsymbol{x}} \text{ on } \boldsymbol{x_j x_k} \text{ s.t. } \|\tilde{\boldsymbol{x}} - \boldsymbol{x}\| \leq 2h\Upsilon(\boldsymbol{x})\}\,.$$

This still leads to the total of roughly 2Υ evaluations of V_s for each mesh point, but it is no longer necessary to search for all *Considered* $\boldsymbol{x}$ such that $\bar{\boldsymbol{x}} \in \mathrm{NF}(\boldsymbol{x})$ each time a new $\bar{\boldsymbol{x}}$ is *Accepted*.[13]

Furthermore, an additional update relaxation can be used with the Eulerian discretization-based methods if the boundary condition for the PDE is nearly constant. In the initial computation of $V(\boldsymbol{x})$ it is often not necessary to consider the entire $\mathrm{NF}(\boldsymbol{x})$: we can stop as soon as we find $\boldsymbol{x_j x_k} \in \mathrm{NF}(\boldsymbol{x})$ such that $V_{\boldsymbol{x_j},\boldsymbol{x_k}}(\boldsymbol{x})$ satisfies the upwinding conditions (see section 8.2.2). The viscosity solution u of the Hamilton–Jacobi PDE is Lipschitz-continuous, and therefore ∇u exists almost everywhere. As shown in [46], if u is differentiable at the point $\boldsymbol{x}$ and the vehicle's speed profile $S_f(\boldsymbol{x})$ is strictly convex, then there exists a unique optimizing control $\boldsymbol{a}(\boldsymbol{x})$. Thus, away from the shocks, there should not be multiple simplexes in $\mathrm{NF}(\boldsymbol{x})$ producing updates which satisfy the upwinding criteria.

For a fixed grid X, the numerical evidence suggests that the "relaxation" significantly improves efficiency of the program. As the grid is refined, the numerical solution obtained by the "relaxed" scheme converges to the viscosity solution—sometimes slower, but often (depending on the type of anisotropy) even faster than the solution computed by the "full-update" scheme. However, the *asymptotic order of accuracy* of the "relaxed" and "full-update" schemes seems to be the same in all of our numerical experiments (e.g., see section 9.4.1).

9.4. Numerical experiments.

In this section we consider several test problems, each of which can be described by a non-Eikonal (anisotropic) Hamilton–Jacobi PDE. In each case, the speed function is assumed to be known from the characterization of a particular application domain. For example, in the optimal-trajectory test problem, $f(\boldsymbol{x}, \boldsymbol{a})$ reflects the assumptions about the speed of the controlled vehicle, while in the seismic imaging test problem, the front-expansion speed F is derived using the assumptions about the elliptical nature of the "impulse-response surface" for the anisotropic medium.

9.4.1. Geodesic distances on manifolds.

The first test problem is to find the geodesic distance on a manifold $z = g(x, y)$. As described in [21] and [36], this can

[13] The numerical experiments indicate that a much smaller "widening" of *NF* is sufficient in practice.

be accomplished by approximating the manifold with a triangulated mesh and then solving the distance equation $\|\nabla u\| = 1$ on that mesh. Since the latter equation is Eikonal, the Fast Marching Method can be used to solve it efficiently. However, if one desires to formulate the problem in the $x-y$ plane instead of the intrinsic manifold coordinates, then the corresponding equation for u is not Eikonal. Indeed, in the $x-y$ plane, the manifold's geodesic distance function u has to satisfy (65) with the speed function F defined as

$$F(x, y, \omega) = \sqrt{\frac{1 + g_y^2 \cos^2(\omega) + g_x^2 \sin^2(\omega) - g_x g_y \sin(2\omega)}{1 + g_x^2 + g_y^2}}, \tag{72}$$

where ω is the angle between $\nabla u(x, y)$ and the positive direction of the x-axis. The degree of anisotropy in this equation is substantial, since the dependence of F upon ω can be pronounced when ∇g is relatively large.[14]

As shown in section 8.1, u can also be considered as a value function for the corresponding min-time optimal-trajectory problem and must, therefore, satisfy (22). The vehicle's speed function $f(x, y, \boldsymbol{a})$ can be defined by applying (67) to the speed of front propagation $F(x, y, \omega)$. However, it is even easier to obtain f from the control-theoretic considerations. If the vehicle moving with speed $f(x, y, \boldsymbol{a})$ in the x-y plane is just a shadow of another vehicle moving with a unit speed on the manifold, then this vehicle's speed profile is just an orthogonal projection of a unit circle from the manifold's tangent plane onto the x-y plane, i.e.,

$$f(x, y, \boldsymbol{a}) = (1 + (\nabla g(x, y) \cdot \boldsymbol{a})^2)^{-\frac{1}{2}}, \tag{73}$$

where $\boldsymbol{a}$ is a vector of unit length and f is the control-theoretic speed of motion in the direction $\boldsymbol{a}$ (see section 3.1 and section 8.1 for details).

As an example, we consider the manifold $g(x, y) = .9 \sin(2\pi x) \sin(2\pi y)$ and compute the geodesic distance on it from the origin. The *anisotropy coefficient* for this problem is $\Upsilon = \frac{F_2}{F_1} = \sqrt{100 + 324\pi^2}/10 \approx 5.7$. Since the analytical solution is not available, we use the results of the tested method on the mesh with 385×385 mesh points as an estimate of the "true" value function. This estimate is used to perform the convergence analysis on coarser regular meshes in the x-y plane for the following:

M1: iterative solution to the first-order semi-Lagrangian scheme (25),
M2: OUM based on the first-order semi-Lagrangian scheme (section 6),
M3: same as M2, but with the "relaxed update" (section 9.3),
M4: OUM based on the first-order finite-differences scheme (section 8).

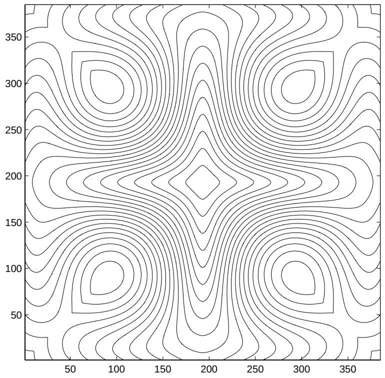
See Figure 5 for the level sets of the value function and for the table of error estimates.

9.4.2. First arrivals in inhomogeneous anisotropic medium. Finally, we include an example of the first arrival travel times computation with applications to seismic imaging. We start with a computational domain which suggests material layering under a sinusoidal profile. The computational domain is the square $[-a, a] \times [-a, a]$, with layer shapes

$$C(x) = A \sin\left(\frac{m\pi x}{a} + \beta\right), \tag{74}$$

where A is the amplitude of the sinusoidal profile, m is the number of periods, and β is the phase offset. The domain is split into n layers by the curves $y_i(x) = C(x) + b_i$, where $i = 1, \ldots, (n - 1)$.

[14] The algorithm presented in [21] using the manifold-approximating mesh is certainly more efficient for this problem; here, it serves as a convenient test problem for the general anisotropic OUMs.



L_∞ Error	25^2	49^2	97^2	193^2
M1	0.15474	0.05902	0.02619	0.00866
M2	0.36131	0.25581	0.13021	0.04195
M3	0.37647	0.25679	0.13070	0.04327
M4	0.36052	0.25940	0.13042	0.04174

L_2 Error	25^2	49^2	97^2	193^2
M1	0.05503	0.02281	0.01073	0.00381
M2	0.13918	0.09901	0.04876	0.01416
M3	0.14022	0.09848	0.04766	0.01374
M4	0.13749	0.09659	0.04626	0.01374

FIG. 5. *The geodesic distance from the origin on the surface* $z = .9\sin(2\pi x)\sin(2\pi y)$ *computed on the square* $[-.5, .5] \times [-.5, .5]$ *in the x-y plane. The error estimates were produced on refined meshes taking the corresponding method on a* 385^2 *mesh to be the true solution (hence the seemingly higher-than-first-order rate at the end).*

In each layer, the anisotropic speed profile S_f is given at every point (x, y) by an ellipse with the bigger axis (of length $2F_2$) tangential to the curve $C(x)$ and the smaller axis (of length $2F_1$) normal to the curve. F_1 and F_2 are constants in each layer. Thus, the ellipse's orientation and shape depend on (x, y).

This leads to an anisotropic Hamilton–Jacobi equation of the form

$$\|\nabla u(x, y)\|F = 1, \qquad u(0, 0) = 0, \tag{75}$$

where the front propagation speed at every point (x, y) is given by the formula

$$F(x, y, u_x, u_y) = F_2 \left(\frac{(1 + q^2)u_x^2 + (1 + p^2)u_y^2 - 2pqu_xu_y}{(1 + p^2 + q^2)(u_x^2 + u_y^2)} \right)^{1/2}, \tag{76}$$

with

$$\begin{bmatrix} p \\ q \end{bmatrix} = \frac{\sqrt{(\frac{F_2}{F_1})^2 - 1}}{\sqrt{1 + \left(\frac{dC}{dx}(x)\right)^2}} \begin{bmatrix} \frac{dC}{dx}(x) \\ -1 \end{bmatrix}.$$

Formula (76) is derived using the the elliptical shape of $S_f(x, y)$ and applying formula (66) of section 8.1.

These calculations are performed using the OUMs with the control-theoretic and finite-difference formulas for computing an update-from-a-single-simplex. Both methods produce numerical solutions converging to the value function of the corresponding min-time optimal trajectory problem.

The equi-arrival curves shown in Figure 6 are obtained on a 193 × 193 regular mesh using the following parameter values:

$$a = .5, \quad A = .1225, \quad m = 2, \quad \beta = 0, \quad \text{and layer offsets } b_i = (-.25, 0, 0.25).$$

The max/min speed pair (F_2, F_1) for each layer is given in the figures. We note that in one of these examples the global *anisotropy coefficient* $\Upsilon = \frac{3}{.2} = 15$.

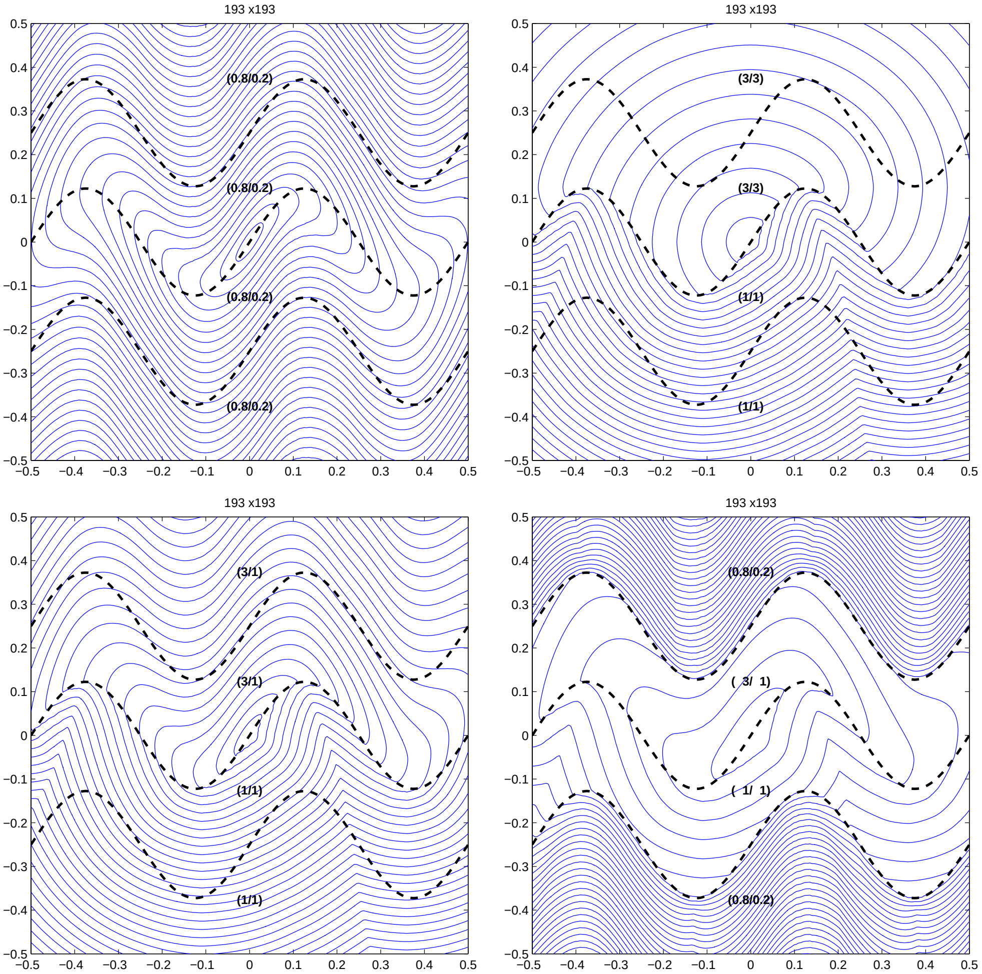


FIG. 6. *Seismic imaging test problem: equi-arrival curves in an inhomogeneous, multilayer medium.*

Remark 9.1. Since the speed function F is discontinuous across the layer boundaries, the standard viscosity solution results for the Hamilton–Jacobi–Bellman equation [10, 9] are not directly applicable. Thus, our proof of convergence in section 7 is not valid in this case either. Nevertheless, the produced numerical solutions seem to converge to the true value function of the corresponding control problem. This is not surprising since our methods are based on approximating Bellman's optimality principle, which is valid for a value function u under much more general assumptions about the speed (or the cost) of motion.

9.5. Conclusions. The methods presented in this paper are applicable for the static Hamilton–Jacobi–Bellman PDEs with convex Hamiltonian and finite speed function bounded away from zero (see [46] for additional background information and details of the proofs). We are currently working on extending these OUMs to a wider class of problems including treatment of nonconvex Hamiltonians, discontinuous speed functions, degenerate speed profiles (i.e., $f_1 = 0$), and stochastic control. We also note that parallelizable single-pass methods based on the same upwinding techniques may be built extending the ideas behind Dial's algorithm for the shortest path on the network.

We believe that similar decoupling techniques hold much promise for the nonlinear problems for which the notion of "information propagation" is well defined.

Appendix. We present a proof of the causality property for the semi-Lagrangian discretization of the Eikonal equation on an unstructured mesh with acute simplexes. We begin by restating the discretization formula (25) for the n-dimensional simplex s with the vertices at $\boldsymbol{x}, \boldsymbol{x_1}, \dots, \boldsymbol{x_n}$. If $\tilde{\boldsymbol{x}}$ is a point on the $\boldsymbol{x_1} \dots \boldsymbol{x_n}$ face of the simplex, we will use $\zeta = (\zeta_1, \dots, \zeta_n)$ for its barycentric coordinates:

$$\zeta_i \in [0,1] \text{ for all } i, \qquad \sum_{i=1}^{n} \zeta_i = 1, \qquad \sum_{i=1}^{n} \zeta_i \boldsymbol{x_i} = \tilde{\boldsymbol{x}}.$$

Using Ξ to denote the set of all possible barycentric coordinates, we can write the isotropic version of the upwinding update formula (25):

$$V_s(\boldsymbol{x}) = \min_{\zeta \in \Xi} \left\{ \frac{\tau(\zeta)}{f(\boldsymbol{x})} + \sum_{i=1}^{n} \zeta_i U(\boldsymbol{x_i}) \right\}, \tag{77}$$

where $\tau(\zeta) = \|\tilde{\boldsymbol{x}} - \boldsymbol{x}\|$. Define the unit directional vectors $\boldsymbol{P_i} = \frac{\boldsymbol{x} - \boldsymbol{x_i}}{\|\boldsymbol{x} - \boldsymbol{x_i}\|}$. To justify using Dijkstra-like decoupling with this discretization, we prove the following version of Property (4.1).

PROPERTY A.1 (Causality). *If $(P_j \cdot P_k) \geq 0$ for all j and k, and if $\zeta = (\zeta_1, \dots, \zeta_n)$ is the minimizer in formula* (77)*, then "$\zeta_i > 0$" implies "$V_s > U(\boldsymbol{x_i})$."*

Proof. Suppose $\zeta_i > 0$ for all $i \in \{1, \dots, n\}$. (If that is not the case, the same argument can be repeated for the lower-dimensional simplex on which all barycentric coordinates are positive.)

Noting that $\frac{\partial \tau}{\partial \zeta_i}(\zeta) = \frac{(\boldsymbol{x} - \boldsymbol{x_i}) \cdot (\boldsymbol{x} - \tilde{\boldsymbol{x}})}{\tau(\zeta)}$, we can write the Kuhn–Tucker optimality conditions for ζ as follows:

$$\frac{(\boldsymbol{x} - \boldsymbol{x_i}) \cdot (\boldsymbol{x} - \tilde{\boldsymbol{x}})}{\tau(\zeta) f(\boldsymbol{x})} + U(\boldsymbol{x_i}) = \lambda \qquad \text{for all } i \in \{1, \dots, n\}, \tag{78}$$

where λ is the Lagrange multiplier. We note that

$$\lambda - U(\boldsymbol{x_i}) = \frac{\|\boldsymbol{x} - \boldsymbol{x_i}\| \sum_{j=1}^{n} \zeta_j \|\boldsymbol{x} - \boldsymbol{x_j}\| (\boldsymbol{P_i} \cdot \boldsymbol{P_j})}{\tau(\zeta) f(\boldsymbol{x})} > 0, \tag{79}$$

by the acuteness of simplex s and since $\zeta_i > 0$. Next, we note that, multiplying (78) by ζ_i and summing over all i's,

(80)
$$\lambda = \lambda \left(\sum_{i=1}^{n} \zeta_i \right) = \frac{\left(\sum_{i=1}^{n} \zeta_i (\boldsymbol{x} - \boldsymbol{x_i}) \right) \cdot (\boldsymbol{x} - \tilde{\boldsymbol{x}})}{\tau(\zeta) f(\boldsymbol{x})} + \sum_{i=1}^{n} \zeta_i U(\boldsymbol{x_i}) = \frac{\tau(\zeta)}{f(\boldsymbol{x})} + \sum_{i=1}^{n} \zeta_i U(\boldsymbol{x_i}) = V_s.$$

Thus, $V_s > U(\boldsymbol{x_i})$ follows from (79). $\square$

We note that the proof holds on an arbitrary "acute" unstructured mesh in R^n and on manifolds. The "splitting section" techniques developed for the Fast Marching Method can also be used to implement Dijkstra-like decoupling of the above discretization on meshes with obtuse simplexes; see [21, 38] for details.

Finally, to explore the connection between semi-Lagrangian and finite-difference schemes, we further consider a column vector $\boldsymbol{w}$, with the entries $w_i = (V_s - U(\boldsymbol{x_i}))/\|\boldsymbol{x} - \boldsymbol{x_i}\|$. Using formula (79) and a matrix P whose rows are $\boldsymbol{P_i}$'s, we see that

$$\boldsymbol{w} = \frac{1}{\tau(\zeta) f(x)} P(\boldsymbol{x} – \tilde{\boldsymbol{x}}) \ \Rightarrow\ P^{-1}\boldsymbol{w} = \frac{\boldsymbol{x} - \tilde{\boldsymbol{x}}}{\tau(\zeta) f(x)} \ \Rightarrow\ \boldsymbol{w}^T \left(PP^T\right)^{-1} \boldsymbol{w} = \frac{\|\boldsymbol{x} - \tilde{\boldsymbol{x}}\|^2}{\left(\tau(\zeta) f(x)\right)^2} = \frac{1}{f^2(x)}.$$

The latter is a quadratic equation in V_s and coincides with the particular first-order finite-difference upwind formula chosen in [38] for the isotropic front-propagation problems.

Furthermore, the analogous correspondence can be demonstrated for the first-order schemes in the anisotropic case. Starting from the general first-order semi-Lagrangian formula

$$V_s(\boldsymbol{x}) = \min_{\zeta \in \Xi} \left\{ \frac{\tau(\zeta)}{f(\boldsymbol{x}, \frac{\tilde{\boldsymbol{x}} - \boldsymbol{x}}{\tau(\zeta)})} + \sum_{i=1}^{n} \zeta_i U(\boldsymbol{x_i}) \right\}, \tag{81}$$

where $\tau(\zeta)$, $\tilde{\boldsymbol{x}}$, $\boldsymbol{P_i}$, and w_i are defined as above, we note that

$$\begin{aligned} 0 &= \min_{\zeta \in \Xi} \left\{ \frac{\tau(\zeta)}{f(\boldsymbol{x}, \frac{\tilde{\boldsymbol{x}} - \boldsymbol{x}}{\tau(\zeta)})} + \sum_{i=1}^{n} \zeta_i \left(U(\boldsymbol{x_i}) - V_s(\boldsymbol{x})\right) \right\} \\ &= \max_{\zeta \in \Xi} \left\{ \sum_{i=1}^{n} \zeta_i \left(V_s(\boldsymbol{x}) - U(\boldsymbol{x_i})\right) - \frac{\tau(\zeta)}{f(\boldsymbol{x}, \frac{\tilde{\boldsymbol{x}} - \boldsymbol{x}}{\tau(\zeta)})} \right\}. \end{aligned}$$

Moreover, since

$$\begin{aligned} \sum_{i=1}^{n} \zeta_i \left(V_s(\boldsymbol{x}) - U(\boldsymbol{x_i})\right) &= \sum_{i=1}^{n} \left(\zeta_i \|\boldsymbol{x} - \boldsymbol{x_i}\|\right) w_i \\ &= \sum_{i=1}^{n} \left(\zeta_i (\boldsymbol{x} - \boldsymbol{x_i})^T P^{-1}\right) w_i = (\boldsymbol{x} - \tilde{\boldsymbol{x}})^T \left(P^{-1}\boldsymbol{w}\right), \end{aligned}$$

we have

$$\max_{\zeta \in \Xi} \left\{ \frac{\tau(\zeta)}{f(\boldsymbol{x}, \frac{\tilde{\boldsymbol{x}} - \boldsymbol{x}}{\tau(\zeta)})} \left[\left(\left(-\frac{\tilde{\boldsymbol{x}} - \boldsymbol{x}}{\tau(\zeta)} \right) \cdot \frac{P^{-1}\boldsymbol{w}}{\|P^{-1}\boldsymbol{w}\|} \right) \|P^{-1}\boldsymbol{w}\| f\left(\boldsymbol{x}, \frac{\tilde{\boldsymbol{x}} - \boldsymbol{x}}{\tau(\zeta)}\right) - 1 \right] \right\} = 0.$$

Since both functions τ and f are strictly positive, the above is equivalent to

$$\max_{\zeta \in \Xi} \left\{ \left[\left(-\frac{\tilde{\boldsymbol{x}} - \boldsymbol{x}}{\tau(\zeta)} \right) \cdot \frac{P^{-1}\boldsymbol{w}}{\|P^{-1}\boldsymbol{w}\|} \right] \|P^{-1}\boldsymbol{w}\| f\left(\boldsymbol{x}, \frac{\tilde{\boldsymbol{x}} - \boldsymbol{x}}{\tau(\zeta)}\right) - 1 \right\} = 0$$

or, more conveniently,

$$\|P^{-1}\boldsymbol{w}\| \max_{\zeta \in \Xi} \left\{ \left[\left(-\frac{\tilde{\boldsymbol{x}} - \boldsymbol{x}}{\tau(\zeta)} \right) \cdot \frac{P^{-1}\boldsymbol{w}}{\|P^{-1}\boldsymbol{w}\|} \right] f\left(\boldsymbol{x}, \frac{\tilde{\boldsymbol{x}} - \boldsymbol{x}}{\tau(\zeta)}\right) \right\} = 1. \tag{82}$$

Let $\boldsymbol{n} = \frac{P^{-1}\boldsymbol{w}}{\|P^{-1}\boldsymbol{w}\|}$. We note that, as ζ varies, $\boldsymbol{a} = \frac{\tilde{\boldsymbol{x}}-\boldsymbol{x}}{\tau(\zeta)}$ covers all directions within this simplex; moreover, the ζ minimizing (81) will also be the maximizer for (82). If all the ζ_i's are positive, then the corresponding $\boldsymbol{a}$ yields a local maximum of the expression $(-\boldsymbol{a}\cdot\boldsymbol{n})f(\boldsymbol{x},\boldsymbol{a})$. As discussed in section 8.1, such a local maximum is unique if S_f is convex; thus, by the formula (66), equation (82) is equivalent to

$$\|P^{-1}\boldsymbol{w}\|F\left(\boldsymbol{x}, \frac{P^{-1}\boldsymbol{w}}{\|P^{-1}\boldsymbol{w}\|}\right) = 1.$$

Finally, the square of this expression is a variant of the finite-difference formula (69) obtained for the first-order Eulerian discretization in section 8.2.1. The upwinding criterion required for the latter scheme is equivalent to verifying that all ζ_i's are positive.

As of right now, we are unaware of any such connections between the higher-order semi-Lagrangian and Eulerian schemes for the Eikonal or general Hamilton–Jacobi–Bellman equations.

Acknowledgments. The authors would like to thank L. C. Evans for his comments and suggestions on the general structure of the proof. The authors would also like to thank O. Hald, M. Falcone, R. Kohn, A. Spitkovsky, and A. Dukhovny.

REFERENCES

[1] D. Adalsteinsson and J.A. Sethian, *A fast level set method for propagating interfaces*, J. Comput. Phys., 118 (1995), pp. 269–277.

[2] D. Adalsteinsson and J.A. Sethian, *The fast construction of extension velocities in level set methods*, J. Comput. Phys., 148 (1999), pp. 2–22.

[3] R.K. Ahuja, T.L. Magnanti, and J.B. Orlin, *Network Flows: Theory, Algorithms, and Applications*, Prentice–Hall, Englewood Cliffs, NJ, 1993.

[4] M. Bardi, M. Falcone, and P. Soravia, *Fully discrete schemes for the value function of pursuit-evasion games*, in Advances in Dynamic Games and Applications (Geneva, 1992), Ann. Internat. Soc. Dynam. Games 1, Birkhäuser-Boston, Cambridge, MA, 1994, pp. 89–105.

[5] R. Bellman, *Dynamic Programming*, Princeton University Press, Princeton, NJ, 1957.

[6] R. Bellman, *Introduction to the Mathematical Theory of Control Processes*, Academic Press, New York, 1967.

[7] I. Capuzzo Dolcetta and M. Falcone, *Discrete dynamic programming and viscosity solutions*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 6 (1989), pp. 161–183.

[8] I. Capuzzo Dolcetta, *On a discrete approximation of the Hamilton–Jacobi equation of dynamic programming*, Appl. Math. Optim., 10 (1983), pp. 367–377.

[9] M.G. Crandall, L.C. Evans, and P.-L. Lions, *Some properties of viscosity solutions of Hamilton-Jacobi equations*, Trans. Amer. Math. Soc., 282 (1984), pp. 487–502.

[10] M.G. Crandall and P.-L. Lions, *Viscosity solutions of Hamilton-Jacobi equations*, Trans. Amer. Math. Soc., 277 (1983), pp. 1–43.

[11] J.A. Dellinger, *Anisotropic Seismic Wave Propagation*, Ph.D. Dissertation, Department of Geophysics, Stanford University, Stanford, CA, 1991.

[12] E.W. Dijkstra, *A note on two problems in connection with graphs*, Numer. Math., 1 (1959), pp. 269–271.

[13] L.C. Evans, *Partial Differential Equations*, American Mathematical Society, Providence, RI, 1998.

[14] L.C. Evans and P.E. Souganidis, *Differential games and representation formulas for solutions of Hamilton-Jacobi-Isaacs equations*, Indiana Univ. Math. J., 33 (1984), pp. 773–797.

[15] M. Falcone, *The minimum time problem and its applications to front propagation*, in Motion by Mean Curvature and Related Topics, Proceedings of the International Conference at Trento, 1992, Walter de Gruyter, New York, 1994, pp. 70–88.

[16] M. Falcone, *A numerical approach to the infinite horizon problem of deterministic control theory*, Appl. Math. Optim., 15 (1987), pp. 1–13; corrigenda, 23 (1991), pp. 213–214.

[17] M. Falcone and R. Ferretti, *Discrete time high-order schemes for viscosity solutions of Hamilton-Jacobi-Bellman equations*, Numer. Math., 67 (1994), pp. 315–344.

[18] R. Gonzales and E. Rofman, *On deterministic control problems: An approximate procedure for the optimal cost* I. *The stationary problem*, SIAM J. Control Optim., 23 (1985), pp. 242–266.

[19] J. Helmsen, E.G. Puckett, P. Colella, and M. Dorr, *Two new methods for simulating photolithography development in three dimensions*, in Proceedings of the SPIE 1996 International Symposium on Microlithography, Santa Clara, CA, SPIE 2726, SPIE, Bellingham, WA, 1996, pp. 253–261.

[20] A. Jain, *Fundamentals of Digital Image Processing*, Prentice-Hall, Englewood Cliffs, NJ, 1989.

[21] R. Kimmel and J.A. Sethian, *Fast marching methods on triangulated domains*, Proc. Nat. Acad. Sci. USA, 95 (1998), pp. 8341–8435.

[22] R. Kimmel and J.A. Sethian, *Fast Marching Methods for Robotic Navigation with Constraints*, Center for Pure and Applied Mathematics Report, University of California, Berkeley, 1996.

[23] R. Kimmel and J.A. Sethian, *Fast Voronoi diagrams and offsets on triangulated surfaces*, in Curve and Surface Design, P.J. Laurent, P. Sablonniere, and L.L. Schumaker, eds., Vanderbilt University Press, Nashville, TN, 2000, pp. 193–202.

[24] S.N. Kruzhkov, *Generalized solutions of the Hamilton-Jacobi Equations of the Eikonal type*, Math. USSR-Sb., 27 (1975), pp. 406–445.

[25] H.J. Kushner, *Probability Methods for Approximations in Stochastic Control and for Elliptic Equations*, Springer-Verlag, New York, 1977.

[26] H.J. Kushner and P.G. Dupuis, *Numerical Methods for Stochastic Control Problems in Continuous Time*, Academic Press, New York, 1992.

[27] R. Malladi and J.A. Sethian, *An $O(NlogN)$ algorithm for shape modeling*, Proc. Nat. Acad. Sci. USA, 93 (1996), pp. 9389–9392.

[28] J. McGulagh, *Geometrical propositions applied to the wave theory of light*, Trans. Royal Irish Acad., 17 (1837), pp. 241–263.

[29] S. Osher and R.P. Fedkiw, *Level set methods: An overview and some recent results*, J. Comput. Phys., 169 (2001), pp. 463–502.

[30] D. Peng, S. Osher, B. Merriman, and H-K. Zhao, *The geometry of Wulff crystal shapes and its relations with Riemann problems*, in Nonlinear Partial Differential Equations, Contemp. Math. 238, G.-Q. Chen and E. di Benedetto, eds., AMS, Providence, RI, 1999, pp. 251–303.

[31] G.W. Postma, *Wave propagation in a stratified medium*, Geophysics, 20 (1955), pp. 780–806.

[32] J.A. Sethian, *Fast Marching Methods for Computing Seismic Travel Times*, manuscript.

[33] J.A. Sethian, *A fast marching level set method for monotonically advancing fronts*, Proc. Nat. Acad. Sci. USA, 93 (1996), pp. 1591–1595.

[34] J.A. Sethian, *Fast marching level set methods for three-dimensional photolithography development*, in Proceedings of the SPIE 1996 International Symposium on Microlithography, Santa Clara, CA, SPIE 2726, SPIE, Bellingham, WA, 1996, pp. 262–272.

[35] J.A. Sethian, *Fast marching methods*, SIAM Rev., 41 (1999), pp. 199–235.

[36] J.A. Sethian, *Level Set Methods and Fast Marching Methods: Evolving Interfaces in Computational Geometry, Fluid Mechanics, Computer Vision and Materials Sciences*, Cambridge University Press, Cambridge, UK, 1996.

[37] J.A. Sethian and M. Popovici, *Three dimensional traveltimes computation using the fast marching method*, Geophysics, 64 (1999), pp. 516–523.

[38] J.A. Sethian and A. Vladimirsky, *Fast methods for the Eikonal and related Hamilton–Jacobi equations on unstructured meshes*, Proc. Nat. Acad. Sci. USA, 97 (2000), pp. 5699–5703.

[39] J.A. Sethian and A. Vladimirsky, *Ordered upwind methods for static Hamilton-Jacobi equations*, Proc. Nat. Acad. Sci. USA, 98 (2001), pp. 11069–11074.

[40] T.P. Schulze and R.V. Kohn, *A geometric model for coarsening during spiral-mode growth of thin films*, Phys. D, 132 (1999), pp. 520–542.

[41] P. Soravia, *Generalized motion of a front propagating along its normal direction: A differential games approach*, Nonlinear Anal., 22 (1994), pp. 1247–1262.

[42] W. Press, B. Flannery, S. Teukolsky, and W. Vetterling, *Numerical Recipes in C*, Cambridge University Press, Cambridge, UK, 1988.

[43] J.E. Taylor, *Crystalline variational problems*, Bull. Amer. Math. Soc., 84 (1978), pp. 568–588.

[44] J.N. Tsitsiklis, *Efficient algorithms for globally optimal trajectories*, in Proceedings of the IEEE 33rd Conference on Decision and Control, Lake Buena Vista, FL, 1994, pp. 1368–1373.

[45] J.N. Tsitsiklis, *Efficient algorithms for globally optimal trajectories*, IEEE Trans. Automat. Control, 40 (1995), pp. 1528–1538.

[46] A. Vladimirsky, *Fast Methods for Static Hamilton-Jacobi Partial Differential Equations*, Ph.D. Dissertation, Department of Mathematics, University of California, Berkeley, CA, 2001.

SIAM J. Numer. Anal.
Vol. 41, No. 1, pp. 364–381

A HYBRID COLLOCATION METHOD FOR VOLTERRA INTEGRAL EQUATIONS WITH WEAKLY SINGULAR KERNELS[*]

YANZHAO CAO[†], TERRY HERDMAN[‡], AND YUESHENG XU[§]

Abstract. The commonly used graded piecewise polynomial collocation method for weakly singular Volterra integral equations may cause serious round-off error problems due to its use of extremely nonuniform partitions and the sensitivity of such time-dependent equations to round-off errors. The singularity preserving (nonpolynomial) collocation method is known to have only local convergence. To overcome the shortcoming of these well-known methods, we introduce a hybrid collocation method for solving Volterra integral equations of the second kind with weakly singular kernels. In this hybrid method we combine a singularity preserving (nonpolynomial) collocation method used near the singular point of the derivative of the solution and a graded piecewise polynomial collocation method used for the rest of the domain. We prove the optimal order of *global* convergence for this method. The convergence analysis of this method is based on a singularity expansion of the exact solution of the equations. We prove that the solutions of such equations can be decomposed into two parts, with one part being a linear combination of some known singular functions which reflect the singularity of the solutions and the other part being a smooth function. A numerical example is presented to demonstrate the effectiveness of the proposed method and to compare it to the graded collocation method.

Key words. Volterra integral equations, hybrid collocation methods, weakly singular kernels

AMS subject classifications. 65R20, 45D05

PII. S0036142901385593

1. Introduction.

We propose in this paper a hybrid collocation method for solving Volterra integral equations of the second kind with weakly singular kernels. By using the singularity expansion of the exact solution, we analyze this method and prove that it has an optimal order of global convergence. Specifically, for given kernels $K, M \in C(I \times I)$ with $I := [0,1]$ and a given parameter $\alpha \in (0,1)$, we define a Volterra integral operator $T_\alpha : C(I) \to C(I)$ by

$$(T_\alpha y)(t) = \int_0^t G_\alpha(t,s) y(s) ds, \quad t \in I,$$

where

$$G_\alpha(t,s) := (t-s)^{\alpha-1} K(t,s) + M(t,s) \quad \text{for} \quad 0 \le s \le t, \quad 0 \le t \le 1,$$

and consider the Volterra integral equation of the second kind

$$y(t) - (T_\alpha y)(t) = f(t), \quad t \in I, \tag{1.1}$$

[*] Received by the editors February 26, 2001; accepted for publication (in revised form) October 14, 2002; published electronically April 9, 2003.
http://www.siam.org/journals/sinum/41-1/38559.html

[†] Department of Mathematics, Florida A & M University, Tallahassee, FL 32307 (yanzhao.cao@mail.famu.edu).

[‡] Department of Mathematics, Virginia Tech, Blacksburg, VA 24061 (herdman@math.vt.edu).

[§] Corresponding author. Department of Mathematics, West Virginia University, Morgantown, WV 26506 and Academy of Mathematics and System Sciences, Academia Sinica, Beijing 100080, China (yxu@math.wvu.edu). This author was supported in part by the National Science Foundation under grant DMS-9973427 and by the Chinese Academy of Sciences under the program of "One Hundred Distinguished Young Scientists."

where $f \in C(I)$ is a given function and $y \in C(I)$ is the unknown function to be determined. The kernel M is of practical importance because it occurs in the applications to aeroelastic modeling problems [7], where a class of neutral delay equations are converted to integral equations in the form of (1.1). Other related references include [4, 11].

Since $0 < \alpha < 1$, the kernel G_α has a singularity along the diagonal. When (1.1) is solved by a numerical method such as a collocation method or a product-integration method using the piecewise polynomial approximation, the accuracy of the approximate solution depends on the order of piecewise polynomials used in the approximation as well as the degree of smoothness of the exact solution. For instance, when $y \in C^r(I)$ and the approximate subspaces are chosen to be piecewise polynomials of order r, the optimal order r of convergence for the approximate solution y_h to y is achieved, that is,

$$\|y - y_h\|_\infty = O(N^{-r}), \tag{1.2}$$

where N is the number of subintervals in the uniform partition associated with the piecewise polynomial spaces. However, the solution of (1.1) exhibits, in general, singularities at the zero in its derivatives even if the forcing term f is a smooth function and the numerical methods mentioned above may not even yield first order accuracy (see, e.g., [2, 5]). In other words, the use of piecewise polynomials of high order does not produce high order convergence for the numerical method.

There have been many attempts to overcome the difficulties caused by the singularity of the solution of (1.1). One of the most commonly used methods [3, 5, 6, 9, 10, 14, 15, 16] is the graded collocation (GC) method using piecewise polynomials with a graded mesh on interval I according to the behavior of the exact solution near the singular point, which was first introduced by Rice in [13]. Specifically, the GC method partitions I by the following knots:

$$t_i = \left(\frac{i}{N}\right)^{\frac{r}{\alpha}}, \quad i = 0, 1, \ldots, N, \tag{1.3}$$

which ensures that the GC method retains the optimal error estimate (1.2). However, as pointed out in [2, 10], the main disadvantage of the GC method is that subintervals near the singular point in the graded mesh have very small length and thus may cause serious round-off error problems for small α and high order polynomials. Since Volterra equations are time-dependent equations, the numerical solutions of these equations are very sensitive to round-off errors.

Another approach for solving (1.1) is to include some nonpolynomial singular functions which reflect the singularity of the exact solution as part of the basis for the finite dimensional subspace in the collocation method (see [2]). We call it the nonpolynomial collocation (NPC) method. For this method, only a *local* convergence result (in [2]) has been seen so far. It does not seem that an optimal order of *global* convergence can be proved for this method. The idea of including some known singular functions in the usual finite element spaces or piecewise polynomial spaces has been explored in [8] to successfully construct Galerkin methods of high convergence order for Fredholm integral equations of the second kind with weakly singular kernels. This idea leads us to the consideration of the present method.

To treat the problems discussed above for the existing methods, we propose a hybrid collocation (HC) method for solving (1.1) which combines the strength of both the GC and NPC methods. In this method, we introduce a graded mesh different from

(1.3) that avoids using small subintervals near the zero and uses the nonpolynomial function approximation *only* in the first subinterval. Specifically, the length of the first subinterval in the HC method is the same as in a *quasi-uniform* partition, that is, there exist positive constants c_1, c_2 such that

$$\frac{c_1}{N} \leq t_1 \leq \frac{c_2}{N},$$

and a graded partition is used only on $[t_1, 1]$ so that the instability problem appearing in the GC method can be avoided. We compensate the use of a large subinterval for the first interval in the partition by employing nonpolynomial functions $t^{i+j\alpha}$, $i+j\alpha < r$, which characterize the singularity of the exact solution y of (1.1), as trial functions in the first subinterval $[t_0, t_1]$. The primary purpose of this paper is to prove that this method provides an optimal order of *global* convergence by taking the strength of both the GC method and the NPC method, while avoiding the problems from which both these methods have suffered.

To prepare for the analysis of this method, we derive a singularity expansion of the exact solution of (1.1). In other words, we decompose the exact solution into two parts, one being a linear combination of singular functions $t^{i+j\alpha}$ which reflect the singularity of the exact solution and the other being a smooth function. This subject has been well studied in [5]. We will make use of the results presented in [5] and construct further a form of expansion that is useful for the development of the HC method.

We organize this paper in five sections. In section 2, we derive the singularity expansion of the exact solution of (1.1). Section 3 is devoted to a study of hybrid interpolation operators which serve as a base for the development of the HC method. In section 4, we describe the HC method which combines the NPC method used near the singular point based on the singularity expansion obtained in section 2 and a GC method elsewhere. We prove the optimal order of global convergence of this method. Furthermore, we present a theoretical result which gives a comparison of the computational cost of the HC method and the GC method, and the length of the smallest subintervals used in both methods. Our theory shows that the HC method is better than the GC method. Finally in section 5, we provide a numerical example to demonstrate the effectiveness of the HC method. We compare the numerical performance of the HC method with that of the GC method. The numerical results confirm the theory presented in section 4.

2. Singularity expansions.

In this section we establish a preliminary result on the singularity decomposition for the solution of (1.1). Singularity of the solution of (1.1) when the kernel M is zero has been systematically studied in [5]. In the next theorem, we make use of the results in [5] and derive the singularity expansion crucial for the development of the HC method for the general case when $M \neq 0$.

THEOREM 2.1. *Let r be a nonnegative integer. Suppose that K, $M \in C^r(I \times I)$ and f has the form*

$$f(t) = \sum_{j+i\alpha<m} f_{ij} t^{j+i\alpha} + f_m(t), \qquad t \in I, \tag{2.1}$$

where f_{ij} are constants and $f_m \in C^m[0,1]$ for some fixed integer m with $0 \leq m \leq r$. Let y denote the solution of (1.1). *Then there exist constants c_{ij} such that*

$$y(t) = \sum_{j+i\alpha<m} c_{ij} t^{j+i\alpha} + v_m(t), \quad t \in I, \tag{2.2}$$

where $v_m \in C^m(I)$.

Proof. When $M \equiv 0$ formula (2.2) follows from Theorem 1.3.15 of [5] with some modification. The modification is necessary to treat the series in the expansion appearing in Theorem 1.3.15 of [5] so that we have form (2.2).

The general case where the kernel M is not zero will be proved by induction on m. The case when $m = 0$ is obvious. We assume that the theorem holds for $m = k$ and proceed to the case $m = k + 1$. By the induction hypothesis, the solution y of (1.1) has a representation

$$y(t) = \sum_{j+i\alpha<k} c_{ij} t^{j+i\alpha} + v_k(t), \quad t \in I, \tag{2.3}$$

where $v_k \in C^k(I)$. For $t \in I$, we let

$$x(t) := \int_0^t M(t,s) y(s) ds.$$

Substituting (2.3) into the expression of function x gives

$$x(t) = \int_0^t M(t,s) \left[\sum_{j+i\alpha<k} c_{ij} s^{j+i\alpha} + v_k(s) \right] ds, \quad t \in I.$$

To simplify the expression of x, we denote

$$w_{k+1}(t) = \int_0^t M(t,s) v_k(s) ds \quad \text{and} \quad m_{ij}(t) = \int_0^1 M(t,st) s^{j+i\alpha} ds.$$

A simplification with a change of variables leads to the following formula:

$$x(t) = \sum_{j+i\alpha<k+1} c_{i,j-1} m_{i,j-1}(t) t^{j+i\alpha} + w_{k+1}(t), \quad t \in I.$$

It is easily seen that w_{k+1}, $m_{ij} \in C^{k+1}(I)$. Applying the Taylor theorem to the functions m_{ij}, we obtain that

$$x(t) = \sum_{j+i\alpha<k+1} d_{ij} t^{j+i\alpha} + u_{k+1}(t), \quad t \in I,$$

where d_{ij} are constants and u_{k+1} is a function in $C^{k+1}(I)$. Let $\tilde{f} := f - x$ and rewrite (1.1) as

$$y(t) + \int_0^t (t-s)^{\alpha-1} K(t,s) y(s) ds = \tilde{f}(t), \quad t \in I.$$

Note that $\tilde{f}$ has the form (2.1) with $m = k + 1$. By the first part of this proof, we conclude the result of the theorem for the case $m = k + 1$, which advances the induction hypothesis and completes the proof. □

3. Nonpolynomial interpolation operators. Motivated by the singularity expansion of the solution of (1.1), in the next section we will develop the HC method for solving (1.1). To prepare for this development, we define a hybrid interpolation operator and study the bound of this operator.

We first define a nonpolynomial finite dimensional subspaces of $C(I)$. As usual, we denote by N_0 the set of nonnegative integers. For $0 < \alpha < 1$ and a positive integer r we introduce an index set by setting

$$W_{\alpha,r} := \{i + j\alpha : i, j \in N_0, i + j\alpha < r\}.$$

Let ℓ denote the cardinality of the set $W_{\alpha,r}$. Clearly, $W_{\alpha,r}$ contains the first r nonnegative integers $i = 0, 1, \ldots, r-1$. For notational convenience, we write

$$W_{\alpha,r} = \{\nu_j : j = 0, 1, \ldots, \ell - 1\}$$

with the convention that $\nu_j = j$ for $j = 0, 1, \ldots, r-1$. Associated with this index set, we define a finite dimensional space V_r of nonpolynomial functions by

$$V_r := \operatorname{span}\{t^{\nu_j} : j = 0, 1, \ldots, \ell - 1\}.$$

We remark that Theorem 2.1 ensures the solution y of (1.1) has the decomposition

$$y = u + v, \quad u \in C^r(I), \quad \text{and} \quad v \in V_r.$$

Also, we denote by P_r the space of polynomials of degree $\leq r-1$. Because the set $W_{\alpha,r}$ contains the integers $i = 0, 1, \ldots, r-1$, it is clear that $P_r \subset V_r$. In addition, space V_r contains nonpolynomial functions t^{ν_j}, $j = r, r+1, \ldots, \ell - 1$, that reflect the singularity of the derivative of the solution of (1.1).

We next describe a finite dimensional space whose elements are piecewise in V_r. For a given positive integer N, we divide the interval I into N subintervals, that is, $0 = t_0 < t_1 < \cdots < t_N = 1$. For a subinterval J of I and a function $f \in C(I)$, we use $f|_J$ for the restriction of f on J and, moreover, for $V \subseteq C(I)$ we let

$$V|_J := \{v|_J : v \in V\}.$$

Let $h_i = t_i - t_{i-1}$ and $h = \max_{1 \leq i \leq N} h_i$. We define a space of functions piecewise in V by

$$(V)_h := \{v : v|_{[t_{i-1}, t_i]} \in V|_{[t_{i-1}, t_i]},\ i = 1, 2, \ldots, N\},$$

and, in particular, we let $V_{r,h} := (V_r)_h$ and $S_{r,h} := (P_r)_h$. Clearly, we have that $V_r \subseteq V_{r,h}$ and $S_{r,h} \subseteq V_{r,h}$.

We now define interpolation operators $P_{h,1}$ from $C(I)$ to $S_{r,h}$ and $P_{h,2}$ from $C(I)$ to $V_{r,h}$, respectively. To this end, we choose ℓ points τ_j in I such that $0 < \tau_1 < \tau_2 < \cdots < \tau_\ell < 1$. The interpolation points on interval $[t_{k-1}, t_k]$ are obtained by setting $t_{kj} := t_{k-1} + \tau_j h_k$, $k = 1, 2, \ldots, N$, $j = 1, 2, \ldots, \ell$. The interpolation operators $P_{h,1} : C(I) \to S_{r,h}$ and $P_{h,2} : C(I) \to V_{r,h}$ are defined as follows. For $f \in C(I)$

$$(P_{h,1}f)(t_{ij}) = f(t_{ij}), \quad j = 1, 2, \ldots, r,\ i = 1, 2, \ldots, N,$$

and

$$(P_{h,2}f)(t_{ij}) = f(t_{ij}), \quad j = 1, 2, \ldots, \ell,\ i = 1, 2, \ldots, N.$$

Note that for the definition of $P_{h,1}$ we use only the first r points τ_j, $j = 1, 2, \dots, r$. It is known (see [2, 5]) that the operators $P_{h,i}$ for $i = 1, 2$ are uniquely defined.

For a function u that is continuous on $[t_{i-1}, t_i]$, $i = 1, 2, \dots, N$, with possible discontinuities at t_i, define its maximum norm on $[t_m, t_n]$ with $0 \le m < n \le N$ by

$$\|u\|_{[t_m,t_n]} = \max_{m \le i \le n} \max_{t_{i-1} \le t \le t_i} |u(t)|.$$

We will simply use $\|u\|$ when $[t_m, t_n] = I$.

Next we show that the norm of the restriction of $P_{h,2}f$ on $[t_0, t_1]$ is bounded by a constant independent of the choice of t_1. For this purpose, we define, for $i = 1, 2, \dots, \ell$, the Lagrange functions

$$L_{1i}|_{[t_0,t_1]} \in V_r|_{[t_0,t_1]}$$

and

$$L_{1i}(t) = 0, \quad t \in [t_1, 1]$$

such that

$$L_{1i}(t_{1j}) = \delta_{ij}, \quad j = 1, 2, \dots, \ell.$$

It is easily verified that

$$(P_{h,2}f)(t) = \sum_{j=1}^{\ell} f(t_{1j}) L_{1j}(t), \quad t \in [t_0, t_1]. \tag{3.1}$$

Lemma 3.1. *There exists a positive constant c such that for all $t_1 \in (0,1)$ and $i = 1, 2, \dots, \ell$*

$$\|L_{1i}\| = \|L_{1i}\|_{[t_0,t_1]} \le c.$$

Proof. For $i = 1, 2, \dots, \ell$, we write

$$L_{1i}(t) = \sum_{p=1}^{\ell} a_{ip} t^{\nu_p}, \quad t \in [t_0, t_1].$$

For $t \in [t_0, t_1]$, there exists a $\tau \in I$ such that $t = t_0 + h_1 \tau$. For $i = 1, 2, \dots, \ell$ and $p = 1, 2, \dots, \ell$ we set $b_{ip} := h_1^{\nu_p} a_{ip}$. Using these notations, we have that

$$L_{1i}(t) = \sum_{p=1}^{\ell} b_{ip} \tau^{\nu_p}, \quad t \in [t_0, t_1]. \tag{3.2}$$

By the definition of L_{1i} we observe that

$$\sum_{p=1}^{\ell} b_{ip} \tau_j^{\nu_p} = \delta_{ij}, \quad j = 1, 2, \dots, \ell. \tag{3.3}$$

We introduce an $\ell \times \ell$ matrix $\mathbf{D}$ by setting $\mathbf{D} := [d_{jp} : j, p = 1, 2, \dots, \ell]$, where $d_{jp} := \tau_j^{\nu_p}$ and a vector $\mathbf{b}_i$ by $\mathbf{b}_i = [b_{ip} : p = 1, 2, \dots, \ell]^T$. It is easy to verify that the matrix $\mathbf{D}$ is invertible and since all entries d_{jp} are independent of the choice of t_1

we conclude that $\|\mathbf{D}^{-1}\|_\infty$ is bounded by a constant independent of the choice of t_1 where the norm used here is the matrix norm induced from the vector norm $\|\cdot\|_\infty$. Thus, it follows from (3.3) that

$$\|\mathbf{b}_i\|_\infty \le \|\mathbf{D}^{-1}\|_\infty.$$

Hence from (3.2) we confirm the result of this lemma with the constant

$$c := \max_{0\le\tau\le 1}\left(\sum_{p=1}^{\ell}\tau^{\nu_p}\right)\|\mathbf{D}^{-1}\|_\infty. \qquad \square$$

The following lemma, which provides a bound of the norm of $P_{h,2}f$ on $[t_0, t_1]$, is a direct consequence of Lemma 3.1.

LEMMA 3.2. *There exists a positive constant c such that for all $t_1 \in (0,1)$ and for any $f \in C(I)$*

$$\|P_{h,2}f\|_{[t_0,t_1]} \le c\|f\|_{[t_0,t_1]}.$$

The next proposition presents order of convergence for the interpolation $P_{h,2}f$ to a function f having form (2.1).

PROPOSITION 3.3. *There exists a positive constant c such that for all $t_1 \in (0,1)$ and for $f = u + v$, where $u \in C^r(I)$ and $v \in V_r$,*

$$\|f - P_{h,2}f\|_{[t_0,t_1]} \le ch_1^r\|u^{(r)}\|_{[t_0,t_1]}.$$

Proof. For all functions f having the form $f = u+v$, where $u \in C^r(I)$ and $v \in V_r$, recalling that $P_{h,2}v = v$ for $v \in V_r$ we have that

$$f - P_{h,2}f = u + v - P_{h,2}(u+v) = u - P_{h,2}u.$$

Noting that

$$P_{h,2}P_{h,1} = P_{h,1}$$

we conclude that

$$f - P_{h,2}f = (I - P_{h,1})u + P_{h,2}P_{h,1}u - P_{h,2}u = (I - P_{h,2})(I - P_{h,1})u.$$

It follows from Lemma 3.2 that there exists a positive constant c such that for all $t_1 \in (0,1)$ and all such f

$$\|f - P_{h,2}f\|_{[t_0,t_1]} \le c\|(I - P_{h,1})u\|_{[t_0,t_1]} \le ch_1^r\|u^{(r)}\|_{[t_0,t_1]},$$

where the last inequality follows from a standard error estimate for polynomial interpolations. $\square$

We next define a *hybrid* interpolation which has a *global* convergence. To this end, we describe a graded partition of I in terms of parameters α and r. Specifically, for $q := \frac{r}{\alpha}$ we let i_0 be an integer such that

$$\left[\left(\frac{N}{i_0}\right)^q\right] = N,$$

where $[a]$ denotes the largest integer less than or equal to a, and set $N' := N - i_0 + 1$. The partition on I is given by

$$t_0 = 0,\ t_i = \left(\frac{i_0 + i - 1}{N}\right)^q, \quad i = 1, 2, \ldots, N'. \tag{3.4}$$

Note that $t_{N'} = 1$ and the integer i_0 satisfies the condition that

$$N^{1-1/q} \leq i_0 \leq N(N-1)^{-1/q}. \tag{3.5}$$

We remark that as far as stability is concerned this partition is better than the partition (1.3) used in a standard GC method. This point will be made clearer in the next section.

Associated with the graded partition (3.4), we define the *hybrid interpolation* operator Q_h by

$$(Q_h f)|_{[0,t_1]} = (P_{h,2} f)|_{[0,t_1]} \text{ and } (Q_h f)|_{[t_1,1]} = (P_{h,1} f)|_{[t_1,1]}. \tag{3.6}$$

That is, on the first subinterval we use singularity preserving (nonpolynomial) interpolation and on the rest of intervals we use the standard piecewise polynomial interpolation. The operator Q_h will be used in the next section for the development of a *hybrid collocation* method. To prepare for this development, we present an expression of projection Q_h in terms of the Lagrange basis functions. To this end, we define the Lagrange polynomial basis $L_i \in P_r$, $i = 1, 2, \ldots, r$, such that

$$L_i(\tau_j) = \delta_{ij}, \quad j = 1, 2, \ldots, r,$$

and for $k = 2, 3, \ldots, N'$ we define the Lagrange piecewise polynomial basis functions by setting

$$L_{ki}(t) = \begin{cases} L_i\left(\frac{t - t_{k-1}}{h_k}\right), & t \in [t_{k-1}, t_k], \\ 0, & \text{otherwise.} \end{cases}$$

Thus, for all $k = 2, 3 \ldots, N'$ there hold the relations that

$$\|L_{ki}\| = \|L_i\|, \quad i = 1, 2, \ldots, r, \tag{3.7}$$

and we have that

$$(P_{h,1} f)(t) = \sum_{k=2}^{N'} \sum_{j=1}^{r} f(t_{kj}) L_{kj}(t), \quad t \in [t_1, 1].$$

To present projection Q_h, we introduce a notation

$$r_k := \begin{cases} \ell, & k = 1, \\ r, & k = 2, 3, \ldots, N'. \end{cases} \tag{3.8}$$

Consequently, we have that

$$(Q_h f)(t) = \sum_{k=1}^{N'} \sum_{j=1}^{r_k} f(t_{kj}) L_{kj}(t), \quad t \in I. \tag{3.9}$$

In the next proposition, we establish a *global* convergence result for the interpolation projection Q_h.

PROPOSITION 3.4. *Let Q_h be the hybrid interpolation operator defined in* (3.6) *associated with the partition* (3.4). *Suppose that f has a decomposition $f = u + v$, where $u \in C^r(I)$ and $v \in V_r$. Then there exists a positive constant c independent of N such that for all such functions f*

$$\|f - Q_h f\| \leq cN^{-r}.$$

Proof. Note that

$$\|f - Q_h f\| = \max\left\{\|f - Q_h f\|_{[0,t_1]}, \|f - Q_h f\|_{[t_1,1]}\right\}.$$

Employing Proposition 3.3 we conclude that there exists a positive constant c such that

$$\|f - Q_h f\|_{[0,t_1]} = \|f - P_{h,2} f\|_{[0,t_1]} \leq ch_1^r \|u^{(r)}\|_{[0,t_1]}.$$

Noting that by the definition of partition (3.4),

$$h_1 = t_1 = \left(\frac{i_0}{N}\right)^q \leq \left(\frac{N(N-1)^{-1/q}}{N}\right)^q = \frac{1}{N-1} \leq \frac{2}{N},$$

we obtain that

$$\|f - Q_h f\|_{[0,t_1]} \leq c\frac{1}{N^r}.$$

We next estimate the error $\|f - Q_h f\|_{[t_1,1]}$ following a well-known argument by Rice [13]. Since $f = u + v$ with $u \in C^r(I)$ and $v \in V_r$, we have that

$$|f^{(r)}(t)| \leq ct^{\alpha - r} \quad \text{for} \quad t \in [t_1, 1],$$

and the function $t^{\alpha - r}$ is decreasing in t. For $i = 2, 3, \ldots, N'$, we find that

$$h_i = \left[(i + i_0 - 1)^q - (i + i_0 - 2)^q\right] N^{-q} = (i + i_0 - 2)^q \left\{\left[1 + (i + i_0 - 2)^{-1}\right]^q - 1\right\} N^{-q}.$$

By the mean value theorem, there exists θ with $0 < \theta < (i + i_0 - 2)^{-1}$ such that

$$h_i = q(i + i_0 - 2)^{q-1}(1 + \theta)^{q-1} N^{-q} \leq c(i + i_0 - 2)^{q-1} N^{-q}. \tag{3.10}$$

Thus, there exists a positive constant c such that for $i = 2, 3, \ldots, N'$

$$\begin{aligned}\|f - Q_h f\|_{[t_{i-1},t_i]} &= \|f - P_{h,2} f\|_{[t_{i-1},t_i]} \leq ch_i^r \|f^{(r)}\|_{[t_{i-1},t_i]} \\ &\leq c(i + i_0 - 2)^{(q-1)r} N^{-qr} \left(\frac{i + i_0 - 2}{N}\right)^{q(\alpha - r)} = c\frac{1}{N^r},\end{aligned}$$

which completes the proof of this proposition. □

4. A hybrid collocation method. In this section, we use the hybrid interpolation operator Q_h introduced in the last section to develop a hybrid collocation method for solving (1.1). We prove that this method has an optimal order of global convergence. Notice that the singularity in the derivative of the exact solution of (1.1) occurs only at the left end point of the interval I. This fact suggests that we use a singularity preserving collocation method near the left end point and use a standard piecewise polynomial collocation method with a graded partition elsewhere.

We now describe the hybrid collocation method for (1.1). We seek y_h such that

$$y_h|_{[0,t_1]} \in V_{r,h}|_{[0,t_1]}, \quad y_h|_{[t_1,1]} \in S_{r,h}|_{[t_1,1]},$$

and

$$y_h - Q_h T_\alpha y_h = Q_h f, \tag{4.1}$$

where Q_h is the hybrid interpolation operator defined by (3.6) associated with the graded partition (3.4).

To analyze the order of convergence for the hybrid collocation method (4.1), we define an integral operator $T_{\alpha,1}$ by

$$(T_{\alpha,1}y)(t) := (T_\alpha y)(t) \quad \text{for} \quad t \in [0, t_1] \tag{4.2}$$

and

$$(T_{\alpha,1}y)(t) := \int_{t_1}^{t} G_\alpha(t,s)y(s)ds \quad \text{for} \quad t \in [t_1, 1].$$

The study of the error $\|y - y_h\|$ demands a bound on the errors

$$e_{kj} := y(t_{kj}) - y_h(t_{kj}), \quad j = 1, 2, \ldots, r_k,$$

where r_k is defined by (3.8). To this end, we introduce vectors

$$\mathbf{e}_k := [e_{kj} : j = 1, 2, \ldots, r_k]^T \quad \text{for} \quad k = 1, 2, \ldots, N'.$$

We also need vectors

$$\epsilon_k := [\epsilon_{kj} : j = 1, 2, \ldots, r_k]^T \quad \text{for} \quad k = 1, 2, \ldots, N',$$

where

$$\epsilon_{kj} := (T_{\alpha,1}(y - Q_h y))(t_{kj}) + ((T_\alpha - T_{\alpha,1})(y - y_h))(t_{kj}). \tag{4.3}$$

Note that when $k = 1$, it becomes that

$$\epsilon_{1j} = (T_\alpha(y - Q_h y))(t_{1j}).$$

We will bound the vectors $\mathbf{e}_k$ by ϵ_k.

Next, we derive a linear system that gives a recursive formula for the vector $\mathbf{e}_k$. Toward this goal, for $k = 1, 2, \ldots, N'$ and $j = 1, 2, \ldots, r_k$ we evaluate (1.1) and (4.1) at t_{kj} to obtain that

$$y(t_{kj}) - (T_\alpha y)(t_{kj}) = f(t_{kj}) \tag{4.4}$$

and

$$y_h(t_{kj}) - (T_\alpha y_h)(t_{kj}) = f(t_{kj}), \tag{4.5}$$

respectively. Subtracting (4.5) from (4.4) yields

$$e_{kj} = (T_{\alpha,1}(y - y_h))(t_{kj}) + ((T_\alpha - T_{\alpha,1})(y - y_h))(t_{kj}). \tag{4.6}$$

Noticing that Q_h is a projection and $y_h = Q_h y_h$, we have that

$$y - y_h = y - Q_h y + Q_h(y - y_h). \tag{4.7}$$

Substituting (4.7) into the first term in the right-hand side of (4.6) and recalling the definition of ϵ_{kj}, we obtain that

$$e_{kj} = (T_{\alpha,1} Q_h(y - y_h))(t_{kj}) + \epsilon_{kj} \quad \text{for} \quad k = 1, 2, \dots, N', \ j = 1, 2, \dots, r_k. \tag{4.8}$$

We are required to study the first term in the right-hand side of (4.8). In (3.9), we replace f by $y - y_h$ and conclude that

$$(Q_h(y - y_h))(t) = \sum_{i=1}^{N'} \sum_{p=1}^{r_i} e_{ip} L_{ip}(t), \quad t \in I.$$

Applying the operator $T_{\alpha,1}$ to both sides of this equation with evaluating at $t = t_{kj}$ yields

$$(T_{\alpha,1} Q_h(y - y_h))(t_{kj}) = \sum_{i=1}^{N'} \sum_{p=1}^{r_i} e_{ip} (T_{\alpha,1} L_{ip})(t_{kj}).$$

We next make use of the property of the Lagrange basis functions L_{ip} to simplify the right-hand side of the equation above. For notational convenience, we define

$$a_{jp}^k := \int_{t_{k-1}}^{t_{kj}} G_\alpha(t_{kj}, s) L_{kp}(s) ds$$

and

$$d_{jp}^{ki} := \int_{t_{i-1}}^{t_i} G_\alpha(t_{kj}, s) L_{ip}(s) ds.$$

Noting that L_{ip} vanishes outside the interval $[t_{i-1}, t_i]$, an elementary computation leads to the formula that

$$(T_{\alpha,1} Q_h(y - y_h))(t_{kj}) = \sum_{p=1}^{r_k} a_{jp}^k e_{kp} + \sum_{i=2}^{k-1} \sum_{p=1}^{r} d_{jp}^{ki} e_{ip},$$

where we have used the relation that $r_i = r$ for $i = 2, 3, \dots, k-1$. Substituting this equation into the right-hand side of (4.8) yields

$$e_{kj} = \sum_{p=1}^{r_k} a_{jp}^k e_{kp} + \sum_{i=2}^{k-1} \sum_{p=1}^{r} d_{jp}^{ki} e_{ip} + \epsilon_{kj}, \quad j = 1, 2, \dots, r_k. \tag{4.9}$$

By introducing an $r_k \times r_k$ matrix

$$\mathbf{A}_k := [a_{jp}^k : j, p = 1, 2, \dots, r_k] \tag{4.10}$$

and $r \times r$ matrices

$$\mathbf{D}_{ki} := [d_{jp}^{ki} : j, p = 1, 2, \dots, r], \quad i = 2, 3, \dots, k-1, \tag{4.11}$$

we write (4.9) in matrix form as

$$\mathbf{e}_k = \mathbf{A}_k\mathbf{e}_k + \sum_{i=2}^{k-1}\mathbf{D}_{ki}\mathbf{e}_i + \epsilon_k, \quad k = 1, 2, \ldots, N'. \tag{4.12}$$

The matrices $\mathbf{A}_k$ and $\mathbf{D}_{ki}$ are all dependent on the mesh sizes h_i, and we next study such a dependence. Recalling the transformations $s = t_{k-1} + h_k\tau$ when $s \in [t_{k-1}, t_k]$ and $\tau \in [0, 1]$ and $t_{kj} = t_{k-1} + h_k\tau_j$ and using the notations

$$\tilde{K}(t,\tau) := K(t, t_{k-1}+h_k\tau), \;\; \tilde{M}(t,\tau) := M(t, t_{k-1}+h_k\tau), \;\; \tilde{L}_{kp}(\tau) := L_{kp}(t_{k-1}+h_k\tau),$$

$$\tilde{a}^k_{jp} := \int_0^{\tau_j}\left[(\tau_j-\tau)^{\alpha-1}\tilde{K}(t_{kj},\tau) + h_k^{1-\alpha}\tilde{M}(t_{kj},\tau)\right]\tilde{L}_{kp}(\tau)d\tau,$$

and

$$\tilde{d}^{ki}_{jp} := \int_0^1\left[\left(\frac{t_{k-1}-t_{i-1}+h_k\tau_j}{h_i} - \tau\right)^{\alpha-1}\tilde{K}(t_{kj},\tau) + h_i^{1-\alpha}\tilde{M}(t_{kj},\tau)\right]\tilde{L}_{ip}(\tau)d\tau,$$

by changes of variables we have that

$$a^k_{jp} = h_k^\alpha\tilde{a}^k_{jp} \;\text{ and }\; d^{ki}_{jp} = h_i^\alpha\tilde{d}^{ki}_{jp}.$$

By introducing new matrices

$$\tilde{\mathbf{A}}_k := [\tilde{a}^k_{jp} : j, p = 1, 2, \ldots, r_k] \;\text{ and }\; \tilde{\mathbf{D}}_{ki} := [\tilde{d}^{ki}_{jp} : j, p = 1, 2, \ldots, r], \;\; i = 2, 3, \ldots, k-1,$$

we observe that

$$\tilde{\mathbf{A}}_k = h_k^{-\alpha}\mathbf{A}_k \;\text{ and }\; \tilde{\mathbf{D}}_{ki} = h_k^{-\alpha}\mathbf{D}_{ki}.$$

Hence, (4.12) becomes

$$\mathbf{e}_k = h_k^\alpha\tilde{\mathbf{A}}_k\mathbf{e}_k + \sum_{i=2}^{k-1}h_i^\alpha\tilde{\mathbf{D}}_{ki}\mathbf{e}_i + \epsilon_k, \quad k = 1, 2, \ldots, N'. \tag{4.13}$$

Since both $\tilde{K}$ and $\tilde{M}$ are continuous functions on $I \times I$, by using Lemma 3.1 and (3.7) we observe that there exists a positive constant c_1 such that for $k = 1, 2, \ldots, N'$ and $j, p = 1, 2, \ldots, r_k$

$$|\tilde{a}^k_{jp}| \le c_1\int_0^{\tau_j}\left[(\tau_j-\tau)^{\alpha-1} + 1\right]d\tau \le c_1\left(1 + \frac{1}{\alpha}\right) \tag{4.14}$$

and

$$|\tilde{d}^{ki}_{jp}| \le c_1\int_0^1\left(\frac{t_{k-1}-t_{i-1}+h_k\tau_j}{h_i} - \tau\right)^{\alpha-1}d\tau, \quad i = 2, 3, \ldots, k-1. \tag{4.15}$$

Estimate (4.14) implies that there exists a positive constant $c_2 := c_1\ell\left(1 + \frac{1}{\alpha}\right)$ for all $k = 1, 2, \ldots, N'$,

$$\|\tilde{\mathbf{A}}_k\|_\infty \le c_2.$$

Let $h := \max\{h_i : 1 \le i \le N'\}$ and choose $h < c_2^{-1/\alpha}$. It follows that for such an h the matrix $I - h_k^\alpha \tilde{\mathbf{A}}_k$ is invertible and there exists a positive constant c such that

$$\|(I - h_k^\alpha \tilde{\mathbf{A}}_k)^{-1}\| \le c. \tag{4.16}$$

Thus, from (4.13) we conclude that

$$\mathbf{e}_k = (I - h_k^\alpha \tilde{\mathbf{A}}_k)^{-1} \left[\sum_{i=2}^{k-1} h_i^\alpha \tilde{\mathbf{D}}_{ki} \mathbf{e}_i + \epsilon_k \right]. \tag{4.17}$$

We next use estimate (4.15) to study the bound of the entries of matrices $\tilde{\mathbf{D}}_{ki}$. Noting that $h_i < h_{i+1} < \cdots < h_{k-1}$, we have that

$$\frac{t_{k-1} - t_{i-1}}{h_i} = \frac{h_i + \cdots + h_{k-1}}{h_i} \ge k - i$$

and conclude that

$$|\tilde{d}_{jp}^{ki}| \le c_1 \int_0^1 (k - i - \tau)^{\alpha - 1} d\tau. \tag{4.18}$$

It can be verified from a direct computation that there exists a positive constant c such that for all $i = 2, 3, \ldots, k-1$

$$\int_0^1 (k - i - \tau)^{\alpha-1} d\tau \le c(k-i)^{\alpha-1}.$$

Using this estimate in inequality (4.18) yields for $i = 2, 3, \ldots, k-1$ and $j, p = 1, 2, \ldots, r_k$ that

$$|\tilde{d}_{jp}^{ki}| \le c(k-i)^{\alpha-1}$$

and thus

$$\|\tilde{\mathbf{D}}_{ki}\|_\infty \le c(k-i)^{\alpha-1}. \tag{4.19}$$

Combining estimates (4.16) and (4.19) with (4.17) gives

$$\|\mathbf{e}_k\|_\infty \le c \sum_{i=2}^{k-1} h_i^\alpha (k-i)^{\alpha-1} \|\mathbf{e}_i\|_\infty + c\|\epsilon_k\|_\infty.$$

Recalling from (3.10) that there exists a positive constant c such that

$$h_i \le c(i + i_0 - 2)^{q-1} N^{-q},$$

since $i + i_0 - 2 \le 2N$ we conclude that there exists a positive constant c such that

$$h_i \le cN^{-1}.$$

Therefore, we have that

$$\|\mathbf{e}_k\|_\infty \le c \left(\frac{1}{N}\right)^\alpha \sum_{i=2}^{k-1} (k-i)^{\alpha-1} \|\mathbf{e}_i\|_\infty + c\|\epsilon_k\|_\infty. \tag{4.20}$$

We next use inequality (4.20) to obtain the error estimate of the hybrid collocation method (4.1). For this purpose, we recall a discrete Gronwall-type inequality (cf. [5]).

LEMMA 4.1. *Let $0 < \alpha < 1$ and $\{z_i : i = 1, 2, \ldots, n\}$ be a sequence of positive numbers and $n \leq N$. Let ρ and β be two positive numbers such that for all $k = 1, 2, \ldots, n$*

$$z_k \leq \left(\frac{1}{N}\right)^{\alpha} \beta \sum_{i=1}^{k-1} (k-i)^{\alpha-1} z_i + \rho.$$

Then there exists a positive constant c depending only on α and β such that $k = 1, 2, \ldots, n$,

$$z_k \leq c\rho.$$

We are now ready to prove the main result of this paper, which gives an optimal order of global convergence for the hybrid collocation method.

THEOREM 4.2. *Let y be the exact solution of* (1.1), *let N be a positive integer, and let Q_h be the hybrid interpolation operator defined by* (3.6) *associated with the graded partition* (3.4). *Suppose that the forcing function f in* (1.1) *has the form* (2.1). *Then, for sufficiently large N,* (4.1) *has a unique solution y_h and there exists a positive constant c independent of N such that*

$$\|y - y_h\| \leq cN^{-r}.$$

Proof. It follows from Theorem 2.1 that the solution of (1.1) has the form $y = u+v$, where $u \in C^r(I)$ and $v \in V_r$. This allows us to use Proposition 3.4 to prove the result. We first estimate the error on $[t_0, t_1]$. From (4.3) there exists a constant c such that

$$\|\epsilon_1\|_\infty \leq \|T_\alpha(y - Q_h y)\|_{[t_0,t_1]} \leq c\|y - Q_h y\|_{[t_0,t_1]}.$$

It follows from (4.20) and Proposition 3.4 that

$$\|\mathbf{e}_1\|_\infty \leq c\|y - Q_h y\|_{[t_0,t_1]} \leq cN^{-r}.$$

Using Lemma 3.1 and the above estimate we obtain that

$$\|Q_h(y - y_h)\|_{[t_0,t_1]} \leq c\|\mathbf{e}_1\|_\infty \leq cN^{-r},$$

which with Proposition 3.4 gives that

$$\|y - y_h\|_{[t_0,t_1]} \leq \|Q_h(y - y_h)\|_{[t_0,t_1]} + \|Q_h y - y\|_{[t_0,t_1]} \leq cN^{-r}. \tag{4.21}$$

Next we estimate the error on $[t_1, 1]$. Using (4.3) we have that for $k = 2, 3, \ldots, N'$

$$\|\epsilon_k\|_\infty \leq \|T_{\alpha,1}(y - Q_h y)\|_{[t_1,1]} + \|(T_\alpha - T_{\alpha,1})(y - y_h)\|_{[t_1,1]}.$$

Now, by using Proposition 3.4 and estimate (4.21) we conclude that there exists a positive constant c such that

$$\|\epsilon_k\|_\infty \leq c(\|y - Q_h y\|_{[t_1,1]} + \|y - y_h\|_{[t_0,t_1]}) \leq cN^{-r}.$$

Combining the above estimate with (4.20) we obtain that

$$\|\mathbf{e}_k\|_\infty \leq cN^{-\alpha} \sum_{i=2}^{k-1} (k-i)^{\alpha-1} \|\mathbf{e}_i\|_\infty + cN^{-r}.$$

Using Lemma 4.1 with $z_1 = 0$ and $z_i = \|\mathbf{e}_i\|_\infty$ for $i = 2, 3, \ldots N'$, we conclude that there exists a positive constant c such that for all $k = 2, 3, \ldots, N'$

$$\|\mathbf{e}_k\|_\infty \leq cN^{-r}.$$

Now, by the uniform boundedness (3.7) of $\|L_{ij}\|$ for $i = 2, 3, \ldots, N'$ there exists a positive constant c such that

$$\|Q_h(y - y_h)\|_{[t_1,1]} \leq c \max\{\|\mathbf{e}_k\|_\infty : k = 2, 3, \ldots, N'\} \leq cN^{-r}.$$

It follows that there exists a positive constant c such that

$$\|y - y_h\|_{[t_1,1]} \leq \|Q_h(y - y_h)\| + \|Q_h y - y\| \leq cN^{-r},$$

which concludes the proof of the theorem. □

We may also use the compactness of operator T_α and the uniform boundedness of Q_h to prove Theorem 4.2 (see, for example, [1]). In fact, such a proof is more concise. We choose the current proof, for it provides guidance for the construction of a numerical algorithm in our numerical experiments.

In the next proposition, we compare the graded collocation (GC) method with the hybrid collocation (HC) method. To this end, we let $\mathcal{N}_{GC}$ and $\mathcal{N}_{HC}$ denote the number of subintervals used in the GC method and the HC method, and we let $\mathcal{L}_{GC}$ and $\mathcal{L}_{HC}$ denote the length of the smallest subinterval used in the GC method and the HC method, respectively. We also consider the ratios of the largest subinterval over the smallest subinterval for the partitions that associate with the GC method and the HC method, which are denoted by $\mathcal{R}_{GC}$ and $\mathcal{R}_{HC}$. Such a ratio is a good measure for the stability of the corresponding collocation method.

PROPOSITION 4.3. *There hold the estimates that*

$$\mathcal{N}_{GC} - \mathcal{N}_{HC} \geq \frac{N}{\sqrt[q]{N}} - 1,$$

$$\mathcal{L}_{GC} = \frac{1}{N^q}, \qquad \mathcal{L}_{HC} \geq \frac{\sqrt[q]{N}}{N^2},$$

and

$$\mathcal{R}_{GC} \geq q(N-1)^{q-1}, \qquad \mathcal{R}_{HC} \leq \frac{qN}{\sqrt[q]{N}}.$$

Proof. Since $\mathcal{N}_{GC} = N$ and $\mathcal{N}_{HC} = N - i_0 + 1 \leq N - N^{1-1/q} + 1$, we have that

$$\mathcal{N}_{GC} - \mathcal{N}_{HC} \geq N - (N - N^{1-1/q} + 1) = \frac{N}{\sqrt[q]{N}} - 1.$$

The smallest interval used in the GC method is $[0, N^{-q}]$ and thus $\mathcal{L}_{GC} = \frac{1}{N^q}$. For the HC method, it is easily verified that $\mathcal{L}_{HC} = \min\{h_1, h_2\}$, where $h_1 = \left(\frac{i_0}{N}\right)^q$ and $h_2 = \left(\frac{i_0+1}{N}\right)^q - \left(\frac{i_0}{N}\right)^q$. Recalling that i_0 satisfies that condition $N^{1-\frac{1}{q}} \leq i_0 \leq N(N-1)^{-\frac{1}{q}}$, we derive that

$$h_1 \geq \left(\frac{N^{1-\frac{1}{q}}}{N}\right)^q = \frac{1}{N} \geq \frac{\sqrt[q]{N}}{N^2}.$$

On the other hand, there exists a constant θ with $0 < \theta < 1/i_0$ such that

$$h_2 = q i_0^{q-1}(1+\theta)^{q-1}N^{-q}.$$

Thus

$$h_2 \geq q i_0^{q-1}N^{-q} \geq qN^{\frac{1}{q}-2} \geq \frac{\sqrt[q]{N}}{N^2}.$$

This concludes the second estimate in this proposition. The third estimate can be similarly obtained. □

We remark that it follows from Proposition 4.3 that the HC method requires less computational cost than the GC method even though they have the same order of convergence. For example, when $\alpha = \frac{1}{2}$, $r = 3$, and $N = 100$, the HC method uses 54 subintervals while the GC method uses 100 subintervals.

Another important point made in the last proposition is that the HC method is more stable than the GC method since the length of the smallest subinterval used in the HC method is larger than the length of the smallest subinterval used in the GC method. Notice that the length of the smallest interval used in the HC method is not as sensitive to r and α as that in the GC method. For instance, when $N = 1000$, $\alpha = \frac{1}{2}$, and $r = 3$, the length of the smallest interval used in the GC method is $h_1 = 10^{-24}$ while the length of the smallest interval used the HC method is 4.217×10^{-5}. In addition, we see from the proposition that for the GC method $\mathcal{R}_{GC}$ grows in the order $\mathcal{O}(N^{q-1})$ while for the HC method $\mathcal{R}_{HC}$ grows slower than $\mathcal{O}(N)$. When p is large, which is the case when α is small or r is large, $\mathcal{R}_{GC}$ is extremely large. This may cause serious instability problems. The result in the proposition shows that the HC method is much more stable than the GC method.

5. Numerical experiments. In this section, we report results of numerical experiments which confirm the theoretical analysis for the HC method presented in the last section and demonstrate the effectiveness of the method.

In (1.1) we choose $K(s,t) = M(s,t) = 1$, $\alpha = 1/2$ and choose f such that the equation has the exact solution

$$y(t) = \sqrt{t^2+t}\cos t + \sin t, \quad t \in I.$$

Note that the first derivative of this solution has a singularity at $t = 0$.

The purpose of these numerical experiments is to compare the numerical performance of the HC method with the GC method. For both of the methods we use piecewise polynomials of degree 2, that is, $r = 3$, and, in addition, for the HC method we use

$$V_3 = \operatorname{span}\{1,\ t,\ t^2,\ t^{\frac{1}{2}},\ t^{\frac{3}{2}},\ t^{\frac{5}{2}}\}$$

on the first interval.

Tables 5.1 and 5.2 are given to compare the numerical performance of the two methods, where "order of conv." stands for the order of convergence. The weakly singular integrals that appear in these methods are computed by a numerical integration scheme presented in [12] specifically designed for weakly singular integrals of this type.

The HC method and the GC method have the same orders of convergence. The computed orders of convergence are consistent with the theoretical order, which is $r =$

Table 5.1
Numerical performance of the HC method.

N	$\|y - y_h\|$	Order of conv.	$\mathcal{N}_{HC}$	$\mathcal{L}_{HC}$
20	3.666e-4	-	8	7.5419e-02
40	4.680e-5	2.9698	19	2.7681e-02
60	1.408e-5	2.9616	30	1.9022e-02
80	5.983e-6	2.9759	42	1.3423e-02
100	3.069e-6	2.9916	54	1.0779e-02

Table 5.2
Numerical performance of the GC method.

N	$\|y - y_h\|$	Order of conv.	$\mathcal{N}_{GC}$	$\mathcal{L}_{GC}$
20	1.993e-4	-	20	1.5625e-08
40	3.446e-5	2.9720	40	2.4414e-10
60	1.240e-5	2.9576	60	2.1433e-11
80	6.011e-6	2.9776	80	3.8147e-12
100	3.430e-6	2.9896	100	1.0000e-12

3 for both methods. In terms of convergence both methods give satisfactory numerical performance. However, the HC method uses much fewer subintervals. Therefore, it requires less computational cost than the GC method. Also, the length of the smallest subinterval used in the HC method is significantly larger than that in the GC method. When N is large and α is small, for the GC method, the length of the first interval is extremely small, which may cause serious round-off errors. The HC method has a rather uniform partition, which avoids the problem of having small subintervals. In these two aspects, the HC method has performed better than the GC method.

Acknowledgment. The authors are grateful to two referees for their suggestions that improve the presentation of this paper.

REFERENCES

[1] K. E. Atkinson, *The Numerical Solution of Integral Equations of the Second Kind*, Cambridge University Press, Cambridge, UK, 1997.

[2] H. Brunner, *Nonpolynomial spline collocation for Volterra equations with weakly singular kernels*, SIAM J. Numer. Anal., 20 (1983), pp. 1106–1119.

[3] H. Brunner, *The numerical solution of weakly singular Volterra integral equations by collocation on graded meshes*, Math. Comp., 45 (1985), pp 417–437.

[4] H. Brunner, *The numerical solution of weakly singular first-kind Volterra integral equations with delay arguments*, Proc. Estonian Acad. Sci. Phys. Math., 48 (1999), pp. 90–100.

[5] H. Brunner and P. J. van der Houwen, *The Numerical Solution of Volterra Equations*, North-Holland, Amsterdam, 1986.

[6] H. Brunner, A. Pedas, and G. Vainikko, *The piecewise polynomial collocation method for nonlinear weakly singular Volterra equations*, Math. Comp., 68 (1999), pp. 1079–1095.

[7] J. Burns, E. Cliff, and T. Herdman, *A state-space model for an aeroelastic system*, in Proceedings of the 22nd IEEE Conference on Decision and Control, San Antonio, TX, 1983, pp. 1174–1177.

[8] Y. Cao and Y. Xu, *Singularity preserving Galerkin methods for weakly singular Fredholm integral equations*, J. Integral Equations Appl., 6 (1994), pp. 303–334.

[9] T. Diogo, S. McKee, and T. Tang, *A Hermite-type collocation method for the solution of integral equations with a certain weakly singular kernels*, IMA J. Numer. Anal., 11 (1991), pp. 595–605.

[10] T. Diogo, S. McKee, and T. Tang, *Collocation methods for second-kind Volterra integral equations with weakly singular kernels*, Proc. Royal Soc. Edinburgh, 124 (1994), pp. 199–210.

[11] B. Jumarhon, W. Lamb, S. McKee, and T. Tang, *A Volterra integral type method for solving a class of nonlinear initial-boundary value problems*, Numer. Methods Partial Differential Equations, 12 (1996), pp. 265–281.

[12] H. Kaneko and Y. Xu, *Gaussian-type quadratures for weakly singular integrals and their applications to the Fredholm integral equation of the second kind*, Math. Comp., 62 (1994), pp. 739–753.

[13] J. Rice, *On the degree of convergence of nonlinear spline approximations*, in Approximation with Special Emphasis on Spline Functions, I. J. Schoenberg, ed., Academic Press, NY, 1969, pp. 349–365.

[14] H. J. Riele, *Collocation methods for weakly singular second-kind Volterra integral equations with non-smooth solution*, IMA J. Numer. Anal., 2 (1982), pp. 437–449.

[15] T. Tang, *Superconvergence of numerical solutions to weakly singular Volterra integral-differential equations*, Numer. Math., 61 (1992), pp. 373–382.

[16] T. Tang, *A note on collocation methods for Volterra integro-differential equations with weakly singular kernels*, IMA J. Numer. Anal., 13 (1993), pp. 93–99.

SIAM J. NUMER. ANAL.
Vol. 41, No. 1, pp. 382–399

ON THE TIME-CONTINUOUS MASS TRANSPORT PROBLEM AND ITS APPROXIMATION BY AUGMENTED LAGRANGIAN TECHNIQUES*

K. GUITTET[†]

Abstract. In [J. D. BENAMOU AND Y. BRENIER, *Numer. Math.*, 84 (2000), pp. 375–393], a computational fluid dynamic approach was introduced for computing the optimal map occurring in the Monge–Kantorovich problem. Though the described augmented Lagrangian method involves a Hilbertian framework, the discussion was purely formal. Taking advantage of the recent progress in optimal transport theory [L. A. CAFFARELLI, *Comm. Pure Appl. Math.*, 45 (1992), pp. 1141–1151], [L. A. CAFFARELLI, *Ann. of Math.* (2), 144 (1996), pp. 453–496], [D. CORDERO-ERAUSQUIN, *C. R. Acad. Sci. Paris Sér.* I *Math.*, 329 (1999), pp. 199–202], [R. J. MCCANN, *Geom. Funct. Anal.*, 11 (2001), pp. 589–608] and despite the lack of coercivity of the Hilbertian problem, we establish an existence result. Then, under a reasonable assumption of positivity for the density, we prove the existence of saddle-points for both Lagrangians defined in Benamou and Brenier, and finally prove the convergence of the numerical method.

Key words. optimal transport, augmented Lagrangian method, Wasserstein distance

AMS subject classifications. 65N12, 49J35

PII. S0036142901386069

Introduction. Given two nonnegative density functions ρ_0 and ρ_T on $\mathbb{R}^d$ satisfying the compatibility condition

$$\int_{\mathbb{R}^d} \rho_0(x)dx = \int_{\mathbb{R}^d} \rho_T(x)dx, \tag{0.1}$$

a map M is said to transport ρ_0 to ρ_T if, for any bounded subset A of $\mathbb{R}^d$,

$$\int_A \rho_T(x)dx = \int_{M(A)} \rho_0(x)dx. \tag{0.2}$$

Then the Monge–Kantorovich problem (MKP) consists of finding a map M transporting ρ_0 to ρ_T and minimizing the cost

$$\int_{\mathbb{R}^d} |M(x) - x|^2 \rho_0(x)dx. \tag{0.3}$$

The problem of the existence and the characterization of the optimal map has been solved in [2] by Brenier, who showed that the optimal map is the gradient of a convex potential. This result has then been extended to the cases of more general cost functions in [8] and more general geometries in [10]. From a numerical point of view, the computation of the optimal map seems to be a challenging problem (see [1] for a brief review of existing methods). In their work [1], Benamou and Brenier used an artificial (time) variable to linearize the constraints (0.2). Then an augmented numerical resolution of the resulting problem was presented. Although the optimal

*Received by the editors March 5, 2001; accepted for publication (in revised form) September 2, 2002; published electronically April 9, 2003.
http://www.siam.org/journals/sinum/41-1/38606.html

[†]INRIA Rocquencourt, Action OTTO, Domaine de Voluceau, Rocquencourt-B.P. 105, 78153 Le Chesnay Cedex, France (kevin.guittet@inria.fr).

mass transport problem is naturally set up in the frame of probability measures and continuous test functions, the augmented method used in [1] is largely of Hilbertian nature. In order to prove the convergence of the method, it is therefore natural to study the optimal mass transport problem from a Hilbertian point of view. This article presents a step-by-step justification of the numerical method in the required Hilbertian framework and finally provides a convergence proof. As in [1], we will consider only the case of the torus, which considerably simplifies the analysis. Now we define the time-continuous MKP (TCMKP) as follows.

Let $\mathbb{T}^d$ be the d-dimensional unit cube with periodic boundary. Define $Q = [0;T] \times \mathbb{T}^d$. In this study, we note

$$H(Q;\mathrm{div}) = \{f \in L^2(Q)^{1+d} \text{ s.t. } \nabla_{t,x}.f \in L^2(Q)\}, \tag{0.4}$$

$$V(Q) = \{f \in L^2(Q)^{1+d} \text{ s.t. } \|\nabla_{t,x}.f\|_{L^2(Q)} = 0\}. \tag{0.5}$$

Given two densities (ρ_0, ρ_T) in $L^2(\mathbb{T}^d)$ satisfying the compatibility condition (0.1), the TCMKP is to minimize

$$K(\rho, m) = \int_0^T \int_{\mathbb{T}^d} \frac{|m|^2}{2\rho} dxdt, \tag{0.6}$$

over all pairs (ρ, m) in $V(Q)$ satisfying the boundary conditions

$$\begin{aligned} \rho(0,.) &= \rho_0 \qquad \text{in } L^2(\mathbb{T}^d), \\ \rho(T,.) &= \rho_T \qquad \text{in } L^2(\mathbb{T}^d). \end{aligned} \tag{0.7}$$

We denote by $E(\rho_0, \rho_T)$ this minimum.

Remark 0.1. The link between the MKP and the TCMKP may be unclear. This link is used as an important tool in the proof of the main theorem of section 1, to which we refer for more explanation.

In this paper, our aim is to derive a rigorous Hilbertian theory for the TCMKP, and to prove the convergence of the augmented Lagrangian method used in [1]. Thus section 1 deals with the well-posedness of the Hilbertian problem. The formal existence result proved in [1] is made rigorous in the Hilbertian framework, and the link between the optimal cost $E(\rho_0, \rho_T)$ and the Wasserstein distance between ρ_0 and ρ_T is recalled. In section 2, we look at the Lagrangian formulation of the TCMKP and prove an abstract existence result of a saddle-point. Since this Lagrangian L is the starting point of the numerical method, we expect this result to be an important step when looking for a convergence result. Next the purpose of section 3 is to get more information on the saddle-point. Specifically, we show that the Lagrange multiplier of the mass conservation constraint is linked to the optimal pair (ρ^*, m^*). Then we recall the second Lagrangian $\mathcal{L}$ introduced in [1] and use the saddle-points of L to characterize the saddle-points of $\mathcal{L}$. We therefore get an existence result for a saddle-point of $\mathcal{L}$. Finally, section 4 presents a convergence result for the numerical algorithm. This result may be unexpected since some of the classical assumptions required for convergence are not fulfilled.

1. Well-posedness of the Hilbertian problem. In this section, the main result is the existence of a minimizer of the TCMKP and the characterization of the optimal cost $E(\rho_0, \rho_T)$. Those results are summarized in the following theorem.

Theorem 1.1. *For any nonnegative* (ρ_0, ρ_T) *in* $(L^2(\mathbb{T}^d))^2$ *satisfying* (0.1), *we have*

$$E(\rho_0, \rho_1) = \frac{1}{2T} d^2_{Wass}(\rho_0, \rho_1). \tag{1.1}$$

Moreover, there exists a minimizer (ρ^*, m^*) *such that*

$$\|\rho^*\|_{L^\infty([0;T];L^p(\mathbb{T}^d))} \leq M_p, \tag{1.2}$$

$$\|m^*\|_{L^\infty([0;T];L^p(\mathbb{T}^d))} \leq M_p \frac{\sqrt{d}}{T}, \tag{1.3}$$

where $M_p = \max(\|\rho_0\|_{L^p(\mathbb{T}^d)}, \|\rho_T\|_{L^p(\mathbb{T}^d)})$ *for any* $p \in]1; \infty]$.

Sketch of the proof. First, we give a precise definition for the kinetic energy $K(\rho, m)$ defined in (0.6). A first inequality is then derived using the nonlinear interpolation between ρ_0 and ρ_1 introduced by McCann in [9]. The proof of the converse inequality is based on a formal argument of Benamou and Brenier. A minimizing sequence is carefully lifted and mollified, so that this argument actually holds.

The rigorous definition of the kinetic energy follows from the observation that for positive ρ

$$\frac{|m|^2}{2\rho} = \sup_{a + \frac{|b|^2}{2} \leq 0} [a\rho + b.m]. \tag{1.4}$$

Then $K(\rho, m)$ can be defined through the following equality:

$$K(\rho, m) = \sup \int_Q a\rho + b.m, \tag{1.5}$$

where $(a, b) \in L^2(Q) \times L^2(Q)^d$ are subject to the constraint that for all nonnegative $f \in L^\infty(Q)$

$$\int_0^T \int_{\mathbb{T}^d} f \left(a + \frac{|b|^2}{2} \right) dx \leq 0. \tag{1.6}$$

We denote by $\tilde{K}$ the set of all pairs $(a, b) \in L^2(Q)^{1+d}$ satisfying (1.6). It is easy to see that $\tilde{K}$ is a closed convex set in $L^2(Q)^{1+d}$. Moreover, for any fixed pair $(a, b) \in \tilde{K}$, the application $(\rho, m) \mapsto \int_Q a\rho + b.m$ is convex and continuous, and then lower semicontinuous. Then the upper envelope K of those functions is still convex and l.s.c. This property of K will prove to be useful in what follows.

Remark 1.2. Such a characterization of the kinetic energy has been used by Brenier in [3]. The test functions were taken in $\mathcal{C}^0(Q)$, so that the interior of $\tilde{K}$ was not empty. This property allowed Brenier to use a duality theorem of Rockafellar to get the existence of a minimizer. However, this minimizer was found in the set of Radon measures. Since we expect more regularity for our minimizer, our definition of $\tilde{K}$ is slightly different, and this set turns out to be of empty interior.

In order to get the first inequality, we use the interpolation defined by McCann in [9]. Specifically, we have the following lemma.

Lemma 1.3. *Let* (ρ_0, ρ_T) *be two nonnegative densities in* $L^2(\mathbb{T}^d)$. *Then there exists some* $(\rho^*, m^*) \in V(Q)$ *satisfying* (0.7), (1.2), *and* (1.3) *and such that*

$$K(\rho^*, m^*) \leq \frac{1}{2T} d^2_{Wass}(\rho_0, \rho_T). \tag{1.7}$$

Proof. Denote by $\nabla_x \phi$ the optimal transport (which is known to be the gradient of a convex potential) from ρ_0 to ρ_T. It means that $(\nabla_x\phi)_\sharp \rho_0 = \rho_T$. Then, following McCann, we can use this transport to define the nonlinear interpolation between ρ_0 and ρ_T:

$$\bar\rho_t = \frac{1}{T}((T-t)Id + t\nabla_x\phi)_\sharp \rho_0. \tag{1.8}$$

For any $p \in]1;\infty]$, the L^p-norm of $\bar\rho_t$ is displacement convex, which means in particular that

$$\|\bar\rho_t\|_{L^p(\mathbb{T}^d)} \leq \max\left(\|\rho_0\|_{L^p(\mathbb{T}^d)}, \|\rho_T\|_{L^p(\mathbb{T}^d)}\right). \tag{1.9}$$

Now define the characteristics associated with $\nabla_x\phi$,

$$X(t,x) = \frac{1}{T}((T-t)Id + t\nabla_x\phi)(x) = \nabla_x\phi_t(x), \tag{1.10}$$

and the velocity field

$$\begin{cases} v(t, X(t,x)) = \frac{1}{T}(\nabla_x\phi(x) - x) & \text{on } X(t, \{x \in \mathbb{T}^d | \rho_0(x) > 0\}), \\ v(t,y) = 0 & \text{elsewhere.} \end{cases} \tag{1.11}$$

Then define $\bar m(t,x) = \bar\rho(t,x)v(t,x)$. Assuming some regularity, this construction ensures that $(\bar\rho, \bar m) \in V(Q)$ and

$$K(\bar\rho, \bar m) = \frac{1}{2T} d^2_{Wass}(\rho_0, \rho_T). \tag{1.12}$$

Moreover, since $\nabla_x\phi$ is the optimal transport on the torus, we get that for any $x \in \mathbb{T}^d$, $\|\nabla_x\phi(x) - x\| \leq d$. As a consequence, we have the following estimate for any $p \in]1;\infty]$:

$$\|\bar m\|_{L^\infty([0;T];L^p(\mathbb{T}^d))} \leq \frac{\sqrt{d}}{T} \max\left(\|\rho_0\|_{L^p(\mathbb{T}^d)}, \|\rho_T\|_{L^p(\mathbb{T}^d)}\right). \tag{1.13}$$

Now, regularizing the boundary densities and using the regularity theory for the MKP, it is possible to build a weakly converging sequence in $H(Q, \mathrm{div})$. Denote the limit by (ρ^*, m^*). By construction, (ρ^*, m^*) is in $V(Q)$ and satisfies (0.7). Moreover, K is convex and l.s.c, and it follows that

$$E(\rho_0, \rho_1) \leq K(\rho^*, m^*) \leq \frac{1}{2T} d^2_{Wass}(\rho_0, \rho_T). \tag{1.14}$$

Finally, (1.2) and (1.3) follow simply from (1.9) and (1.13). This achieves the proof of Lemma 1.3. □

Now we have to get the converse inequality. Let $(\rho_n, m_n)_{n\in\mathbb{N}}$ be a minimizing sequence for the TCMKP. For any n in $\mathbb{N}^*$, we define

$$\begin{cases} \rho_n^{(1)} = \rho_n + \dfrac{1}{n}, \\ m_n^{(1)} = m_n. \end{cases} \tag{1.15}$$

By construction, we have $(\rho_n^{(1)}, m_n^{(1)}) \in V(Q)$ and

$$\forall (t,x) \in Q, \qquad \rho_n^{(1)}(t,x) \geq \frac{1}{n}. \tag{1.16}$$

Moreover, the definition of K leads to

$$K(\rho_n^{(1)}, m_n^{(1)}) \leq K(\rho_n, m_n). \tag{1.17}$$

Now we want to build a smooth "outer" (in the sense that the time-boundary data are not satisfied, but only carefully approximated) minimizing sequence. This regularization will be necessary in order to define a smooth velocity field and some characteristics, which will then allow us to apply Benamou and Brenier's argument. We therefore define f in $\mathcal{C}(\mathbb{R} \times \mathbb{T}^d)$ as follows:

$$\begin{cases} f(t,x) = (0,x) & \text{if} \quad t < 0, \\ f(t,x) = (t,x) & \text{if} \quad t \in [0;T], \\ f(t,x) = (T,x) & \text{if} \quad t > T. \end{cases} \tag{1.18}$$

Define $\rho_n^* = \rho_n^{(1)} \circ f$ and $m_n^* = \chi_{[0;T]}[m_n^{(1)} \circ f]$. We have $(\rho_n^*, m_n^*) \in V(Q)$. This extension of the functions allows a good convergence of the time-boundary values. Fix $n \in \mathbb{N}^*$. For any k in $\mathbb{N}^*$ we define

$$g_k(t,x) = \left(-\frac{T}{k} + \left(1 + \frac{2}{k}\right)t, x\right). \tag{1.19}$$

The sequence (ρ_n^k, m_n^k) is then defined as follows:

$$\begin{cases} \rho_n^k(t,x) = \xi_{\frac{T}{k}} * [\rho_n^* \circ g_k], \\ m_n^k(t,x) = \left(1 + \dfrac{2}{k}\right) \xi_{\frac{T}{k}} * [m_n^* \circ g_k], \end{cases} \tag{1.20}$$

where ξ is a positive mollifier with support contained in the unit ball of $\mathbb{R}^{d+1}$. It is easy to see that for any k in $\mathbb{N}^*$, (ρ_n^k, m_n^k) is in $V(Q)$. Moreover, the pair (ρ_n^k, m_n^k) satisfies the following properties,

$$\begin{cases} \rho_n^k \in \mathcal{C}^\infty(Q), \\ m_n^k \in \mathcal{C}^\infty(Q)^d \\ \forall (t,x) \in Q, \quad \dfrac{1}{n} \leq \rho_n^k(t,x), \end{cases} \tag{1.21}$$

and some subsequence (still labeled by k) satisfies

$$\begin{cases} \lim_{k\to\infty} \|(\rho_n^k, m_n^k) - (\rho_n^{(1)}, m_n^{(1)})\|_{H(Q;div)} = 0, \\ \lim_{k\to\infty} \|\rho_n^k(0,.) - \rho_0^n\|_{L^2(\mathbb{T}^d)} = 0, \\ \lim_{k\to\infty} \|\rho_n^k(T,.) - \rho_T^n\|_{L^2(\mathbb{T}^d)} = 0. \end{cases} \tag{1.22}$$

Moreover, the sequence $(1/\rho_n^k)_{k\in\mathbb{N}}$ converges weakly* in $L^\infty(\mathbb{T}^d)$ towards $1/\rho_n^{(1)}$. This, together with (1.22), ensures the convergence of $K(\rho_n^k, m_n^k)$ towards $K(\rho_n^{(1)}, m_n^{(1)})$. We then get that there exists a k_n in $\mathbb{N}$ such that

$$K(\rho_n^{k_n}, m_n^{k_n}) \leq K(\rho_n^{(1)}, m_n^{(1)}) + \frac{1}{n}, \tag{1.23}$$

$$\|\rho_n^{k_n}(0,.) - \rho_0^n\|_{L^2(\mathbb{T}^d)} \leq \frac{1}{n}, \tag{1.24}$$

$$\|\rho_n^{k_n}(T,.) - \rho_T^n\|_{L^2(\mathbb{T}^d)} \leq \frac{1}{n}. \tag{1.25}$$

Define $(\rho_n^{(2)}, m_n^{(2)}) = (\rho_n^{k_n}, m_n^{k_n})$. By construction, we have

$$\begin{cases} (\rho_n^{(2)}, m_n^{(2)}) \in \mathcal{C}^\infty(Q)^{d+1} \\ \forall (t,x) \in Q, \ \rho_n^{(2)} \geq \dfrac{1}{n}. \end{cases} \tag{1.26}$$

Moreover, (1.23) implies that

$$K(\rho_n^{(2)}, m_n^{(2)}) \leq K(\rho_n^{(1)}, m_n^{(1)}) + \frac{1}{n}. \tag{1.27}$$

Hence $(\rho_n^{(2)}, m_n^{(2)})$ is an "outer" minimizing sequence. It is now possible to use Benamou and Brenier's argument to get the required inequality. We therefore define the velocity field

$$v_n^{(2)} = \frac{m_n^{(2)}}{\rho_n^{(2)}}. \tag{1.28}$$

$v_n^{(2)}$ is in $\mathcal{C}^\infty(Q)^d$. We are now able to define the characteristics. We look at the differential system

$$\begin{cases} \partial_t X(t,x) = v_n^{(2)}(t, X(t,x)), \\ X(0,x) = x. \end{cases} \tag{1.29}$$

Since $v_n^{(2)}$ is a $\mathcal{C}^\infty$ function, this system is well defined, and the solution is uniquely defined on $[0;T]$. Now we recall the computations in [1] to show that the "optimal displacement" between $X(0,.)$ and $X(T,.)$ follows straight lines. Indeed, we have

$$\begin{aligned} T\int_{\mathbb{T}^d}\int_0^T \rho_n^{(2)}(t,x)|v_n^{(2)}(t,x)|^2 dxdt &= T\int_{\mathbb{T}^d}\int_0^T \rho_n^{(2)}(0,x)|v_n^{(2)}(t,X(t,x))|^2 dxdt \\ &= T\int_{\mathbb{T}^d}\int_0^T \rho_n^{(2)}(0,x)|\partial_t X(t,x)|^2 dxdt \\ &\geq \int_{\mathbb{T}^d} \rho_n^{(2)}(0,x)|X(T,x) - X(0,x)|^2 dx \\ &= \int_{\mathbb{T}^d} \rho_n^{(2)}(0,x)|X(T,x) - x|^2 dx. \end{aligned}$$

It is then sufficient to consider the MKP between the densities $\rho_n^{(2)}(0,.)$ and $\rho_n^{(2)}(T,.)$. From [6], we get the existence of a convex function ϕ_n such that $\nabla_x \phi_n$ minimizes

$$\int_{\mathbb{T}^d} \rho_n^{(2)}(0,x)|M(x)-x|^2 dx \tag{1.30}$$

in the set of application M pushing $\rho_n^{(2)}(0,.)$ forward to $\rho_n^{(2)}(T,.)$. This minimum is the definition of the Wasserstein distance between $\rho_n^{(2)}(0,.)$ and $\rho_n^{(2)}(T,.)$. We then get that

$$d_{Wass}^2(\rho_n^{(2)}(0,.),\rho_n^{(2)}(T,.)) \leq 2T\ K(\rho_n^{(2)}, m_n^{(2)}). \tag{1.31}$$

Summarizing (1.17), (1.27), (1.31), we get

$$\begin{aligned} \frac{1}{2T} d_{Wass}^2(\rho_n^{(2)}(0,.),\rho_n^{(2)}(T,.)) &\leq K(\rho_n^{(2)}, m_n^{(2)}) \\ &\leq K(\rho_n^{(1)}, m_n^{(1)}) + \frac{1}{n} \\ &\leq K(\rho_n, m_n) + \frac{1}{n}. \end{aligned}$$

The Wasserstein distance is continuous with respect to the L^2 distance on the torus (thanks to its finite diameter). Hence we get that $d_{Wass}^2(\rho_n^{(2)}(0,.),\rho_n^{(2)}(T,.))$ converges to $d_{Wass}^2(\rho_0,\rho_T)$ as n goes to infinity. Passing to the limit in the previous inequality, we get

$$\frac{1}{2T} d_{Wass}^2(\rho_0,\rho_T) \leq E(\rho_0,\rho_T). \tag{1.32}$$

This, together with (1.14), achieves the proof of Theorem 1.1.

2. Existence of a saddle-point for the Lagrangian. From now on, we will assume that the optimal density ρ^* satisfies the following assumption:

$$\exists \alpha_1 > 0 \text{ s.t. } \forall (t,x) \in Q,\ \rho^*(x,t) \geq \alpha_1. \tag{2.1}$$

Notice that this assumption is always satisfied for strictly positive smooth boundary data. Indeed, when ρ_0 and ρ_1 are of class $\mathcal{C}^{\alpha,l}(\mathbb{T}^d)$, the regularity theory developed for the optimal mass transport problem by Caffarelli in the case of convex bounded domains (see [4], [5]) and recently applied by Cordero-Erausquin [6] in the case of the flat torus $\mathbb{T}^d = \mathbb{R}^d/\mathbb{Z}^d$ ensures that ϕ is of class $\mathcal{C}^{\alpha,l+2}(\mathbb{T}^d)$. Then the Monge–Ampère equation

$$\rho^*(t, \nabla_x \phi_t(x)) \det(\nabla_x^2 \phi_t(x)) = \rho_0(x) \tag{2.2}$$

holds, and ρ^* is smooth. Moreover, we have $(1-t)^d \leq \det(\nabla_x^2 \phi_t(x)) \leq C$ for a constant C depending only on ρ_0 and ρ_1. This implies that the optimal density ρ^* is bounded from below away from zero, and (2.1) is proved. However, the existence of the first saddle-point defined in [1] may be obtained without the regularity assumption on the boundary densities.

Theorem 2.1. *Assume that the solution* (ρ^*, m^*) *of the TCMKP satisfies* (2.1). *Then there exists a* $\lambda^* \in L_0^2(Q)$ *such that* (ρ^*, m^*, λ^*) *is a saddle-point of the Lagrangian L, defined as follows:*

$$L(\rho, m, \lambda) = K(\rho, m) + \int_0^T \int_{\mathbb{T}^d} (\partial_t \rho + \nabla_x . m)\lambda \; dxdt. \tag{2.3}$$

The proof of this theorem relies mainly on an application of the Hahn–Banach theorem. The two convex sets we want to separate are defined as follows:

$$S = \left\{ (K(\rho, m) - K(\rho^*, m^*) + s, \partial_t \rho + \nabla_x . m) \;\middle|\; \begin{array}{l} (\rho, m) \in H(Q; \mathrm{div}), \\ (0.7) \text{ holds}, \\ s \geq 0, \end{array} \right\}$$

$$T = \left\{ (-t, 0) \in \mathbb{R} \times L_0^2(Q), t > 0 \right\}.$$

The next three lemmas show that S and T satisfy the required assumptions for the Hahn–Banach theorem.

Lemma 2.2. *S and T are convex.*

Proof. The convexity of T is obvious. Let (r_1, ψ_1) and (r_2, ψ_2) be in S. There exists (ρ_1, m_1, s_1) and (ρ_2, m_2, s_2) such that

$$(K(\rho_i, m_i) - K(\rho^*, m^*) + s_i, \partial_t \rho + \nabla_x . m) = (r_i, \psi_i). \tag{2.4}$$

From the convexity of K, we get

$$K\left(\frac{1}{2}(\rho_1 + \rho_2, m_1 + m_2)\right) \leq \frac{1}{2}(K(\rho_1, m_1) + K(\rho_2, m_2)). \tag{2.5}$$

Let $s_3 = \frac{1}{2}(s_1 + s_2) + \frac{1}{2}(K(\rho_1, m_1) + K(\rho_2, m_2)) - K(\frac{1}{2}(\rho_1 + \rho_2, m_1 + m_2))$. We have $s_3 \geq 0$. Define then

$$(\rho_3, m_3) = \frac{1}{2}(\rho_1 + \rho_2, m_1 + m_2),$$

$$\psi_3 = \frac{1}{2}(\psi_1 + \psi_2),$$

$$r_3 = \frac{1}{2}(r_1 + r_2).$$

From the linearity of the divergence operator, we get $\psi_3 = \nabla_{t,x}.(\rho_3, m_3)$. Then we have that $s_3 \geq 0$ and $\psi_3 \in H(Q; \mathrm{div})$ such that

$$(K(\rho_3, m_3) - K(\rho^*, m^*) + s_3, \partial_t \rho + \nabla_x . m) = (r_3, \psi_3).$$

Hence (r_3, ψ_3) is in S. This proves the convexity of S. □

Lemma 2.3. $S \cap T = \emptyset$.

Proof. Let (r, ψ) be in $S \cap T$. We have $\psi = 0$. Let (ρ, m, s) such that

$$(K(\rho, m) - K(\rho^*, m^*) + s, \partial_t \rho + \nabla_x . m) = (r, 0). \tag{2.6}$$

Using the definition of (ρ^*, m^*), we have $r \geq s$. Hence $r \geq 0$. But we should have $r < 0$ since (r, ψ) is in T. This is a contradiction. We conclude that $S \cap T = \emptyset$. □

Lemma 2.4. *The interior of S is not empty.*

Proof. Let $s_0 > 0$. As we will show, $(s_0, 0)$ is an interior point of S. Let $0 < \epsilon < 1/2$. Take (r, g) in a neighborhood of $(s_0, 0)$, that is, such that

$$|r - s_0| + \|g\|_{L^2} < \epsilon. \tag{2.7}$$

We want to prove that (r, g) is in S for ϵ small enough. We are therefore looking for a (ρ, m, s) such that

$$\begin{cases} K(\rho, m) - K(\rho^*, m^*) + s = r, \\ \\ \nabla_{t,x}.(\rho, m) = g. \end{cases}$$

One of the difficulties comes from the fact that ρ has to satisfy some positivity property. In order to control the L^∞-norm of the new density, we integrate the mass production induced by g. We then define $h(t) = \int_0^t \int_{\mathbb{T}^d} g(u, x)dxdu$. Since g is in $L^2_0(Q)$, we have $h(T) = 0$. This condition is necessary to allow the recovery of the boundary conditions for ρ. Our strategy is then to split ρ into two parts. The first part has to stay close to the optimal density, while the second has to track the mass production. We therefore define $\rho_2(t, x) = \theta\alpha_1 + h(t)$ for some θ. Then we must have the following equality:

$$\nabla_x.m(t, x) = g(t, x) - \int_{\mathbb{T}^d} g(t, y)dy. \tag{2.8}$$

For almost every t in $[0; T]$, we solve the system

$$\begin{cases} \Delta_x(\psi_t) = g(t, x) - \displaystyle\int_{\mathbb{T}^d} g(t, y)dy, \\ \\ \psi_t \in H^1(\mathbb{T}^d) \cap L^2_0(\mathbb{T}^d), \end{cases}$$

under periodic boundary conditions. We then take $\tilde{m}(t, x) = \nabla_x \psi_t(x)$. By construction, we have

$$\|\tilde{m}(t, .)\|_{L^2(\mathbb{T}^d)} \leq C\|g(t, .)\|_{L^2(\mathbb{T}^d)}. \tag{2.9}$$

Integrating in t, we get $\|\tilde{m}\|_{L^2(Q)} \leq C\|g\|_{L^2(Q)}$.

Remark 2.5. This construction does not take care of the measurability of the resulting function $\tilde{m}$, which could be easily stated. Anyway, the bound (2.9) allows some regularization process ...

Now we consider $K(\rho^* + h(t), m^* + \tilde{m})$. We want to prove that this action is close to the minimum. Therefore, we define

$$(\rho_1, m_1) = (\rho^* - \theta\alpha_1, m^*), \tag{2.10}$$

$$(\rho_2, m_2) = (h(t) + \theta\alpha_1, \tilde{m}). \tag{2.11}$$

The inequality $K((\rho_1, m_1) + (\rho_2, m_2)) \leq K(\rho_1, m_1) + K(\rho_2, m_2)$ follows from the convexity and the homogeneity property of K.

$$\begin{aligned} K(\rho_1, m_1) &= \int_{\mathbb{T}^d} \int_0^T \frac{|m^*|^2}{2(\rho^* - \theta\alpha_1)} dxdt \\ &= \int_{\mathbb{T}^d} \int_0^T \frac{|m^*|^2}{2\rho^*} \frac{\rho^*}{\rho^* - \theta\alpha_1} dxdt. \end{aligned}$$

But we have $\alpha_1 \leq \rho^*$, so that $\rho^* - \theta\alpha_1 \geq (1-\theta)\rho^*$. Hence we have

$$K(\rho_1, m_1) \leq \frac{1}{1-\theta} K(\rho^*, m^*).$$

Finally, we get

$$K(\rho_1, m_1) - K(\rho^*, m^*) \leq \frac{\theta}{1-\theta} K(\rho^*, m^*). \tag{2.12}$$

Now we have to estimate $K(\rho_2, m_2)$. We have

$$\begin{aligned} \rho_2(t,x) &= \int_0^t \int_{\mathbb{T}^d} g(u,x) dx du + \theta\alpha_1 \\ &= \int_0^T \int_{\mathbb{T}^d} g(u,x)\chi_{[0;t]} dx du + \theta\alpha_1. \end{aligned} \tag{2.13}$$

Then for all $(t,x) \in Q$

$$\theta\alpha_1 - \|g\|_{L^2(Q)}\sqrt{t} \leq \rho_2(t,x) \leq \theta\alpha_1 + \|g\|_{L^2(Q)}\sqrt{t}.$$

Hence for $\|g\|_{L^2(Q)} \leq \frac{1}{\sqrt{T}}\theta\alpha_1$ we have

$$\forall (t,x) \in Q, \qquad \rho_2(t,x) \geq 0.$$

Moreover, taking $2\epsilon\sqrt{T} \leq \alpha_1\theta$, we get

$$\begin{aligned} K(\rho_2, m_2) &\leq \int_0^T \int_{\mathbb{T}^d} \frac{1}{\theta\alpha_1} |m_2|^2 \\ &\leq \frac{1}{\theta\alpha_1} \|m_2\|^2_{L^2(Q)} \\ &\leq \frac{C}{\theta\alpha_1} \|g\|^2_{L^2(Q)}. \end{aligned}$$

Finally, we get

$$\begin{aligned} K(\rho, m) - K(\rho^*, m^*) &\leq \frac{\theta}{1-\theta} K(\rho^*, m^*) + \frac{C}{\theta\alpha_1} \|g\|^2_{L^2(Q)} \\ &\leq \frac{\theta}{1-\theta} K(\rho^*, m^*) + \frac{C}{\theta\alpha_1} \epsilon^2 \\ &\leq 4\epsilon \frac{\sqrt{T}}{\alpha_1} K(\rho^*, m^*) + \frac{C}{2\sqrt{T}} \epsilon \quad \text{when taking } \theta = 2\epsilon \frac{\sqrt{T}}{\alpha_1}. \end{aligned}$$

We want $K(\rho, m) - K(\rho^*, m^*) \leq r$. It is sufficient to take a small ϵ, since r is of the same order as s_0. It is then possible to take $s = r - (K(\rho, m) - K(\rho^*, m^*))$, and we finally get (ρ, m, s) such that

$$(K(\rho, m) - K(\rho^*, m^*) + s, \partial_t\rho + \nabla_x.m) = (r, g). \tag{2.14}$$

We deduce that the interior of S is not empty. $\square$

It is now possible to finish the proof of the theorem. From the Hahn–Banach theorem, there exists a nonzero linear form separating S and T. We then have $(\alpha_0, \phi_0) \in \mathbb{R} \times L^2_0$ such that

$$\forall(\rho, m, s, t),\ \alpha_0(K(\rho, m) - K(\rho^*, m^*) + s)\ + \langle \nabla_{t,x}.(\rho, m), \phi_0\rangle\ \geq\ -\alpha_0 t. \tag{2.15}$$

First we take $(\rho, m) = (\rho^*, m^*)$. We have

$$\forall(s, t), \qquad s \geq 0, t > 0, \alpha_0 s \geq \alpha_0 t.$$

We deduce that α_0 is nonnegative. Assume now that $\alpha_0 = 0$. Then (2.15) becomes

$$\forall(\rho, m, s), \qquad \langle \nabla_{t,x}.(\rho, m), \phi_0\rangle \geq 0. \tag{2.16}$$

Let ψ be the solution in L^2_0 of the following system:

$$\begin{cases} \Delta\psi = -\phi_0, \\ \partial_t\psi(0, .) = \rho_0, \\ \partial_t\psi(T, .) = \rho_T, \end{cases}$$

with periodic boundary conditions in space.

We define $(\rho, m) = \nabla_{t,x}\psi$. We then get $\|\phi_0\|^2_{L^2} \leq 0$, and therefore $\phi_0 = 0$, which is a contradiction. We deduce that $\alpha_0 > 0$. Finally, define $\lambda^* = \frac{\phi_0}{\alpha_0}$. We have to check that the triplet (ρ^*, m^*, λ^*) is indeed a saddle-point of L. Taking $s = 0$ and letting t go to 0, we get

$$\forall(\rho, m) \in H(Q; \text{div}), \qquad L(\rho^*, m^*, \lambda^*) \leq L(\rho, m, \lambda^*). \tag{2.17}$$

Moreover, we have $\nabla_{t,x}.(\rho^*, m^*) = 0$, and therefore

$$\forall\lambda \in L^2_0, \qquad L(\rho^*, m^*, \lambda) \leq L(\rho^*, m^*, \lambda^*). \tag{2.18}$$

Hence, the triplet (ρ^*, m^*, λ^*) is a saddle-point of L, and Theorem 2.1 is proved.

3. More on the saddle-point.

Now we prove some properties of the saddle-point. Indeed, we have

$$K(\rho^*, m^*) = \sup_{(a,b)\in\tilde{K}} \int_{[0;T]} \int_{\mathbb{T}^d} a\rho^* + b.m^* dxdt. \tag{3.1}$$

As a consequence of the uniform boundedness of the velocity fields in the proof of Lemma 1.3, we see that the optimal pair (a^*, b^*) is actually reached and bounded in $L^\infty(Q)$. Indeed, this pair is given by the following equalities:

$$\begin{cases} a^* = -\dfrac{|m^*|^2}{2{\rho^*}^2}, \\ b^* = \dfrac{m^*}{\rho^*}. \end{cases} \tag{3.2}$$

Let $(\tilde{\rho}, \tilde{m})$ in $\mathcal{C}^\infty(Q)^{1+d}$ and δ in $\mathbb{R}$. Assume furthermore that

$$\tilde{\rho}(0,.) = 0, \qquad \tilde{\rho}(T,.) = 0. \tag{3.3}$$

Then the pair $(\rho = \rho^* + \tilde{\rho}, m = m^* + \tilde{m})$ satisfies

$$0 \leq \delta \int_{[0;T]} \int_{\mathbb{T}^d} (a^* \tilde{\rho} + b^* . \tilde{m}) + \lambda^* \nabla_{t,x} . (\tilde{\rho}, \tilde{m}) dx dt + O(\delta^2). \tag{3.4}$$

Letting δ go to 0 and then using the density of $\mathcal{C}^\infty(Q)^{1+d}$ in $H(Q; \mathrm{div})$, we get that for any $(\tilde{\rho}, \tilde{m})$ in $H(Q; \mathrm{div})$ satisfying (3.3)

$$\int_{\mathbb{T}^d} (a^* \tilde{\rho} + b^* . \tilde{m}) + \lambda^* \nabla_{t,x} . (\tilde{\rho}, \tilde{m}) dx dt = 0. \tag{3.5}$$

We recall now that any function u in $L^2(Q)^{1+d}$ can be uniquely (and continuously) decomposed in $L^2(Q)^{1+d}$ as a sum $u = u_1 + u_2$ such that

$$\begin{gathered} u_1 = \nabla_{t,x} \phi \quad \text{for some } \phi \in H^1(Q), \\ u_2 \text{ satisfies } \left\{ \begin{array}{ll} \nabla_{t,x} . u_2 = 0 & \text{in } L^2(Q), \\ u_2 . n = 0 & \text{on } \partial Q. \end{array} \right. \end{gathered} \tag{3.6}$$

We then write $(a^*, b^*) = \nabla \phi^* + v^*$. Inserting into (3.5) and integrating by part, we get that for all $(\tilde{\rho}, \tilde{m})$ in $H(Q; \mathrm{div})$ satisfying (3.3)

$$\int_{[0;T]} \int_{\mathbb{T}^d} (\lambda^* - \phi^*) \nabla_{t,x} . (\tilde{\rho}, \tilde{m}) dx dt + \int_{[0;T]} \int_{\mathbb{T}^d} v^* . (\tilde{\rho}, \tilde{m}) dx dt = 0. \tag{3.7}$$

Hence we get that if $\nabla_{t,x} . (\tilde{\rho}, \tilde{m}) = 0$,

$$\int_{[0;T]} \int_{\mathbb{T}^d} v^* . (\tilde{\rho}, \tilde{m}) dx dt = 0. \tag{3.8}$$

Integrating by parts the product $\langle v^*, \nabla_{t,x} \phi \rangle$ for any ϕ in $H^1(Q)$ and using the definition of v^*, we get 0. Then using the decomposition property of $L^2(Q)^{1+d}$ for any v in $L^2(Q)^{1+d}$, we conclude that $v^* = 0$. Then inserting into (3.7) and using the definition of a saddle-point, we see that (ρ^*, m^*, ϕ^*) is a saddle-point of L.

As in [1], we are now ready to define a new Lagrangian $\mathcal{L}$. We also take the same notations

$$\left\{ \begin{array}{l} \mu = (\rho, m), \\ q = (a, b), \\ F(q) = \left\{ \begin{array}{l} 0 \text{ if } q \in \tilde{K}, \\ +\infty \text{ else}, \end{array} \right. \\ G(\phi) = \displaystyle\int_{\mathbb{T}^d} [\phi(0,.)\rho_0 - \phi(T,.)\rho_T], \\ \langle \mu, q \rangle = \displaystyle\int_{[0;T]} \int_{\mathbb{T}^d} \mu . q \; dx dt \end{array} \right.$$

to get that for all $(\mu, q, \phi) \in L^2(Q)^{d+1} \times L^2(Q)^{d+1} \times H^1(Q)$ (from now on, this space will be denoted by $E(Q)$)

$$\mathcal{L}(\mu, q, \phi) = -F(q) - G(\phi) + \langle \mu, q - \nabla_{t,x} \phi \rangle. \tag{3.9}$$

We now have the following theorem.

THEOREM 3.1. *(μ^*, q^*, ϕ^*) is a saddle-point of $\mathcal{L}$ in $E(Q)$. Moreover, any saddle-point of $\mathcal{L}$ in $E(Q)$ is of the form $(\tilde{\mu}, q^*, \phi^* + \tilde{C})$, where $\tilde{C}$ is a constant, and $\tilde{\mu}$ is a solution of the time-continuous mass transport problem.*

Proof. Let μ be in $L^2(Q)^{1+d}$. We have $\langle \mu, q^* - \nabla_{t,x}\phi^* \rangle = 0 = \langle \mu, q^* - \nabla_{t,x}\phi^* \rangle$. Hence we have

$$\mathcal{L}(\mu^*, q^*, \phi^*) \leq \mathcal{L}(\mu, q^*, \phi^*).$$

Let (q, ϕ) be in $\tilde{K} \times H^1(Q)$. First we observe that a simple integration by parts gives $\langle \mu^*, \nabla_{t,x}\phi \rangle = -G(\phi)$, since μ^* is in $V(Q)$ and satisfies (0.7). Then we have

$$\begin{aligned}
\mathcal{L}(\mu^*, q, \phi) &= \langle \mu^*, q - \nabla_{t,x}\phi \rangle - G(\phi) \\
&= \langle \mu^*, q \rangle \\
&\leq \langle \mu^*, q^* \rangle \\
&\leq \langle \mu^*, q^* - \nabla_{t,x}\phi^* \rangle - G(\phi^*) \\
&\leq \mathcal{L}(\mu^*, q^*, \phi^*).
\end{aligned}$$

Then, summarizing these inequalities, we get that for any (μ, q, ϕ) in $L^2(Q)^{1+d} \times L^2(Q)^{1+d} \times H^1(Q)$ we have

$$\mathcal{L}(\mu^*, q, \phi) \leq \mathcal{L}(\mu^*, q^*, \phi^*) \leq \mathcal{L}(\mu, q^*, \phi^*), \tag{3.10}$$

which means that (μ^*, q^*, ϕ^*) is a saddle-point of $\mathcal{L}$.

Let $(\tilde{\mu}, \tilde{q}, \tilde{\phi})$ in $L^2(Q)^{1+d} \times L^2(Q)^{1+d} \times H^1(Q)$ be another saddle-point of $\mathcal{L}$. We have $\tilde{q} \in \tilde{K}$. Assume next that $\tilde{q} \neq \nabla_{t,x}\tilde{\phi}$, and define $\mu_n = n(\nabla_{t,x}\tilde{\phi} - \tilde{q})$. From the definition of a saddle-point, we get that

$$\begin{aligned}
\mathcal{L}(\tilde{\mu}, \tilde{q}, \tilde{\phi}) &\leq \langle \mu_n, \tilde{q} - \nabla_{t,x}\tilde{\phi} \rangle - G(\tilde{\phi}) \\
&\leq -n\|\tilde{q} - \nabla_{t,x}\tilde{\phi}\|^2_{L^2(Q)} - G(\tilde{\phi}).
\end{aligned}$$

We obtain a contradiction by letting n go to infinity. Then $\tilde{q}$ is a gradient. Using the decomposition property if $L^2(Q)^{1+d}$, we write $\tilde{\mu} = \mu^* + \bar{\mu} + \nabla_{t,x}\bar{\phi}$. From the definition of a saddle-point, we know that for any (q, ϕ) in $L^2(Q)^{1+d} \times H^1(Q)$ we have

$$\mathcal{L}(\tilde{\mu}, q, \phi) \leq \mathcal{L}(\tilde{\mu}, \tilde{q}, \tilde{\phi}). \tag{3.11}$$

Assume that there exists some ϕ_1 such that $\langle \tilde{\mu}, \nabla_{t,x}\phi_1 \rangle + G(\phi_1) < 0$. Then taking $(q, \phi) = (0, n\phi_1)$ in (3.11), we get

$$-n\left(\langle \tilde{\mu}, \nabla_{t,x}\phi_1 \rangle + G(\phi_1)\right) \leq \mathcal{L}(\tilde{\mu}, \tilde{q}, \tilde{\phi}). \tag{3.12}$$

Letting n go to infinity, we obtain a contradiction. Hence we see that for all ϕ in $H^1(Q)$, $\langle \tilde{\mu}, \nabla_{t,x}\phi \rangle + G(\phi) = 0$. In particular, this has to be true with $\bar{\phi}$. Using our decomposition of $\tilde{\mu}$, we get

$$\begin{aligned}
0 &= \langle \mu^* + \bar{\mu} + \nabla_{t,x}\bar{\phi}, \bar{\phi} \rangle + G(\bar{\phi}) \\
&= \langle \mu^*, \nabla_{t,x}\bar{\phi} \rangle + G(\bar{\phi}) + \langle \bar{\mu}, \nabla_{t,x}\bar{\phi} \rangle + \|\nabla_{t,x}\bar{\phi}\|^2_{L^2(Q)}.
\end{aligned}$$

Integrating by parts and using the properties of μ^* and $\bar{\mu}$, we deduce that $\nabla_{t,x}\bar{\phi}$ is null. Hence we see that $\tilde{\mu}$ is in fact in $V(Q)$ and satisfies the boundary conditions (0.7).

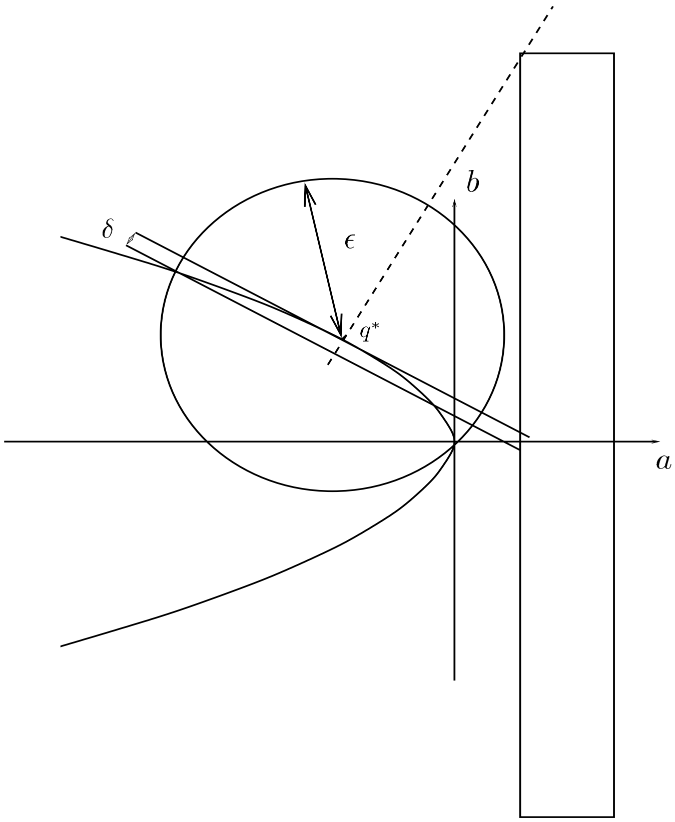


FIG. 3.1. *Geometrical intuition.*

We now have to prove that $\tilde{q} = q^*$. From the fact that both pairs have to define saddle-points, we get

$$\langle \tilde{\mu}, q^* - \tilde{q} \rangle = 0, \tag{3.13}$$

$$\langle \mu^*, q^* - \tilde{q} \rangle = 0, \tag{3.14}$$

$$\langle \mu^*, q^* \rangle = \langle \tilde{\mu}, \tilde{q} \rangle. \tag{3.15}$$

The identity $\tilde{q} = q^*$ follows quite obviously from (3.14). Indeed, the set $\tilde{K}$ is strictly convex, and the L^∞ bounds on q^* ensure that we have some kind of uniform strict convexity. Specifically, we have the following result.

LEMMA 3.2. *Let $\tilde{q}$ be in $L^2(Q)$ such that $\langle \mu^*, q^* - \tilde{q} \rangle = 0$. Then $\tilde{q} = q^*$ a.e. on Q.*

Proof. Since q^* is bounded in $L^\infty(Q)$, we can have uniform estimates on the strict convexity of $\tilde{K}$. The geometric intuition is easily understood when looking at Figure 3.1.

Let $\epsilon > 0$. There exists some $\delta > 0$ such that for any (t, x) in Q

$$\|q^* - \tilde{q}\| > \epsilon \Rightarrow \mu^*.(q^* - \tilde{q}) < -\delta \|\mu^*\|. \tag{3.16}$$

Let $A_\epsilon = \{(t, x) \in Q \text{ s.t. } \|q^* - \tilde{q}\| > \epsilon\}$. We have

$$\begin{aligned}0 &= \langle \mu^*, q^* - q\rangle \\ &\leq \int_{A_\epsilon} \mu^*.(q^* - q)dtdx \\ &\leq -\delta \int_{A_\epsilon} \|\mu^*\| \\ &\leq -\delta\alpha_1 |A_\epsilon|.\end{aligned}$$

Then we get that $|A_\epsilon| = 0$. This achieves the proof of the lemma. □

Finally, since $(\tilde{\mu}, \tilde{q}, \tilde{\phi})$ is a saddle-point of $\mathcal{L}$, we have

$$\langle \tilde{\mu}, \tilde{q}\rangle = K(\tilde{\rho}, \tilde{m}). \tag{3.17}$$

Then we get from (3.15) (and from the fact that $\tilde{\mu}$ is in $V(Q)$ and satisfies (0.7)) that $\tilde{\mu}$ is a solution of the time-continuous mass transport problem. This achieves the proof of Theorem 3.1. □

This theorem answers a question that was left open in [1] on the existence of the saddle-point for $\mathcal{L}$ in the infinite-dimensional case. Moreover, we give a precise functional background for searching this saddle-point.

Remark 3.3. In Theorem 3.1, the variable μ is taken in $L^2(Q)^{1+d}$ rather than in $H(Q; \mathrm{div})$. This will remove a constraint when the question turns to the effective search for the saddle-point.

4. On the algorithm of [1]. In [1], the authors defined an augmented Lagrangian as a preliminary for their numerical method. To get some more coercivity, they perturbed the functional F. Here, since we have already made an assumption of boundedness away from zero on the saddle-point, such a perturbation is not necessary. We then define on $E(Q)$ the augmented Lagrangian

$$\begin{aligned}L_r(\mu, q, \phi) = F(q) + G(\phi) &+ \langle \mu, \nabla_{t,x}\phi - q\rangle \\ &+ \frac{r}{2}\langle \nabla_{t,x}\phi - q, \nabla_{t,x}\phi - q\rangle,\end{aligned} \tag{4.1}$$

where r is a positive parameter.

Remark 4.1. We can change the signs since we proved the existence of a saddle-point of $\mathcal{L}$. In this formulation, the constraint is to get q as a gradient. This constraint has been augmented, instead of the old constraint that μ is in $V(Q)$ and satisfies (0.7).

It is a classical result (see [7]) that if (μ, q, ϕ) is a saddle-point of L_r, it is also a saddle-point of $\mathcal{L}$ (and conversely). Benamou and Brenier then used a numerical algorithm ALG2 to solve the problem. We recall here this algorithm and refer to [1] for more explanation on the steps and some numerical results. Here we are now concerned with the convergence of the method in the continuous case.

ALGORITHM ALG2.

- $(\phi^{n-1}, q^{n-1}, \mu^n)$ are given.
- Step A. Find ϕ^n in $H^1(Q) \cap L^2_0(Q)$ such that

$$L_r(\phi^n, q^{n-1}, \mu^n) \leq L_r(\phi, q^{n-1}, \mu^n) \ \ \forall \phi. \tag{4.2}$$

- Step B. Find q^n in $L^2(Q)^{1+d}$ such that

$$L_r(\phi^n, q^n, \mu^n) \leq L_r(\phi^n, q, \mu^n) \ \ \forall q. \tag{4.3}$$

- Step C. Do

$$\mu^{n+1} = \mu^n + \delta(\nabla_{t,x}\phi^n - q^n) \tag{4.4}$$

(where $r > 0$ is the parameter of the augmented Lagrangian).

- Go back to Step A.

Remark 4.2. In Step A, the minimization is performed over $H^1(Q) \cap L^2_0(Q)$ in order to have a unique solution. Moreover, it is a way to fix the additive constant from Theorem 3.1.

The convergence of the sequence constructed by ALG2 is proved in [7] under some quite general assumptions, which are unfortunately not fully satisfied here. However, their proof can be adapted in order to deal with the problem under consideration.

THEOREM 4.3. *Assume that*

$$0 < \delta < \delta_M = \frac{1+\sqrt{5}}{2} r. \tag{4.5}$$

*Then the sequence constructed by ALG*2 *satisfies the following convergence properties:*

$$\phi^n \to \phi^* \textit{ strongly in } H^1(Q), \tag{4.6}$$

$$q^n \to q^* \textit{ strongly in } L^2(Q)^{1+d}, \tag{4.7}$$

$$\mu^{n+1} - \mu^n \to 0 \textit{ strongly in } L^2(Q)^{1+d}, \tag{4.8}$$

$$\mu^n \textit{ is bounded in } L^2(Q)^{1+d}. \tag{4.9}$$

Moreover, if $\tilde{\mu}$ *is a (weak) cluster point of* (μ^n) *in* $L^2(Q)^{1+d}$, *then* $(\tilde{\mu}, q^*, \phi^*)$ *is a saddle-point of* L_r *on* $E(Q)$.

Proof. The proof of the convergence of ALG2 in [7] requires some uniform convexity properties for F, which are not satisfied here. However, due to the particular form of our function F, the first part of their proof simplifies (since any term containing F vanishes). Hence a simple rewriting of their (intricate) calculations leads to ($|f|$ denotes the L^2-norm of f)

$$\begin{cases} (|\bar{\mu}^n|^2 + \delta r|\bar{q}^{n-1}|^2) - (|\bar{\mu}^{n+1}|^2 + \delta r|\bar{q}^n|^2) \geq \delta(2r-\delta)|\nabla_{t,x}\bar{\phi}^n - \bar{q}^n|^2 \\ \quad + \delta r|\bar{q}^n - \bar{q}^{n-1}|^2 - \delta|r-\delta| \left(\dfrac{1}{\alpha}|\nabla_{t,x}\bar{\phi}^{n-1} - \bar{q}^{n-1}|^2 + \alpha|\bar{q}^n - \bar{q}^{n-1}|^2\right), \end{cases} \tag{4.10}$$

where $\bar{\mu}^n = \mu^n - \mu^*$, $\bar{q}^n = q^n - q^*$, $\bar{\phi}^n = \phi^n - \phi^*$, and $\alpha > 0$ is a parameter.

If $0 < \delta \leq r$, taking $\alpha = 1$ and observing that $|r - \delta| = r - \delta$, we get

$$v_{n-1} - v_n \geq \delta r|\nabla_{t,x}\bar{\phi}^n - \bar{q}^n|^2 + \delta^2|\bar{q}^n - \bar{q}^{n-1}|^2, \tag{4.11}$$

with $v_n = (|\bar{\mu}^{n+1}|^2 + \delta r|\bar{q}^n|^2 + \delta(r-\delta)|\nabla_{t,x}\bar{\phi}^n - \bar{q}^n|^2)$. If $r < \delta < \delta_M$, taking $\alpha = \frac{1+\sqrt{5}}{2}$, we have

$$w_{n-1} - w_n \geq \frac{\delta_M \delta}{r}(\delta_M - \delta)|\nabla_{t,x}\bar{\phi}^n - \bar{q}^n|^2 + \delta(\delta_M - \delta)|\bar{q}^n - \bar{q}^{n-1}|^2, \tag{4.12}$$

with $w_n = (|\bar{\mu}^{n+1}|^2 + \delta r|\bar{q}^n|^2 + \frac{\delta r}{\delta_M}(\delta - r)|\nabla_{t,x}\bar{\phi}^n - \bar{q}^n|^2)$.

In (4.11) and (4.12), the right-hand sides are nonnegative. Then the sequences (v_n) and (w_n) are decreasing. Hence we have that (μ_n) and (q_n) are uniformly bounded in $L^2(Q)^{1+d}$. Moreover, we see from the right-hand sides that $\sum_{n=1}^{\infty} |\nabla_{t,x}\bar{\phi}^n - \bar{q}^n|^2$ and $\sum_{n=1}^{\infty} |\bar{q}^n - \bar{q}^{n-1}|^2$ are finite. It implies that

$$\begin{cases} \nabla_{t,x}\bar{\phi}^n - \bar{q}^n \to 0 \quad \text{strongly in } L^2(Q)^{1+d}, \\ \\ (\bar{q}^n) \quad \text{is a Cauchy sequence in } L^2(Q)^{1+d}. \end{cases} \tag{4.13}$$

We then have that (q_n) converges strongly in $L^2(Q)^{1+d}$ to some $\tilde{q}$. Hence we have that $(\nabla_{t,x}\phi_n)$ also converges strongly to $\tilde{q}$ in $L^2(Q)^{1+d}$. Since ϕ_n is in $L^2_0(Q)$, we see then that the sequence (ϕ_n) converges strongly in $H^1(Q) \cap L^2_0(Q)$ to some $\tilde{\phi}$. Moreover, the sequence (μ_n) is bounded in $L^2(Q)^{1+d}$. We can then extract a subsequence (still denoted by n) weakly converging in $L^2(Q)^{1+d}$ to some $\tilde{\mu}$. We now have to prove that $\tilde{q} = q^*$ and $\tilde{\phi} = \phi^*$.

Step A means that for any ϕ in $H^1(Q)$

$$\begin{aligned} 0 \le G(\phi) - G(\phi^n) &+ \langle \mu^n, \nabla_{t,x}\phi - \nabla_{t,x}\phi^n \rangle \\ &+ r\langle \nabla_{t,x}\phi^n - q^{n-1}, \nabla_{t,x}\phi - \nabla_{t,x}\phi^n \rangle. \end{aligned} \tag{4.14}$$

Step B means that for any q in $L^2(Q)^{1+d}$

$$0 \le F(q) - \langle \mu^n, q - q^n \rangle + r\langle q^n - \nabla_{t,x}\phi^n, q - q^n \rangle. \tag{4.15}$$

Taking $\phi = \phi^*$ in (4.14) and $q = q^*$ in (4.15), and letting n tend to infinity, we get

$$0 \le G(\phi^*) - G(\tilde{\phi}) + \langle \tilde{\mu}, \nabla_{t,x}\phi^* - \nabla_{t,x}\tilde{\phi} \rangle, \tag{4.16}$$

$$0 \le -\langle \tilde{\mu}, q^* - \tilde{q} \rangle. \tag{4.17}$$

Then, adding the two inequalities and using that $\nabla_{t,x}\tilde{\phi} = \tilde{q}$ and $\nabla_{t,x}\phi^* = q^*$, we get

$$G(\tilde{\phi}) \le G(\phi^*). \tag{4.18}$$

Moreover, since (μ^*, q^*, ϕ^*) is a saddle-point of L_r, we have

$$G(\phi^*) \le G(\tilde{\phi}). \tag{4.19}$$

We deduce that $G(\phi^*) = G(\tilde{\phi})$. We recall now that for any ϕ in $H^1(Q)$, we have $G(\phi) + \langle \mu^*, \nabla_{t,x}\phi \rangle = 0$. Using this equality with ϕ^* and $\tilde{\phi}$, we get

$$\langle \mu^*, q^* \rangle = \langle \mu^*, \tilde{q} \rangle. \tag{4.20}$$

We now recall Lemma 3.2 to get $\tilde{q} = q^*$. Hence $\tilde{\phi} = \phi^*$. To end the proof of Theorem 4.3, it remains to show that $(\tilde{\mu}, q^*, \phi^*)$ is a saddle-point of L_r. Letting n go to infinity in (4.14) and (4.15) and adding the resulting inequalities, we get that for any (q, ϕ) in $L^2(Q)^{1+d} \times H^1(Q)$,

$$G(\phi^*) \le F(q) + G(\phi) + \langle \tilde{\mu}, \nabla_{t,x}\phi - q \rangle. \tag{4.21}$$

We conclude that $(\tilde{\mu}, q^*, \phi^*)$ is a saddle-point of $\mathcal{L}$ and then also a saddle-point of L_r. This achieves the proof of Theorem 4.3. □

Acknowledgments. The author would like to thank J. D. Benamou and Y. Brenier, whose careful reading of the manuscript substantially improved the paper.

REFERENCES

[1] J. D. Benamou and Y. Brenier, *A computational fluid mechanics solution to the Monge–Kantorovich mass transfer problem*, Numer. Math., 84 (2000), pp. 375–393.

[2] Y. Brenier, *Polar factorization and monotone rearrangement of vector-valued functions*, Comm. Pure Appl. Math., 44 (1991), pp. 375–417.

[3] Y. Brenier, *A homogenized model for vortex sheets*, Arch. Rational Mech. Anal., 138 (1997), pp. 319–353.

[4] L. A. Caffarelli, *Boundary regularity of maps with convex potentials*, Comm. Pure Appl. Math., 45 (1992), pp. 1141–1151.

[5] L. A. Caffarelli, *Boundary regularity of maps with convex potentials.* II, Ann. of Math. (2), 144 (1996), pp. 453–496.

[6] D. Cordero-Erausquin, *Sur le transport de mesures périodiques*, C. R. Acad. Sci. Paris Sér. I Math., 329 (1999), pp. 199–202.

[7] M. Fortin and R. Glowinski, *Augmented Lagrangian Methods. Applications to the Numerical Solution of Boundary Value Problems*, Stud. Math. Appl. 15, North–Holland, Amsterdam, 1983.

[8] W. Gangbo and R. J. McCann, *The geometry of optimal transportation*, Acta Math., 177 (1996), pp. 113–161.

[9] R. J. McCann, *A convexity principle for interacting gases*, Adv. Math., 128 (1997), pp. 153–179.

[10] R. J. McCann, *Polar factorization of maps on Riemannian manifolds*, Geom. Funct. Anal., 11 (2001), pp. 589–608.

SIAM J. NUMER. ANAL.
Vol. 41, No. 2, pp. 401–435

COMPUTATION OF PERIODIC SOLUTION BIFURCATIONS IN ODES USING BORDERED SYSTEMS*

E. J. DOEDEL†, W. GOVAERTS‡, AND YU. A. KUZNETSOV§

Abstract. We consider numerical methods for the computation and continuation of the three generic secondary periodic solution bifurcations in autonomous ODEs, namely the fold, the period-doubling (or flip) bifurcation, and the torus (or Neimark–Sacker) bifurcation. In the fold and flip cases we append one scalar equation to the standard periodic BVP that defines the periodic solution; in the torus case four scalar equations are appended. Evaluation of these scalar equations and their derivatives requires the solution of linear BVPs, whose sparsity structure (after discretization) is identical to that of the linearization of the periodic BVP. Therefore the calculations can be done using existing numerical linear algebra techniques, such as those implemented in the software AUTO and COLSYS.

Key words. bifurcations, periodic solutions, continuation, boundary value problems

AMS subject classifications. 65P30, 65L10, 37M20, 37G15

PII. S0036142902400779

1. Introduction.

We consider parameterized ODEs of the form

$$\frac{dx}{dt} \equiv x' = f(x, \alpha), \tag{1.1}$$

where $x \in \mathbf{R}^n$ is the *state variable*, where $\alpha \in \mathbf{R}^m$ represents *parameters*, and where $f(x, \alpha) \in \mathbf{R}^n$ is a (usually nonlinear) smooth function of x and α. Examples of systems of the form (1.1) are ubiquitous in mathematical models in physics, engineering, chemistry, economics, finance, etc.

The simplest solutions of (1.1) are the *equilibria*, that is, solutions of the equation

$$f(x, \alpha) = 0.$$

An equilibrium (x_0, α_0) is asymptotically stable if all eigenvalues of the Jacobian matrix $f_x(x_0, \alpha_0)$ have a strictly negative real part; it is unstable if there is at least one eigenvalue with a strictly positive real part. In generic one-parameter problems, i.e., when $m = 1$, eigenvalues on the imaginary axis appear in two ways: as a simple zero eigenvalue, or as a conjugate pair $\pm i\omega$, $\omega > 0$, of purely imaginary eigenvalues. The first singularity corresponds generically to a *limit point bifurcation*, where two solutions coalesce and annihilate each other under parameter variation. The second singularity corresponds generically to a *Hopf bifurcation*, from which a family of periodic solutions emerges. Early papers on the numerical computation of bifurcations of equilibria are [16], [22], and [20].

*Received by the editors January 9, 2002; accepted for publication (in revised form) October 16, 2002; published electronically April 23, 2003.

http://www.siam.org/journals/sinum/41-2/40077.html

†Applied Mathematics 217-50, California Institute of Technology, Pasadena CA 91125 (on leave from Concordia University, Montreal, Canada) (doedel@cs.concordia.ca).

‡Department of Applied Mathematics and Computer Science, University of Gent, Krijgslaan 281-S9, B-9000 Gent, Belgium (willy.govaerts@rug.ac.be).

§Mathematical Institute, Utrecht University, Budapestlaan 6, P.O. Box 80010, 3508 TA Utrecht, The Netherlands and Institute of Mathematical Problems of Biology, Russian Academy of Sciences, Pushchino, Moscow Region, 142290 Russia (kuznetsov@math.uu.nl).

Periodic solutions are solutions for which $x(T) = x(0)$ for some number $T > 0$. The minimal such T is called the *period*. In generic one-parameter problems, periodic solutions can bifurcate in several ways that can be characterized by the properties of the *monodromy matrix*. The monodromy matrix is the linearized T-shift along orbits of (1.1), evaluated at the point $x(0)$ on the periodic solution. The eigenvalues of this matrix are called the *Floquet multipliers* of the periodic solution [14], [17].

A periodic solution always has a multiplier equal to 1. If this multiplier has geometric multiplicity 1, then we call the periodic solution *regular*. The corresponding eigenvector of the monodromy matrix is the tangent vector to the periodic solution at the point where the monodromy matrix is computed. If all other multipliers are strictly inside the unit circle in the complex plane, then the periodic solution is asymptotically stable. If at least one multiplier has modulus greater than 1, then the periodic solution is unstable. In all other cases, one should take into account higher-order derivatives of the T-shift to decide whether or not the periodic orbit is stable.

Three singularities, determined by the monodromy matrix, can occur along a one-parameter family ("curve" or "branch") of periodic solutions, namely (1) *a fold singularity*, when the multiplier 1 has algebraic multiplicity equal to or greater than 2; (2) *a flip singularity*, when there is a multiplier equal to -1; (3) *a Neimark–Sacker singularity*, when there is a conjugate pair of complex multipliers with modulus 1.

Under some genericity conditions, each of these singularities implies a certain bifurcation scenario. These conditions always include some *spectral conditions* on the critical multipliers, i.e., multiplicity restrictions and the absence of other critical multipliers. Furthermore, there are *nondegeneracy conditions* that can be formulated in terms of the system at the critical parameter values, and *transversality conditions* that are determined by the system's dependence on the parameter (see [17]). We shall list all relevant genericity conditions in the following sections.

Generically, the first critical case (fold) corresponds to a point on the periodic solution family where the curve turns quadratically with respect to the free parameter. This phenomenon is called a *limit point* (*fold*) *bifurcation*: Two periodic solutions collide and disappear when the parameter passes the critical value. The second case (flip) indicates generically a *period-doubling* of the periodic solution; i.e., there are nearby periodic solutions of approximately double period. It is also called the *flip bifurcation*. Finally, the third case (Neimark–Sacker) corresponds generically to a bifurcation of an *invariant torus*, on which the flow contains periodic or quasi-periodic motions. This phenomenon is often called the *Neimark–Sacker bifurcation*. There is some ambiguity in calling a bifurcation by the same name as the corresponding singularity. However, this is a common practice in the applied literature.

The aim of this paper is to formulate the computation and continuation of the three generic periodic solution bifurcation curves as *minimally extended BVPs* to which standard numerical approximation methods as well as convergence theory apply. *Fully extended BVPs* for continuing periodic solution bifurcations have been implemented in AUTO [6] (see also [7], [15]). The latter approach doubles the number of function components in the case of the period-doubling and fold bifurcations, and triples it in the case of the torus bifurcation. Fully extended BVPs also yield a more complicated Jacobian sparsity structure (after discretization) than that corresponding to the underlying periodic BVP. There are efficient solution techniques for such sparse linear systems; see, for example, [10]. However, these are not very easy to implement and they are specific for each bifurcation. By contrast, the minimal BVPs presented in this paper for the period-doubling and fold bifurcations have the same number of function components as the periodic solution problem. In the torus case the number of BVP function components is only doubled, but the resulting system is overdetermined. The most important numerical advantage is that only one type of sparse system needs to be solved, namely the one corresponding to the underlying periodic BVP. Conceptually, the approach used in this paper is similar to the *bordering technique* for equilibrium bifurcations [5], [12], [13], [17].

The paper is organized as follows. Section 2 is devoted to the computation of one-parameter families of periodic solutions to (1.1). Classical results on the regularity of BVPs defining families of periodic solutions are proved here for completeness. Sections 3 and 4 present the main results of the paper. Here we construct functionals that vanish at bifurcation points of periodic solutions and we prove that they are well-defined and regular. As is usual, only some of the nondegeneracy conditions that appear in bifurcation analysis are necessary for regularity. Section 5 deals with various computational issues, including efficient computation of the defining systems and their derivatives. A numerical example is given in section 6.

2. Computation and continuation of periodic solutions. Numerical continuation is a technique to compute solution curves to an underdetermined system of equations. Details can be found, for example, in [1], [3], [12], and [16]. It is a basic ingredient of the numerical bifurcation algorithms implemented in AUTO [6] and CONTENT [18]. In this case only one parameter is free, so for practical purposes the parameter vector reduces to a scalar. In this paper we restrict our discussion to issues that are specific to the case of periodic orbits.

To compute a periodic solution of period T of (1.1), one first fixes the period by rescaling time. Then (1.1) becomes

$$x'(t) = Tf(x(t), \alpha), \tag{2.1}$$

and we look for solutions of period 1, that is,

$$x(0) = x(1). \tag{2.2}$$

The period T is one of the unknowns of the problem. In a continuation context, we assume that a solution $(x_{k-1}(\cdot), T_{k-1}, \alpha_{k-1})$ is known, and we want to find $(x_k(\cdot), T_k, \alpha_k)$, which we denote by $(x(\cdot), T, \alpha)$. Equations (2.1) and (2.2) together do not fix the solution completely, since any solution can be translated freely in time; that is, if $x(t)$ is a solution, then so is $x(t+\sigma)$ for any σ. To fix the solution it is necessary to add a "phase condition." In AUTO [6] and CONTENT [18] the integral constraint

$$\int_0^1 x^*(\tau)x'_{k-1}(\tau)\; d\tau = 0 \tag{2.3}$$

is used to fix the phase. (We use "*" to denote transpose.)

The periodic solution is now determined by (2.1), (2.2), (2.3), which together form a BVP with an integral constraint.

In our continuation context, the periodic orbit $x(t)$ and the scalars T and α vary along the solution family. In the setting of Keller's pseudoarclength continuation method [16] the continuation equation is

$$\int_0^1 (x(\tau) - x_{k-1}(\tau))^* \dot{x}_{k-1}(\tau)\; d\tau \;\; + (T - T_{k-1})\dot{T}_{k-1} + (\alpha - \alpha_{k-1})\dot{\alpha}_{k-1} = \Delta s, \tag{2.4}$$

where the derivatives are taken with respect to arclength in the function space, and should not be confused with the time derivatives in, for example, (2.3).

A widely used method to discretize the above BVP is the method of orthogonal collocation with piecewise polynomials. It is used in COLSYS [2], as well as in AUTO and CONTENT. The method is known for its high accuracy [4], and it is particularly suitable for difficult problems, due to its known optimal mesh adaptation techniques [21]. The numerical continuation of the discretized equations leads to structured, sparse linear systems [9]. To describe these systems it is convenient to formulate the BVP in terms of operators on function spaces.

Denote by $\mathcal{C}^k([a,b],\mathbf{R}^n)$ the space of k times continuously differentiable functions defined on $[a,b]$ and with values in $\mathbf{R}^n$. Let D be the differentiation operator acting from $\mathcal{C}^1([a,b],\mathbf{R}^n)$ to $\mathcal{C}^0([a,b],\mathbf{R}^n)$. Any $n \times n$ matrix $M(t)$ smoothly depending on $t \in [a,b]$ defines an operator from $\mathcal{C}^1([a,b],\mathbf{R}^n)$ into itself by the matrix multiplication $(M\psi)(t) = M(t)\psi(t)$. The Dirac evaluation operator at the point t is denoted δ_t.

For a given $\phi \in \mathcal{C}^0([0,1],\mathbf{R}^n)$ we denote by Int_ϕ the linear functional from $\mathcal{C}^0([0,1],\mathbf{R}^n)$ into $\mathbf{R}$ defined by

$$\mathrm{Int}_\phi(v) = \langle \phi, v\rangle = \int_0^1 \phi^*(\tau)v(\tau)\, d\tau.$$

Suppose we want to compute a periodic solution of (1.1); i.e., we want to solve the system (2.1), (2.2), (2.3), and (2.4) for $(x(t), T, \alpha)$ by a Newton-like method. The Fréchet derivative operator corresponding to this problem has the form

$$\begin{pmatrix} D - Tf_x(x(t),\alpha) & -f(x(t),\alpha) & -Tf_\alpha(x(t),\alpha) \\ \delta_0 - \delta_1 & 0 & 0 \\ \mathrm{Int}_{x'_{k-1}}(\cdot) & 0 & 0 \\ \mathrm{Int}_{\dot{x}_{k-1}}(\cdot) & \dot{T}_{k-1} & \dot{\alpha}_{k-1} \end{pmatrix}. \tag{2.5}$$

The discrete version of these linear operators is a square matrix that has a large matrix corresponding to $D - Tf_x(x(t),\alpha)$ in the upper left corner, bordered on the right by two extra columns and at the bottom by $n+2$ extra rows. The big matrix in the upper left corner is a block band matrix. Systems of this form are solved in AUTO by a specially adapted elimination algorithm that computes the multipliers as a by-product [9].

Consider the fundamental variational equation

$$X' - Tf_x(x(t),\alpha)X = 0 \tag{2.6}$$

and the adjoint equation

$$X' + Tf_x^*(x(t),\alpha)X = 0. \tag{2.7}$$

Denote by $\Phi(t)$ the fundamental matrix solution of (2.6), for which $\Phi(0) = I$, where $I = I_{n\times n}$ is the n-dimensional identity matrix. Then $\Phi(1)$ is the monodromy matrix of the periodic solution. The eigenvalues of $\Phi(1)$ are the Floquet multipliers, and there is always at least one multiplier that is equal to 1. A corresponding eigenvector is $x'(0)$. For a *regular periodic solution* the multiplier 1 has geometric multiplicity 1. Similarly, denote by $\Psi(t)$ the fundamental matrix solution to (2.7) for which $\Psi(0) = I$. One has $\Psi(t) = [(\Phi(t))^{-1}]^*$.

If $v(t)$ is a vector solution to (2.6) with initial values $v(0) = v_0$ and $w(t)$ is a vector solution to (2.7) with initial values $w(0) = w_0$, then the inner product satisfies $w^*(t)v(t) = w_0^* v_0$; i.e., it is independent of time t.

The left and right eigenvectors of the monodromy matrix $\Phi(1)$ for a geometrically simple eigenvalue 1 will be denoted p_0, q_0, respectively. It is easily seen that p_0 (respectively, q_0) is also the right (respectively, left) eigenvector of $\Psi(1)$ for the eigenvalue 1. Furthermore, q_0 is a scalar multiple of $x'(0)$.

We now state some basic facts about the linear operator (2.5) when linearized about a regular periodic solution $(x(t), T, \alpha)$.

PROPOSITION 1. *If* $(x(t), T, \alpha)$ *is a regular periodic solution of* (2.1), *then the operator*

$$\begin{bmatrix} D - Tf_x(x(t), \alpha) \\ \delta_1 - \delta_0 \end{bmatrix} : \mathcal{C}^1([0,1], \mathbf{R}^n) \to \mathcal{C}^0([0,1], \mathbf{R}^n) \times \mathbf{R}^n \tag{2.8}$$

has a one-dimensional kernel spanned by Φq_0. *Its range has codimension* 1; *if* $\zeta \in \mathcal{C}^0([0,1], \mathbf{R}^n)$, $r \in \mathbf{R}^n$, *then* $(\zeta, r)^*$ *is in the range if and only if* $\langle \Psi p_0, \zeta \rangle = p_0^* r$. *In particular, if* $r = 0$, *then* $(\zeta, 0)^*$ *is in the range if and only if* $\langle \Psi p_0, \zeta \rangle = 0$.

Proof. First, let $v(t)$ be in the kernel of (2.8). Then v must have the form $v(t) = \Phi(t)v_0$ for a vector v_0. Since $0 = (\delta_1 - \delta_0)v = v(1) - v(0) = (\Phi(1) - I)v_0$, we infer that v_0 must be a right eigenvector of $\Phi(1)$ for the eigenvalue 1.

Next, let $\zeta \in \mathcal{C}^0([0,1], \mathbf{R}^n)$, $r \in \mathbf{R}^n$, be given. If $(\zeta, r)^*$ is in the range of (2.8), then there must exist a $v \in \mathcal{C}^1([0,1], \mathbf{R}^n)$ for which

$$v'(t) - Tf_x(x(t), \alpha)v(t) = \zeta(t).$$

The general solution of this linear differential equation is

$$v(t) = \Phi(t) \left[v_0 + \int_0^t \Psi^*(\tau)\zeta(\tau)\, d\tau \right],$$

where $v_0 = v(0)$ is an initial vector. Also, we must have $v(1) - v(0) = r$, that is,

$$(\Phi(1) - I)v_0 + \Phi(1) \int_0^1 \Psi^*(\tau)\zeta(\tau)\, d\tau = r.$$

Such a vector v_0 can be found if and only if

$$p_0^* \left(\Phi(1) \int_0^1 \Psi^*(\tau)\zeta(\tau)\, d\tau \ - r \right) = 0,$$

that is, if

$$p_0^* \int_0^1 \Psi^*(\tau)\zeta(\tau)\, d\tau \ - p_0^* r = 0,$$

from which the second result follows. □

COROLLARY 1. *If* $(x(t), T, \alpha)$ *is a regular periodic solution of* (2.1), *then the operator*

$$\begin{bmatrix} D - Tf_x(x(t), \alpha) \\ \delta_1 - \delta_0 \\ \mathrm{Int}_\phi \end{bmatrix} : \mathcal{C}^1([0,1], \mathbf{R}^n) \ \to \ \mathcal{C}^0([0,1], \mathbf{R}^n) \times \mathbf{R}^n \times \mathbf{R} \tag{2.9}$$

is one-to-one if and only if $\langle \phi, \Phi q_0 \rangle \neq 0$.

PROPOSITION 2. *If* $(x(t), T, \alpha)$ *is a regular periodic solution of* (2.1), *then the operator*

$$\begin{bmatrix} D + T f_x^*(x(t), \alpha) \\ \delta_1 - \delta_0 \end{bmatrix} \; : \; \mathcal{C}^1([0,1], \mathbf{R}^n) \;\to\; \mathcal{C}^0([0,1], \mathbf{R}^n) \times \mathbf{R}^n$$

has a one-dimensional kernel spanned by Ψp_0. *Its range has codimension* 1; *if* $\zeta \in \mathcal{C}^0([0,1], \mathbf{R}^n)$, $r \in \mathbf{R}^n$, *then* $(\zeta, r)^*$ *is in the range if and only if* $\langle \Phi q_0, \zeta \rangle = q_0^* r$. *In particular, if* $r = 0$, *then* $(\zeta, 0)^*$ *is in the range if and only if* $\langle \Phi q_0, \zeta \rangle = 0$.

Proof. The proof is similar to the proof of Proposition 1. □

COROLLARY 2. *If* $(x(t), T, \alpha)$ *is a regular periodic solution of* (2.1), *then the operator*

$$\begin{bmatrix} D + T f_x^*(x(t), \alpha) \\ \delta_1 - \delta_0 \\ \mathrm{Int}_\psi \end{bmatrix} \; : \; \mathcal{C}^1([0,1], \mathbf{R}^n) \;\to\; \mathcal{C}^0([0,1], \mathbf{R}^n) \times \mathbf{R}^n \times \mathbf{R} \tag{2.10}$$

is one-to-one if and only if $\langle \psi, \Psi p_0 \rangle \neq 0$.

PROPOSITION 3. *Let* $(x(t), T, \alpha)$ *be a regular periodic solution of* (2.1), *and let* $\phi_0, \psi_0 \in \mathcal{C}^0([0,1], \mathbf{R}^n)$ *be such that* $\langle \phi_0, \Phi q_0 \rangle \neq 0$, $\langle \psi_0, \Psi p_0 \rangle \neq 0$. *Then the operator*

$$\begin{bmatrix} D - T f_x(x(t), \alpha) & \psi_0 \\ \delta_1 - \delta_0 & 0 \\ \mathrm{Int}_{\phi_0} & 0 \end{bmatrix} \; : \; \mathcal{C}^1([0,1], \mathbf{R}^n) \times \mathbf{R} \;\to\; \mathcal{C}^0([0,1], \mathbf{R}^n) \times \mathbf{R}^n \times \mathbf{R}$$

is one-to-one and onto.

Proof. To prove that the operator is one-to-one, suppose that

$$\begin{bmatrix} D - T f_x(x(t), \alpha) & \psi_0 \\ \delta_1 - \delta_0 & 0 \\ \mathrm{Int}_{\phi_0} & 0 \end{bmatrix} \begin{pmatrix} v \\ G \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

for $v \in \mathcal{C}^1([0,1], \mathbf{R}^n), G \in \mathbf{R}$. In particular, it follows that

$$\begin{bmatrix} D - T f_x(x(t), \alpha) \\ \delta_0 - \delta_1 \end{bmatrix} v = \begin{pmatrix} -G\psi_0 \\ 0 \end{pmatrix}.$$

Since $\langle \psi_0, \Psi p_0 \rangle \neq 0$, it follows from the last statement in Proposition 1 that $G = 0$. By Corollary 1 and the assumption that $\langle \phi_0, \Phi q_0 \rangle \neq 0$, it follows that $v = 0$ as well.

To prove that the operator is onto we consider the equation

$$\begin{bmatrix} D - T f_x(x(t), \alpha) & \psi_0 \\ \delta_1 - \delta_0 & 0 \\ \mathrm{Int}_{\phi_0} & 0 \end{bmatrix} \begin{pmatrix} v \\ G \end{pmatrix} = \begin{pmatrix} \zeta \\ r \\ s \end{pmatrix}, \tag{2.11}$$

where $\zeta \in \mathcal{C}^0([0,1], \mathbf{R}^n), r \in \mathbf{R}^n, s \in \mathbf{R}$. In particular, the first two equations can be written

$$\begin{bmatrix} D - T f_x(x(t), \alpha) \\ \delta_1 - \delta_0 \end{bmatrix} v = \begin{pmatrix} \zeta - G\psi_0 \\ r \end{pmatrix}. \tag{2.12}$$

By Proposition 1 this equation is solvable for v, say, $v = v_p$, if

$$\langle \Psi p_0, \zeta - G\psi_0 \rangle = p_0^* r,$$

that is, if we choose

$$G = G_p \equiv \frac{\langle \Psi p_0, \zeta \rangle - p_0^* r}{\langle \Psi p_0, \psi_0 \rangle},$$

where, by assumption, the denominator does not vanish. Now

$$v(t) = v_p(t) + c\Phi(t)q_0$$

is also a solution of (2.12) for any constant c. The third equation in (2.11) can now be written as

$$\int_0^1 \phi_0^*(\tau)[v_p(\tau) + c\Phi(\tau)q_0]d\tau = s.$$

By the assumption that $\langle \phi_0, \Phi q_0 \rangle \neq 0$ it follows that the third equation is satisfied if we take

$$c = \frac{s - \int_0^1 \phi_0^*(\tau)v_p(\tau)d\tau}{\int_0^1 \phi_0^*(\tau)\Phi(\tau)q_0 \ d\tau} . \qquad \square$$

PROPOSITION 4. *Let* $(x(t), T, \alpha)$ *be a regular periodic solution of* (2.1)*, and let* $\phi_0, \psi_0 \in \mathcal{C}^0([0,1], \mathbf{R}^n)$ *be such that* $\langle \phi_0, \Phi q_0 \rangle \neq 0$*,* $\langle \psi_0, \Psi p_0 \rangle \neq 0$*. Then the operator*

$$\begin{bmatrix} D + Tf_x^*(x(t), \alpha) & \phi_0 \\ \delta_1 - \delta_0 & 0 \\ \mathrm{Int}_{\psi_0} & 0 \end{bmatrix} : \mathcal{C}^1([0,1], \mathbf{R}^n) \times \mathbf{R} \ \rightarrow \ \mathcal{C}^0([0,1], \mathbf{R}^n) \times \mathbf{R}^n \times \mathbf{R}$$

is one-to-one and onto.

Proof. The proof is similar to the proof of Proposition 3. $\square$

3. Test functionals for bifurcations of periodic solutions.

For the fold and Hopf singularities of equilibria, several test functions, and corresponding minimally extended defining systems, are discussed in [12] and incorporated in CONTENT [11]. To obtain similar systems for the case of periodic orbits, we define *simple singularities* of periodic solutions, specifically the limit point, the period-doubling bifurcation, and the torus bifurcation, and we then construct functionals that vanish at these singularities.

3.1. A test functional for the fold bifurcation.

Let $(x(t), T, \alpha)$ define a periodic solution of (1.1); i.e., it satisfies (2.1), (2.2), and (2.3). We say that the solution has a *simple fold singularity* if the monodromy matrix $\Phi(1)$ has an eigenvalue $+1$ with algebraic multiplicity 2 and geometric multiplicity 1, while there are no other critical multipliers.[1]

Let p_0 and q_0 denote the corresponding left and right eigenvectors, which satisfy

$$(\Phi(1) - I)q_0 = 0, \qquad (\Psi(1) - I)p_0 = 0,$$

$$(\Phi(1) - I)^* p_0 = 0, \qquad (\Psi(1) - I)^* q_0 = 0,$$

with

$$p_0^* p_0 = q_0^* q_0 = 1.$$

[1] A geometrically double eigenvalue +1 corresponds to a higher degeneracy. Recall that by definition a *regular* periodic solution has a geometrically simple multiplier +1.

At a simple fold, where the multiplier 1 has algebraic multiplicity 2, we also have generalized eigenvectors p_1 and q_1 satisfying

$$(\Phi(1) - I)q_1 = q_0, \qquad (\Psi(1) - I)p_1 = p_0,$$

where q_1 and p_1 can be chosen so that

$$q_1^* q_0 \; = \; p_1^* p_0 \; = \; 0.$$

Note that in the multiplicity-2 case we also have $p_0^* q_0 = p_1^*(\Psi(1) - I)^* q_0 = 0$.

PROPOSITION 5. *If $(x(t), T, \alpha)$ is a regular periodic solution of* (2.1)*, then the operator*

$$\begin{bmatrix} D - Tf_x(x(t), \alpha) & -f(x(t), \alpha) \\ \delta_1 - \delta_0 & 0 \\ \mathrm{Int}_{f(x(\cdot), \alpha)} & 0 \end{bmatrix} \tag{3.1}$$

from $\mathcal{C}^1([0,1], \mathbf{R}^n) \times \mathbf{R}$ into $\mathcal{C}^0([0,1], \mathbf{R}^n) \times \mathbf{R}^n \times \mathbf{R}$ is one-to-one if the multiplier 1 *has algebraic multiplicity* 1*. If the multiplier* 1 *has algebraic multiplicity* 2*, i.e., at a simple fold, then the operator has a one-dimensional kernel, spanned by the vector*

$$\begin{pmatrix} v \\ 1 \end{pmatrix} \in \mathcal{C}^1([0,1], \mathbf{R}^n) \times \mathbf{R}, \tag{3.2}$$

where $v(t) = \frac{c_0}{T}\Phi(t)(c_2 q_0 - (q_1 - tq_0))$, where c_2 is determined by the condition that

$$q_0^* \int_0^1 \Phi^*(\tau)\Phi(\tau)[c_2 q_0 - (q_1 - \tau q_0)] \; d\tau = 0,$$

and where c_0 is determined by the condition that $x'(0) = c_0 q_0$.

Proof. Consider the homogeneous equations

$$\begin{bmatrix} D - Tf_x(x(t), \alpha) & -f(x(t), \alpha) \\ \delta_1 - \delta_0 & 0 \\ \mathrm{Int}_{f(x(\cdot), \alpha)} & 0 \end{bmatrix} \begin{pmatrix} v \\ S \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}. \tag{3.3}$$

From the first equation in (3.3) we have

$$\begin{aligned} v(t) \;&=\; \Phi(t)\left[v_0 + S\int_0^t \Psi^*(\tau)f(x(\tau), \alpha) \; d\tau\right] \;=\; \Phi(t)\left[v_0 + \tfrac{S}{T}\int_0^t \Psi^*(\tau)x'(\tau) \; d\tau\right] \\ &=\; \Phi(t)\left[v_0 + \tfrac{S}{T}\int_0^t \Psi^*(\tau)\Phi(\tau) \; d\tau \; x'(0)\right] \;=\; \Phi(t)\left[v_0 + \tfrac{St}{T}x'(0)\right], \end{aligned}$$

where we used the facts that $\Psi^*(\tau)\Phi(\tau) = I$ and $x'(t) = \Phi(t)x'(0)$. Above, $v_0 = v(0)$ is an initial vector. By the second equation in (3.3) we have

$$0 = v(1) - v(0) = (\Phi(1) - I)v_0 + \frac{S}{T}x'(0),$$

that is,

$$(\Phi(1) - I)v_0 = -\frac{S}{T}x'(0).$$

Now $(\Phi(1) - I)x'(0) = 0$, so that $x'(0) = c_0 q_0$, for some $c_0 \in \mathbf{R}$, $c_0 \neq 0$. Thus we must solve

$$(\Phi(1) - I)v_0 = -c_0 \frac{S}{T} q_0, \tag{3.4}$$

where q_0 spans the kernel of $\Phi(1) - I$.

If the multiplier 1 has algebraic multiplicity 1, then we must have $S = 0$, $v_0 = c_1 q_0$, and hence $v(t) = c_1 \Phi(t) q_0$. By the third equation in (3.3)

$$0 = \int_0^1 f^*(x(\tau), \alpha) v(\tau)\ d\tau = \frac{1}{T}\int_0^1 x'^*(\tau) v(\tau)\ d\tau = \frac{1}{T}\int_0^1 \left[\Phi(\tau)x'(0)\right]^* c_1 \Phi(\tau) q_0\ d\tau$$

or

$$c_0 c_1\ q_0^* \left(\int_0^1 \Phi^*(\tau)\Phi(\tau)\ d\tau\right)\ q_0 = 0,$$

from which it follows that $c_1 = 0$. Thus $v(t) \equiv 0$. It follows that the operator (3.1) is one-to-one.

At a simple fold the multiplier 1 has algebraic multiplicity 2. In this case (3.4) is also solvable if S is nonzero, namely

$$v_0 = -c_0 \frac{S}{T} q_1 + c_2 q_0,$$

where $c_2 \in \mathbf{R}$ is arbitrary. The third equation in (3.3) then implies

$$\begin{aligned} 0 &= \int_0^1 x'^*(\tau) v(\tau)\ d\tau \\ &= \int_0^1 x'^*(\tau)\Phi(\tau)[v_0 + \tfrac{S\tau}{T} x'(0)]\ d\tau \\ &= \int_0^1 x'^*(\tau)\Phi(\tau)[-c_0 \tfrac{S}{T} q_1 + c_2 q_0 + \tfrac{S\tau}{T} c_0 q_0]\ d\tau \\ &= \int_0^1 [\Phi(\tau)x'(0)]^* \Phi(\tau)[-c_0 \tfrac{S}{T} q_1 + c_2 q_0 + \tfrac{S\tau}{T} c_0 q_0]\ d\tau \\ &= c_0 q_0^* \int_0^1 \Phi^*(\tau)\Phi(\tau)[-c_0 \tfrac{S}{T} q_1 + c_2 q_0 + \tfrac{S\tau}{T} c_0 q_0]\ d\tau, \end{aligned}$$

from which it follows that

$$c_2 = \frac{c_0 S q_0^* \int_0^1 \Phi^*(\tau)\Phi(\tau)[q_1 - \tau q_0]\ d\tau}{T\ q_0^* \int_0^1 \Phi^*(\tau)\Phi(\tau)\ d\tau\ q_0}. \qquad \square$$

PROPOSITION 6. *Let* $(x(t), T, \alpha)$ *be a regular periodic solution of* (2.1) *and consider the operator*

$$M_1 = \begin{bmatrix} D - T f_x(x(t), \alpha) & -f(x(t), \alpha) \\ \delta_1 - \delta_0 & 0 \\ \mathrm{Int}_{f(x(\cdot), \alpha)} & 0 \end{bmatrix} \tag{3.5}$$

from $\mathcal{C}^1([0,1], \mathbf{R}^n) \times \mathbf{R}$ *into* $\mathcal{C}^0([0,1], \mathbf{R}^n) \times \mathbf{R}^n \times \mathbf{R}$. *If the multiplier* 1 *has algebraic multiplicity* 1, *then* M_1 *is onto. If it has algebraic multiplicity* 2, *i.e., at a simple fold,* *then the range of M_1 has codimension* 1 *and the vector*

$$\left(\begin{array}{c} \Psi p_0 \\ -p_0 \\ 0 \end{array} \right) \in \mathcal{C}^0([0,1], \mathbf{R}^n) \times \mathbf{R}^n \times \mathbf{R} \tag{3.6}$$

is complementary to the range space.

Proof. Consider a vector $(\xi, \eta, \omega)^*$ in $\mathcal{C}^0([0,1], \mathbf{R}^n) \times \mathbf{R}^n \times \mathbf{R}$. This vector is in the range of M_1 if and only if there exist $(v, S)^*$ in $\mathcal{C}^1([0,1], \mathbf{R}^n) \times \mathbf{R}$ such that

$$M_1 \left(\begin{array}{c} v \\ S \end{array} \right) = \left(\begin{array}{c} \xi \\ \eta \\ \omega \end{array} \right). \tag{3.7}$$

The first equation in (3.7) implies that

$$v(t) = \Phi(t) \left[v(0) + \int_0^t \Psi^*(\tau)(\xi(\tau) + Sf(x(\tau), \alpha))\ d\tau \right].$$

The second equation in (3.7) then implies

$$\eta = v(1) - v(0) = (\Phi(1) - I)v(0) + \Phi(1) \int_0^1 \Psi^*(\tau)(\xi(\tau) + Sf(x(\tau), \alpha))\ d\tau.$$

Now

$$\int_0^1 \Psi^*(\tau) f(x(\tau), \alpha) d\tau = \frac{1}{T} \int_0^1 \Psi^*(\tau) x'(\tau)\ d\tau = \frac{1}{T} \int_0^1 \Psi^*(\tau) c_0 \Phi(\tau) q_0\ d\tau = \frac{c_0}{T} q_0.$$

So

$$\eta = (\Phi(1) - I)v(0) + \frac{Sc_0}{T} q_0 + \Phi(1) \int_0^1 \Psi^*(\tau)\xi(\tau)\ d\tau. \tag{3.8}$$

If 1 is an algebraically simple eigenvalue of $\Phi(1)$, then q_0 is not in the range of $(\Phi(1) - I)$. For given ξ and η, (3.8) can be solved for $v(0)$ and S. Moreover, the solution is unique up to the addition of a scalar multiple of q_0 to $v(0)$. Since

$$\int_0^1 (x'(\tau))^* \Phi(\tau) q_0\ d\tau = c_0 \int_0^1 (\Phi(\tau) q_0)^* \Phi(\tau) q_0\ d\tau \neq 0,$$

the scalar is determined uniquely by the third equation in (3.7).

If 1 is an algebraically double eigenvalue of $\Phi(1)$, i.e., at a fold point, then (3.8) is solvable if and only if

$$p_0^* \eta = p_0^* \int_0^1 \Psi^*(\tau) \xi(\tau)\ d\tau.$$

If so, the third equation in (3.7) again determines the solution uniquely. □

PROPOSITION 7. *If $(x(t), T, \alpha)$ is a regular periodic solution of* (2.1)*, then the operator*

$$\left[\begin{array}{cc} D + Tf_x^*(x(t), \alpha) & -f(x(t), \alpha) \\ \delta_1 - \delta_0 & 0 \\ \mathrm{Int}_{f(x(\cdot), \alpha)} & 0 \end{array} \right] \tag{3.9}$$

from $\mathcal{C}^1([0,1],\mathbf{R}^n)\times\mathbf{R}\to\mathcal{C}^0([0,1],\mathbf{R}^n)\times\mathbf{R}^n\times\mathbf{R}$ *is one-to-one if the multiplier* 1 *has algebraic multiplicity* 1*. If the multiplier* 1 *has algebraic multiplicity* 2*, i.e., at a simple fold, then the operator has a one-dimensional kernel, spanned by* $(\Psi^* p_0, 0)^* \in \mathcal{C}^1([0,1],\mathbf{R}^n)\times\mathbf{R}$.

Proof. Consider the homogeneous equations

$$\begin{bmatrix} D+Tf_x^*(x(t),\alpha) & -f(x(t),\alpha) \\ \delta_1-\delta_0 & 0 \\ \mathrm{Int}_{f(x(\cdot),\alpha)} & 0 \end{bmatrix}\begin{pmatrix} w \\ R\end{pmatrix} = \begin{pmatrix} 0\\0\\0\end{pmatrix}. \tag{3.10}$$

From the first equation in (3.10) we have

$$w(t)=\Psi(t)\left[w_0+\frac{R}{T}\int_0^t \Phi^*(\tau)x'(\tau)\,d\tau\right],$$

where $w_0 = w(0)$ is an initial vector. The second equation in (3.10) implies

$$0 = w(1)-w(0) = (\Psi(1)-I)w_0 + \frac{R}{T}\Psi(1)\int_0^1 \Phi^*(\tau)x'(\tau)\,d\tau$$

or

$$(\Psi(1)-I)w_0 = -\frac{R}{T}\Psi(1)\int_0^1 \Phi^*(\tau)\Phi(\tau)\,d\tau\; x'(0).$$

Given R, this equation is solvable for w_0 if

$$-Rq_0^*\Psi(1)\int_0^1 \Phi^*(\tau)\Phi(\tau)\,d\tau\; x'(0)=0,$$

that is, recalling that $x'(0)=c_0q_0$, $c_0\neq 0$, and $q_0^*\Psi(1)=q_0^*$ if

$$c_0Rq_0^*\int_0^1 \Phi^*(\tau)\Phi(\tau)\,d\tau\; q_0 = 0.$$

It follows that $R=0$, independently of the algebraic multiplicity of the eigenvalue 1. Thus $w(t)=\Psi(t)w_0$, where $(\Psi(1)-I)w_0=0$, so that $w_0=c_3p_0$ for some $c_3\in\mathbf{R}$.

From the third equation in (3.10) it follows that

$$\begin{aligned} 0 = \int_0^1 w^*(\tau)x'(\tau)\,d\tau &= \int_0^1 [c_3\Psi(\tau)p_0]^*\Phi(\tau)x'(0)\,d\tau \\ &= c_0\;c_3\;p_0^*\int_0^1 \Psi^*(\tau)\Phi(\tau)\,d\tau\; q_0 = c_0\;c_3\;p_0^*q_0. \end{aligned}$$

If the multiplier 1 has algebraic multiplicity 1, then $p_0^*q_0\neq 0$. In this case $c_3=0$ and hence $w(t)\equiv 0$; that is, the operator (3.9) is one-to-one.

If the multiplier 1 has algebraic multiplicity 2, then $p_0^*q_0=0$, and we can choose $c_3\neq 0$. In this case $w_0\neq 0$; hence $w(t)\not\equiv 0$. It follows that the operator (3.9) has a one-dimensional kernel. □

PROPOSITION 8. *If* $(x(t),T,\alpha)$ *is a regular periodic solution of* (2.1)*, then the operator*

$$M_2 = \begin{bmatrix} D+Tf_x^*(x(t),\alpha) & -f(x(t),\alpha) \\ \delta_1-\delta_0 & 0 \\ \mathrm{Int}_{f(x(\cdot),\alpha)} & 0 \end{bmatrix} \tag{3.11}$$

from $\mathcal{C}^1([0,1],\mathbf{R}^n)\times\mathbf{R}\to\mathcal{C}^0([0,1],\mathbf{R}^n)\times\mathbf{R}^n\times\mathbf{R}$ *is onto if the multiplier* 1 *has algebraic multiplicity* 1*. If the multiplier* 1 *has algebraic multiplicity* 2*, i.e., at a simple fold, then the range has codimension* 1*, and the vector* $(0,0,1)^*\in\mathcal{C}^0([0,1],\mathbf{R}^n)\times\mathbf{R}^n\times\mathbf{R}$ *is complementary to the range space.*

Proof. Consider a vector $(\xi,\eta,\omega)^*$ in $\mathcal{C}^0([0,1],\mathbf{R}^n)\times\mathbf{R}^n\times\mathbf{R}$. This vector is in the range of M_2 if and only if there exist $(w,R)^*$ in $\mathcal{C}^1([0,1],\mathbf{R}^n)\times\mathbf{R}$ such that

$$M_2\begin{pmatrix} w\\ R\end{pmatrix}=\begin{pmatrix}\xi\\ \eta\\ \omega\end{pmatrix}. \tag{3.12}$$

The first equation in (3.12) implies that

$$w(t)=\Psi(t)\left[w(0)+\int_0^t\Phi^*(\tau)(\xi(\tau)+Rc_0\Phi(\tau)q_0)\;d\tau\right].$$

The second equation in (3.12) then implies

$$\eta=w(1)-w(0)=(\Psi(1)-I)w(0)+\Psi(1)\int_0^1\Phi^*(\tau)(\xi(\tau)+Rc_0\Phi(\tau)q_0)d\tau.$$

We thus obtain the equation

$$(\Psi(1)-I)w(0)=\eta-Rc_0\Psi(1)\int_0^1\Phi^*(\tau)\Phi(\tau)q_0\;d\tau-\Psi(1)\int_0^1\Phi^*(\tau)\xi(\tau)\;d\tau.$$

This equation is solvable for $w(0)$ if and only if

$$q_0^*\eta=Rc_0q_0^*\int_0^1\Phi^*(\tau)\Phi(\tau)q_0\;d\tau+q_0^*\int_0^1\Phi^*(\tau)\xi(\tau)d\tau.$$

The latter equation is solvable uniquely for R, so the previous one is solvable for $w(0)$ and defines it up to the addition of a scalar multiple of p_0.

Now suppose that $(w,R)^*$ solve the first two equations in (3.12), where $w(0)=w_0+rp_0$ and r is arbitrary. The third equation in (3.12) then requires

$$c_0q_0^*(w_0+rp_0)=\omega+\text{two integral terms which are linear in }\xi(t)\text{ and }R.$$

If the eigenvalue 1 of $\Phi(1)$ has algebraic multiplicity 1, then this equation has a unique solution in r and thus M_2 is one-to-one and onto. If the eigenvalue has algebraic multiplicity 2, then the range of M_2 has codimension at most 1. If we set $\xi(t)\equiv0$, $\eta=0$, $\omega=1$, then necessarily $R=0$, $\omega=0$ as well, and thus the third equation in (3.12) cannot be solved. So the range of M_2 has codimension 1, and $(0,0,1)^*$ is a vector complementary to the range. □

PROPOSITION 9. *Let* $(x(t),T,\alpha)$ *be a regular periodic solution of* (2.1) *that has a simple fold singularity; i.e.,* $\Phi(1)$ *has eigenvalue* 1 *with algebraic multiplicity* 2*. Then there exist* $v_{01},w_{01},\ v_{11},w_{11}\in\mathcal{C}^0([0,1],\mathbf{R}^n)$*,* $w_{02},v_{12}\in\mathbf{R}^n$*,* $w_{03},v_{02},v_{13},w_{12}\in\mathbf{R}$ *such that*

$$N_1=\begin{bmatrix} D-Tf_x(x(t),\alpha) & -f(x(t),\alpha) & w_{01}\\ \delta_1-\delta_0 & 0 & w_{02}\\ \mathrm{Int}_{f(x(\cdot),\alpha)} & 0 & w_{03}\\ \mathrm{Int}_{v_{01}} & v_{02} & 0\end{bmatrix}$$

and

$$N_2 = \begin{bmatrix} D + Tf_x^*(x(t),\alpha) & -f(x(\cdot),\alpha) & v_{11} \\ \delta_1 - \delta_0 & 0 & v_{12} \\ \mathrm{Int}_{f(x(\cdot),\alpha)} & 0 & v_{13} \\ \mathrm{Int}_{w_{11}} & w_{12} & 0 \end{bmatrix}$$

from $\mathcal{C}^1([0,1],\mathbf{R}^n)\times\mathbf{R}^n\times\mathbf{R}$ *to* $\mathcal{C}^0([0,1],\mathbf{R}^n)\times\mathbf{R}^n\times\mathbf{R}\times\mathbf{R}$ *are one-to-one and onto.*

For any such choice of the bordering elements we define $v, w \in \mathcal{C}^1([0,1],\mathbf{R}^n)$ *and* $S, G, H, R \in \mathbf{R}$ *by the equations*

$$N_1 \begin{pmatrix} v \\ S \\ G \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix} \tag{3.13}$$

and

$$N_2 \begin{pmatrix} w \\ R \\ H \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ -1 \end{pmatrix}. \tag{3.14}$$

Then in a neighborhood of $(x(t),T,\alpha)$, $G = 0$ *if and only if* $H = 0$. *Moreover, this happens if and only if the regular periodic solution has a simple fold singularity.*

Proof. We choose

$$\begin{pmatrix} v_{01}(t) \\ v_{02} \end{pmatrix} = \begin{pmatrix} v(t) \\ 1 \end{pmatrix},$$

where v is given in the statement of Proposition 5. Further we set

$$\begin{pmatrix} w_{01}(t) \\ w_{02} \\ w_{03} \end{pmatrix} = \begin{pmatrix} \Psi^*(t)p_0 \\ 0 \\ 0 \end{pmatrix}.$$

By Propositions 5 and 6, N_1 is one-to-one and onto. We further set

$$\begin{pmatrix} w_{11}(t) \\ w_{12} \end{pmatrix} = \begin{pmatrix} \Psi^*(t)p_0 \\ 0 \end{pmatrix},$$

$$\begin{pmatrix} v_{11}(t) \\ v_{12} \\ v_{13} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}.$$

By Propositions 7 and 8, N_2 is one-to-one and onto. The last statement in the proposition is proved by standard arguments. □

3.2. A test functional for the period-doubling bifurcation. By definition, at a *simple flip singularity* there is an algebraically simple Floquet multiplier equal to -1 and no other multipliers with unit modulus, except for an algebraically simple multiplier $+1$. The left and right eigenvectors of the monodromy matrix $\Phi(1)$ for the eigenvalue -1 will be denoted by p_2 and q_2, respectively. They are also the right and left eigenvector, respectively, of $\Psi(1)$ for the eigenvalue -1.

PROPOSITION 10. *If $(x(t), T, \alpha)$ corresponds to a simple flip singularity, then the operator*

$$\begin{bmatrix} D - Tf_x(x(t), \alpha) \\ \delta_0 + \delta_1 \end{bmatrix} : \mathcal{C}^1([0,1], \mathbf{R}^n) \to \mathcal{C}^0([0,1], \mathbf{R}^n) \times \mathbf{R}^n$$

has a one-dimensional kernel spanned by Φq_2. Its range has codimension 1; if $\zeta \in \mathcal{C}^0([0,1], \mathbf{R}^n)$, $r \in \mathbf{R}^n$, then $(\zeta, r)^*$ is in the range if and only if $\langle \Psi p_2, \zeta \rangle = -p_2^* r$. In particular, if $r = 0$, then $(\zeta, 0)^*$ is in the range if and only if $\langle \Psi p_2, \zeta \rangle = 0$.

Proof. The proof is similar to the proof of Proposition 1. □

COROLLARY 3. *If $(x(t), T, \alpha)$ corresponds to a simple flip singularity, then the operator*

$$\begin{bmatrix} D - Tf_x(x(t), \alpha) \\ \delta_0 + \delta_1 \\ \mathrm{Int}_\phi \end{bmatrix} : \mathcal{C}^1([0,1], \mathbf{R}^n) \ \to \ \mathcal{C}^0([0,1], \mathbf{R}^n) \times \mathbf{R}^n \times \mathbf{R} \tag{3.15}$$

is one-to-one if and only if $\langle \phi, \Phi q_2 \rangle \neq 0$.

PROPOSITION 11. *If $(x(t), T, \alpha)$ corresponds to a simple flip singularity, then the operator*

$$\begin{bmatrix} D + Tf_x^*(x(t), \alpha) \\ \delta_0 + \delta_1 \end{bmatrix} : \mathcal{C}^1([0,1], \mathbf{R}^n) \ \to \ \mathcal{C}^0([0,1], \mathbf{R}^n) \times \mathbf{R}^n$$

has a one-dimensional kernel spanned by Ψp_2. Its range has codimension 1; if $\zeta \in \mathcal{C}^0([0,1], \mathbf{R}^n)$, $r \in \mathbf{R}^n$, then $(\zeta, r)^*$ is in the range if and only if $\langle \Phi q_2, \zeta \rangle = -q_2^* r$. In particular, if $r = 0$, then $(\zeta, 0)^*$ is in the range if and only if $\langle \Phi q_2, \zeta \rangle = 0$.

Proof. The proof is similar to the proof of Proposition 2. □

COROLLARY 4. *If $(x(t), T, \alpha)$ corresponds to a simple flip singularity, then the operator*

$$\begin{bmatrix} D + Tf_x^*(x(t), \alpha) \\ \delta_0 + \delta_1 \\ \mathrm{Int}_\psi \end{bmatrix} : \mathcal{C}^1([0,1], \mathbf{R}^n) \ \to \ \mathcal{C}^0([0,1], \mathbf{R}^n) \times \mathbf{R}^n \times \mathbf{R} \tag{3.16}$$

is one-to-one if and only if $\langle \psi, \Psi p_2 \rangle \neq 0$.

PROPOSITION 12. *Let $(x(t), T, \alpha)$ correspond to a simple flip singularity, and let $\phi_0, \psi_0 \in \mathcal{C}^0([0,1], \mathbf{R}^n)$ be such that $\langle \phi_0, \Phi q_2 \rangle \neq 0$, $\langle \psi_0, \Psi p_2 \rangle \neq 0$. Then the operator*

$$\begin{bmatrix} D - Tf_x(x(t), \alpha) & \psi_0 \\ \delta_0 + \delta_1 & 0 \\ \mathrm{Int}_{\phi_0} & 0 \end{bmatrix} : \mathcal{C}^1([0,1], \mathbf{R}^n) \times \mathbf{R} \ \to \ \mathcal{C}^0([0,1], \mathbf{R}^n) \times \mathbf{R}^n \times \mathbf{R}$$

is one-to-one and onto.

Proof. The proof is similar to the proof of Proposition 3. □

PROPOSITION 13. *Let $(x(t), T, \alpha)$ correspond to a simple flip singularity, and let $\phi_0, \psi_0 \in \mathcal{C}^0([0,1], \mathbf{R}^n)$ be such that $\langle \phi_0, \Phi q_2 \rangle \neq 0$, $\langle \psi_0, \Psi p_2 \rangle \neq 0$. Then the operator*

$$\begin{bmatrix} D + Tf_x^*(x(t), \alpha) & \phi_0 \\ \delta_0 + \delta_1 & 0 \\ \mathrm{Int}_{\psi_0} & 0 \end{bmatrix} : \mathcal{C}^1([0,1], \mathbf{R}^n) \times \mathbf{R} \ \to \ \mathcal{C}^0([0,1], \mathbf{R}^n) \times \mathbf{R}^n \times \mathbf{R}$$

is one-to-one and onto.

Proof. The proof is similar to the proof of Proposition 4. □

PROPOSITION 14. *Let $(x(t), T, \alpha)$ be a periodic solution close to a simple flip singularity, and let $\phi_0, \psi_0 \in \mathcal{C}^0([0,1], \mathbf{R}^n)$ be such that $\langle \phi_0, \Phi q_2 \rangle \neq 0$, $\langle \psi_0, \Psi p_2 \rangle \neq 0$, so that the operators M_3 and M_4 (defined below) from $\mathcal{C}^1([0,1], \mathbf{R}^n) \times \mathbf{R}$ into $\mathcal{C}^0([0,1], \mathbf{R}^n) \times \mathbf{R}^n \times \mathbf{R}$ are both one-to-one and onto. Let $v, w \in \mathcal{C}^1([0,1], \mathbf{R}^n), G$, $H \in \mathbf{R}$ be defined by the equations*

$$M_3 \begin{pmatrix} v \\ G \end{pmatrix} \equiv \begin{bmatrix} D - Tf_x(x(t), \alpha) & \psi_0 \\ \delta_0 + \delta_1 & 0 \\ \mathrm{Int}_{\phi_0} & 0 \end{bmatrix} \begin{pmatrix} v \\ G \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, \tag{3.17}$$

$$M_4 \begin{pmatrix} w \\ H \end{pmatrix} \equiv \begin{bmatrix} D + Tf_x^*(x(t), \alpha) & \phi_0 \\ \delta_0 + \delta_1 & 0 \\ \mathrm{Int}_{\psi_0} & 0 \end{bmatrix} \begin{pmatrix} w \\ H \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ -1 \end{pmatrix}. \tag{3.18}$$

Then $G = H$. Furthermore, $G = 0$ if and only if the periodic solution corresponds to a simple flip singularity. If so, then $v(0)$ is the right eigenvector of the monodromy matrix for the eigenvalue -1.

Proof. Multiplying the first equation in (3.17) on the left with $w^*(t)$, integrating over the interval $[0,1]$, and using the last equation in (3.18) we obtain

$$\int_0^1 w^* v'(\tau)\, d\tau - T \int_0^1 w^*(\tau) f_x(x(\tau), \alpha) v(\tau)\, d\tau - G = 0.$$

Integrating the first term by parts, using the second equations in (3.17) and (3.18), we obtain

$$-\int_0^1 v^*(\tau) w'(\tau)\, d\tau - T \int_0^1 v^*(\tau) f_x^*(x(\tau), \alpha) w(\tau)\, d\tau - G = 0.$$

Using the first equation in (3.18) we get

$$-\langle v, (-H\phi_0) \rangle - G = 0.$$

Using the third equation in (3.17) we obtain $G = H$. The other statements in the proposition are now obvious. □

3.3. A test functional for the torus bifurcation. We say that a periodic solution has a *simple Neimark–Sacker singularity* if the monodromy matrix $\Phi(1)$ has a conjugate pair of simple complex multipliers with modulus 1 (i.e., $e^{\pm i\theta}$, $0 < \theta < \pi$) and no other multipliers with unit modulus, except an algebraically simple eigenvalue $+1$. Furthermore, let $p_1, p_2 \in \mathbf{R}^n$ (respectively, $q_1, q_2 \in \mathbf{R}^n$) be such that $p_1 + ip_2$ (respectively, $q_1 + iq_2$) is a left (respectively, right) complex eigenvector of the monodromy matrix $\Phi(1)$. Thus

$$\begin{gathered} (p_1 + ip_2)^H \Phi(1) = e^{i\theta}(p_1 + ip_2)^H, \\ \Phi(1)(q_1 + iq_2) = e^{i\theta}(q_1 + iq_2), \\ \Psi(1)(p_1 + ip_2) = e^{i\theta}(p_1 + ip_2), \\ (q_1 + iq_2)^H \Psi(1) = e^{i\theta}(q_1 + iq_2)^H, \end{gathered}$$

where $(p_1 + ip_2)^H = p_1^* - ip_2^*$, $(q_1 + iq_2)^H = q_1^* - iq_2^*$.

In this section it is convenient to extend the definition of $x(t), \Phi(t)$, and $\Psi(t)$ to the interval $[0,2]$ by periodicity with period 1 and to redefine

$$\mathrm{Int}_{\phi}(v) = \langle \phi, v \rangle = \int_0^2 \phi^*(\tau) v(\tau)\ d\tau.$$

We start with the following result.

PROPOSITION 15. *Let $(x(t), T, \alpha)$ define a periodic solution; i.e., it satisfies* (2.1), (2.2), *and* (2.3). *Let $(x(t), T, \alpha)$ correspond to a simple Neimark–Sacker singularity with multipliers $e^{\pm i\theta}$, $0 < \theta < \pi$. Let $\kappa = \cos\theta$ and consider the operator*

$$\left[\begin{array}{c} D - T f_x(x(t), \alpha) \\ \delta_0 - 2\kappa\delta_1 + \delta_2 \end{array} \right] \ :\ \mathcal{C}^1([0,2], \mathbf{R}^n) \ \rightarrow\ \mathcal{C}^0([0,2], \mathbf{R}^n) \times \mathbf{R}^n. \tag{3.19}$$

Then we have the following:

(i) *The operator* (3.19) *has a two-dimensional kernel spanned by $\Phi(t)q_1$ and $\Phi(t)q_2$.*

(ii) *The operator* (3.19) *has a range with* codim 2. *The vectors*

$$\left(\begin{array}{c} \Psi p_1 \\ 0 \end{array} \right), \left(\begin{array}{c} \Psi p_2 \\ 0 \end{array} \right) \in \mathcal{C}^0([0,2], \mathbf{R}^n) \times \mathbf{R}^n$$

span a two-dimensional subspace that is complementary to the range of (3.19).

Proof. Let v be in the kernel of (3.19). Then v must have the form $v(t) = \Phi(t)v_0$ with $v_0 \in \mathbf{R}^n$. We further have

$$0 = (\delta_0 - 2\kappa\delta_1 + \delta_2)v = v(0) - 2\kappa v(1) + v(2) = (\Phi(1) - e^{i\theta}I)(\Phi(1) - e^{-i\theta}I)v_0.$$

We infer that it is necessary and sufficient that v_0 is in the span of q_1, q_2.

As a first step in the proof of (ii) we consider $\zeta \in \mathcal{C}^0([0,2], \mathbf{R}^n)$, $r \in \mathbf{R}^n$, and we give a necessary and sufficient condition in order that $(\zeta, r)^*$ be in the range of (3.19). First, there must exist a $v \in \mathcal{C}^1([0,2], \mathbf{R}^n)$ for which

$$v'(t) - T f_x(x(t), \alpha) v(t) = \zeta(t).$$

The general solution of this linear differential equation is

$$v(t) = \Phi(t) \left[v_0 + \int_0^t \Psi^*(\tau)\zeta(\tau)\ d\tau \right],$$

where $v_0 = v(0)$ is an initial vector. Also, we must have $v(0) - 2\kappa v(1) + v(2) = r$, that is,

$$(\Phi(1) - e^{i\theta}I)(\Phi(1) - e^{-i\theta}I)v_0 - 2\kappa\Phi(1) \int_0^1 \Psi^*(\tau)\zeta(\tau)\ d\tau + \Phi(1)^2 \int_0^2 \Psi^*(\tau)\zeta(\tau) d\tau = r.$$

This is an equation for v_0 which is solvable if and only if

$$-2\kappa p^H \Phi(1) \int_0^1 \Psi^*(\tau)\zeta(\tau)\ d\tau + p^H \Phi(1)^2 \int_0^2 \Psi^*(\tau)\zeta(\tau)\ d\tau = p^H r$$

or, equivalently,

$$-2\kappa e^{i\theta} \int_0^1 p^H \Psi^*(\tau)\zeta(\tau)\ d\tau + e^{2i\theta} \int_0^2 p^H \Psi^*(\tau)\zeta(\tau)\ d\tau = p^H r.$$

If we define the linear functional L by setting

$$L(\zeta) = -2\kappa e^{i\theta} \int_0^1 p^H \Psi^*(\tau)\zeta(\tau)\, d\tau + e^{2i\theta} \int_0^2 p^H \Psi^*(\tau)\zeta(\tau)\, d\tau, \tag{3.20}$$

then we infer that $(\zeta, r)^*$ is in the range of (3.19) if and only if $L(\zeta) = p^H r$.

As a second step in the proof of (ii) we compute $L(\Psi p_1)$ and $L(\Psi p_2)$. We have

$$L(\Psi p_1) = -2\cos\theta e^{i\theta} \int_0^1 p^H \Psi^*(\tau)\Psi(\tau)p_1\, d\tau + e^{2i\theta} \int_0^2 p^H \Psi^*(\tau)\Psi(\tau)p_1\, d\tau$$

$$= e^{i\theta}(-2\cos\theta+\cos\theta+i\sin\theta) \int_0^1 p^H \Psi^*(\tau)\Psi(\tau)p_1\, d\tau + e^{2i\theta} \int_0^1 p^H \Psi^*(1+\tau)\Psi(1+\tau)p_1\, d\tau.$$

Now we note that

$$\Psi(1+\tau)p_1 = \Psi(\tau)\Psi(1)p_1 = \Psi(\tau)(\cos\theta p_1 - \sin\theta p_2)$$

and

$$p^H \Psi^*(1+\tau) = [\Psi(\tau)\Psi(1)p]^H = [e^{i\theta}\Psi(\tau)p]^H = e^{-i\theta} p^H \Psi^*(\tau).$$

Hence

$$L(\Psi p_1) = i\sin\theta e^{i\theta} \int_0^1 p^H \Psi^*(\tau)\Psi(\tau)p\, d\tau = (-\sin\theta + i\cos\theta)\sin\theta \int_0^1 \|\Psi(\tau)p\|^2\, d\tau.$$

By a similar argument we find that

$$L(\Psi p_2) = (\cos\theta + i\sin\theta)\sin\theta \int_0^1 \|\Psi(\tau)p\|^2 d\tau.$$

As a third step in the proof of (ii) we show that the range of (3.19) has codimension 2 by proving that every $(\xi, r)^*$ can be written in a unique way as

$$\begin{pmatrix} \xi \\ r \end{pmatrix} = \begin{pmatrix} \xi_0 \\ r_0 \end{pmatrix} + \alpha \begin{pmatrix} 0 \\ p_1 \end{pmatrix} + \beta \begin{pmatrix} 0 \\ p_2 \end{pmatrix}, \tag{3.21}$$

with $(\xi_0, r_0)^*$ in the range of (3.19) and $\alpha, \beta \in \mathbf{R}$.

Obviously $\xi_0 = \xi$, and r_0 has to satisfy the conditions

$$p^H r_0 = L(\xi), \quad r_0 = r - \alpha p_1 - \beta p_2.$$

These conditions imply

$$\begin{pmatrix} p_1^* p_1 & p_1^* p_2 \\ p_2^* p_1 & p_2^* p_2 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} p_1^* r - \text{Re}\,[L(\xi)] \\ p_2^* r + \text{Im}\,[L(\xi)] \end{pmatrix}.$$

This nonsingular linear system defines α, β in a unique way. Next, r_0 is defined by the requirement $r_0 = r - \alpha p_1 - \beta p_2$, and with this choice we have $p^H r_0 = L(\xi)$.

As the fourth and last step to prove (ii) we will show that

$$\begin{pmatrix} \Psi p_1 \\ 0 \end{pmatrix}, \begin{pmatrix} \Psi p_2 \\ 0 \end{pmatrix},$$

and we will also span a two-dimensional space complementary to the range of (3.19). To this end we decompose

$$\begin{pmatrix} \Psi p_1 \\ 0 \end{pmatrix} = \begin{pmatrix} \Psi p_1 \\ r_1 \end{pmatrix} + \alpha_1 \begin{pmatrix} 0 \\ p_1 \end{pmatrix} + \beta_1 \begin{pmatrix} 0 \\ p_2 \end{pmatrix},$$

$$\begin{pmatrix} \Psi p_2 \\ 0 \end{pmatrix} = \begin{pmatrix} \Psi p_2 \\ r_2 \end{pmatrix} + \alpha_2 \begin{pmatrix} 0 \\ p_1 \end{pmatrix} + \beta_2 \begin{pmatrix} 0 \\ p_2 \end{pmatrix}$$

in the decomposition of (3.21). Then $\alpha_1, \beta_1, \alpha_2, \beta_2$ are defined by the matrix equation

$$\begin{pmatrix} p_1^* p_1 & p_1^* p_2 \\ p_2^* p_1 & p_2^* p_2 \end{pmatrix} \begin{pmatrix} \alpha_1 & \alpha_2 \\ \beta_1 & \beta_2 \end{pmatrix} = \begin{pmatrix} -\mathrm{Re}\,[L(\Psi p_1)] & -\mathrm{Re}\,[L(\Psi p_2)] \\ \mathrm{Im}\,[L(\Psi p_1)] & \mathrm{Im}\,[L(\Psi p_2)] \end{pmatrix}.$$

The proof of (ii) is complete if we show that

$$\begin{pmatrix} \alpha_1 & \alpha_2 \\ \beta_1 & \beta_2 \end{pmatrix}$$

is a nonsingular matrix or, equivalently, that

$$\begin{pmatrix} -\mathrm{Re}\,[L(\Psi p_1)] & -\mathrm{Re}\,[L(\Psi p_2)] \\ \mathrm{Im}\,[L(\Psi p_1)] & \mathrm{Im}\,[L(\Psi p_2)] \end{pmatrix}$$

is nonsingular. By the second step this matrix is equal to

$$\begin{pmatrix} \sin\theta & -\cos\theta \\ \cos\theta & \sin\theta \end{pmatrix} \sin\theta \int_0^1 \|\Psi(\tau)p\|^2 d\tau. \tag{3.22}$$

Since $\sin\theta \neq 0$ in (3.22) the proof is complete. □

PROPOSITION 16. *Let $(x(t), T, \alpha)$ define a periodic solution; that is, it satisfies* (2.1), (2.2), *and* (2.3). *Let $(x(t), T, \alpha)$ correspond to a simple Neimark–Sacker singularity with multipliers $e^{\pm i\theta}$, $0 < \theta < \pi$. Set $\kappa = \cos\theta$ and consider the operator*

$$\begin{bmatrix} D + T f_x^*(x(t), \alpha) \\ \delta_0 - 2\kappa\delta_1 + \delta_2 \end{bmatrix} : \mathcal{C}^1([0,2], \mathbf{R}^n) \to \mathcal{C}^0([0,2], \mathbf{R}^n) \times \mathbf{R}^n. \tag{3.23}$$

Then we have the following:

(i) *The operator* (3.23) *has a two-dimensional kernel spanned by $\Psi(t)p_1$ and $\Psi(t)p_2$.*

(ii) *The operator* (3.23) *has a range of codimension* 2. *The vectors*

$$\begin{pmatrix} \Phi q_1 \\ 0 \end{pmatrix}, \begin{pmatrix} \Phi q_2 \\ 0 \end{pmatrix} \in \mathcal{C}^0([0,2], \mathbf{R}^n) \times \mathbf{R}^n$$

span a two-dimensional subspace that is complementary to the range of (3.23).

Proof. The proof is similar to the proof of the preceding proposition. □

COROLLARY 5. *Let $(x(t), T, \alpha)$ correspond to a simple Neimark–Sacker singularity of a periodic solution. If $\kappa = \cos\theta$, then the operators*

$$\begin{bmatrix} D - T f_x(x(t), \alpha) & \Psi p_1 & \Psi p_2 \\ \delta_0 - 2\kappa\delta_1 + \delta_2 & 0 & 0 \\ \mathrm{Int}_{\Phi(\cdot)q_1} & 0 & 0 \\ \mathrm{Int}_{\Phi(\cdot)q_2} & 0 & 0 \end{bmatrix} : \mathcal{C}^1([0,2], \mathbf{R}^n) \times \mathbf{R}^2 \to \mathcal{C}^0([0,2], \mathbf{R}^n) \times \mathbf{R}^n \times \mathbf{R}^2$$

and

$$\begin{bmatrix} D+Tf_x^*(x(t),\alpha) & \Phi q_1 & \Phi q_2 \\ \delta_0-2\kappa\delta_1+\delta_2 & 0 & 0 \\ \mathrm{Int}_{\Psi(\cdot)p_1} & 0 & 0 \\ \mathrm{Int}_{\Psi(\cdot)p_2} & 0 & 0 \end{bmatrix} : \mathcal{C}^1([0,2],\mathbf{R}^n)\times\mathbf{R}^2 \rightarrow \mathcal{C}^0([0,2],\mathbf{R}^n)\times\mathbf{R}^n\times\mathbf{R}^2$$

are both one-to-one and onto.

Proof. The proof is standard. □

Proposition 17. *Let $(x(t),T,\alpha)$ be close to a simple Neimark–Sacker singularity of periodic solutions and κ close to the value $\cos\theta$ at the singular point. Furthermore, let (ψ_0,ψ_1) span a space sufficiently close to the span of $(\Psi p_1,\Psi p_2)$, and let (ϕ_0,ϕ_1) span a space sufficiently close to $(\Phi q_1,\Phi q_2)$, so that the operators*

$$M_5 = \begin{bmatrix} D-Tf_x(x(t),\alpha) & \psi_0 & \psi_1 \\ \delta_0-2\kappa\delta_1+\delta_2 & 0 & 0 \\ \mathrm{Int}_{\phi_0} & 0 & 0 \\ \mathrm{Int}_{\phi_1} & 0 & 0 \end{bmatrix} : \mathcal{C}^1([0,2],\mathbf{R}^n)\times\mathbf{R}^2 \rightarrow \mathcal{C}^0([0,2],\mathbf{R}^n)\times\mathbf{R}^n\times\mathbf{R}^2$$

and

$$M_6 = \begin{bmatrix} D+Tf_x^*(x(t),\alpha) & \phi_0 & \phi_1 \\ \delta_0-2\kappa\delta_1+\delta_2 & 0 & 0 \\ \mathrm{Int}_{\psi_0} & 0 & 0 \\ \mathrm{Int}_{\psi_1} & 0 & 0 \end{bmatrix} : \mathcal{C}^1([0,2],\mathbf{R}^n)\times\mathbf{R}^2 \rightarrow \mathcal{C}^0([0,2],\mathbf{R}^n)\times\mathbf{R}^n\times\mathbf{R}^2$$

are both one-to-one and onto. Let $v_1,v_2,w_1,w_2 \in \mathcal{C}^1([0,2],\mathbf{R}^n), G,H \in \mathbf{R}^{2\times 2}$ be defined by the equations

$$M_5\begin{pmatrix} v_1 & v_2 \\ G_{11} & G_{12} \\ G_{21} & G_{22} \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}, \tag{3.24}$$

$$M_6\begin{pmatrix} w_1 & w_2 \\ H_{11} & H_{21} \\ H_{12} & H_{22} \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ -1 & 0 \\ 0 & -1 \end{pmatrix}. \tag{3.25}$$

If $(x(t),T,\alpha)$ is a periodic solution, then $G=0$ if and only if $H=0$. Moreover, this happens if and only if $(x(t),T,\alpha)$ corresponds to a simple Neimark–Sacker singularity of periodic solutions with the multipliers $e^{\pm i\theta}$, where $\kappa=\cos(\theta)$.

Proof. The proof is standard. □

4. Regularity of the defining systems. In this section we prove that, under natural nondegeneracy and transversality conditions, the test functionals constructed in the previous section are regular (with respect to the arclength parameter along the periodic solution family). This implies regularity of defining systems consisting of the periodic BVP (2.1), (2.2), (2.3), and the condition for the corresponding functional to vanish, for the two-parameter continuation of the bifurcation.

4.1. Regularity at a fold bifurcation. To prove the regularity of the test functional G for the simple fold singularity in Proposition 9, we proceed as in the case of the fold singularity of equilibria [12], [3].

The computation of periodic orbits is based on the equation

$$F(X,\alpha)=0, \tag{4.1}$$

where $X \equiv (x(\cdot),T) \in \mathcal{C}^1([0,1],\mathbf{R}) \times \mathbf{R}$, and $F(X) \in \mathcal{C}^0([0,1],\mathbf{R}) \times \mathbf{R}^n \times \mathbf{R}$ is given by

$$F(X) \equiv \begin{pmatrix} x'(t) - Tf(x(t),\alpha) \\ x(1)-x(0) \\ \displaystyle\int_0^1 x^*(\tau)x'_{k-1}(\tau)\, d\tau \end{pmatrix}$$

(see (2.1), (2.2), and (2.3)). The Fréchet derivative $F_X(X,\alpha)$ of this operator (with x_{k-1} substituted by x upon differentiation) is M_1 as defined in (3.1). By Propositions 5 and 6, if the periodic orbit has a simple fold singularity, then F_X is singular. Moreover, the left and right singular vectors are then

$$\begin{pmatrix} \Psi p_0 \\ -p_0 \\ 0 \end{pmatrix}$$

and

$$\begin{pmatrix} v \\ 1 \end{pmatrix},$$

given in (3.2) and (3.6), respectively. By definition, a simple fold point is *nondegenerate* if

$$\begin{pmatrix} \Psi p_0 \\ -p_0 \\ 0 \end{pmatrix}^* F_{XX} \begin{pmatrix} v \\ 1 \end{pmatrix} \begin{pmatrix} v \\ 1 \end{pmatrix} \neq 0. \tag{4.2}$$

Let α be a scalar parameter in (1.1). A nondegenerate fold point is called *regular* if $[F_X \;\; F_\alpha]$ is onto at the singularity. This is the usual *transversality condition* for the limit point bifurcation, which can be equivalently expressed as

$$\begin{pmatrix} \Psi p_0 \\ -p_0 \\ 0 \end{pmatrix}^* F_\alpha \neq 0. \tag{4.3}$$

Let s denote arclength along the family of periodic orbits. We think of X and α as functions of s so that (4.1) is an identity in s. By (3.13) this also defines G as a function of s. Suppose that a fold singularity occurs at $s = s_0$. We will prove that $G_s(s_0) \neq 0$ near a regular fold point, i.e., a simple fold singularity where both (4.2) and (4.3) hold.

Taking derivatives of (3.13) with respect to s we find

$$N_1 \begin{pmatrix} v_s \\ S_s \\ G_s \end{pmatrix} = \begin{pmatrix} (F_{XX}X_s + F_{X\alpha}\alpha_s)\begin{pmatrix} v \\ S \end{pmatrix} \\ 0 \end{pmatrix}. \tag{4.4}$$

In this expression

$$\begin{pmatrix} v \\ S \end{pmatrix}$$

is a right singular vector of F_X. Furthermore, at the fold singularity $\alpha_s = 0$. Since $F_X X_s + F_\alpha \alpha_s \equiv 0$ it follows that X_s is also a right singular vector of F_X. Now by (4.4) we have $G_s(s_0) \neq 0$ if and only if

$$F_{XX} \begin{pmatrix} v \\ 1 \end{pmatrix} \begin{pmatrix} v \\ 1 \end{pmatrix}$$

is not in the range of M_1; under our assumptions this is equivalent to (4.2).

4.2. Regularity at a period-doubling bifurcation. We have seen that locally, near a simple flip singularity, the system consisting of (2.1), (2.2), (2.3), and $G = 0$ (where G is given by (3.17)) defines the set of simple flips in $(x(\cdot), T, \alpha)$-space if the conditions $\langle \phi_0, \Phi q_2 \rangle \neq 0$, $\langle \psi_0, \Psi p_2 \rangle \neq 0$ hold. We will now prove that this is a regular system if an appropriate transversality condition for the period-doubling bifurcation holds.

Let s denote arclength along the family of periodic orbits so that $(x(s)(t), T(s), \alpha(s))$ is a solution of (2.1), (2.2), and (2.3) for all s near the bifurcation value s_0. The simplicity of the flip singularity implies that -1 is the algebraically simple eigenvalue of $\Phi(s_0)(1)$ so that it can be continued smoothly, together with its left and right eigenvectors, for nearby values of s. Specifically, we denote by $\lambda(s)$ an eigenvalue of $\Phi(s)(1)$, with left and right eigenvectors $p(s), q(s)$, that is,

$$\begin{array}{ll} \Phi(s)(1)q(s) = \lambda(s)q(s), & p^*(s)\Phi(s)(1) = \lambda(s)p^*(s), \\ \Psi(s)(1)p(s) = \lambda^{-1}(s)p(s), & q^*(s)\Psi(s)(1) = \lambda^{-1}(s)q^*(s), \\ p(s_0) = p_2, & q(s_0) = q_2, \\ \lambda(s_0) = -1. & \end{array} \tag{4.5}$$

The simplicity condition implies that

$$p^*(s)q(s) \neq 0$$

for all s sufficiently close to s_0. By standard arguments, (4.5) implies

$$p^*(s)q(s)\lambda_s(s) = p^*(s)\Phi_s(s)(1)q(s). \tag{4.6}$$

To get an explicit formula for $\Phi_s(s_0)(1)$ we start from the observation that

$$(D - T(s)f_x(x(s), \alpha(s)))\Phi = 0.$$

Taking derivatives, and using somewhat simplied notation, we obtain

$$(D - Tf_x)\Phi_s = (Tf_x)_s\Phi.$$

Multiplying on the right by an arbitrary vector $\xi \in \mathbf{R}^n$, we have

$$(D - Tf_x)\Phi_s\xi = (Tf_x)_s\Phi\xi.$$

This is a linear differential equation for $\Phi_s\xi$ with solution

$$\Phi_s(s)(t)\xi = \Phi(s)(t)\left[\zeta + \int_0^t \Psi^*(s)(\tau)(Tf_x)_s(s)(\tau)\Phi(s)(\tau)\xi \, d\tau\right]$$

for some $\zeta \in \mathbf{R}^n$. For $t = 0$ this reduces to

$$\Phi_s(s)(0)\xi = \Phi(s)(0)\zeta.$$

Since $\Phi(s)(0) = I$, $\Phi_s(s)(0) = 0$, this implies that $\zeta = 0$, so that

$$\Phi_s(s)(t)\xi = \Phi(s)(t) \int_0^t \Psi^*(s)(\tau)(Tf_x)_s(s)(\tau)\Phi(s)(\tau)\xi \, d\tau \tag{4.7}$$

for all $\xi \in \mathbf{R}^n$. From (4.6) we get

$$p^*(s)q(s)\lambda_s(s) = \lambda(s)p^*(s) \int_0^1 \Psi^*(\tau)(Tf_x)_s(s)(\tau)\Phi(s)(\tau)q(s) \, d\tau. \tag{4.8}$$

The natural transversality condition for the period-doubling bifurcation is $\lambda_s(s_0) \neq 0$. We now show that this is equivalent to $G_s(s_0) \neq 0$, thus establishing regularity.

Proposition 18. *The conditions $\lambda_s(s_0) \neq 0$ and $G_s(s_0) \neq 0$ are equivalent near a simple flip singularity.*

Proof. The equations (3.17) are to be considered as identities in s; by taking derivatives we obtain

$$(D - Tf_x)v_s = (Tf_x)_s v - \psi_0 G_s, \tag{4.9}$$

$$(\delta_0 + \delta_1)v_s = 0, \tag{4.10}$$

$$\mathrm{Int}_{\phi_0} v_s = 0.$$

The solution of (3.17) at $s = s_0$ is given by $G(s_0) = 0$, $v(s_0)(t) = \Phi(s_0)(t)q_2$. Now, at $s = s_0$ (4.9) is a linear differential equation for $v_s(s_0)(t)$ with solution

$$v_s(s_0)(t) = \Phi(s_0)(t) \left[\zeta + \int_0^t \Psi^*(s_0)(\tau)((Tf_x)_s(s_0)(\tau)v(s_0)(\tau) - \psi_0 G_s(s_0)) \, d\tau\right]$$

for some vector $\zeta \in \mathbf{R}^n$. Using (4.10) we find

$$0 = (I + \Phi(s_0)(1))\zeta + \Phi(s_0)(1) \int_0^1 \Psi^*(s_0)(\tau)((Tf_x)_s(s_0)(\tau)\Phi(s_0)(\tau)q_2 - \psi_0 G_s(s_0)) \, d\tau.$$

This equation in ζ has a solution if and only if

$$p^*(s_0)\Phi(s_0)(1) \int_0^1 \Psi^*(s_0)(\tau)((Tf_x)_s(s_0)(\tau)\Phi(s_0)(\tau)q_2 - \psi_0 G_s(s_0)) \, d\tau = 0,$$

that is,

$$p_2^* \int_0^1 \Psi^*(s_0)(\tau)(Tf_x)_s(s_0)(\tau)\Phi(s_0)(\tau)q_2 \, d\tau = \langle \psi_0, \Psi p_2 \rangle G_s(s_0).$$

By (4.8) this implies

$$-(p_2^* q_2)\lambda_s(s_0) = \langle \psi_0, \Psi p_2 \rangle G_s(s_0).$$

Since $p_2^* q_2$ and $\langle \psi_0, \Psi p_2 \rangle$ are nonzero, this completes the proof. □

4.3. Regularity at a torus bifurcation. Again, let s denote arclength along the family of periodic orbits so that $(x(s)(t), T(s), \alpha(s))$ is a solution of (2.1), (2.2), and (2.3) for all s near the critical value s_0 corresponding to a simple Neimark–Sacker singularity. Thus $\Phi(s_0)(1)$ has algebraically simple eigenvalues $e^{\pm i\theta}$. Let $\lambda(s) = \lambda_1(s) + i\lambda_2(s)$, $p(s) = p_1(s) + ip_2(s)$, $q(s) = q_1(s) + iq_2(s)$ be the smooth continuations of the critical multiplier $e^{i\theta}$ and the corresponding left and right eigenvectors. The natural transversality condition for the torus bifurcation is the requirement that $\lambda(s)$ crosses the unit circle in the complex plane at nonzero velocity, i.e.,

$$\lambda_1(s_0)\lambda_{1s}(s_0) + \lambda_2(s_0)\lambda_{2s}(s_0) \neq 0. \tag{4.11}$$

Proposition 19. *The system consisting of* (2.1), (2.2), (2.3), *and the conditions*

$$\begin{array}{rcl} G_{11} & = & 0, \\ G_{12} & = & 0, \\ G_{21} & = & 0, \\ G_{22} & = & 0, \end{array} \tag{4.12}$$

where the G_{ij} are defined in Proposition 17, *together form a regular defining system for periodic solutions having a simple Neimark–Sacker singularity if the natural transversality condition* (4.11) *is satisfied.*

Proof. To prove that the system (2.1), (2.2), (2.3), (4.12) is a regular defining system (i.e., has full linear rank), we consider the implicit solution $(x(s)(t), T(s), \alpha(s))$ of (2.1), (2.2), (2.3). So $G_{11}, G_{12}, G_{21}, G_{22}$ are functions of s, κ only, and we have to prove that

$$\begin{pmatrix} G_{11s} & G_{11\kappa} \\ G_{12s} & G_{12\kappa} \\ G_{21s} & G_{21\kappa} \\ G_{22s} & G_{22\kappa} \end{pmatrix}$$

has rank 2. Assume that $c_1, c_2 \in \mathbf{R}$ are such that

$$c_1 G_{ijs} + c_2 G_{ij\kappa} = 0, \quad (i, j = 1, 2). \tag{4.13}$$

We start by noting that $p^H(s)q(s) \neq 0$ in a neighborhood of $s = s_0$. By standard arguments

$$(p^H q)\lambda_s = p^H \Phi_s(1) q, \tag{4.14}$$

where for simplicity of notation we have suppressed the dependence on s. To get an expression for $\Phi_s(1)$ we start from the identity

$$(D - Tf_x)\Phi \equiv 0.$$

Taking derivatives with respect to s and multiplying with any vector $\zeta \in \mathbf{R}^n$ we find

$$(D - Tf_x)\Phi_s \zeta = (Tf_x)_s \Phi \zeta.$$

The solution of this linear differential equation in $\Phi_s \zeta$ is

$$\Phi_s \zeta(t) = \Phi(s)(t) \left[\xi + \int_0^t \Psi^*(s)(\tau)(Tf_x)_s(s)(\tau)\Phi(s)(\tau)\zeta \, d\tau \right],$$

where ξ is determined by the initial conditions. Since for $t = 0$ we have $\Phi(0) = I, \Phi_s(0) = 0$, it follows that $\xi = 0$. Choosing $\zeta = q$ we obtain from (4.14) that

$$(p^H q)\lambda_s = \lambda p^H \int_0^1 \Psi^*(s)(\tau)(Tf_x)_s(s)(\tau)\Phi(s)(\tau)q \; d\tau. \tag{4.15}$$

From (3.24) we infer that

$$M_5 \begin{bmatrix} v_{1s} & v_{2s} \\ G_{11s} & G_{12s} \\ G_{21s} & G_{22s} \end{bmatrix} = \begin{bmatrix} (Tf_x)_s v_1 & (Tf_x)_s v_2 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}, \tag{4.16}$$

$$M_5 \begin{bmatrix} v_{1\kappa} & v_{2\kappa} \\ G_{11\kappa} & G_{12\kappa} \\ G_{21\kappa} & G_{22\kappa} \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 2v_1(1) & 2v_2(1) \\ 0 & 0 \\ 0 & 0 \end{bmatrix}. \tag{4.17}$$

Combining (4.13), (4.16), and (4.17) we obtain

$$M_5 \begin{bmatrix} c_1 v_{1s} + c_2 v_{1\kappa} & c_1 v_{2s} + c_2 v_{2\kappa} \\ 0 & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} c_1 (Tf_x)_s v_1 & c_1 (Tf_x)_s v_2 \\ 2c_2 v_1(1) & 2c_2 v_2(1) \\ 0 & 0 \\ 0 & 0 \end{bmatrix}.$$

Hence

$$\begin{pmatrix} c_1 (Tf_x)_s v_1 \\ 2c_2 v_1(1) \end{pmatrix}, \begin{pmatrix} c_1 (Tf_x)_s v_2 \\ 2c_2 v_2(1) \end{pmatrix}$$

are both in the range of (3.19). As an essential step in the proof of Proposition 15 it was shown that this implies

$$c_1 L((Tf_x)_s v_1) = 2c_2 p^H v_1(1),$$

$$c_1 L((Tf_x)_s v_2) = 2c_2 p^H v_2(1),$$

where the linear operator L is defined in (3.20). Since v_1, v_2 are in the kernel of (3.19) we have

$$v_1(\tau) = \Phi(\tau) v_1(0), \quad v_2(\tau) = \Phi(\tau) v_2(0).$$

Combining the last four formulae we find

$$c_1 L((Tf_x)_s \Phi q) = 2c_2 p^H \Phi(1) q = 2c_2 e^{i\theta}(p^H q). \tag{4.18}$$

Now,

$$\begin{aligned} L((Tf_x)_s \Phi q) &= -2\kappa e^{i\theta} \int_0^1 p^H \Psi^*(\tau)(Tf_x)_s \Phi(\tau) q \; d\tau + e^{2i\theta} \int_0^2 p^H \Psi^*(\tau)(Tf_x)_s \Phi(\tau) q \; d\tau \\ &= e^{i\theta}(\cos\theta + i\sin\theta - 2\cos\theta) \int_0^1 p^H \Psi^*(\tau)(Tf_x)_s \Phi(\tau) q \; d\tau \\ &\quad + e^{2i\theta} \int_0^1 p^H \Psi^*(1+\tau)(Tf_x)_s \Phi(1+\tau) q \; d\tau. \end{aligned}$$

Also,

$$p^H\Psi^*(1+\tau) = (\Psi(1+\tau)p)^H = (\Psi(\tau)\Psi(1)p)^H = p^H\Phi^{-1}(1)\Psi^*(\tau) = e^{-i\theta}p^H\Psi^*(\tau)$$

and

$$\Phi(1+\tau)q = \Phi(\tau)\Phi(1)q = e^{i\theta}\Phi(\tau)q.$$

Hence

$$L((Tf_x)_s\Phi q) = e^{i\theta}2i\sin\theta\int_0^1 p^H\Psi^*(\tau)(Tf_x)_s\Phi(\tau)q\ d\tau.$$

By (4.15) this implies

$$L((Tf_x)_s\Phi q) = 2i\sin\theta(p^Hq)\lambda_s.$$

Using (4.18) we further obtain

$$2ic_1\sin\theta(p^Hq)\lambda_s = 2c_2e^{i\theta}(p^Hq).$$

Dividing by $2(p^Hq)$ we obtain

$$(-\sin\theta\lambda_{2s} + i\sin\theta\lambda_{1s})c_1 = (\cos\theta + i\sin\theta)c_2.$$

Taking real and imaginary parts of this complex equality we find

$$\begin{pmatrix} -\sin\theta\lambda_{2s} & -\cos\theta \\ \sin\theta\lambda_{1s} & -\sin\theta \end{pmatrix}\begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

The determinant of the 2×2 matrix in this expression is equal to

$$\sin\theta(\cos\theta\lambda_{1s} + \sin\theta\lambda_{2s}) = \sin\theta(\lambda_1\lambda_{1s} + \lambda_2\lambda_{2s}).$$

By (4.11) and $\sin\theta \neq 0$ this implies that $c_1 = c_2 = 0$, which completes the proof. □

5. Computational issues. In this section we discuss computational issues related to the implementation of our defining systems, namely the computation of the derivatives of the test functionals with respect to the unknowns of the system, $x(t), \alpha, T$, as well as the problem of adapting the defining systems along the bifurcation branch. We also explicitly show the BVPs that must be solved.

5.1. Fold bifurcation. Proposition 9 implies that locally, near a simple fold singularity of periodic solutions, the system consisting of (2.1), (2.2), (2.3), and

$$G = 0$$

defines the set of simple folds in $(x(\cdot), T, \alpha)$-space; here G is defined by (3.13). Under natural nondegeneracy and transversality conditions, the regularity of this system was proved in section 4.1.

We need the derivatives of G with respect to the unknowns of the system, i.e., with respect to $x(\cdot), \alpha, T$.

Denoting by z any component of α or T we infer from (3.13) that

$$N_1 \begin{pmatrix} v_z \\ S_z \\ G_z \end{pmatrix} = \begin{pmatrix} [Tf_x(x(t),\alpha]_z v + [f(x(t),\alpha)]_z S \\ 0 \\ -\mathrm{Int}_{[f(x(\cdot),\alpha)]_z} v \\ 0 \end{pmatrix}.$$

Numerically we solve a discretized version of this equation, say

$$N_1^d \begin{pmatrix} v_z \\ S_z \\ G_z \end{pmatrix} = \begin{pmatrix} ([Tf_x(x(t),\alpha]_z v + [f(x(t),\alpha)]_z S)_d \\ 0 \\ -(\mathrm{Int}_{[f(x(\cdot),\alpha)]_z} v)_d \\ 0 \end{pmatrix}, \tag{5.1}$$

where N_1^d is the discretized version of N_1, i.e., a large square matrix with a structure that can be efficiently factorized, for example, as in AUTO [9].

Note that a large number of linear systems having the same structured matrix N_1^d must be solved. Moreover, all right-hand sides are known before the factorization. Thus the solution can be done in a single factorization process, without storing the factors.

$(N_1^d)^T$ has a block structure that is very similar to N_1^d. If an efficient solution strategy for $(N_1^d)^T$ is also developed, then it is possible to avoid solving (5.1) for all relevant z. Instead, a single system with $(N_1^d)^T$ is to be solved. In transposed form it is given by

$$(w_1^*, w_2^*, w_3, w_4) N_1^d = (0, 0, 1). \tag{5.2}$$

Combining (5.1) and (5.2) we find

$$G_z = w_1^*([Tf_x(x(t),\alpha]_z v + [f(x(t),\alpha)]_z S)_d - w_3(\mathrm{Int}_{[f(x(\cdot),\alpha)]_z} v)_d.$$

Notice that (3.13) is equivalent to the system

$$\left\{ \begin{array}{rcl} v'(t) - Tf_x(x(t),\alpha)v(t) - Sf(x(t),\alpha) + Gw_{01}(t) &=& 0, \\ v(1) - v(0) + Gw_{02} &=& 0, \\ \displaystyle\int_0^1 v^*(\tau) f(x(\tau),\alpha)\, d\tau + Gw_{03} &=& 0, \\ \displaystyle\int_0^1 v^*(\tau) v_{01}(\tau)\, d\tau + Sv_{02} &=& 1, \end{array} \right. \tag{5.3}$$

while (3.14) can be explicitly written as

$$\left\{ \begin{array}{rcr} w'(t) + Tf_x^*(x(t),\alpha)w(t) - Rf(x(t),\alpha) + Hv_{11}(t) &=& 0, \\ w(1) - w(0) + Hv_{12} &=& 0, \\ \displaystyle\int_0^1 w^*(\tau) f(x(\tau),\alpha)\, d\tau + Hv_{13} &=& 0, \\ \displaystyle\int_0^1 w^*(\tau) w_{11}(\tau)\, d\tau + Rw_{12} &=& -1. \end{array} \right. \tag{5.4}$$

Discretizations of these systems, for example by orthogonal collocation, result in linearized Newton systems having the same sparsity as the linear systems arising from

(2.5). They can therefore be solved using the same numerical linear algebra algorithms.

In practice we need to adapt the auxiliary variables (i.e., $w_{01}, w_{02}, w_{03}, v_{01}, v_{02}, v_{11}, v_{12}, v_{13}$, w_{11}, and w_{12}) along a computed branch of fold bifurcations of periodic orbits. For the bordering rows in N_1 and N_2, the natural choice is to take the kernel vectors of M_1 and M_2, respectively, at a previously computed solution point. These kernel vectors are obtained as a by-product of solving (5.3) and (5.4). For the column bordering of N_1 we need a vector that is not in the range of M_1. By Proposition 6, a possible choice is

$$\begin{pmatrix} w_{01} \\ w_{02} \\ w_{03} \end{pmatrix} = \begin{pmatrix} \Psi p_0 \\ 0 \\ 0 \end{pmatrix},$$

which by Proposition 7 can be derived from the solution of (5.4). Finally, a bordering column for N_2 is given in Proposition 8:

$$\begin{pmatrix} v_{11} \\ v_{12} \\ v_{13} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}.$$

Therefore, problems (5.3) and (5.4) actually take the following simplified forms:

$$\begin{cases} v'(t) - Tf_x(x(t),\alpha)v(t) - Sf(x(t),\alpha) + Gw_{01}(t) &= 0, \\ v(1) - v(0) &= 0, \\ \displaystyle\int_0^1 v^*(\tau) f(x(\tau),\alpha)\, d\tau &= 0, \\ \displaystyle\int_0^1 v^*(\tau) v_{01}(\tau)\, d\tau + S &= 1 \end{cases}$$

and

$$\begin{cases} w'(t) + Tf_x^*(x(t),\alpha)w(t) - Rf(x(t),\alpha) &= 0, \\ w(1) - w(0) &= 0, \\ \displaystyle\int_0^1 w^*(\tau) f(x(\tau),\alpha)\, d\tau + H &= 0, \\ \displaystyle\int_0^1 w^*(\tau) w_{11}(\tau)\, d\tau &= -1. \end{cases}$$

5.2. Period-doubling. By Proposition 14, simple flips are determined by (2.1), (2.2), (2.3), and the condition $G = 0$, where G is given by (3.17), assuming the conditions $\langle \phi_0, \Phi q_2 \rangle \neq 0$, $\langle \psi_0, \Psi p_2 \rangle \neq 0$ hold. To solve such systems numerically, we need the derivatives of G with respect to the unknowns of the system, i.e., with respect to $x(t), \alpha, T$. These can be approximated by finite differences, using (3.17). As in the fold case, they can be obtained exactly by solving an "adjoint problem" to (3.17). In this case the adjoint problem is (3.18).

Proposition 20. *Let z denote a component of the problem parameter vector α, or let z denote the period T, on both of which the quantity G in* (3.17) *depends. Let v and w be obtained from* (3.17) *and* (3.18), *respectively. Then the derivative of G with respect to z can be written as*

$$G_z = -\int_0^1 w^*(\tau)[Tf_x(x(\tau),\alpha)]_z v(\tau)\, d\tau,$$

while the linear part of the variation of G *with respect to* $x \mapsto x + \delta x$ *is given by*

$$\delta G = -\int_0^1 w^*(\tau) T f_{xx}(x(\tau), \alpha)) v(\tau)(\delta x)(\tau)\ d\tau.$$

Proof. By differentiating (3.17) we obtain

$$M_1 \begin{pmatrix} v_z \\ G_z \end{pmatrix} = \begin{pmatrix} [Tf_x(x(t), \alpha)]_z v \\ 0 \\ 0 \end{pmatrix}. \tag{5.5}$$

Multiplying the first equation in (5.5) from the left with w^*, integrating over the interval $[0, 1]$, and using the third equation in (3.18) we get

$$\begin{aligned} &\int_0^1 w^*(\tau) v_z'(\tau)\ d\tau - \int_0^1 w^*(\tau) T f_x(x(\tau), \alpha) v_z(\tau)\ d\tau - G_z \\ &= \int_0^1 w^*(\tau)[Tf_x(x(\tau), \alpha)]_z v(\tau)\ d\tau. \end{aligned}$$

Integrating the first term in this expression by parts, and using the second equations in (3.18) and (5.5), we obtain

$$\begin{aligned} &-\int_0^1 v_z^*(\tau) w'(\tau)\ d\tau - \int_0^1 v_z^*(\tau) T f_x^*(x(\tau), \alpha) w(\tau)\ d\tau - G_z \\ &= \int_0^1 w^*(\tau)[Tf_x(x(\tau), \alpha)]_z v(\tau)\ d\tau. \end{aligned}$$

Using the first equation in (3.18) we get

$$-\int_0^1 v_z^*(\tau)(-\phi_0(\tau) H)\ d\tau - G_z = \int_0^1 w^*(\tau)[Tf_x(x(\tau), \alpha)]_z v(\tau)\ d\tau.$$

By the last equation in (5.5) the first part of the proposition follows.

The linear parts of the variations of G and v under variation of x satisfy

$$M_1 \begin{pmatrix} \delta v \\ \delta G \end{pmatrix} = \begin{pmatrix} T f_{xx}(x(t), \alpha) v\ \delta x \\ 0 \\ 0 \end{pmatrix}.$$

Similar to the derivation above, this implies the second part of the proposition. □

Notice that (3.17) is equivalent to the system

$$\left\{ \begin{array}{rcl} v'(t) - T f_x(x(t), \alpha) v(t) + G\psi_0(t) & = & 0, \\ v(0) + v(1) & = & 0, \\ \displaystyle\int_0^1 \phi_0^*(\tau) v(\tau)\ \ d\tau & = & 1, \end{array} \right. \tag{5.6}$$

while (3.18) can be explicitly written as

$$\left\{ \begin{array}{rcr} w'(t) + T f_x^*(x(t), \alpha) w(t) + H\phi_0(t) & = & 0, \\ w(0) + w(1) & = & 0, \\ \displaystyle\int_0^1 \psi_0^*(\tau) w(\tau)\ \ d\tau & = & -1. \end{array} \right. \tag{5.7}$$

Discretizations of these systems, for example by orthogonal collocation, result in linearized Newton systems having the same sparsity as the linear systems arising from (2.5). They can therefore be solved using the same numerical linear algebra algorithms.

The natural choice for starting values of ϕ_0, ψ_0 is

$$\phi_0(t) = \Phi(t)q_2, \quad \psi_0(t) = \Psi(t)p_2.$$

In a continuation context, it is necessary to regularly update ϕ_0 and ψ_0. Specifically, v obtained from (3.17) can be used to update ϕ_0, and w obtained from (3.18) can be used to update ψ_0. Indeed, after convergence to a period-doubling bifurcation, v spans the kernel of

$$\begin{pmatrix} D - Tf_x(x(t), \alpha) \\ \delta_0 + \delta_1 \end{pmatrix},$$

and, similarly, w spans the kernel of

$$\begin{pmatrix} D + Tf_x^*(x(t), \alpha) \\ \delta_0 + \delta_1 \end{pmatrix}.$$

5.3. Torus bifurcation. We have proved in Proposition 17 that the matrix equation $G = 0$ can be used to continue numerically curves of periodic solutions having Neimark–Sacker singularities, in particular, torus bifurcation points. Some issues require further attention.

First of all, we mention that the BVP for G is defined on the interval $[0, 2]$ and that 3-*point boundary conditions* are involved (at $t = 0, 1$, and 2).

To solve the system (2.1), (2.2), (2.3), (4.12) efficiently by a Newton-like method, one needs the derivatives G_{ijz}, where z is T or a component of α. From (3.24) we infer that

$$M_5 \begin{pmatrix} v_{1z} & v_{2z} \\ G_{11z} & G_{12z} \\ G_{21z} & G_{22z} \end{pmatrix} = \begin{pmatrix} [Tf_x(x(t), \alpha]_z v_1 & [Tf_x(x(t), \alpha)]_z v_2 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix}.$$

One also needs the derivatives with respect to κ; for this we find

$$M_5 \begin{pmatrix} v_{1\kappa} & v_{2\kappa} \\ G_{11\kappa} & G_{12\kappa} \\ G_{21\kappa} & G_{22\kappa} \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 2v_1(1) & 2v_2(1) \\ 0 & 0 \\ 0 & 0 \end{pmatrix}.$$

Numerically we solve the discretized versions of these equations, say

$$M_5^d \begin{pmatrix} v_{1z} & v_{2z} \\ G_{11z} & G_{12z} \\ G_{21z} & G_{22z} \end{pmatrix} = \begin{pmatrix} [Tf_x(x(t), \alpha]_z v_1 & [Tf_x(x(t), \alpha)]_z v_2 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix}. \tag{5.8}$$

One also needs the derivatives with respect to κ; for this we find

$$M_5^d \begin{pmatrix} v_{1\kappa} & v_{2\kappa} \\ G_{11\kappa} & G_{12\kappa} \\ G_{21\kappa} & G_{22\kappa} \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 2v_1(1) & 2v_2(1) \\ 0 & 0 \\ 0 & 0 \end{pmatrix}, \tag{5.9}$$

where M_5^d is the discretized version of M_5, i.e., a large square matrix of the same structure as that factored efficiently in AUTO.

We again note that a large number of linear systems with the same structured matrix M_5^d has to be solved. All right-hand sides are known when the factorization is done. Thus the solution of all systems can be done during a single factorization process of M_5^d without storing the factors.

$(M_5^d)^*$ has a block structure that is very similar to that of M_5^d. If an efficient solution strategy for $(M_5^d)^*$ is also developed, then it is possible to avoid solving (5.8) for all relevant z and (5.9). Instead, a single system with $(M_5^d)^*$ is to be solved. In transposed form it is given by

$$\begin{pmatrix} w_1^{1*} & w_1^{2*} & G_{11} & G_{12} \\ w_2^{1*} & w_2^{2*} & G_{21} & G_{22} \end{pmatrix} M_5^d = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}. \tag{5.10}$$

Combining (5.8) and (5.10) we find

$$\begin{pmatrix} G_{11z} & G_{12z} \\ G_{21z} & G_{22z} \end{pmatrix} = \begin{pmatrix} w_1^{1*}[Tf_x(x(t),\alpha)]_z v_1 & w_1^{1*}[Tf_x(x(t),\alpha)]_z v_2 \\ w_2^{1*}[Tf_x(x(t),\alpha)]_z v_1 & w_2^{1*}[Tf_x(x(t),\alpha)]_z v_2 \end{pmatrix}$$

if z is T or one of the components of x, α. For κ we find

$$\begin{pmatrix} G_{11\kappa} & G_{12\kappa} \\ G_{21\kappa} & G_{22\kappa} \end{pmatrix} = \begin{pmatrix} 2w_1^{2*}v_1(1) & 2w_1^{2*}v_2(1) \\ 2w_2^{2*}v_1(1) & 2w_2^{2*}v_2(1) \end{pmatrix}.$$

Next notice that (3.24) is equivalent to the system

$$\left\{ \begin{array}{lcl} v_1' - Tf_x(x(t),\alpha)v_1 + G_{11}\psi_0 + G_{21}\psi_1 & = & 0, \\ v_2' - Tf_x(x(t),\alpha)v_2 + G_{12}\psi_0 + G_{22}\psi_1 & = & 0, \\ v_1(0) - 2\kappa v_1(1) + v_1(2) & = & 0, \\ v_2(0) - 2\kappa v_2(1) + v_2(2) & = & 0, \\ \displaystyle\int_0^2 \phi_0^*(\tau)v_1(\tau)\ d\tau & = & 1, \\ \displaystyle\int_0^2 \phi_1^*(\tau)v_2(\tau)\ d\tau & = & 0, \\ \displaystyle\int_0^2 \phi_0^*(\tau)v_1(\tau)\ d\tau & = & 0, \\ \displaystyle\int_0^2 \phi_1^*(\tau)v_2(\tau)\ d\tau & = & 1, \end{array} \right. \tag{5.11}$$

while (3.25) can be explicitly written as

$$(5.12)\qquad \begin{cases} w_1' + Tf_x^*(x(t),\alpha)w_1 + H_{11}\phi_0 + H_{21}\phi_1 & = \quad 0, \\ w_2' + Tf_x^*(x(t),\alpha)w_2 + H_{12}\phi_0 + H_{22}\phi_1 & = \quad 0, \\ w_1(0) - 2\kappa w_1(1) + w_1(2) & = \quad 0, \\ w_2(0) - 2\kappa w_2(1) + w_2(2) & = \quad 0, \\ \displaystyle\int_0^2 \psi_0^*(\tau)w_1(\tau) \quad d\tau & = \ -1, \\ \displaystyle\int_0^2 \psi_1^*(\tau)w_2(\tau) \quad d\tau & = \quad 0, \\ \displaystyle\int_0^2 \psi_0^*(\tau)w_1(\tau) \quad d\tau & = \quad 0, \\ \displaystyle\int_0^2 \psi_1^*(\tau)w_2(\tau) \quad d\tau & = \ -1. \end{cases}$$

Discretizations of these systems, for example by orthogonal collocation, result in linearized Newton systems having the same sparsity as the linear systems arising from (2.5). They can therefore be solved using the same numerical linear algebra algorithms.

In a continuation context, the vector-functions $\phi_0, \phi_1, \psi_0, \psi_1$ should be updated. This can be done by solving both (5.11) and (5.12). Indeed, v_1, v_2 span the two-dimensional space in which ϕ_0, ϕ_1 should be chosen and w_1, w_2 similarly span the space in which ψ_0, ψ_1 should be chosen (some orthogonalization and scaling may be appropriate).

Finally, recall that we compute the Neimark–Sacker points by using essentially an overdetermined system. This necessitates some changes in the elimination strategy when solving the linear systems.

6. Numerical example.

In this section we illustrate our new techniques on a test example, a simple feedback control system of Lur'e type:

$$(6.1)\qquad \begin{cases} \dot{x}_1 & = & x_2, \\ \dot{x}_2 & = & x_3, \\ \dot{x}_3 & = & -\alpha x_3 - \beta x_2 - x_1 + x_1^2, \end{cases}$$

where α and β are positive parameters. It is well known (see, for example [17, section 5.4]) that the equilibrium $x_1 = x_2 = x_3 = 0$ of (6.1) has a supercritical Hopf bifurcation at

$$\alpha_0 = \frac{1}{\beta},$$

generating a stable periodic solution that exists for $\alpha < \alpha_0$. This periodic solution has a supercritical period-doubling bifurcation at $\alpha_1 \approx 0.630302$.

The discretized continuation problem (2.1), (2.2), and (2.3) for the periodic solution has been programmed in the MATLAB Continuation Toolbox [19]. The method of orthogonal collocation with piecewise polynomials is used, similar to the one implemented in AUTO. It is characterized by the number `NTST` of mesh points and the number `NCOL` of collocation points. At each computed point on the solution curve, a discrete version of (5.6) is set up and solved. This gives a value of the test function G to detect a flip singularity. A constant bordering function ψ_0 is used, while the

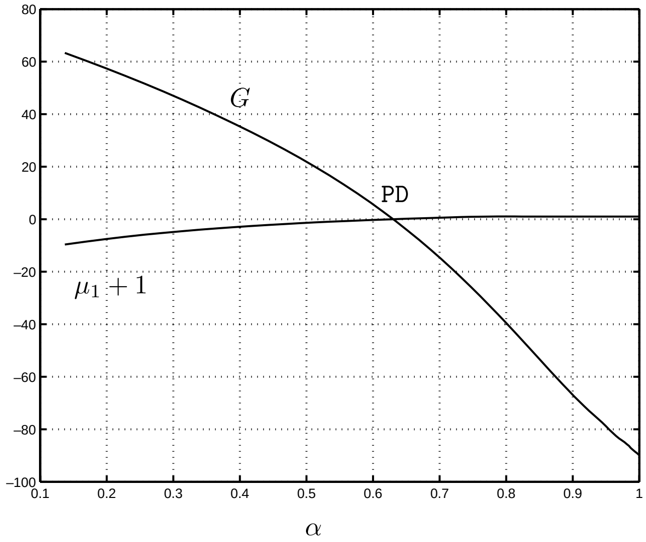


FIG. 1. *Test function $G(\alpha)$ and $\mu_1(\alpha) + 1$ for $\beta = 1$.*

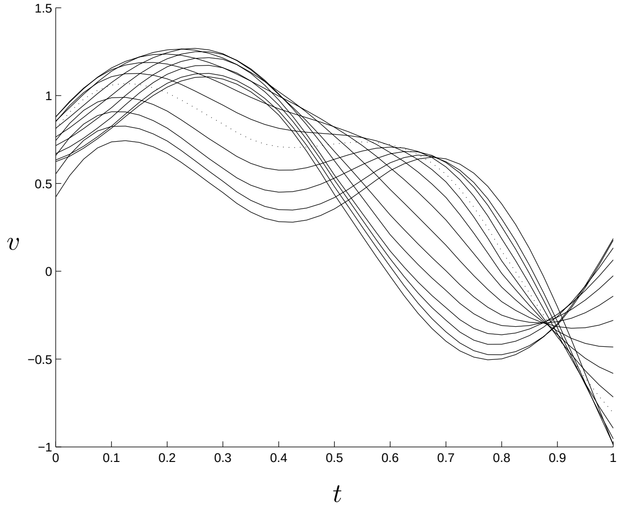


FIG. 2. *Solutions $v(t)$ at different α-values for $\beta = 1$.*

computed approximation to v is used to update the bordering function ϕ_0. Figures 1 and 2 are produced with `NTST=10` and `NCOL=4`.

Figure 1 shows the behavior of G as a function of α for $\beta = 1$. For this value of β, Hopf bifurcation occurs at $\alpha_0 = 1$. In the same figure, the function $\mu_1 + 1$ is plotted, where μ_1 is a nontrivial Floquet multiplier of the periodic solution for which $\mu_1(\alpha_1) = -1$. The multipliers are computed via a specially adapted elimination algorithm from AUTO. As can be seen, G vanishes together with $\mu_1 + 1$. Moreover, close examination of numerical data gives the above bifurcation value α_1 with seven correct decimal places.

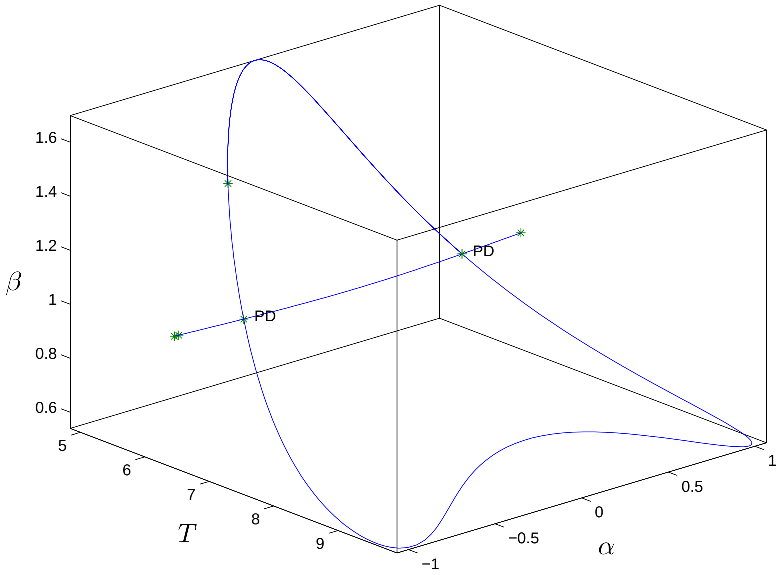


FIG. 3. *Cycle and period-doubling branches.*

Figure 2 shows a family of computed profiles $v(t)$ along the solution curve. The dashed solution corresponds to the bifurcation parameter value α_1. Finally, Figure 3 shows the two-parameter continuation of the period-doubling bifurcation curve, which corresponds to a close curve. The continuation is started at one of the PD points in the one-parameter path of periodic solutions discussed above.

We now briefly address the important issue of comparing our new method for continuing period-doubling bifurcations to the algorithm based on a fully extended system, i.e., (2.1), (2.2), and (2.3), augmented by

$$\left\{\begin{array}{rcl} v'(t) - Tf_x(x(t),\alpha)v(t) & = & 0, \\ v(0)+v(1) & = & 0, \\ \displaystyle\int_0^1 \phi_0^*(\tau)v(\tau) \;\; d\tau & = & 1, \end{array}\right.$$

as implemented in AUTO. The corresponding discretized system is nearly twice the size as the discretized minimally extended system composed of (2.1), (2.2), (2.3), and $G = 0$, where G is to be computed from (5.6). However, for the minimally extended system one has to solve the extra BVP (5.7) in order to calculate the Jacobian matrix of the discretized bordered system. For comparison, both methods were implemented in a similar fashion, using the standard sparse matrix solver in the Continuation Toolbox [19], and tested using different choices for the number of mesh points and the number of collocation points. Table 1 shows the execution times required by the two methods for computing the same number (300) of solution points along the period-doubling curve shown in Figure 3. Computations were done on a 350 Mhz PC.

Clearly the bordered system of this paper is faster, and its advantage widens as the number of mesh points and the number of collocation points increases. In the computations we used an adaptive step length, and the bordered system actually resulted in larger steps than the fully extended system. Details of the implementation and more extensive comparisons will be reported elsewhere.

Table 1

NTST	NCOL	Minimally extended system	Fully extended system
10	4	101,8 s	122,3 s
10	5	134,9 s	159,4 s
20	4	269,9 s	358,6 s
20	5	371,9 s	558,2 s
30	4	529,8 s	808,0 s
30	5	751,0 s	1260,3 s
40	4	886,0 s	1528,8 s
40	5	1376,8 s	2528,6 s

REFERENCES

[1] E. L. Allgower and K. Georg, *Numerical Path Following*, Handb. Numer. Anal. 5, P. G. Ciarlet and J. L. Lions, eds., North–Holland, Amsterdam, 1996.

[2] U. M. Ascher, J. Christiansen, and R. D. Russell, *A collocation solver for mixed order systems of boundary value problems*, Math. Comp., 33 (1979), pp. 659–679.

[3] W. J. Beyn, A. Champneys, E. Doedel, W. Govaerts, Yu. A. Kuznetsov, and B. Sandstede, *Numerical continuation, and computation of normal forms*, in Handbook of Dynamical Systems, Vol. 2, B. Fiedler, ed., Elsevier, Amsterdam, 2002, pp. 149–219.

[4] C. De Boor and B. Swartz, *Collocation at Gaussian points*, SIAM J. Numer. Anal., 10 (1973), pp. 582–606.

[5] D. W. Decker and H. B. Keller, *Multiple limit point bifurcation*, J. Math. Anal., 75 (1980), pp. 417–430.

[6] E. J. Doedel, A. R. Champneys, T. F. Fairgrieve, Yu. A. Kuznetsov, B. Sandstede, and X. J. Wang, auto97: *Continuation and Bifurcation Software for Ordinary Differential Equations (with HomCont)*, Concordia University, Montreal, QC, Canada, 1997. Available via ftp from ftp.cs.concordia.ca/pub/doedel/auto.

[7] E. J. Doedel, A. D. Jepson, and H. B. Keller, *Numerical methods for Hopf bifurcation and continuation of periodic solution paths*, in Computing Methods in Applied Sciences and Engineering VI, R. Glowinski, and J. L. Lions, eds., North–Holland, Amsterdam, 1984, pp. 127–136.

[8] E. J. Doedel, H. B. Keller, and J. P. Kernévez, *Numerical analysis and control of bifurcation problems: Part* I, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 1 (1991), pp. 493–520.

[9] E. J. Doedel, H. B. Keller, and J. P. Kernévez, *Numerical analysis and control of bifurcation problems: Part* II, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 1 (1991), pp. 745–772.

[10] T. F. Fairgrieve, *The Computations and Use of Floquet Multipliers for Bifurcation Analysis*, Ph.D. Thesis, University of Toronto, Toronto, ON, Canada, 1994.

[11] W. Govaerts, Yu. A. Kuznetsov, and B. Sijnave, *Implementation of Hopf and double Hopf continuation using bordering methods*, ACM Trans. Math. Software, 24 (1998), pp. 418–436.

[12] W. J. F. Govaerts, *Numerical Methods for Bifurcations of Dynamical Equilibria*, SIAM, Philadelphia, 2000.

[13] A. Griewank and G. W. Reddien, *Characterization and computation of generalized turning points*, SIAM J. Numer. Anal., 21 (1984), pp. 176–185.

[14] J. Guckenheimer and P. Holmes, *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*, Appl. Math. Sci. 42, Springer-Verlag, New York, 1983.

[15] A. D. Jepson, *Numerical Hopf Bifurcation*, Ph.D. thesis, California Institute of Technology, Pasadena, CA, 1981.

[16] H. B. Keller, *Numerical solution of bifurcation and nonlinear eigenvalue problems*, in Applications of Bifurcation Theory, P. H. Rabinowitz, ed., Academic Press, New York, 1977, pp. 359–384.

[17] Yu. A. Kuznetsov, *Elements of Applied Bifurcation Theory*, 2nd ed., Springer-Verlag, New York, 1998.

[18] Yu. A. Kuznetsov and V. V. Levitin, content: *A Multiplatform Environment for Analyzing Dynamical Systems*, Dynamical Systems Laboratory, CWI, Amsterdam, 1995–1997. Available via ftp from ftp.cwi.nl/pub/content.

[19] Yu. A. Kuznetsov, W. Mestrom, and A. M. Riet, *A Continuation Toolbox in* matlab, Mathematical Institute, Utrecht University, Utrecht, The Netherlands, 2001,

http://www.math.uu.nl/people/kuznet/cm.

[20] G. Moore and A. Spence, *The calculation of turning points of nonlinear equations*, SIAM J. Numer. Anal., 17 (1980), pp. 567–576.

[21] R. D. Russell and J. Christiansen, *Adaptive mesh selection strategies for solving boundary value problems*, SIAM J. Numer. Anal., 15 (1978), pp. 59–80.

[22] R. Seydel, *Numerical computation of branch points in nonlinear equations*, Numer. Math., 33 (1979), pp. 339–352.

SIAM J. NUMER. ANAL.
Vol. 41, No. 2, pp. 436–456

THE P_1^{mod} ELEMENT: A NEW NONCONFORMING FINITE ELEMENT FOR CONVECTION-DIFFUSION PROBLEMS*

PETR KNOBLOCH† AND LUTZ TOBISKA‡

Abstract. We consider a nonconforming streamline diffusion finite element method for solving convection-diffusion problems. The loss of the Galerkin orthogonality of the streamline diffusion method when applied to nonconforming finite element approximations results in an additional error term which cannot be estimated uniformly with respect to the perturbation parameter for the standard piecewise linear or rotated bilinear elements. Therefore, starting from the Crouzeix–Raviart element, we construct a modified nonconforming first order finite element space on shape regular triangular meshes satisfying a patch test of higher order. A rigorous error analysis of this P_1^{mod} element applied to a streamline diffusion discretization is given. The numerical tests show the robustness and the high accuracy of the new method.

Key words. convection-diffusion problems, streamline diffusion method, nonconforming finite elements, error estimates

AMS subject classifications. 65N30, 65N15

PII. S0036142902402158

1. Introduction.

We consider the convection-diffusion equation

$$-\varepsilon\,\Delta\,u+\boldsymbol{b}\cdot\nabla u+c\,u=f\quad\text{in }\Omega,\qquad\qquad u=u_b\quad\text{on }\partial\Omega, \tag{1.1}$$

where $\Omega\subset\mathbb{R}^2$ is a bounded domain with a polygonal boundary $\partial\Omega$, $\varepsilon\in(0,1)$ is constant, $\boldsymbol{b}\in W^{1,\infty}(\Omega)^2$, $c\in L^\infty(\Omega)$, $f\in L^2(\Omega)$, and $u_b\in H^{3/2}(\partial\Omega)$. We assume that

$$c-\frac{1}{2}\operatorname{div}\boldsymbol{b}\geq c_0\,, \tag{1.2}$$

where c_0 is a positive constant. This assumption guarantees that (1.1) admits a unique solution for all positive values of the parameter ε.

In the convection dominated case, in which $\varepsilon\ll 1$, the standard Galerkin finite element method produces unphysical oscillations if the local mesh size is not small enough. Among several possible remedies for this undesirable behavior, the streamline diffusion method [8], [15] attracted considerable attention over the last decade, in particular because of its structural simplicity, generality, and the quality of the numerical solution. Summarizing the existing literature we come to the conclusion that in the case of conforming finite element approximations the convergence properties of the streamline diffusion methods are well understood; see, e.g., [6], [10], [14], [15], [18]. Particularly, using piecewise polynomial approximations of degree k in the convection dominated regime ($\varepsilon\leq h$), one can prove the error estimate

$$|||u-u_h|||\leq C\,h^{k+1/2}\,\|u\|_{k+1,\Omega}\,, \tag{1.3}$$

*Received by the editors February 5, 2002; accepted for publication (in revised form) September 27, 2002; published electronically April 23, 2003.

http://www.siam.org/journals/sinum/41-2/40215.html

†Faculty of Mathematics and Physics, Institute of Numerical Mathematics, Charles University, Sokolovská 83, 186 75 Praha 8, Czech Republic (knobloch@karlin.mff.cuni.cz). The research of this author was supported by the European Science Foundation (Programme AMIF—Applied Mathematics for Industrial Flow Problems), by the Czech Grant Agency under grants 201/99/P029 and 201/99/0267, and by grant MSM 113200007.

‡Institut für Analysis und Numerik, Otto–von–Guericke–Universität Magdeburg, Postfach 4120, 39016 Magdeburg, Germany (tobiska@mathematik.uni-magdeburg.de).

where $|||\cdot|||$ denotes the streamline diffusion norm defined in section 3.

The situation changes dramatically if nonconforming finite element approximations are used. Finite element methods of nonconforming type are attractive in computational fluid dynamics since they easily fulfill the Babuška–Brezzi condition. Moreover, because of their edge-oriented degrees of freedom they result in cheap local communication when implementing the method on a MIMD-machine (cf. [5], [9], [16]). Unfortunately, compared to conforming approximations much less is known about the convergence properties of streamline diffusion-type methods for nonconforming finite element approximations.

It has been shown in [12] that special care is necessary to prove the error estimate (1.3) in the nonconforming case. Indeed, when considering nonconforming approximation spaces we lose the continuity property over inner element edges, and the coercivity of the corresponding bilinear form depends on the type of discretization for the convective term. Our assumptions guarantee that the bilinear form with the so-called skew-symmetric discretization of the convective term (cf. the bilinear form a_h^{skew} in section 3) is always coercive in contrast to the convective form (cf. the bilinear form a_h^{conv} in subsection 4.1). On the other hand, the skew-symmetric form leads to an additional term in the consistency error which is difficult to estimate uniformly in ε. In [11], [12] these difficulties have been overcome by adding some special jump terms and thus modifying the standard streamline diffusion finite element method. However, a drawback of these jump terms is that they decrease the sparsity of the stiffness matrix and that they are difficult to implement. So we would like to avoid the jump terms, but then the coercivity of the convective bilinear form is open in general. Recently, it has been discovered in [17] that this coercivity can be guaranteed for the nonconforming rotated bilinear element on rectangular meshes if $|\boldsymbol{b}|_{1,\infty,\Omega}$ is small compared to c_0. Unfortunately, a similar result is not true for the nonconforming linear triangular Crouzeix–Raviart element [4], not even on three-directional meshes. However, also in cases when the convective bilinear form is coercive, the optimal order $O(h^{k+1/2})$ cannot be shown in general. For example, in [17] a superconvergence property on uniform meshes was necessary to prove an ε-uniform convergence result of optimal order $O(h^{3/2})$. Thus, summarizing the known results we see that in general, without using jump terms and on general meshes, we cannot guarantee the same optimal convergence results as in the conforming case.

Particularly, our numerical experiences show that, in the convection dominated regime, it is often not possible to obtain an acceptable accuracy using the mentioned Crouzeix–Raviart element combined with the standard streamline diffusion discretization. In fact, this method is—even for smooth functions—not ε-uniformly convergent. Therefore, the aim of this paper is to develop a first order nonconforming method on general triangular meshes which guarantees the same optimal convergence properties as in the conforming case but does not employ any modifications (such as the above jump terms) of the standard streamline diffusion method. Let us mention that our ideas are not restricted to the first order of accuracy and that an extension to higher order methods is straightforward.

Our method is based on using the standard streamline diffusion discretization with the skew-symmetric form of the convective term and on introducing a new nonconforming finite element space. The theoretical analysis presented in this paper shows that the optimal convergence order known from the conforming finite element method can be recovered if the nonconforming space satisfies a patch test of order 3 since then a better estimate of the consistency error can be obtained. We shall construct such a space by enriching the Crouzeix–Raviart space by suitable nonconforming bubble functions and by restricting the enlarged space to its subspace of functions satisfying the patch test of order 3. The finite element space obtained in such a way contains *modified* Crouzeix–Raviart functions and therefore we call this new element the P_1^{mod} element. This new element not only guarantees the optimal convergence order but also leads to very robust discretizations and much more accurate results than the Crouzeix–Raviart element. In addition, the iterative solver used to compute the discrete solution converges much faster than for the Crouzeix–Raviart element. Let us also mention that the P_1^{mod} element satisfies a discrete Korn inequality (cf. [13]), which is not true for most first order nonconforming finite elements, including the Crouzeix–Raviart element.

The enrichment of the Crouzeix–Raviart space by bubble functions may resemble the techniques where the bubble functions are used to recover various stabilized methods and to find a reasonable rule for the choice of the stabilizing parameters (cf., e.g., [1], [2]). However, our approach is completely different since we start from a stabilized method and the bubble functions are added not to replace the stabilization but to provide an additional stability. In addition, the bubble functions are coupled with the Crouzeix–Raviart functions so that they cannot be eliminated from the discrete problem.

The paper is organized in the following way. Section 2 introduces various notation which will be used in the subsequent sections. In section 3, we recall the weak formulation of (1.1) and describe a nonconforming streamline diffusion finite element discretization. Then the error analysis is presented in section 4. Section 5 is devoted to the construction of the P_1^{mod} element. Section 6 shows that the piecewise linear part of a P_1^{mod} discrete solution asymptotically behaves in the same way as the discrete solution itself, which is useful for postprocessing. Finally, in section 7, we present numerical results which demonstrate the good behavior of discretizations employing the P_1^{mod} element.

2. Notation. We assume that we are given a family $\{\mathcal{T}_h\}$ of triangulations of the domain Ω parametrized by a positive parameter $h \to 0$. Each triangulation $\mathcal{T}_h$ consists of a finite number of closed triangular elements K such that $h_K \equiv \operatorname{diam}(K) \leq h$, $\overline{\Omega} = \bigcup_{K\in\mathcal{T}_h} K$, and any two different elements K, $\widetilde{K} \in \mathcal{T}_h$ are either disjoint or possess either a common vertex or a common edge. In order to prevent the elements from degenerating when h tends to zero, we assume that the family of triangulations is regular; i.e., there exists a constant C independent of h such that

$$\frac{h_K}{\varrho_K} \leq C \qquad \forall\ K \in \mathcal{T}_h,\ h > 0\,,$$

where ϱ_K is the maximum diameter of circles inscribed into K.

We denote by $\mathcal{E}_h$ the set of edges E of $\mathcal{T}_h$. The set of inner edges will be denoted by $\mathcal{E}_h^i$ and the set of boundary edges by $\mathcal{E}_h^b$. Further, we denote by h_E the length of the edge E and by S_E the union of the elements adjacent to E (i.e., S_E consists of one or two elements). For any edge E, we choose a fixed unit normal vector $\boldsymbol{n}_E$ to E. If $E \in \mathcal{E}_h^b$, then $\boldsymbol{n}_E$ coincides with the outer normal vector to $\partial\Omega$. Consider any $E \in \mathcal{E}_h^i$, and let K, $\widetilde{K}$ be the two elements possessing the edge E denoted in such a way that $\boldsymbol{n}_E$ points into $\widetilde{K}$. If v is a function belonging to the space

$$H^{1,h}(\Omega) = \{v \in L^2(\Omega)\,;\ v|_K \in H^1(K)\ \ \forall\ K \in \mathcal{T}_h\}\,,$$

then we define the jump of v across E by

$$[|v|]_E = (v|_K)|_E - (v|_{\widetilde{K}})|_E\,. \tag{2.1}$$

If $E \in \mathcal{E}_h^b$, then we set $[|v|]_E = v|_E$, which is the jump defined by (2.1) with v extended by zero outside Ω.

To formulate a streamline diffusion method for (1.1), we need finite element functions which are piecewise H^2. We assume this regularity with respect to subdivisions of the elements of the triangulation only, which allows more flexibility in the construction of finite element spaces approximating $H_0^1(\Omega)$ (cf. Remark 5.1). The subdivisions can be defined using a triangulation $\widehat{\mathcal{G}}$ of the standard reference element $\widehat{K}$, and we assume that the set $\widehat{\mathcal{G}}$ is invariant under affine regular mappings of $\widehat{K}$ onto $\widehat{K}$. Then, for any element $K \in \mathcal{T}_h$, we can introduce a subdivision

$$\mathcal{G}_K = \{F_K(\widehat{G})\,;\ \widehat{G} \in \widehat{\mathcal{G}}\}\,,$$

where $F_K : \widehat{K} \to K$ is any affine regular mapping which maps $\widehat{K}$ onto K. In view of the invariance of the triangulation $\widehat{\mathcal{G}}$, the set $\mathcal{G}_K$ is independent of the choice of F_K. The space of piecewise H^2 functions with respect to the above subdivision of $\mathcal{T}_h$ will be denoted by

$$H^{2,h}_{\widehat{\mathcal{G}}}(\Omega) = \left\{v \in L^2(\Omega)\,;\ v|_G \in H^2(G)\ \ \forall\, G \in \mathcal{G}_K,\, K \in \mathcal{T}_h\right\}.$$

In the following sections, we shall also need the spaces

$$\widetilde{\mathrm{V}}_h^{conf} = \{v_h \in C(\overline{\Omega})\,;\ v_h|_K \in P_1(K)\ \ \forall\, K \in \mathcal{T}_h\}\,, \qquad \mathrm{V}_h^{conf} = \widetilde{\mathrm{V}}_h^{conf} \cap H_0^1(\Omega)\,,$$
$$\mathrm{V}_h^{nc} = \left\{v_h \in L^2(\Omega)\,;\ v_h|_K \in P_1(K)\ \ \forall\, K \in \mathcal{T}_h\,,\ \ \int_E [|v_h|]_E\,\mathrm{d}\sigma = 0\ \ \forall\, E \in \mathcal{E}_h\right\},$$

and we shall denote by $i_h : H^2(\Omega) \to \widetilde{\mathrm{V}}_h^{conf}$ the Lagrange interpolation operator.

Throughout the paper we use standard notation $L^p(\Omega)$, $W^{k,p}(\Omega)$, $H^k(\Omega) = W^{k,2}(\Omega)$, $C(\overline{\Omega})$, etc. for the usual function spaces; see, e.g., [3]. The norm and seminorm in the Sobolev space $W^{k,p}(\Omega)$ will be denoted by $\|\cdot\|_{k,p,\Omega}$ and $|\cdot|_{k,p,\Omega}$, respectively, and we set $\|\cdot\|_{k,\Omega} = \|\cdot\|_{k,2,\Omega}$ and $|\cdot|_{k,\Omega} = |\cdot|_{k,2,\Omega}$. For the space $H^{1,h}(\Omega)$, we define an analogue of $|\cdot|_{1,\Omega}$ by

$$|v|_{1,h} = \left(\sum_{K\in\mathcal{T}_h} |v|^2_{1,K}\right)^{1/2}, \qquad v \in H^{1,h}(\Omega)\,.$$

The inner product in the space $L^2(G)$ will be denoted by $(\cdot,\cdot)_G$, and we set $(\cdot,\cdot) = (\cdot,\cdot)_\Omega$. Finally, we denote by C a generic constant independent of h and ε.

3. Weak formulation and discrete problem. Denoting by $\widetilde{u}_b \in H^2(\Omega)$ an extension of u_b, a natural weak formulation of the convection-diffusion equation (1.1) reads as follows:

Find $u \in H^1(\Omega)$ such that $u - \widetilde{u}_b \in H_0^1(\Omega)$ and

$$a(u,v) = (f,v) \qquad \forall\, v \in H_0^1(\Omega)\,,$$

where

$$a(u,v) = \varepsilon\,(\nabla u, \nabla v) + (\boldsymbol{b}\cdot\nabla u, v) + (c\,u, v)\,.$$

This weak formulation has a unique solution.

We intend to approximate the space $H_0^1(\Omega)$ by a nonconforming finite element space V_h and at this stage we assume only that

$$\mathrm{V}_h^{conf} \subset \mathrm{V}_h \subset H^{1,h}(\Omega) \cap H^{2,h}_{\widehat{\mathcal{G}}}(\Omega)\,. \tag{3.1}$$

The inclusion $\mathrm{V}_h^{conf} \subset \mathrm{V}_h$ ensures first order approximation properties of V_h with respect to $|\cdot|_{1,h}$ when $h \to 0$.

A finite element discretization of (1.1) could be simply obtained by using the bilinear forms

$$a_h^d(u,v) = \varepsilon \sum_{K\in\mathcal{T}_h} (\nabla u, \nabla v)_K\,, \quad a_h^c(u,v) = \sum_{K\in\mathcal{T}_h} (\boldsymbol{b}\cdot\nabla u, v)_K\,, \qquad u,v \in H^{1,h}(\Omega)\,,$$

instead of the first two terms in $a(u,v)$ and by replacing the space $H_0^1(\Omega)$ in the weak formulation by the finite element space V_h. However, the bilinear form corresponding to the discrete problem generally would not be coercive and therefore, before passing from the weak formulation to the discrete problem, we first apply integration by parts to the convective term $(\boldsymbol{b}\cdot\nabla u, v)$ to obtain

$$(\boldsymbol{b}\cdot\nabla u, v) = \frac{1}{2}\left[(\boldsymbol{b}\cdot\nabla u, v) - (\boldsymbol{b}\cdot\nabla v, u) - (\operatorname{div}\boldsymbol{b}, u\,v)\right], \qquad u \in H^1(\Omega),\ v \in H_0^1(\Omega)\,.$$

Thus, a discrete analogue of the second term in the bilinear form a also is

$$a_h^s(u,v) = \frac{1}{2}\sum_{K\in\mathcal{T}_h}\left[(\boldsymbol{b}\cdot\nabla u, v)_K - (\boldsymbol{b}\cdot\nabla v, u)_K - (\operatorname{div}\boldsymbol{b}, u\,v)_K\right], \qquad u,v \in H^{1,h}(\Omega)\,.$$

This bilinear form is skew-symmetric if $\operatorname{div}\boldsymbol{b} = 0$. That gives rise to the notation a_h^{skew} below. For $u \in H^{2,h}_{\widehat{\mathcal{G}}}(\Omega)$ and $v \in H^{1,h}(\Omega)$, we define a streamline diffusion term by

$$a_h^{sd}(u,v) = \sum_{K\in\mathcal{T}_h}\sum_{G\in\mathcal{G}_K} (-\varepsilon\,\Delta\,u + \boldsymbol{b}\cdot\nabla u + c\,u,\ \delta_K\,\boldsymbol{b}\cdot\nabla v)_G\,,$$

where $\delta_K \geq 0$ is a control parameter. Now, denoting

$$\begin{aligned} a_h^{skew}(u,v) &= a_h^d(u,v) + a_h^s(u,v) + (c\,u, v) + a_h^{sd}(u,v)\,,\\ l_h(v) &= (f,v) + \sum_{K\in\mathcal{T}_h} (f,\ \delta_K\,\boldsymbol{b}\cdot\nabla v)_K\,, \end{aligned}$$

the streamline diffusion finite element method investigated in this paper reads as follows:

Find $u_h \in H^{1,h}(\Omega)$ such that $u_h - i_h\widetilde{u}_b \in \mathrm{V}_h$ and

$$a_h^{skew}(u_h, v_h) = l_h(v_h) \qquad \forall\ v_h \in \mathrm{V}_h\,. \tag{3.2}$$

A natural norm for investigating the properties of the problem (3.2) is the streamline diffusion norm

$$|||v||| = \left(\sum_{K\in\mathcal{T}_h}\{\varepsilon\,|v|_{1,K}^2 + c_0\,\|v\|_{0,K}^2 + \delta_K\,\|\boldsymbol{b}\cdot\nabla v\|_{0,K}^2\}\right)^{1/2}.$$

Using standard arguments (cf. [3, Chapter III]), we deduce that there exist constants μ_1, μ_2 independent of h such that

$$\|\Delta\, v_h\|_{0,G} \le \mu_1\, h_K^{-1}\, |v_h|_{1,G} \qquad \forall\; v_h \in \mathrm{V}_h,\, G \in \mathcal{G}_K,\, K \in \mathcal{T}_h\,, \tag{3.3}$$

$$|v_h|_{1,K} \le \mu_2\, h_K^{-1}\, \|v_h\|_{0,K} \qquad \forall\; v_h \in \mathrm{V}_h,\, K \in \mathcal{T}_h\,. \tag{3.4}$$

Assuming that the control parameter δ_K satisfies

$$0 \le \delta_K \le \min\left\{\frac{c_0}{2\,\|c\|^2_{0,\infty,K}}, \frac{h_K^2}{2\,\varepsilon\,\mu_1^2}\right\}\,, \tag{3.5}$$

one can prove (cf. [12]) that the bilinear form a_h^{skew} is coercive, i.e.,

$$a_h^{skew}(v_h, v_h) \ge \frac{1}{2}\,|||v_h|||^2 \qquad \forall\; v_h \in \mathrm{V}_h\,. \tag{3.6}$$

This implies that the discrete problem (3.2) has a unique solution and that this solution does not depend on the choice of the extension $\widetilde{u}_b$ of u_b (cf. also Remark 5.2).

REMARK 3.1. *We admit $\delta_K = 0$ in (3.5) since the streamline diffusion stabilization is important in convection dominated regions only.*

4. Error analysis. If the weak solution of (1.1) satisfies $u \in H^2(\Omega)$, then it fulfills (1.1) almost everywhere in Ω. Multiplying (1.1) by $v_h \in \mathrm{V}_h$ and integrating by parts, we infer that

$$a_h^{skew}(u, v_h) = l_h(v_h) + r_h^d(u, v_h) + r_h^s(u, v_h) \qquad \forall\; v_h \in \mathrm{V}_h\,, \tag{4.1}$$

where the consistency errors r_h^d and r_h^s are given by

$$r_h^d(u, v_h) = \varepsilon \sum_{K \in \mathcal{T}_h} \int_{\partial K} \frac{\partial u}{\partial \boldsymbol{n}_{\partial K}}\, v_h \,\mathrm{d}\sigma = \varepsilon \sum_{E \in \mathcal{E}_h} \int_E \frac{\partial u}{\partial \boldsymbol{n}_E}\, [|v_h|]_E \,\mathrm{d}\sigma\,,$$

$$r_h^s(u, v_h) = -\frac{1}{2} \sum_{K \in \mathcal{T}_h} \int_{\partial K} (\boldsymbol{b} \cdot \boldsymbol{n}_{\partial K})\, u\, v_h \,\mathrm{d}\sigma = -\frac{1}{2} \sum_{E \in \mathcal{E}_h} \int_E (\boldsymbol{b} \cdot \boldsymbol{n}_E)\, u\, [|v_h|]_E \,\mathrm{d}\sigma$$

with $\boldsymbol{n}_{\partial K}$ denoting the unit outer normal vector to the boundary of K. For estimating the consistency errors, we shall use the following lemma.

LEMMA 4.1. *For any edge $E \in \mathcal{E}_h$ and any integer $k \ge 0$, let $\mathcal{M}_E^k$ be the projection operator from $L^2(E)$ onto $P_k(E)$ defined by*

$$\int_E q\, \mathcal{M}_E^k\, v \,\mathrm{d}\sigma = \int_E q\, v \,\mathrm{d}\sigma \qquad \forall\; q \in P_k(E),\, v \in L^2(E)\,.$$

Then there exists a constant C independent of E and h such that

$$\left|\int_E \varphi\, (v - \mathcal{M}_E^k\, v) \,\mathrm{d}\sigma\right| \le C\, h_E^{k+1}\, |\varphi|_{1,K}\, |v|_{k+1,K} \tag{4.2}$$

for all $K \in \mathcal{T}_h$, $E \subset \partial K$, $\varphi \in H^1(K)$, and $v \in H^{k+1}(K)$.

Proof. See [4, Lemma 3]. □

Now we are in a position to prove a convergence result for the discrete problem (3.2).

Theorem 4.2. *Let the assumptions* (3.1) *and* (3.5) *be fulfilled, and let the space* V_h *satisfy the patch test of order* $k+1$*, i.e.,*

$$\int_E [|v_h|]_E\, q\, \mathrm{d}\sigma = 0 \qquad \forall\ v_h \in \mathrm{V}_h,\ q \in P_k(E),\ E \in \mathcal{E}_h\,, \tag{4.3}$$

where $k \geq 0$ *is a given integer. Let the weak solution of* (1.1) *belong to* $H^m(\Omega)$*, let* $m = \max\{2, k+1\}$*, and let* $\boldsymbol{b} \in W^{k+1,\infty}(\Omega)^2$*. Then the discrete solution* u_h *satisfies*

$$|||u-u_h||| \leq C\,h \left(\sum_{K\in\mathcal{T}_h} \gamma_K\, |u|^2_{2,K}\right)^{1/2} + C\,h^k \left(\sum_{E\in\mathcal{E}_h} \gamma_E\, \|u\|^2_{m,S_E}\right)^{1/2}, \tag{4.4}$$

where

$$\gamma_K = \varepsilon + h_K^2 + \delta_K + (\max\{\varepsilon,\delta_K\})^{-1}\, h_K^2\,, \qquad \gamma_E = \min\left\{\frac{h_E^2}{\varepsilon}, 1\right\}.$$

Proof. Denoting $w = i_h u - u$ and $w_h = i_h u - u_h$, we have $w_h \in \mathrm{V}_h$ and it follows from (3.2) and (4.1) that

$$a_h^{skew}(w_h, v_h) = a_h^{skew}(w, v_h) + r_h^d(u, v_h) + r_h^s(u, v_h) \qquad \forall\ v_h \in \mathrm{V}_h\,. \tag{4.5}$$

Integrating by parts, we obtain for any $v_h \in \mathrm{V}_h$

$$a_h^s(w, v_h) = -a_h^c(v_h, w) - (\operatorname{div}\boldsymbol{b}, w\,v_h) + n_h^s(w, v_h)\,,$$

where

$$n_h^s(w, v_h) = \frac{1}{2} \sum_{E\in\mathcal{E}_h} \int_E (\boldsymbol{b}\cdot\boldsymbol{n}_E)\, w\, [|v_h|]_E\, \mathrm{d}\sigma\,.$$

Hence denoting

$$a_h(w, v_h) = a_h^d(w, v_h) - a_h^c(v_h, w) + (c - \operatorname{div}\boldsymbol{b}, w\,v_h) + a_h^{sd}(w, v_h)\,,$$

we have

$$a_h^{skew}(w, v_h) = a_h(w, v_h) + n_h^s(w, v_h)\,. \tag{4.6}$$

Combining (4.5), (4.6), (3.6), and the triangular inequality, we infer that

$$\begin{aligned}\frac{1}{2}\,|||u-u_h||| \leq \frac{1}{2}\,|||w||| &+ \sup_{v_h\in\mathrm{V}_h} \frac{a_h(w, v_h)}{|||v_h|||} \\ &+ \sup_{v_h\in\mathrm{V}_h} \frac{n_h^s(w, v_h)}{|||v_h|||} + \sup_{v_h\in\mathrm{V}_h} \frac{r_h^d(u, v_h)}{|||v_h|||} + \sup_{v_h\in\mathrm{V}_h} \frac{r_h^s(u, v_h)}{|||v_h|||}\,.\end{aligned}$$

The first two terms on the right-hand side are well known from the conforming analysis of the problem (3.2) (cf., e.g., [15]) and can be estimated by

$$\frac{1}{2}\,|||w||| + \sup_{v_h\in\mathrm{V}_h} \frac{a_h(w, v_h)}{|||v_h|||} \leq C\,h \left(\sum_{K\in\mathcal{T}_h} \gamma_K\, |u|^2_{2,K}\right)^{1/2}. \tag{4.7}$$

The remaining three terms are purely nonconforming terms. The estimation of $r_h^d(u,v_h)$ is the easiest one: In view of (4.3), we have for any $E \in \mathcal{E}_h$

$$\int_E \frac{\partial u}{\partial \boldsymbol{n}_E}\,[|v_h|]_E\,\mathrm{d}\sigma = \int_E \left(\frac{\partial u}{\partial \boldsymbol{n}_E} - \mathcal{M}_E^0\,\frac{\partial u}{\partial \boldsymbol{n}_E}\right)[|v_h|]_E\,\mathrm{d}\sigma,$$

and hence, applying (4.2), we deduce that

$$r_h^d(u,v_h) \le C\,h\,\varepsilon^{1/2}\,|u|_{2,\Omega}\,|||v_h|||\,.$$

To estimate $r_h^s(u,v_h)$, we apply (4.3) and Lemma 4.1, and we obtain

$$\begin{aligned}\int_E (\boldsymbol{b}\cdot\boldsymbol{n}_E)\,u\,[|v_h|]_E\,\mathrm{d}\sigma &= \int_E [(\boldsymbol{b}\cdot\boldsymbol{n}_E)\,u - \mathcal{M}_E^k((\boldsymbol{b}\cdot\boldsymbol{n}_E)\,u)]\,[|v_h|]_E\,\mathrm{d}\sigma\\ &\le C\,h_E^{k+1}\,\|u\|_{k+1,S_E}\,|v_h|_{1,S_E}\,,\end{aligned}$$

where the norms over S_E are considered to be defined elementwise. Using (3.4), we derive

$$\int_E (\boldsymbol{b}\cdot\boldsymbol{n}_E)\,u\,[|v_h|]_E\,\mathrm{d}\sigma \le C\,h_E^k\,\|u\|_{k+1,S_E}\,\gamma_E^{1/2}\,(\varepsilon\,|v_h|_{1,S_E}^2 + c_0\,\|v_h\|_{0,S_E}^2)^{1/2}\,,$$

which implies that

$$r_h^s(u,v_h) \le C\,h^k\left(\sum_{E\in\mathcal{E}_h}\gamma_E\,\|u\|_{k+1,S_E}^2\right)^{1/2}|||v_h|||\,.$$

The term $n_h^s(w,v_h)$ can be estimated analogously. The only difference is that we also use the estimate $\|w\|_{k+1,S_E} \le C\,h_E\,|u|_{2,S_E} + \min\{1,k\}\,\|u\|_{k+1,S_E}$. So, we get

$$n_h^s(w,v_h) \le C\,h^{\max\{1,k\}}\left(\sum_{E\in\mathcal{E}_h}\gamma_E\,\|u\|_{m,S_E}^2\right)^{1/2}|||v_h|||\,. \tag{4.8}$$

As we see, for $k=0$, the consistency error $r_h^s(u,v_h)$ behaves worse than the term $n_h^s(w,v_h)$ and does not allow any ε-uniform convergence. Summing up all the estimates, we obtain the theorem. □

REMARK 4.1. *The above estimate together with the condition* (3.5) *suggests setting*

$$\delta_K = \begin{cases}\kappa_K\,h_K & \textit{if } h_K > \varepsilon,\\ 0 & \textit{if } h_K \le \varepsilon,\end{cases}$$

where κ_K is bounded independently of h and satisfies

$$0 < \kappa_0 \le \kappa_K \le \min\left\{\frac{c_0}{2\,\|c\|_{0,\infty,K}^2\,h_K},\frac{h_K}{2\,\varepsilon\,\mu_1^2}\right\}\,.$$

Then $(\max\{\varepsilon,\delta_K\})^{-1}\,h_K^2 \le (\min\{1,\kappa_0\})^{-1}\,h_K$*, and hence* $\gamma_K \le C\,(\varepsilon + h_K)$*.*

Let us consider the convection dominated case $\varepsilon \leq h$, and let δ_K be defined as in Remark 4.1, which implies that $\gamma_K \leq C\,h$. Since the sum over edges in (4.4) stems from the nonconformity only, we obtain for $\mathrm{V}_h = \mathrm{V}_h^{conf}$ the well-known estimate

$$|||u - u_h||| \leq C\,h^{3/2}\,|u|_{2,\Omega}\,,$$

where the constant C is independent of u, h, and ε. Therefore, the estimate is called ε-uniform. It is known that this estimate is optimal on general meshes.

For a general nonconforming space V_h satisfying the assumptions of Theorem 4.2, the estimate (4.4) leads to the ε-uniform estimate

$$|||u - u_h||| \leq C\,h^{3/2}\,|u|_{2,\Omega} + C\,h^k\|u\|_{\max\{2,k+1\},\Omega}\,. \tag{4.9}$$

Thus, if we use the space $\mathrm{V}_h = \mathrm{V}_h^{nc}$, which satisfies (4.3) for $k = 0$ only, the ε-uniform convergence order is 0. Numerical experiments really confirm this pessimistic prediction (see section 7), which suggests that it is generally a property of the method and not a consequence of an inaccurate estimation. On the other hand, the estimate (4.9) shows that the optimal ε-uniform convergence order 3/2 can be recovered if the space V_h satisfies the patch test of order 3, i.e., $k = 2$. This is an unusual requirement for a nonconforming first order finite element space, but we shall show in section 5 that such spaces can easily be constructed.

4.1. Remarks on the convective discretization. In numerical computations, one also often considers the discrete problem (3.2) with a_h^{skew} replaced by the convective bilinear form a_h^{conv} defined by

$$a_h^{conv}(u,v) = a_h^d(u,v) + a_h^c(u,v) + (c\,u,v) + a_h^{sd}(u,v)\,. \tag{4.10}$$

Note that a result similar to (3.6) does not hold for this bilinear form. Indeed,

$$a_h^{conv}(v_h,v_h) \geq \frac{1}{2}\,|||v_h|||^2 + \frac{1}{2}\sum_{E\in\mathcal{E}_h}\int_E (\boldsymbol{b}\cdot\boldsymbol{n}_E)\,[|v_h^2|]_E\,\mathrm{d}\sigma \qquad \forall\ v_h \in \mathrm{V}_h\,,$$

where the additional term is of order $O(\|v_h\|_{0,\Omega}^2/h)$ in general (cf. [17]). Of course, the coercivity is not necessary to prove the unique solvability and to establish error estimates. It would be sufficient if an inf-sup condition were satisfied, precisely, if the constants

$$\alpha_h = \inf_{w_h\in V_h^{nc}}\ \sup_{v_h\in V_h^{nc}} \frac{a_h^{conv}(w_h,v_h)}{|||v_h|||\ |||w_h|||} \tag{4.11}$$

could be bounded from below by some positive constant independent of h or at least with a known dependence on h. Unfortunately, this is an open problem.

Let us consider the discrete problem (3.2) with a_h^{skew} replaced by a_h^{conv}. We again set $w = i_h u - u$ and $w_h = i_h u - u_h$. To estimate the error $u - u_h = w_h - w$ it suffices to investigate w_h since w can be estimated by (4.7). Since there is no consistency error induced by the convective term, we obtain

$$\alpha_h\,|||w_h||| \leq \sup_{v_h\in\mathrm{V}_h}\frac{a_h(w,v_h)}{|||v_h|||} + 2\sup_{v_h\in\mathrm{V}_h}\frac{n_h^s(w,v_h)}{|||v_h|||} + \sup_{v_h\in\mathrm{V}_h}\frac{r_h^d(u,v_h)}{|||v_h|||}\,.$$

Hence $\alpha_h\,|||w_h|||$ can be estimated by the right-hand side of (4.4). However, for $k = 0$, we can apply (4.8) and hence, for δ_K defined as in Remark 4.1, we always get at least

$$|||u - i_h u||| + \alpha_h\,|||i_h u - u_h||| \leq C\,h\,|u|_{2,\Omega}\,.$$

Moreover, for $\mathrm{V}_h = \mathrm{V}_h^{nc}$ and $u \in H^3(\Omega)$, we have the estimate

$$\int_E (\boldsymbol{b} \cdot \boldsymbol{n}_E)\, w\, [|v_h|]_E \,\mathrm{d}\sigma \le C\, h_E^3\, \|u\|_{3,S_E}\, |v_h|_{1,S_E} \tag{4.12}$$

so that in the convection dominated case $\varepsilon \le h$ we even obtain

$$|||u - i_h u||| + \alpha_h\, |||i_h u - u_h||| \le C\, h^{3/2}\, \|u\|_{3,\Omega}\,.$$

Let us mention how to prove (4.12). We denote by j_h the piecewise quadratic Lagrange interpolation operator and, for any edge $E \in \mathcal{E}_h$, we set $\boldsymbol{b}_E = \mathcal{M}_E^0\, \boldsymbol{b}$. Then we have for any $E \in \mathcal{E}_h$

$$\begin{aligned}\int_E (\boldsymbol{b} \cdot \boldsymbol{n}_E)\, w\, [|v_h|]_E \,\mathrm{d}\sigma &= \int_E ((\boldsymbol{b} - \boldsymbol{b}_E) \cdot \boldsymbol{n}_E)\, w\, [|v_h|]_E \,\mathrm{d}\sigma \\ &+ \int_E (\boldsymbol{b}_E \cdot \boldsymbol{n}_E)\, (j_h u - u)\, [|v_h|]_E \,\mathrm{d}\sigma + \int_E (\boldsymbol{b}_E \cdot \boldsymbol{n}_E)\, (i_h u - j_h u)\, [|v_h|]_E \,\mathrm{d}\sigma\,.\end{aligned} \tag{4.13}$$

The last term on the right-hand side vanishes since $i_h u - j_h u$ is even on E and $[|v_h|]_E$ is odd on E. Using Lemma 4.1, we derive for any $z \in H^1(\Omega)$

$$\int_E z\, [|v_h|]_E \,\mathrm{d}\sigma = \int_E (z - \mathcal{M}_E^0\, z)\, [|v_h|]_E \,\mathrm{d}\sigma \le C\, h_E\, |z|_{1,S_E}\, |v_h|_{1,S_E}\,.$$

This implies that the first two terms on the right-hand side of (4.13) can be estimated by $C\, h_E^3\, (|u|_{2,S_E} + |u|_{3,S_E})\, |v_h|_{1,S_E}$, which proves (4.12).

The above considerations suggest that, in some cases, the bilinear form a_h^{conv} may lead to better results than a_h^{skew}, particularly in the case that $\alpha_h \ge \alpha_0 > 0$ could be verified.

5. Definition of the P_1^{mod} element.

We have seen above that it is desirable to construct nonconforming first order finite element spaces satisfying the patch test of a higher order than usual. In this section, we present a possible way of constructing such spaces. The idea is to enrich the space V_h^{nc} by suitable supplementary functions and then to restrict the enlarged space to its subspace of functions satisfying the patch test of a given order. Our basic requirement is that this procedure must not destroy the edge-oriented structure of the space V_h^{nc}. This construction will lead to a new finite element space containing as a subspace modified functions from V_h^{nc}. Therefore, we denote the new space V_h^{mod}, and we call the corresponding finite element the P_1^{mod} element.

We introduce the P_1^{mod} element by describing the respective shape functions on the standard reference triangle $\widehat{K}$. It turns out that independently of the required order of the patch test it suffices to enrich the space $P_1(\widehat{K})$ corresponding to V_h^{nc} by three functions $\widehat{b}_1$, $\widehat{b}_2$, and $\widehat{b}_3$ associated, respectively, with the edges $\widehat{E}_1$, $\widehat{E}_2$, and $\widehat{E}_3$ of the element $\widehat{K}$. This gives the space

$$P_1^{mod}(\widehat{K}) = P_1(\widehat{K}) \oplus \mathrm{span}\{\widehat{b}_1, \widehat{b}_2, \widehat{b}_3\}\,.$$

We assume for $i \in \{1, 2, 3\}$ that

$$\widehat{b}_i \in H^1(\widehat{K})\,, \qquad \widehat{b}_i|_{\partial\widehat{K}\setminus\widehat{E}_i} = 0\,, \tag{5.1}$$

$$\widehat{b}_i|_{\widehat{E}_i} \text{ is odd with respect to the midpoint of } \widehat{E}_i, \tag{5.2}$$

$$\int_{\widehat{E}_i} [(1 - 2\,\widehat{\lambda}_{i+1}) + \widehat{b}_i]\, \widehat{q} \,\mathrm{d}\widehat{\sigma} = 0 \qquad \forall\, \widehat{q} \in P_1(\widehat{E}_i)\,, \tag{5.3}$$

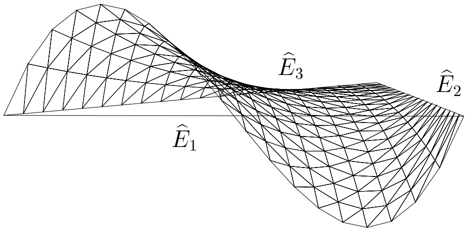


FIG. 5.1. *Function* $\widehat{\lambda}_2^2\,\widehat{\lambda}_3 - \widehat{\lambda}_2\,\widehat{\lambda}_3^2$.

where $\widehat{\lambda}_i$ is the barycentric coordinate on $\widehat{K}$ with respect to the vertex of $\widehat{K}$ opposite the edge $\widehat{E}_i$. (We set $\widehat{\lambda}_4 \equiv \widehat{\lambda}_1$.) In addition, because of the streamline diffusion method, we suppose that on the triangulation $\widehat{\mathcal{G}}$ of $\widehat{K}$

$$\widehat{b}_i|_{\widehat{G}} \in H^2(\widehat{G}) \qquad \forall\, \widehat{G} \in \widehat{\mathcal{G}}\,. \tag{5.4}$$

Note that to verify (5.3), it suffices to prove its validity for $\widehat{q} = \widehat{\lambda}_{i+1}|_{\widehat{E}_i}$. A simple example of $\widehat{b}_i$ satisfying the assumptions (5.1)–(5.4) is the function (cf. Figure 5.1)

$$\widehat{b}_i = 10\,(\widehat{\lambda}_{i+1}^2\,\widehat{\lambda}_{i+2} - \widehat{\lambda}_{i+1}\,\widehat{\lambda}_{i+2}^2)\,, \tag{5.5}$$

where the indices are to be considered modulo 3.

For any element $K \in \mathcal{T}_h$, we introduce a regular affine mapping $F_K : \widehat{K} \to K$ such that $F_K(\widehat{K}) = K$ and, using this mapping, we transform the shape functions from $\widehat{K}$ onto K. In this way, we obtain the spaces

$$P_1^{mod}(K) = P_1(K) \oplus \operatorname{span}\{b_{K,E}|_K\}_{E\in\mathcal{E}_h,\, E\subset\partial K}\,,$$

where

$$b_{K,E} = \begin{cases} \widehat{b}_i \circ F_K^{-1} & \text{in } K, \\ 0 & \text{in } \Omega \setminus K \end{cases}$$

for $E = F_K(\widehat{E}_i)$, $i = 1, 2, 3$. For each element K, we introduce six local nodal functionals

$$I_{K,E}(v) = \frac{1}{h_E}\int_E v\,\mathrm{d}\sigma\,, \quad J_{K,E}(v) = \frac{3}{h_E}\int_E v\,(2\,\lambda_E - 1)\,\mathrm{d}\sigma\,, \quad E \in \mathcal{E}_h,\, E \subset \partial K\,,$$

where $\lambda_E \in P_1(E)$ equals 1 at one endpoint of E and 0 at the other endpoint of E. It is easy to verify that these functionals are unisolvent with the space $P_1^{mod}(K)$. Of course, we could also use other local nodal functionals. However, we prefer the above functionals since they lead to dual basis functions having nice properties.

Now, the finite element space V_h^{mod} approximating the space $H_0^1(\Omega)$ is defined in a standard way: It consists of all functions which belong to the space $P_1^{mod}(K)$ on any element $K \in \mathcal{T}_h$, which are continuous on all inner edges in the sense of the equality of nodal functionals and for which all nodal functionals associated with boundary edges vanish. This means that

$$\mathrm{V}_h^{mod} = \Bigg\{ v_h \in L^2(\Omega)\,;\; v_h|_K \in P_1^{mod}(K) \;\; \forall\, K \in \mathcal{T}_h\,, \qquad \int_E [|v_h|]_E\, q\,\mathrm{d}\sigma = 0 \;\; \forall\, q \in P_1(E),\, E \in \mathcal{E}_h \Bigg\}\,.$$

For any inner edge $E \in \mathcal{E}_h^i$, we define global nodal functionals

$$I_E(v) = I_{K,E}(v)\,, \qquad\qquad J_E(v) = J_{K,E}(v)\,,$$

where K is any element adjacent to E. (Note that, for $v \in \mathrm{V}_h^{mod}$, the choice of K has no influence on the values of $I_{K,E}(v)$ and $J_{K,E}(v)$.) We denote by $\{\psi_E, \chi_E\}_{E\in\mathcal{E}_h^i}$ a basis of V_h^{mod} which is dual to the functionals I_E, J_E; i.e., for any $E, E' \in \mathcal{E}_h^i$, we have

$$I_E(\psi_{E'}) = \delta_{E,E'}\,, \qquad J_E(\psi_{E'}) = 0\,, \qquad I_E(\chi_{E'}) = 0\,, \qquad J_E(\chi_{E'}) = \delta_{E,E'}\,,$$

where $\delta_{E,E'} = 1$ for $E = E'$ and $\delta_{E,E'} = 0$ for $E \neq E'$. To establish formulas for ψ_E and χ_E, we denote by K, $\widetilde{K}$ the two elements adjacent to E; by E, E_1, E_2 the edges of K; by E, E_3, E_4 the edges of $\widetilde{K}$; and by ζ_E the standard basis function of V_h^{nc} associated with the edge E (i.e., ζ_E is piecewise linear, equals 1 on E, and vanishes at the midpoints of all edges different from E). Then

$$\psi_E = \zeta_E + \beta_{E,1}\, b_{K,E_1} + \beta_{E,2}\, b_{K,E_2} + \beta_{E,3}\, b_{\widetilde{K},E_3} + \beta_{E,4}\, b_{\widetilde{K},E_4}\,, \tag{5.6}$$

$$\chi_E = \beta_{E,5}\, b_{K,E} + \beta_{E,6}\, b_{\widetilde{K},E}\,, \tag{5.7}$$

where the coefficients $\beta_{E,1}, \dots, \beta_{E,6}$ are uniquely determined and equal 1 or -1. If the functions $\widehat{b}_1, \widehat{b}_2, \widehat{b}_3$ are chosen in a suitable way (e.g., $\widehat{b}_i = \widehat{b}_1 \circ \widehat{F}_i$, where $\widehat{F}_i$ is an affine transformation of $\widehat{K}$ onto $\widehat{K}$), then $\chi_E \in H_0^1(\Omega)$, and hence the functions χ_E generate a conforming subspace of V_h^{mod}. (This is also the case for the functions χ_E presented in subsection 5.2 below.) The functions ψ_E are always purely nonconforming functions since they have jumps across the edges $E_1, \dots, E_4$, and they can be viewed as modified basis functions of V_h^{nc}. In addition, from (5.6) and (5.7), it follows that, for any $v_h \in \mathrm{V}_h^{mod}$ and any $E \in \mathcal{E}_h$, the jump $[|v_h|]_E$ is odd with respect to the midpoint of E. Therefore,

$$\int_E [|v_h|]_E\, q\, \mathrm{d}\sigma = 0 \tag{5.8}$$

for any even function $q \in L^1(E)$. Particularly, (5.8) holds for any $q \in P_2(E)$ vanishing at the endpoints of E. This together with the definition of V_h^{mod} implies that (5.8) holds for any $q \in P_2(E)$; i.e., the space V_h^{mod} satisfies the patch test of order 3. Moreover, if (5.3) holds for any $\widehat{q} \in P_k(\widehat{E}_i)$ with some $k > 1$, then it is easy to show that the basis functions ψ_E and χ_E satisfy the patch test of order $k+1$. Consequently, the whole space V_h^{mod} then satisfies the patch test of order at least $k+1$.

Let us mention that, denoting

$$\mathrm{B}_h = \operatorname{span}\{b_{K,E}\}_{K\in\mathcal{T}_h,\, E\in\mathcal{E}_h,\, E\subset\partial K}\,, \tag{5.9}$$

the space V_h^{mod} can also be written as

$$\mathrm{V}_h^{mod} = \left\{ v_h \in \mathrm{V}_h^{nc} \oplus \mathrm{B}_h\,;\; \int_E [|v_h|]_E\, q\, \mathrm{d}\sigma = 0 \;\; \forall\, q \in P_2(E),\, E \in \mathcal{E}_h \right\}.$$

Therefore, the space V_h^{mod} can be regarded as the space V_h^{nc} enriched by the nonconforming bubble functions $b_{K,E}$ and then restricted to the subspace of functions satisfying the patch test of order 3.

5.1. Properties of the modified method. As we required at the beginning, the space V_h^{mod} is an edge-oriented nonconforming finite element space possessing first order approximation properties with respect to $|\cdot|_{1,h}$. The supports of the basis functions ψ_E, χ_E are contained in the supports of the basis functions ζ_E of V_h^{nc}, and hence the space V_h^{mod} can be implemented using the same data structures as the space V_h^{nc}. In addition, owing to (5.4), the space V_h^{mod} consists of piecewise continuous functions which are continuous in the midpoints of inner edges and vanish in the midpoints of boundary edges. This is a further feature common with the space V_h^{nc}.

However, as we have shown above, there is an immense difference in the behavior of the solutions to the discrete problem (3.2) for these two spaces: Whereas no ε-uniform convergence can be shown for the space V_h^{nc}, the space V_h^{mod} guarantees the ε-uniform estimate (cf. (4.9))

$$|||u - u_h||| \le C\, h^{\min\{l,3/2\}}\, \|u\|_{l+1,\Omega}\,, \qquad l = 1, 2\,. \tag{5.10}$$

Thus, for $u \in H^3(\Omega)$, we get the optimal ε-uniform convergence order 3/2. Moreover, numerical tests indicate that discretizations using the space V_h^{mod} are much more accurate than those ones using the space V_h^{nc} (cf. section 7).

The price we pay for the ε-uniform estimate (5.10) is that $\dim \mathrm{V}_h^{mod} = 2 \dim \mathrm{V}_h^{nc}$ and that, consequently, the stiffness matrix corresponding to V_h^{mod} is generally four times larger than the one corresponding to V_h^{nc}. However, this does not mean that using the space V_h^{mod} is more expensive than using the space V_h^{nc} since typically a prescribed accuracy can be attained with the space V_h^{mod} on much coarser meshes than with the space V_h^{nc}.

The number of nonzero entries of the stiffness matrix corresponding to the space V_h^{mod} can be reduced to about 80 % by using functions $\widehat{b}_1, \widehat{b}_2, \widehat{b}_3$ with disjoint interiors of their supports (cf. Remark 5.1 below). In this case, the functions χ_E can be easily eliminated from the discrete problem by static condensation. That halves the number of unknowns and reduces the number of nonzero entries to about 65 %.

The dimension of the space V_h^{mod} is asymptotically the same as for the nonconforming piecewise quadratic element [7]. Since this element has second order approximation properties with respect to $|\cdot|_{1,h}$ one would expect a faster convergence than for the P_1^{mod} element. However, the element of [7] satisfies the patch test of order 2 only, and hence the corresponding consistency error tends to zero with the ε-uniform convergence order 1 (cf. the second term in (4.4)). Consequently, the ε-uniform convergence order of the discrete solution is at most 1 in the convection dominated case, whereas we have 3/2 for the P_1^{mod} element. Note also that the P_1^{mod} element is more suitable for a parallel implementation than the element of [7].

Remark 5.1. *Functions $\widehat{b}_1, \widehat{b}_2, \widehat{b}_3$ with disjoint interiors of their supports mentioned above can be obtained in the following way. We divide the reference triangle $\widehat{K}$ into three subtriangles by connecting the barycenter of $\widehat{K}$ with the vertices of $\widehat{K}$ and denote by $\widehat{K}_i$ the subtriangle adjacent to the edge $\widehat{E}_i$, $i = 1, 2, 3$. Then we require that $\widehat{b}_i$ vanishes outside the subtriangle $\widehat{K}_i$. On $\widehat{K}_i$, the function $\widehat{b}_i$ can be defined, e.g., by* (5.5)*, where $\widehat{\lambda}_{i+1}$ and $\widehat{\lambda}_{i+2}$ are now considered as barycentric coordinates on $\widehat{K}_i$ with respect to the endpoints of $\widehat{E}_i$. If we set $\widehat{\mathcal{G}} = \{\widehat{K}_1, \widehat{K}_2, \widehat{K}_3\}$, then all the assumptions on $\widehat{b}_i$ made above are satisfied. Note that generally $\widehat{b}_i \notin H^2(\widehat{K})$ so that the assumption that finite element functions are piecewise H^2 only with respect to a subdivision of $\mathcal{T}_h$ really has a practical importance.*

Remark 5.2. *In the discrete problem* (3.2)*, inhomogenous Dirichlet boundary conditions are represented by the condition $u_h - i_h\widetilde{u}_b \in \mathrm{V}_h$. This is equivalent to*

$u_h - \widetilde{u}_{bh} \in \mathrm{V}_h$, where $\widetilde{u}_{bh} \in H^{1,h}(\Omega)$ is any function satisfying $\widetilde{u}_{bh} - i_h\widetilde{u}_b \in \mathrm{V}_h$. Now let us consider the P_1^{mod} element. If we extend the definitions of the global nodal functionals I_E, J_E and the basis functions ψ_E, χ_E to boundary edges, then

$$i_h\widetilde{u}_b = \sum_{E\in\mathcal{E}_h} I_E(i_h\widetilde{u}_b)\,\psi_E + J_E(i_h\widetilde{u}_b)\,\chi_E\,.$$

Thus, the inhomogenous Dirichlet boundary conditions can be implemented by setting

$$\widetilde{u}_{bh} = \sum_{E\in\mathcal{E}_h^b} I_E(i_h\widetilde{u}_b)\,\psi_E + J_E(i_h\widetilde{u}_b)\,\chi_E\,.$$

It is easy to see that then $\widetilde{u}_{bh}$ does not depend on the choice of the extension $\widetilde{u}_b$ of u_b.

5.2. Simple representation of the basis functions $\boldsymbol{\psi_E}$ and $\boldsymbol{\chi_E}$. Let us close this section by returning to the example of $\widehat{b}_i$ given in (5.5) and rewriting the formulas (5.6), (5.7) for this particular case. We denote by K and $\widetilde{K}$ the two elements adjacent to an edge $E \in \mathcal{E}_h^i$ and by λ_1, λ_2 and $\widetilde{\lambda}_1$, $\widetilde{\lambda}_2$ the barycentric coordinates on K and $\widetilde{K}$ with respect to the endpoints of E. Further, we respectively denote by λ_3 and $\widetilde{\lambda}_3$ the remaining barycentric coordinates on K and $\widetilde{K}$. Then

$$\psi_E = \begin{cases} 1 - 2\,\lambda_3 - 10\,(\lambda_1^2\,\lambda_3 - \lambda_1\,\lambda_3^2) - 10\,(\lambda_2^2\,\lambda_3 - \lambda_2\,\lambda_3^2) & \text{in } K,\\ 1 - 2\,\widetilde{\lambda}_3 - 10\,(\widetilde{\lambda}_1^2\,\widetilde{\lambda}_3 - \widetilde{\lambda}_1\,\widetilde{\lambda}_3^2) - 10\,(\widetilde{\lambda}_2^2\,\widetilde{\lambda}_3 - \widetilde{\lambda}_2\,\widetilde{\lambda}_3^2) & \text{in } \widetilde{K}\setminus E,\\ 0 & \text{in } \Omega\setminus\{K\cup\widetilde{K}\}, \end{cases}$$

and, after dividing by 10,

$$\chi_E = \begin{cases} \lambda_1^2\,\lambda_2 - \lambda_1\,\lambda_2^2 & \text{in } K,\\ \widetilde{\lambda}_1^2\,\widetilde{\lambda}_2 - \widetilde{\lambda}_1\,\widetilde{\lambda}_2^2 & \text{in } \widetilde{K}\setminus E,\\ 0 & \text{in } \Omega\setminus\{K\cup\widetilde{K}\}. \end{cases}$$

These basis functions were used in the numerical calculations presented in section 7.

6. Convergence of the piecewise linear part $\boldsymbol{u_h^{lin}}$ of $\boldsymbol{u_h}$. Let us consider the discrete problem (3.2) with $\mathrm{V}_h = \mathrm{V}_h^{mod}$. Then the discrete solution u_h belongs to $\widetilde{\mathrm{V}}_h^{nc} \oplus \mathrm{B}_h$, where

$$\widetilde{\mathrm{V}}_h^{nc} = \left\{ v_h \in L^2(\Omega)\,;\; v_h|_K \in P_1(K) \;\;\forall\, K \in \mathcal{T}_h\,,\;\; \int_E [|v_h|]_E\,\mathrm{d}\sigma = 0 \;\;\forall\, E \in \mathcal{E}_h^i \right\}$$

and B_h was defined in (5.9). Thus, u_h can be uniquely decomposed into its piecewise linear part $u_h^{lin} \in \widetilde{\mathrm{V}}_h^{nc}$ and its bubble part $u_h^{bub} \in \mathrm{B}_h$, i.e.,

$$u_h = u_h^{lin} + u_h^{bub}\,.$$

We shall show that u_h^{lin} converges to the weak solution with the same convergence order as u_h. First, let us prove the following orthogonality result.

LEMMA 6.1. *The spaces $\widetilde{\mathrm{V}}_h^{nc}$ and B_h are orthogonal with respect to the $H_0^1(\Omega)$ inner product, i.e.,*

$$\sum_{K\in\mathcal{T}_h} (\nabla v_h, \nabla b_h)_K = 0 \qquad \forall\, v_h \in \widetilde{\mathrm{V}}_h^{nc},\, b_h \in \mathrm{B}_h\,. \tag{6.1}$$

Consequently,

$$|v_h|_{1,h}^2 + |b_h|_{1,h}^2 = |v_h + b_h|_{1,h}^2 \qquad \forall\, v_h \in \widetilde{\mathrm{V}}_h^{nc},\ b_h \in \mathrm{B}_h\,. \tag{6.2}$$

Moreover, for any element $K \in \mathcal{T}_h$ and any $\boldsymbol{a} \in \mathbb{R}^2$, we have

$$\|\boldsymbol{a}\cdot\nabla v_h\|_{0,K}^2 + \|\boldsymbol{a}\cdot\nabla b_h\|_{0,K}^2 = \|\boldsymbol{a}\cdot\nabla(v_h + b_h)\|_{0,K}^2 \qquad \forall\, v_h \in \widetilde{\mathrm{V}}_h^{nc},\ b_h \in \mathrm{B}_h\,. \tag{6.3}$$

Proof. For any $v_h \in \widetilde{\mathrm{V}}_h^{nc}$, $b_h \in \mathrm{B}_h$, $i, j \in \{1, 2\}$, and $K \in \mathcal{T}_h$, we derive

$$\int_K \frac{\partial v_h}{\partial x_i}\,\frac{\partial b_h}{\partial x_j}\,\mathrm{d}x = -\int_K \frac{\partial^2 v_h}{\partial x_i\,\partial x_j}\,b_h\,\mathrm{d}x + \int_{\partial K} \frac{\partial (v_h|_K)}{\partial x_i}\,(\boldsymbol{n}_{\partial K})_j\,b_h|_K\,\mathrm{d}\sigma = 0\,.$$

Hence we obtain (6.1) and also

$$(\boldsymbol{a}\cdot\nabla v_h, \boldsymbol{a}\cdot\nabla b_h)_K = 0 \qquad \forall\, v_h \in \widetilde{\mathrm{V}}_h^{nc},\ b_h \in \mathrm{B}_h,\ K \in \mathcal{T}_h,\ \boldsymbol{a} \in \mathbb{R}^2\,.$$

The validity of (6.2) and (6.3) is then obvious. □

With respect to the $L^2(\Omega)$ norm, an analogous orthogonality result is generally not available. Nevertheless, transforming the functions v_h, b_h onto the reference element and using the equivalence of norms on finite-dimensional spaces, we can prove that

$$\|v_h\|_{0,\Omega} + \|b_h\|_{0,\Omega} \le C\,\|v_h + b_h\|_{0,\Omega} \qquad \forall\, v_h \in \widetilde{\mathrm{V}}_h^{nc},\ b_h \in \mathrm{B}_h\,. \tag{6.4}$$

Let the weak solution of (1.1) belong to $H^2(\Omega)$. Then it follows from (6.2) that

$$|u_h^{lin} - i_h u|_{1,h} \le |u_h - i_h u|_{1,h}, \tag{6.5}$$

and hence, with respect to $|\cdot|_{1,h}$, the function u_h^{lin} approximates the piecewise linear interpolate of u at least as well as u_h. Moreover, we obtain the following result.

Theorem 6.2. *Let $u \in H^2(\Omega)$. Then*

$$|u - u_h^{lin}|_{1,h} \le |u - u_h|_{1,h} + 2\,|u - i_h u|_{1,\Omega}\,, \tag{6.6}$$

$$\|u - u_h^{lin}\|_{0,\Omega} \le C\,\|u - u_h\|_{0,\Omega} + C\,\|u - i_h u\|_{0,\Omega}\,, \tag{6.7}$$

$$|||u - u_h^{lin}||| \le C\left(1 + \max_{K\in\mathcal{T}_h} \delta_K^{1/2}\right)(|||u - u_h||| + |||u - i_h u|||)\,. \tag{6.8}$$

Proof. Inequality (6.6) is a direct consequence of (6.5). Analogously, using (6.4), we get (6.7). To prove (6.8), let us consider any $K \in \mathcal{T}_h$ and any $\boldsymbol{a} \in \mathbb{R}^2$. Applying (6.3), we deduce that

$$\begin{aligned}\|\boldsymbol{b}\cdot\nabla(i_h u - u_h^{lin})\|_{0,K} &\le \|(\boldsymbol{b} - \boldsymbol{a})\cdot\nabla(i_h u - u_h^{lin})\|_{0,K} + \|\boldsymbol{a}\cdot\nabla(i_h u - u_h)\|_{0,\Omega}\\ &\le \|\boldsymbol{b} - \boldsymbol{a}\|_{0,\infty,K}\,(|i_h u - u_h^{lin}|_{1,K} + |i_h u - u_h|_{1,K}) + \|\boldsymbol{b}\cdot\nabla(i_h u - u_h)\|_{0,K}\,.\end{aligned}$$

Since $\inf_{\boldsymbol{a}\in\mathbb{R}^2} \|\boldsymbol{b} - \boldsymbol{a}\|_{0,\infty,K} \le C\,h_K\,|\boldsymbol{b}|_{1,\infty,K}$, it follows from (3.4) that

$$\|\boldsymbol{b}\cdot\nabla(i_h u - u_h^{lin})\|_{0,K} \le C\,(\|i_h u - u_h^{lin}\|_{0,K} + \|i_h u - u_h\|_{0,K}) + \|\boldsymbol{b}\cdot\nabla(i_h u - u_h)\|_{0,K}\,.$$

Now, using (6.6), (6.7), and the triangular inequality, we obtain (6.8). □

The above estimates show that u_h^{lin} converges to the weak solution with the same convergence orders as u_h and that the estimate of Theorem 4.2 remains valid for u_h^{lin}. Therefore, it is possible and for practical reasons sensible to consider the piecewise linear part of u_h as a discrete solution of (1.1).

7. Numerical results. In this section, we present numerical results computed using either the discretization (3.2) or a discretization obtained from (3.2) by replacing a_h^{skew} by a_h^{conv} defined in (4.10). We used the P_1^{nc} element ($\mathrm{V}_h = \mathrm{V}_h^{nc}$) and the P_1^{mod} element ($\mathrm{V}_h = \mathrm{V}_h^{mod}$) defined using $\widehat{b}_i$ given in (5.5). (We considered the basis functions presented in subsection 5.2.) For the P_1^{mod} element, we obtained almost identical results for a_h^{conv} and a_h^{skew}, and therefore we show only results obtained for the following three discretizations: a_h^{conv}/P_1^{nc}, a_h^{skew}/P_1^{nc}, and a_h^{skew}/P_1^{mod}.

The bilinear forms a_h^{skew} and a_h^{conv} were computed exactly, whereas the right-hand side l_h was evaluated using a quadrature formula which is exact for piecewise cubic f. The arising linear systems were solved applying the GMRES method with ILU preconditioning. The computations were terminated if the ratio of the norms of the residuum and the right-hand side was smaller than 10^{-8}.

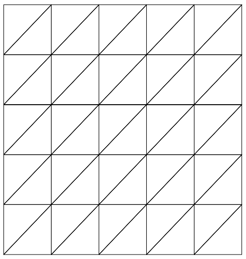


FIG. 7.1. *Type of triangulations used in numerical computations.*

All presented computational results were obtained for $\Omega = (0,1)^2$ discretized using Friedrichs–Keller triangulations of the type depicted in Figure 7.1. We present results obtained for $h \doteq 7.07 \cdot 10^{-2}$, $h \doteq 3.54 \cdot 10^{-2}$, $h \doteq 1.77 \cdot 10^{-2}$, and $h \doteq 8.84 \cdot 10^{-3}$, which corresponds to 800, 3200, 12800, and 51200 elements, respectively. The errors of the discrete solutions were measured in the norms $\|\cdot\|_{0,\Omega}$, $|\cdot|_{1,h}$, $|||\cdot|||$ and in the discrete L^∞ norm $\|\cdot\|_{0,\infty,h}$ which is defined as the maximum of the errors in the midpoints of edges. The evaluation of $\|\cdot\|_{0,\Omega}$ (resp., $|\cdot|_{1,h}$) was exact for piecewise quadratic (resp., cubic) functions. For the P_1^{mod} element, we give the errors of the piecewise linear part u_h^{lin} of u_h. (See section 6.) The convergence orders were always computed using values from triangulations with $h \doteq 1.77 \cdot 10^{-2}$ and $h \doteq 8.84 \cdot 10^{-3}$.

The three discretizations were used to solve the convection-diffusion equation (1.1) for three types of solutions specified in Examples 7.1–7.3 below. The parameter δ_K was defined as in Remark 4.1 with $\kappa_K = 1$, $\kappa_K = 0.25$, and $\kappa_K = 0.2$, respectively. Examples 7.1 and 7.2 are the same as in [11] and [12].

EXAMPLE 7.1 (smooth polynomial solution). *Let* $\boldsymbol{b} = (3,2)$, $c = 2$, *and* $u_b = 0$. *For a given* $\varepsilon > 0$, *the right-hand side* f *is chosen such that*

$$u(x,y) = 100\,x^2\,(1-x)^2\,y\,(1-y)\,(1-2\,y)$$

is the exact solution of (1.1); *see Figure* 7.2.

For $\varepsilon = 1$, we observed optimal convergence orders for all three discretizations, and the errors of the discrete solutions were very similar. To investigate whether the methods are ε-uniform, i.e., whether an estimate of the type

$$|||u - u_h||| \le C\,h^\nu\,\|u\|$$

holds with C and ν independent of ε, we considered $\varepsilon = h^\alpha$ for various values of α. Tables 7.1–7.3 show results obtained for $\alpha = 4$. We remark that the values of h

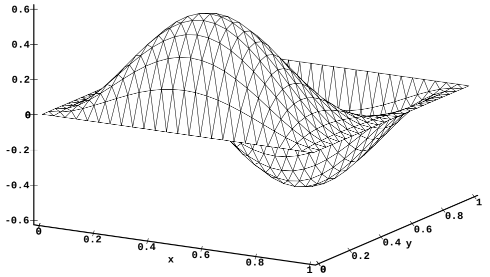


FIG. 7.2. *Exact solution of Example* 7.1.

and ε are rounded in all the tables. The solutions of the discretization a_h^{conv}/P_1^{nc} converge with the optimal order 3/2 in the streamline diffusion norm $|||\cdot|||$, which is in correspondence with subsection 4.1. Note, however, that on unstructured meshes this optimal convergence order was not observed, which indicates that the constants α_h in (4.11) generally cannot be bounded from below by some $\alpha_0 > 0$ independent of h. The influence of the consistency error r_h^s with respect to ε can clearly be seen from Table 7.2: the solution of (3.2) with the P_1^{nc} element does not converge in $|||\cdot|||$, which is in agreement with Theorem 4.2. Table 7.3 shows that the P_1^{mod} element leads to best possible convergence orders which can be expected from a first order finite element space. Particularly, we observe the convergence order 3/2 in the streamline

TABLE 7.1
Example 7.1; *errors for* a_h^{conv} *with the* P_1^{nc} *element and* $\varepsilon = h^4$.

h	ε	$\|\cdot\|_{0,\Omega}$	$\vert\cdot\vert_{1,h}$	$\vert\vert\vert\cdot\vert\vert\vert$	$\|\cdot\|_{0,\infty,h}$
7.07−2	2.50−5	1.49−2	1.40+0	1.43−1	6.87−2
3.54−2	1.56−6	5.86−3	1.09+0	5.10−2	3.88−2
1.77−2	9.77−8	2.07−3	7.57−1	1.80−2	2.20−2
8.84−3	6.10−9	6.94−4	4.98−1	6.36−3	1.20−2
conv. order		1.58	0.60	1.50	0.88

TABLE 7.2
Example 7.1; *errors for* a_h^{skew} *with the* P_1^{nc} *element and* $\varepsilon = h^4$.

h	ε	$\|\cdot\|_{0,\Omega}$	$\vert\cdot\vert_{1,h}$	$\vert\vert\vert\cdot\vert\vert\vert$	$\|\cdot\|_{0,\infty,h}$
7.07−2	2.50−5	4.56−1	4.29+1	7.79−1	1.89+0
3.54−2	1.56−6	4.32−1	8.66+1	7.43−1	1.71+0
1.77−2	9.77−8	4.27−1	1.78+2	7.09−1	1.47+0
8.84−3	6.10−9	4.37−1	3.72+2	6.86−1	1.53+0
conv. order		−0.03	−1.06	0.05	−0.06

TABLE 7.3
Example 7.1; *errors for* a_h^{skew} *with the* P_1^{mod} *element and* $\varepsilon = h^4$.

h	ε	$\|\cdot\|_{0,\Omega}$	$\vert\cdot\vert_{1,h}$	$\vert\vert\vert\cdot\vert\vert\vert$	$\|\cdot\|_{0,\infty,h}$
7.07−2	2.50−5	2.19−3	2.14−1	1.48−1	7.76−3
3.54−2	1.56−6	5.53−4	1.07−1	5.24−2	2.03−3
1.77−2	9.77−8	1.40−4	5.37−2	1.85−2	5.12−4
8.84−3	6.10−9	3.53−5	2.69−2	6.56−3	1.28−4
conv. order		1.99	1.00	1.50	2.00

TABLE 7.4
Example 7.1; *comparison between* a_h^{conv} *with* P_1^{nc} *and* a_h^{skew} *with* P_1^{mod} *for* $h \doteq 8.84 \cdot 10^{-3}$.

ε	$\Vert \cdot \Vert_{0,\Omega}$		$\vert \cdot \vert_{1,h}$		$\vert\vert\vert \cdot \vert\vert\vert$		$\Vert \cdot \Vert_{0,\infty,h}$	
	P_1^{nc}	P_1^{mod}	P_1^{nc}	P_1^{mod}	P_1^{nc}	P_1^{mod}	P_1^{nc}	P_1^{mod}
1−04	4.14−5	3.61−5	2.94−2	2.69−2	6.29−3	6.56−3	1.90−4	1.27−4
1−06	4.83−4	3.52−5	3.46−1	2.69−2	6.33−3	6.56−3	8.31−3	1.28−4
1−08	6.93−4	3.53−5	4.98−1	2.69−2	6.36−3	6.56−3	1.20−2	1.28−4
1−10	6.96−4	3.53−5	5.00−1	2.69−2	6.36−3	6.56−3	1.20−2	1.28−4

diffusion norm, which is again in agreement with our theory in section 4. The convergence orders are better than for a_h^{conv}/P_1^{nc} and a_h^{skew}/P_1^{nc}, and the discrete solutions obtained using the P_1^{mod} element are in all cases more accurate than P_1^{nc} solutions. Table 7.4 shows results obtained for various values of ε on a fixed triangulation. The errors for a_h^{skew}/P_1^{mod} are almost independent of ε in all norms, whereas the errors for a_h^{conv}/P_1^{nc} increase when ε decreases.

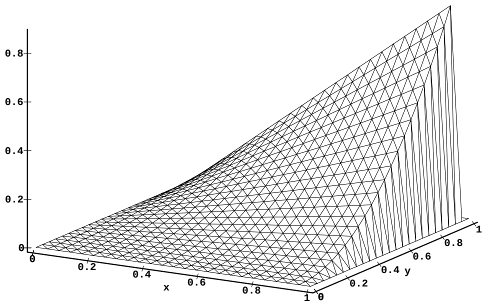


FIG. 7.3. *Exact solution of Example* 7.2.

EXAMPLE 7.2 (layers at the outflow part of the boundary). *Let* $\boldsymbol{b} = (2,3)$ *and* $c = 1$. *For a given* $\varepsilon > 0$, *the right-hand side* f *and the boundary condition* u_b *are chosen such that*

$$u(x,y) = x\,y^2 - y^2 \exp\left(\frac{2\,(x-1)}{\varepsilon}\right) - x \exp\left(\frac{3\,(y-1)}{\varepsilon}\right) + \exp\left(\frac{2\,(x-1)+3\,(y-1)}{\varepsilon}\right)$$

is the exact solution of (1.1). *This function has boundary layers at* $x = 1$ *and* $y = 1$; *see Figure* 7.3.

All three discretizations gave identical errors in $\vert\vert\vert \cdot \vert\vert\vert$ with convergence order 1.00 and in $\vert \cdot \vert_{1,h}$ with convergence order 0.50. The reduction of the convergence order

TABLE 7.5
Example 7.2; *comparison between all three discretizations for* $\varepsilon = 10^{-8}$.

h	$\Vert \cdot \Vert_{0,\Omega}$			$\Vert \cdot \Vert_{0,\infty,h}$		
	a_h^{conv} P_1^{nc}	a_h^{skew} P_1^{nc}	a_h^{skew} P_1^{mod}	a_h^{conv} P_1^{nc}	a_h^{skew} P_1^{nc}	a_h^{skew} P_1^{mod}
7.07−2	1.32+0	7.54−1	8.72−2	9.21+0	3.65+0	6.08−1
3.54−2	1.92+0	8.23−1	6.22−2	1.89+1	4.74+0	6.37−1
1.77−2	2.74+0	8.70−1	4.42−2	3.84+1	5.72+0	6.52−1
8.84−3	3.89+0	8.98−1	3.13−2	7.72+1	6.50+0	6.60−1
order	−0.50	−0.05	0.50	−1.01	−0.18	−0.02

Table 7.6

Example 7.2; *errors for* a_h^{conv} *with the* P_1^{nc} *element and* $\varepsilon = 10^{-8}$.

h	$\|\cdot\|^*_{0,\Omega}$	$\lvert\cdot\rvert^*_{1,h}$	$\lvert\lvert\lvert\cdot\rvert\rvert\rvert^*$	$\|\cdot\|^*_{0,\infty,h}$
7.07−2	2.53−2	2.83+0	2.99−2	1.93−1
3.54−2	9.20−4	2.03−1	2.87−3	9.07−3
1.77−2	9.75−5	4.02−2	9.62−4	2.93−4
8.84−3	2.42−5	1.99−2	3.39−4	7.14−5
order	2.01	1.01	1.50	2.04

Table 7.7

Example 7.2; *errors for* a_h^{skew} *with the* P_1^{nc} *element and* $\varepsilon = 10^{-8}$.

h	$\|\cdot\|^*_{0,\Omega}$	$\lvert\cdot\rvert^*_{1,h}$	$\lvert\lvert\lvert\cdot\rvert\rvert\rvert^*$	$\|\cdot\|^*_{0,\infty,h}$
7.07−2	3.09−1	3.47+1	3.36−1	1.31+0
3.54−2	3.13−1	6.98+1	3.22−1	1.33+0
1.77−2	3.14−1	1.40+2	3.19−1	1.31+0
8.84−3	3.15−1	2.80+2	3.18−1	1.31+0
order	0.00	−1.00	0.00	0.00

Table 7.8

Example 7.2; *errors for* a_h^{skew} *with the* P_1^{mod} *element and* $\varepsilon = 10^{-8}$.

h	$\|\cdot\|^*_{0,\Omega}$	$\lvert\cdot\rvert^*_{1,h}$	$\lvert\lvert\lvert\cdot\rvert\rvert\rvert^*$	$\|\cdot\|^*_{0,\infty,h}$
7.07−2	1.69−3	3.54−2	1.48−2	1.74−2
3.54−2	4.05−5	8.80−3	2.78−3	4.37−4
1.77−2	8.63−6	4.37−3	9.79−4	2.93−5
8.84−3	2.16−6	2.19−3	3.46−4	7.37−6
order	2.00	1.00	1.50	1.99

is caused by the interpolation error in the boundary layer region since the thickness of the layers is smaller than h for all triangulations used. The errors in $\|\cdot\|_{0,\Omega}$ and $\|\cdot\|_{0,\infty,h}$ are shown in Table 7.5. The errors for a_h^{conv}/P_1^{nc} increase for decreasing h and the errors for a_h^{skew}/P_1^{nc} do not change significantly. For a_h^{skew}/P_1^{mod}, the discrete solution converges in $\|\cdot\|_{0,\Omega}$ with order 0.50. Since the boundary layer is not resolved by the mesh, no convergence is observed in the maximum norm.

The streamline diffusion method with conforming finite element approximations is known to approximate solutions with layers on nonlayer-adapted meshes at least outside the layers very precisely. Tables 7.6–7.8 show the behavior of the discrete solutions outside the boundary layers in the domain $\Omega^* = (0, 0.8)^2$. The discretizations a_h^{conv}/P_1^{nc} and a_h^{skew}/P_1^{mod} give optimal convergence orders, but a_h^{skew}/P_1^{mod} is about 10 times more accurate than a_h^{conv}/P_1^{nc} in all norms except for $|||\cdot|||$. Table 7.7 shows that the discretization a_h^{skew}/P_1^{nc} is completely useless.

Example 7.3 (inner and boundary layers). *We set* $\boldsymbol{b} = (1/2, \sqrt{3}/2)$, $c = 0$, $f = 0$, *and*

$$u_b(x,y) = \begin{cases} 0 & \textit{for } x \geq 1/2 \textit{ or } y = 1, \\ 1 & \textit{else.} \end{cases}$$

For $\varepsilon \to 0$, *the solution* u *of* (1.1) *tends to the function*

$$u^0(x,y) = \begin{cases} 0 & \textit{for } y \leq \sqrt{3}\,(x - 1/2), \\ 1 & \textit{else,} \end{cases}$$

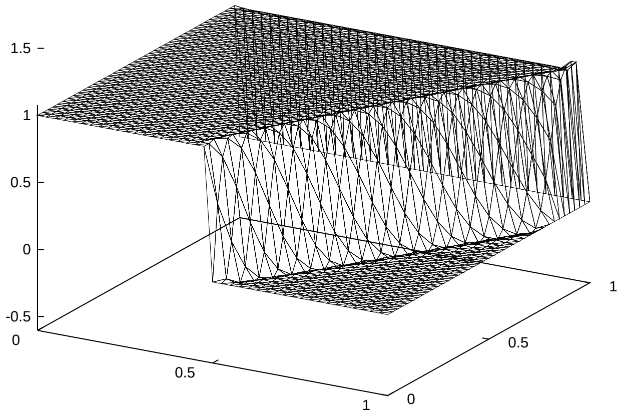


FIG. 7.4. *Solution of Example* 7.3 *for* $h \doteq 3.54 \cdot 10^{-2}$.

which is the solution of the hyperbolic limit of (1.1). *Thus, for small* ε, *the solution* u *has an inner layer along the line* $y = \sqrt{3}\,(x - 1/2)$ *and boundary layers along* $y = 1$ *and* $x = 1$, $y > \sqrt{3}/2$. *We consider* $\varepsilon = 10^{-6}$ *below.*

This example does not fit into the theory presented in this paper, particularly since $u_b \notin H^{3/2}(\partial\Omega)$. However, it is a challenging test case which can indicate the quality of numerical methods for solving (1.1).

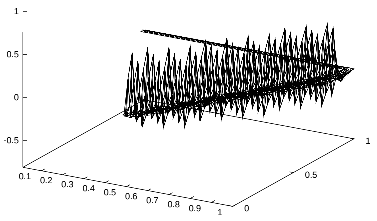


FIG. 7.5. *Example* 7.3; *errors larger than* 0.01 *for* $h \doteq 1.77 \cdot 10^{-2}$.

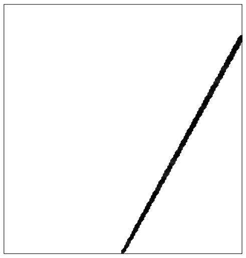


FIG. 7.6. *Example* 7.3; *region of errors larger than* 0.1 *for* $h \doteq 8.84 \cdot 10^{-3}$.

Figures 7.4–7.6 show results computed using the discrete problem (3.2) with the P_1^{mod} element. Instead of showing the discontinuous solution u_h directly, we present a corresponding conforming function $\widetilde{u}_h \in \widetilde{\mathrm{V}}_h^{conf}$ such that the value of $\widetilde{u}_h$ at any inner vertex is equal to the arithmetic mean value of the values of u_h at the midpoints of edges connected with this vertex. The errors of $\widetilde{u}_h$ in Figures 7.5 and 7.6 were computed using the limit solution u^0. We see that inner and boundary layers are detected very well and that the method behaves in a robust way, although the assumptions made in section 1 are not satisfied.

We can conclude that in all numerical tests we have performed, the P_1^{mod} element always led to optimal convergence orders and behaved very robustly with respect to ε. The accuracy of solutions obtained using the P_1^{mod} element was always better than for the P_1^{nc} element and, moreover, the iterative solver used to compute the discrete solutions converged much faster for the P_1^{mod} element than for discretizations using the P_1^{nc} element. Thus, the P_1^{mod} element not only improves the stability of the discrete solution but also the convergence properties of the solver.

REFERENCES

[1] F. Brezzi, D. Marini, and E. Süli, *Residual-free bubbles for advection-diffusion problems: The general error analysis*, Numer. Math., 85 (2000), pp. 31–47.

[2] F. Brezzi, T. J. R. Hughes, L. D. Marini, A. Russo, and E. Süli, *A priori error analysis of residual-free bubbles for advection-diffusion problems*, SIAM J. Numer. Anal., 36 (1999), pp. 1933–1948.

[3] P. G. Ciarlet, *Basic Error Estimates for Elliptic Problems*, in Handbook of Numerical Analysis, Vol. 2—Finite Element Methods (pt. 1), P. G. Ciarlet and J. L. Lions, eds., North–Holland, Amsterdam, 1991, pp. 17–351.

[4] M. Crouzeix and P.-A. Raviart, *Conforming and nonconforming finite element methods for solving the stationary Stokes equations* I, Rev. Franc. Automat. Inform. Rech. Operat., 7 (1973), R-3, pp. 33–76.

[5] O. Dorok, V. John, U. Risch, F. Schieweck, and L. Tobiska, *Parallel finite element methods for the incompressible Navier–Stokes equations*, in Flow Simulation with High-Performance Computers II, Notes Numer. Fluid Mech. 52, E. H. Hirschel, ed., Vieweg–Verlag, Braunschweig, Wiesbaden, 1996, pp. 20–33.

[6] K. Eriksson and C. Johnson, *Adaptive streamline diffusion finite element methods for stationary convection-diffusion problems*, Math. Comp., 60 (1993), pp. 167–188.

[7] M. Fortin and M. Soulie, *A non-conforming piecewise quadratic finite element on triangles*, Internat. J. Numer. Methods Engrg., 19 (1983), pp. 505–520.

[8] T. J. R. Hughes and A. N. Brooks, *A multi-dimensional upwind scheme with no crosswind diffusion*, in Finite Element Methods for Convection Dominated Flows, AMD 34, T. J. R. Hughes, ed., ASME, New York, 1979, pp. 19–35.

[9] V. John, *Parallele Lösung der inkompressiblen Navier–Stokes Gleichungen auf adaptiv verfeinerten Gittern*, Ph.D. thesis, Otto–von–Guericke–Universität Magdeburg, Magdeburg, Germany, 1997.

[10] C. Johnson and J. Saranen, *Streamline diffusion methods for the incompressible Euler and Navier–Stokes equations*, Math. Comp., 47 (1986), pp. 1–18.

[11] V. John, G. Matthies, F. Schieweck, and L. Tobiska, *A streamline-diffusion method for nonconforming finite element approximations applied to convection-diffusion problems*, Comput. Methods Appl. Mech. Engrg., 166 (1998), pp. 85–97.

[12] V. John, J. M. Maubach, and L. Tobiska, *Nonconforming streamline-diffusion-finite-element-methods for convection-diffusion problems*, Numer. Math., 78 (1997), pp. 165–188.

[13] P. Knobloch, *On Korn's inequality for nonconforming finite elements*, Technische Mechanik, 20 (2000), pp. 205–214 and 375 (errata).

[14] K. Nijima, *Pointwise error estimates for a streamline diffusion finite element scheme*, Numer. Math., 56 (1990), pp. 707–719.

[15] H.-G. Roos, M. Stynes, and L. Tobiska, *Numerical Methods for Singularly Perturbed Differential Equations. Convection-Diffusion and Flow Problems*, Springer–Verlag, Berlin, 1996.

[16] F. Schieweck, *Parallele Lösung der stationären inkompressiblen Navier–Stokes Gleichungen*, Habilitationsschrift, Otto–von–Guericke–Universität Magdeburg, Magdeburg, Germany, 1997.

[17] M. Stynes and L. Tobiska, *The streamline-diffusion method for nonconforming Q_1^{rot} elements on rectangular tensor-product meshes*, IMA J. Numer. Anal., 21 (2001), pp. 123–142.

[18] L. Tobiska and R. Verfürth, *Analysis of a streamline diffusion finite element method for the Stokes and Navier–Stokes equations*, SIAM J. Numer. Anal., 33 (1996), pp. 107–127.

SIAM J. NUMER. ANAL.
Vol. 41, No. 2, pp. 457–486

APPROXIMATION OF TIME-DEPENDENT VISCOELASTIC FLUID FLOW: SUPG APPROXIMATION*

VINCENT J. ERVIN† AND WILLIAM W. MILES†

Abstract. In this article we consider the numerical approximation to the time-dependent viscoelasticity equations with an Oldroyd B constitutive equation. The approximation is stabilized by using a streamline upwind Petrov–Galerkin (SUPG) approximation for the constitutive equation. We analyze both the semidiscrete and fully discrete numerical approximations. For both discretizations we prove the existence of, and derive a priori error estimates for, the numerical approximations.

Key words. viscoelasticity, finite element method, fully discrete, SUPG

AMS subject classification. 65N30

PII. S003614290241177X

1. Introduction.

Accurate numerical simulations of time-dependent viscoelastic flows are important to the understanding of many phenomena in non-Newtonian fluid mechanics, particularly those associated with flow instabilities. Aside from [3], previous numerical analysis in this area has been for steady-state flows.

In the case of Newtonian fluid flow, the assumption that the extra stress tensor is proportional to the deformation tensor allows the stress to be eliminated from the modeling equations, giving the Navier–Stokes equations. In viscoelasticity, assuming an Oldroyd B-type fluid, the stress is defined by a (hyperbolic) differential constitutive equation. Very different from computational fluid dynamics simulations, in viscoelasticity, because of a "slow flow" assumption, the nonlinearity in the momentum equation is often neglected. The difficulty in performing accurate numerical computations arises from the hyperbolic character of the constitutive equation, which does not contain a dissipative (stabilizing) term for the stress. Care must be used in discretizing the constitutive equation to avoid the introduction of spurious oscillations into the approximation.

The first error analysis for the steady-state finite element (FE) approximation of viscoelastic fluid was presented by Baranger and Sandri [2]. In [2] a discontinuous FE formulation was used for the discretization of the constitutive equation, with the approximation for the stress being discontinuous. Motivated by implementation considerations, Najib and Sandri in [12] modified the discretization in [2] to obtain a decoupled system of two equations, showed the algorithm was convergent, and derived a priori error estimates. In [14], Sandri presented an analysis of an FE approximation to this problem, wherein the constitutive equation was discretized using a streamline upwind Petrov–Galerkin (SUPG) method. For the constitutive equation discretized using the method of characteristics, Baranger and Machmoum in [1] analyzed this approach and gave error estimates for the approximations.

For the analysis of the time-dependent problem, Baranger and Wardi [3] studied a discontinuous Galerkin (DG) approximation to inertialess flow in $\mathbb{R}^2$, using techniques similar to those used for the steady-state problem. With the Hood–Taylor

*Received by the editors July 8, 2002; accepted for publication (in revised form) October 9, 2002; published electronically April 23, 2003. This work was partially supported by the ERC Program of the National Science Foundation under award ERC-9731680.

http://www.siam.org/journals/sinum/41-2/41177.html

†Department of Mathematical Sciences, Clemson University, Clemson, SC 29634-0975 (vjervin@clemson.edu, wmiles@ces.clemson.edu).

FE pair used to approximate the velocity and pressure and a discontinuous linear approximation for the stress, they showed, under the assumption $\Delta t \leq C_1 h^{3/2}$, that the discrete H^1 and L^2 errors for the velocity and stress, respectively, were bounded by $C(\Delta t + h^{3/2})$.

In this paper we analyze the SUPG approximation to the time-dependent equations in $\mathbb{R}^{\acute{d}}$, $\acute{d} = 2, 3$. For the fully discrete analysis we extend the approach used in [11] for compressible Navier–Stokes equations to non-Newtonian flow. For ν denoting the SUPG coefficient, and assuming Hood–Taylor FE pair approximation for the velocity and pressure, and a continuous FE approximation for the viscoelastic stress, under the assumption $\Delta t,\ \nu \leq C_1 h^{\acute{d}/2}$, we obtain that the discrete H^1 and L^2 errors for the velocity and stress, respectively, are bounded by $C(\Delta t + \nu + h^2)$.

This paper is organized as follows. A description of the modeling equations is presented in section 2. Section 3 contains a description of the mathematical notation and several lemmas used in the analysis. The semidiscrete and fully discrete approximations are then presented and analyzed in sections 4 and 5, respectively.

2. The Oldroyd B model and the approximating system.

In this section we describe the modeling equations for viscoelastic fluid flow (see also [2]).

2.1. The problem.

Consider a fluid flowing in a bounded, connected domain $\Omega \in \mathcal{R}^{\acute{d}}$. The boundary of Ω, $\partial\Omega$ is assumed to be Lipschitzian. The vector $\mathbf{n}$ represents the outward unit normal to $\partial\Omega$. The velocity vector is denoted by $\mathbf{u}$, pressure by p, total stress by $\mathbf{T}$, and extra stress by τ. For ease of notation, we use the convention of summation on repeated indices and denote differentiation with a comma. For example, $\frac{\partial \mathbf{u}}{\partial x_i}$ is written $u_{,i}$. Then for a tensor τ and a vector $\mathbf{w}$, $\nabla \cdot \tau$ denotes $\tau_{ij,j}$, and $\mathbf{w} \cdot \nabla$ denotes the operator $w_i \frac{\partial}{\partial x_i}$. The deformation tensor, $D(\mathbf{u})$, and the vorticity tensor, $W(\mathbf{u})$, are given by

$$D(\mathbf{u}) \ = \ \frac{1}{2}\left(\nabla\mathbf{u} + (\nabla\mathbf{u})^T\right), \qquad W(\mathbf{u}) \ = \ \frac{1}{2}\left(\nabla\mathbf{u} - (\nabla\mathbf{u})^T\right).$$

The Oldroyd model can be described using an *objective derivative* [2], denoted by $\hat{\partial}\sigma/\partial t$, where

$$\frac{\hat{\partial}\sigma}{\partial t} := \frac{\partial \sigma}{\partial t} + \mathbf{u}\cdot\nabla\sigma + g_a(\sigma, \nabla\mathbf{u}), \qquad\qquad a \in [-1, 1],$$

and

$$\begin{aligned} g_a(\sigma, \nabla\mathbf{u}) &:= \sigma W(\mathbf{u}) - W(\mathbf{u})\sigma - a(D(\mathbf{u})\sigma + \sigma D(\mathbf{u})) \\ &= \frac{1-a}{2}\left(\sigma\nabla\mathbf{u} + (\nabla\mathbf{u})^T\sigma\right) - \frac{1+a}{2}\left((\nabla\mathbf{u})\sigma + \sigma(\nabla\mathbf{u})^T\right). \end{aligned}$$

Oldroyd's model for stress employs a decomposition of the extra stress into two parts: a Newtonian part and a viscoelastic part. Thus $\tau = \tau_N + \tau_V$. The Newtonian part is given by $\tau_N = 2(1-\alpha)D(\mathbf{u})$. The $(1-\alpha)$ represents that part of the total viscosity which is considered Newtonian. Hence $\alpha \in (0, 1)$ represents the proportion of the total viscosity that is considered to be viscoelastic in nature. For example, if a polymer is immersed within a Newtonian carrier fluid, α is related to the percentage of polymer in the mix. The constitutive law is (see [2])

$$\tau_V + \lambda\frac{\hat{\partial}\tau_V}{\partial t} - 2\alpha D(\mathbf{u}) = 0, \tag{2.1}$$

where λ is the Weissenberg number, which is a dimensionless constant defined as the product of the relaxation time and a characteristic strain rate [4]. For notational simplicity, the subscript V is dropped, and τ will be used below to denote the viscoelastic component of the extra stress.

The momentum balance for the fluid is given by

$$Re\left(\frac{d\mathbf{u}}{dt}\right) = -\nabla p + \nabla\cdot(2(1-\alpha)D(\mathbf{u}) + \tau) + \mathbf{f}, \tag{2.2}$$

where Re is the Reynolds number, $\mathbf{f}$ the body forces acting on the fluid, and $d\mathbf{u}/dt$ is the material derivative. Recall that

$$Re = \frac{LV\rho}{\mu}, \qquad \begin{aligned} L &= \text{characteristic length scale,}\\ V &= \text{characteristic velocity scale,}\\ \rho &= \text{fluid density,}\\ \mu &= \text{fluid viscosity.}\end{aligned}$$

In addition to (2.1) and (2.2) we also have the incompressibility condition

$$\nabla\cdot\mathbf{u} = 0 \qquad \text{in } \Omega.$$

To fully specify the problem, appropriate boundary conditions must also be given. The simplest such condition is the homogeneous Dirichlet condition for velocity. In this case, there is no inflow boundary, and, thus, no boundary condition is required for stress. Summarizing, the modeling equations are

$$Re\left(\frac{\partial\mathbf{u}}{\partial t} + \mathbf{u}\cdot\nabla\mathbf{u}\right) + \nabla p - 2(1-\alpha)\nabla\cdot D(\mathbf{u}) - \nabla\cdot\tau = \mathbf{f} \qquad \text{in } \Omega, \tag{2.3}$$

$$\tau + \lambda\left(\frac{\partial\tau}{\partial t} + \mathbf{u}\cdot\nabla\tau + g_a(\tau,\nabla\mathbf{u})\right) - 2\alpha D(\mathbf{u}) = 0 \qquad \text{in } \Omega, \tag{2.4}$$

$$\nabla\cdot\mathbf{u} = 0 \qquad \text{in } \Omega, \tag{2.5}$$

$$\mathbf{u} = 0 \qquad \text{on } \partial\Omega, \tag{2.6}$$

$$\mathbf{u}(0,\mathbf{x}) = \mathbf{u}_0(\mathbf{x}) \qquad \text{in } \Omega, \tag{2.7}$$

$$\tau(0,\mathbf{x}) = \tau_0(\mathbf{x}) \qquad \text{in } \Omega. \tag{2.8}$$

In [8], Guillope and Saut proved the following for the “slow-flow” model of (2.3)–(2.8) (i.e., the $\mathbf{u}\cdot\nabla\mathbf{u}$ term in (2.3) is ignored):

1. local existence, in time, of a unique, regular solution, and
2. under a small data assumption on $\mathbf{f}, \mathbf{f}', \mathbf{u}_0, \tau_0$, the global existence (in time) of a unique solution for $\mathbf{u}$ and τ.

In contrast to the Navier–Stokes equations, well-posedness for general models in viscoelasticity is still not well understood. Results which are known fall into one of three types [13]:

1. for initial value problems, solutions have been shown to exist locally in time,
2. global existence (in time) of solutions if the initial conditions are small perturbations of the rest state, and
3. for steady-state problems, existence of solutions which are small perturbations of the analogous Newtonian case.

2.2. The variational formulation. In this section, we develop the variational formulation of (2.3)–(2.6). The following notation will be used. The $L^2(\Omega)$ norm and inner product will be denoted by $\|\cdot\|$ and $(\cdot,\cdot)$. Likewise, the $L^p(\Omega)$ norms and the Sobolev $W^k_p(\Omega)$ norms are denoted by $\|\cdot\|_{L^p}$ and $\|\cdot\|_{W^k_p}$, respectively. For the semi-norm in $W^k_p(\Omega)$ we use $|\cdot|_{W^k_p}$. H^k is used to represent the Sobelev space W^k_2, and $\|\cdot\|_k$ denotes the norm in H^k. The following function spaces are used in the analysis:

$$\begin{aligned}
\text{Velocity space} : X &:= H^1_0(\Omega) := \left\{\mathbf{u} \in H^1(\Omega) : \mathbf{u} = 0 \ \text{ on } \ \partial\Omega\right\},\\
\text{Stress space} : S &:= \left\{\tau = (\tau_{ij}) : \tau_{ij} = \tau_{ji}; \tau_{ij} \in L^2(\Omega); 1 \le i,j \le 3\right\}\\
&\quad \cap \left\{\tau = (\tau_{ij}) : \mathbf{u}\cdot\nabla\tau \ \in L^2(\Omega), \forall \mathbf{u} \in X\right\},\\
\text{Pressure space} : Q &:= L^2_0(\Omega) = \left\{q \in L^2(\Omega) : \int_\Omega q \, dx = 0\right\},\\
\text{Divergence-free space} : Z &:= \left\{v \in X : \int_\Omega q(\nabla\cdot v)\, dx = 0, \ \forall\, q \in Q\right\}.
\end{aligned}$$

The variational formulation of (2.3)–(2.6) proceeds in the usual manner. Taking the inner product of (2.3), (2.4), and (2.5) with a velocity test function, a stress test function, and a pressure test function, respectively, we obtain

(2.9)
$$Re\left(\frac{\partial\mathbf{u}}{\partial t} + \mathbf{u}\cdot\nabla\mathbf{u}, \mathbf{v}\right) - (p, \nabla\cdot\mathbf{v}) + (2(1-\alpha)D(\mathbf{u}) + \tau, D(\mathbf{v})) = (\mathbf{f}, \mathbf{v}) \qquad \forall\, \mathbf{v} \in X,$$

$$\left(\tau + \lambda\left(\frac{\partial\tau}{\partial t} + \mathbf{u}\cdot\nabla\tau + g_a(\tau, \nabla\mathbf{u})\right) - 2\alpha D(\mathbf{u}), \psi\right) = 0 \qquad \forall\, \psi \in S, \tag{2.10}$$

$$(\nabla\cdot\mathbf{u}, q) = 0 \qquad \forall\, q \in Q. \tag{2.11}$$

The space Z is the space of weakly divergence-free functions. Note that the condition

$$(\nabla\cdot\mathbf{u}, q) = 0 \qquad \forall\, q \in Q, \ \mathbf{u} \in X,$$

is equivalent in a “distributional” sense to

$$(\mathbf{u}, \nabla q) = 0 \qquad \forall\, q \in Q, \ \mathbf{u} \in X, \tag{2.12}$$

where in (2.12), $(\cdot,\cdot)$ denotes the duality pairing between H^{-1} and H^1_0 functions. In addition, note that the velocity and pressure spaces X and Q satisfy the *inf-sup* condition

$$\inf_{q\in Q} \sup_{\mathbf{v}\in X} \frac{(q, \nabla\cdot\mathbf{v})}{\|q\|\,\|\mathbf{v}\|_1} \ge \beta > 0. \tag{2.13}$$

Since the inf-sup condition (2.13) holds, an equivalent variational formulation to (2.9)–(2.11) is the following: *Find* $(\mathbf{u}, \tau) : [0, T] \to X \times S$ *such that*

$$Re\left(\frac{\partial\mathbf{u}}{\partial t} + \mathbf{u}\cdot\nabla\mathbf{u}, \mathbf{v}\right) + (2(1-\alpha)D(\mathbf{u}) + \tau, D(\mathbf{v})) = (\mathbf{f}, \mathbf{v}) \ \ \forall\, \mathbf{v} \in Z, \tag{2.14}$$

$$\left(\tau + \lambda\left(\frac{\partial\tau}{\partial t} + \mathbf{u}\cdot\nabla\tau + g_a(\tau, \nabla\mathbf{u})\right) - 2\alpha D(\mathbf{u}), \psi\right) = 0 \qquad \forall\, \psi \in S. \tag{2.15}$$

Before discussion of the numerical approximation of (2.14), (2.15), we summarize the mathematical notation and interpolation properties used in the analysis.

3. Mathematical notation. In this section the mathematical framework and approximation properties are summarized.

Let $\Omega \subset \mathbb{R}^{\acute{d}} (\acute{d} = 2, 3)$ be a polygonal domain, and let T_h be a triangulation of Ω made of triangles (in $\mathbb{R}^2$) or tetrahedrals (in $\mathbb{R}^3$). Thus, the computational domain is defined by

$$\Omega = \bigcup K, \qquad K \in T_h.$$

We assume that there exist constants c_1, c_2 such that

$$c_1 h \le h_K \le c_2 \rho_K,$$

where h_K is the diameter of triangle (tetrahedral) K, ρ_K is the diameter of the greatest ball (sphere) included in K, and $h = \max_{K \in T_h} h_K$. Let $P_k(A)$ denote the space of polynomials on A of degree no greater than k. Then we define the FE spaces as follows:

$$\begin{aligned} X_h &:= \left\{ \mathbf{v} \in X \cap C(\bar{\Omega})^{\acute{d}} : \mathbf{v}|_K \in P_k(K)\ \forall K \in T_h \right\}, \\ S_h &:= \left\{ \sigma \in S \cap C(\bar{\Omega})^{\acute{d} \times \acute{d}} : \sigma|_K \in P_m(K)\ \forall K \in T_h \right\}, \\ Q_h &:= \left\{ q \in Q \cap C(\bar{\Omega}) : q|_K \in P_q(K)\ \forall K \in T_h \right\}, \\ Z_h &:= \left\{ \mathbf{v} \in X_h : (q, \nabla \cdot \mathbf{v}) = 0\ \forall q \in Q_h \right\}, \end{aligned}$$

where $C(\bar{\Omega})^{\acute{d}}$ denotes a vector valued function with $\acute{d}$ components continuous on $\bar{\Omega}$. Analogous to the continuous spaces, we assume that X_h and Q_h satisfy the discrete *inf-sup* condition

$$\inf_{q \in Q_h} \sup_{\mathbf{v} \in X_h} \frac{(q, \nabla \cdot \mathbf{v})}{\|q\| \, \|\mathbf{v}\|_1} \ge \beta > 0. \tag{3.1}$$

We summarize several properties of FE spaces and Sobolev's spaces which we will use in our subsequent analysis. For $(\mathbf{u}, p) \in H^{k+1}(\Omega)^{\acute{d}} \times H^{q+1}(\Omega)$ we have (see [7]) that there exists $(\mathcal{U}, \mathcal{P}) \in Z_h \times Q_h$ such that

$$\|\mathbf{u} - \mathcal{U}\| \le C_I h^{k+1} |\mathbf{u}|_{W_2^{k+1}}, \tag{3.2}$$

$$\|\mathbf{u} - \mathcal{U}\|_{W_2^1} \le C_I h^{k} |\mathbf{u}|_{W_2^{k+1}}, \tag{3.3}$$

$$\|p - \mathcal{P}\| \le C_I h^{q+1} |p|_{W_2^{q+1}}. \tag{3.4}$$

Let $\mathcal{T} \in S_h$ be a P_1 continuous interpolant of τ. For $\tau \in H^{m+1}(\Omega)^{\acute{d} \times \acute{d}}$ we have that

$$\|\tau - \mathcal{T}\| \ + \ h|\tau - \mathcal{T}|_{W_2^1} \le C_I h^{m+1} \|\tau\|_{W_2^{m+1}}, \tag{3.5}$$

$$\|\tau - \mathcal{T}\|_{L^4} \ + \ h|\tau - \mathcal{T}|_{W_4^1} \le C_I h^{m+1-\acute{d}/4} \|\tau\|_{W_2^{m+1}}. \tag{3.6}$$

From [5], we have the following results.

LEMMA 3.1. *Let $\{T_h\}$, $0 < h \le 1$, denote a quasi-uniform family of subdivisions of a polyhedral domain $\Omega \subset \mathbb{R}^{\acute{d}}$. Let $(\hat{K}, P, N)$ be a reference finite element such that $P \subset W_p^l(\hat{K}) \cap W_q^m(\hat{K})$, where $1 \le p \le \infty$, $1 \le q \le \infty$, and*

$0 \leq m \leq l$. For $K \in T_h$, let (K, P_K, N_K) be the affine equivalent element, and $V_h = \{v : v$ is measurable and $v|_K \in P_K \forall K \in T_h\}$. Then there exists $C = C(l, p, q)$ such that

$$\left[\sum_{K \in T_h} \|v\|^p_{W^l_p(K)}\right]^{1/p} \leq C\, h^{m-l+\min(0,\, \frac{\acute{d}}{p} - \frac{\acute{d}}{q})} \left[\sum_{K \in T_h} \|v\|^q_{W^m_q(K)}\right]^{1/q} \tag{3.7}$$

for all $v \in V_h$. □

LEMMA 3.2. *Let I_h denote the interpolant of v. Then for all $v \in W^m_p(\Omega) \cap C^r(\Omega)$ and $0 \leq s \leq \min\{m, r+1\}$,*

$$\|v - I_h\|_{s,\infty} \leq C\, h^{m-s-\acute{d}/p}\, |v|_{W^m_p}. \quad □ \tag{3.8}$$

When $v(\mathbf{x}, t)$ is defined on the entire time interval $(0, T)$, we define

$$\|v\|_{\infty,k} := \sup_{0<t<T} \|v(\cdot, t)\|_k,$$

$$\|v\|_{0,k} := \left(\int_0^T \|v(\cdot, t)\|^2_k \, dt\right)^{1/2}.$$

For the analysis of the fully discrete approximation, we use Δt to denote the step size for t so that $t_n = n\Delta t$, $n = 0, 1, 2, \ldots, N$, and define

$$f^n := f(n\Delta t) \quad \text{and} \quad d_t f := \frac{f(t_n) - f(t_{n-1})}{\Delta t}. \tag{3.9}$$

We also use the following additional norms:

$$|||v|||_{\infty,k} := \max_{1 \leq n \leq N} \|v^n\|_k,$$

$$|||v|||_{0,k} := \left[\sum_{n=1}^N \Delta t\, \|v^n\|^2_k\right]^{1/2}.$$

4. Semidiscrete approximation. In this section we present the analysis of a semidiscrete approximation to (2.14), (2.15). We begin by introducing some notation specific to the semidiscrete approximation and cite some lemmas used in the analysis.

For $\sigma_u := \sigma + \nu h\, \mathbf{u} \cdot \nabla\sigma$ we define

$$A(\mathbf{w}, (\mathbf{u}, \tau), (\mathbf{v}, \psi)) := (\tau, \psi_w) - 2\alpha(D(\mathbf{u}), \psi_w) + 2\alpha(\tau, D(\mathbf{v})) + \alpha(1-\alpha)(\nabla\mathbf{u}, \nabla\mathbf{v}), \tag{4.1}$$

$$B(\mathbf{u}, \mathbf{v}, \tau, \sigma) := (\mathbf{u} \cdot \nabla\tau, \sigma_v) + \frac{1}{2}(\nabla \cdot \mathbf{u}\, \tau, \sigma), \tag{4.2}$$

$$c(\mathbf{w}, \mathbf{u}, \mathbf{v}) := (\mathbf{w} \cdot \nabla\mathbf{u}, \mathbf{v}), \tag{4.3}$$

$$\tilde{c}(\mathbf{w}, \mathbf{u}, \mathbf{v}) := \frac{1}{2}\left(c(\mathbf{w}, \mathbf{u}, \mathbf{v}) - c(\mathbf{w}, \mathbf{v}, \mathbf{u})\right). \tag{4.4}$$

LEMMA 4.1 (see [10]). *For $\mathbf{u}, \mathbf{v}, \mathbf{w} \in X$, there exists a constant C_1 such that*

$$|\tilde{c}(\mathbf{u}, \mathbf{v}, \mathbf{w})| \leq C_1 \|\mathbf{u}\|^{1/2} \|\nabla\mathbf{u}\|^{1/2} \|\nabla\mathbf{v}\|^{1/2} \|\nabla\mathbf{w}\|^{1/2}. \quad □ \tag{4.5}$$

Note the following:

(i) $\tilde{c}(\mathbf{u},\mathbf{v},\mathbf{w}) = c(\mathbf{u},\mathbf{v},\mathbf{w})$ when $\nabla\cdot\mathbf{u} = 0$ in Ω, and $\mathbf{u} = 0$ on $\partial\Omega$.

(ii) $\tilde{c}(\mathbf{u},\mathbf{v},\mathbf{v}) = 0$, even when $\nabla\cdot\mathbf{u} \neq 0$.

(iii) For $\mathbf{u} \in X$, from the Poincaré–Friedrichs inequality we have that there exists a constant $C_{PF} = C(\Omega)$ such that $\|\mathbf{u}\|^2 \leq C_{PF}^2 \|\nabla\mathbf{u}\|^2$.

The operators $A(\cdot,(\cdot,\cdot),(\cdot,\cdot)) : X \times (X \times H^1(\Omega)^{n\times n}) \times (X \times H^1(\Omega)^{n\times n}) \to \mathbb{R}$ and $B(\cdot,\cdot,\cdot,\cdot) : X \times X \times H^1(\Omega)^{n\times n} \times H^1(\Omega)^{n\times n} \to \mathbb{R}$ are the same as those used in [2], [14]. When $\mathbf{u} = \mathbf{v}$, we omit the second variable in $B(\cdot,\cdot,\cdot,\cdot)$.

Lemma 4.2. *We have that*

$$B(\mathbf{u},\tau,\tau) \;=\; \nu h(\mathbf{u}\cdot\nabla\tau, \mathbf{u}\cdot\nabla\tau). \tag{4.6}$$

Proof. On integrating $(\mathbf{u}\cdot\nabla\tau,\sigma)$ by parts we have

$$B(\mathbf{u},\mathbf{v},\tau,\sigma) \;:=\; -(\mathbf{u}\cdot\nabla\sigma,\tau) \;+\; \nu\, h\,(\mathbf{u}\cdot\nabla\tau, \mathbf{v}\cdot\nabla\sigma) \;-\; \frac{1}{2}(\nabla\cdot\mathbf{u}\,\sigma,\tau). \tag{4.7}$$

Setting $\mathbf{v} = \mathbf{u}$, $\sigma = \tau$, and combining (4.2) and (4.7), the stated result follows. □

Lemma 4.3. *For $\mathbf{w} \in X$, $(\mathbf{u},\tau) \in X \times S$ and h sufficiently small, we have*

$$A(\mathbf{w},(\mathbf{u},\tau),(\mathbf{u},\tau)) \;+\; \lambda B(\mathbf{w},\tau,\tau) \geq C_A(\|\tau\|^2 + \|\mathbf{u}\|_1^2).$$

Proof. Using the definitions of A and B, we obtain

$$\begin{aligned}
A(\mathbf{w},(\mathbf{u},\tau),(\mathbf{u},\tau)) + \lambda B(\mathbf{w},\tau,\tau) \;&=\; \|\tau\|^2 \;+\; (\tau, \nu h\,\mathbf{w}\cdot\nabla\tau) \;-\; 2\alpha(D(\mathbf{u}),\tau) \\
&\quad - 2\alpha(D(\mathbf{u}), \nu h\,\mathbf{w}\cdot\nabla\tau) + 2\alpha(\tau, D(\mathbf{u})) + \alpha(1-\alpha)\|\nabla\mathbf{u}\|^2 \\
&\quad + \lambda\nu h\|\mathbf{w}\cdot\nabla\tau\|^2 \\
&\geq \|\tau\|^2 + \alpha(1-\alpha)\|\nabla\mathbf{u}\|^2 + \lambda\nu h\,\|\mathbf{w}\cdot\nabla\tau\|^2 - \frac{1}{2}\|\tau\|^2 \\
&\quad - \frac{1}{2}\nu^2 h^2\,\|\mathbf{w}\cdot\nabla\tau\|^2 \;-\; \frac{1}{2}\alpha(1-\alpha)\|\nabla\mathbf{u}\|^2 - \frac{\alpha\nu^2 h^2}{2(1-\alpha)}\|\mathbf{w}\cdot\nabla\tau\|^2 \\
&\geq \frac{1}{2}\|\tau\|^2 + \frac{\alpha(1-\alpha)}{2}\|\nabla\mathbf{u}\|^2 + \left(\lambda\nu h \;-\; \frac{\nu^2 h^2}{2} \;-\; \frac{\alpha\nu^2 h^2}{2(1-\alpha)}\right)\|\mathbf{w}\cdot\nabla\tau\|^2 \qquad (4.8) \\
&\geq C_A\left(\|\tau\|^2 + \|\mathbf{u}\|_1^2\right)
\end{aligned}$$

for h sufficiently small, using (iii). □

Now we define the semidiscrete approximation of (2.14), (2.15) as

Find $(\mathbf{u}_h,\tau_h) : [0,T] \;\to\; X_h \times S_h$ such that

(4.9)

$$Re\,(\mathbf{u}_{h\,t},\mathbf{v}) + Re\,\tilde{c}(\mathbf{u}_h,\mathbf{u}_h,\mathbf{v}) \;+\; (1-\alpha)(\nabla\mathbf{u}_h,\nabla\mathbf{v}) \;+\; (\tau_h, D(\mathbf{v})) \;=\; (\mathbf{f},\mathbf{v}) \;\;\forall\,\mathbf{v}\in Z_h,$$

(4.10)

$$\lambda\,(\tau_{h\,t},\sigma) + \lambda\,B(\mathbf{u}_h,\tau_h,\sigma) + \lambda(g_a(\tau_h,\nabla\mathbf{u}_h),\sigma_{u_h}) + (\tau_h,\sigma_{u_h}) - 2\alpha(D(\mathbf{u}_h),\sigma_{u_h}) = 0 \quad \forall\,\sigma\in S_h.$$

4.1. Analysis of the semidiscrete approximation. In this section, we show that, under suitable conditions, a unique solution to the discretized system exists. Fixed point theory is used to establish the desired result. The proof is established using the following four steps:

1. Define an iterative map in such a way that a fixed point of the map is a solution to (4.9), (4.10).
2. Show that the map is well defined and bounded on bounded sets.
3. Show that there exists an invariant ball on which the map is a contraction.
4. Apply Schauder's fixed point theorem to establish the existence and uniqueness of the discrete approximation.

THEOREM 4.4. *Assume that the system* (2.3)–(2.8) *(and thus* (2.14)–(2.15)*) has a solution* $(\mathbf{u}, \tau, \mathbf{p}) \in L^2(0, T; H^{k+1}) \times L^\infty(0, T; H^{m+1}) \times L^2(0, T; H^{q+1})$*. In addition assume that* $k, m \geq \acute{d}/2$*, and*

$$\|\nabla \mathbf{u}\|_\infty, \|\tau\|_\infty, \|\nabla \tau\|_\infty, \|\mathbf{u}\|_{k+1}, \|\tau\|_{m+1}, \|p\|_{q+1} \leq D_0 \quad \forall t \in [0, T]. \tag{4.11}$$

Then for D_0 *and* h *sufficiently small, there exists a unique solution to* (4.9)–(4.10) *satisfying*

$$\int_0^T \left(\|\mathbf{u} - \mathbf{u}_h\|^2 \, + \, \|\nabla(\mathbf{u} - \mathbf{u}_h)\|^2 \right) \, dt \leq C h^{\min\{k, m, q+1\}}, \tag{4.12}$$

$$\sup_{0 \leq t \leq T} \|\tau - \tau_h\| \leq C h^{\min\{k, m, q+1\}}. \tag{4.13}$$

Proof. Step 1. *The iterative map.* A mapping $\xi : L^2(0, T; Z_h) \times L^\infty(0, T; S_h) \rightarrow L^2(0, T; Z_h) \times L^\infty(0, T; S_h)$ is defined via $(\mathbf{u}_2, \tau_2) = \xi(\mathbf{u}_1, \tau_1)$, where $(\mathbf{u}_2, \tau_2)$ satisfies

(4.14)
$$Re\,(\mathbf{u}_{2\,t}, \mathbf{v}) + \; Re\, \tilde{c}(\mathbf{u}_1, \mathbf{u}_2, \mathbf{v}) + (1 - \alpha)(\nabla \mathbf{u}_2, \nabla \mathbf{v}) + (\tau_2, D(\mathbf{v})) = (\mathbf{f}, \mathbf{v}) \;\; \forall \, \mathbf{v} \in Z_h,$$

(4.15)
$$\lambda(\tau_{2\,t}, \sigma) + \lambda B(\mathbf{u}_1, \tau_2, \sigma) + (\tau_2, \sigma_{u_1}) - 2\alpha(D(\mathbf{u}_h), \sigma_{u_1}) = -\lambda(g_a(\tau_1, \nabla \mathbf{u}_1), \sigma_{u_1}) \qquad \forall \, \sigma \in S_h.$$

Thus, given an initial guess $(\mathbf{u}_h, \tau_h) \approx (\mathbf{u}_1, \tau_1)$, solving (4.14), (4.15) for $(\mathbf{u}_2, \tau_2)$ gives a new approximation to the solution. Also, it is clear that a fixed point of (4.14), (4.15) is a solution to the approximating system (4.9), (4.10) (i.e., $\xi(\mathbf{u}_1, \tau_1) = (\mathbf{u}_1, \tau_1)$ implies that $(\mathbf{u}_1, \tau_1)$ is a solution to (4.9), (4.10)).

Step 2. *Show that* ξ *is well defined and bounded on bounded sets.* Note that (4.14), (4.15) corresponds to a first order system of ODEs for the FEM (finite element method) coefficients $\mathbf{c}_{\mathbf{u}_2}$ and $\mathbf{c}_{\tau_2}$ of $\mathbf{u}_2$ and τ_2, respectively. That is, (4.14), (4.15) is equivalent to

$$\begin{bmatrix} A_{11} & 0 \\ 0 & A_{22} \end{bmatrix} \begin{bmatrix} \mathbf{c}_{\mathbf{u}_2} \\ \mathbf{c}_{\tau_2} \end{bmatrix}' \; = \; \mathbf{F}(t, \mathbf{c}_{\mathbf{u}_2}, \mathbf{c}_{\tau_2}),$$

where

$$\mathbf{F}(t, \mathbf{c}_{\mathbf{u}_2}, \mathbf{c}_{\tau_2}) = \begin{bmatrix} (\mathbf{f}, \mathbf{v}) \, - \, Re\, \tilde{c}(\mathbf{u}_1, \mathbf{u}_2, \mathbf{v}) \, - \, (1 - \alpha)(\nabla \mathbf{u}_2, \nabla \mathbf{v}) \, - \, (\tau_2, D(\mathbf{v})) \\ -\lambda(g_a(\tau_h, \nabla \mathbf{u}_h), \sigma_{u_1}) \, - \, \lambda\, B(\mathbf{u}_1, \tau_2, \sigma) \, - \, (\tau_2, \sigma_{u_1}) \, + \, 2\alpha(D(\mathbf{u}_h), \sigma_{u_1}) \end{bmatrix},$$

and A_{11} and A_{22} are "mass" (invertible) matrices.

Note that $\mathbf{F} : [0, T] \times \mathbb{R}^{dim(\mathbf{c}_{\mathbf{u}_2})} \times \mathbb{R}^{dim(\mathbf{c}_{\tau_2})} \rightarrow \mathbb{R}^{dim(\mathbf{c}_{\mathbf{u}_2})} \times \mathbb{R}^{dim(\mathbf{c}_{\tau_2})}$ is a linear function with respect to the FEM coefficients $\mathbf{c}_{\mathbf{u}_2}, \mathbf{c}_{\tau_2}$. Thus, for $f(t)$ a continuous function of t, we have that $\mathbf{F}$ is Lipschitz continuous. Then, from ODE theory (see

[6]), we are guaranteed that there exists a unique local solution for $(\mathbf{c}_{\mathbf{u}_2}, \mathbf{c}_{\tau_2})$, and hence for $(\mathbf{u}_2, \tau_2)$.

Next, to establish the existence of $(\mathbf{u}_2, \tau_2)$ on $[0, T]$, we show that it remains bounded in the appropriate norms on that interval.

Multiplying (4.14) through by 2α and adding the result to (4.15), $(\mathbf{u}_2, \tau_2)$ is equivalently determined via

$$\begin{aligned} &2\alpha Re(\mathbf{u}_{2\,t}, \mathbf{v}) + 2\alpha Re\tilde{c}(\mathbf{u}_1, \mathbf{u}_2, \mathbf{v}) + A(\mathbf{u}_1, (\mathbf{u}_2, \tau_2), (\mathbf{v}, \sigma)) + \lambda(\tau_{2\,t}, \sigma) + \lambda B(\mathbf{u}_1, \tau_2, \sigma) \\ &\qquad = 2\alpha(\mathbf{f}, \mathbf{v}) - \lambda(g_a(\tau_1, \nabla \mathbf{u}_1), \sigma_{u_1}) \ \ \forall\ (\mathbf{v}, \sigma) \in Z_h \times S_h. \end{aligned} \tag{4.16}$$

Choosing $\mathbf{v} = \mathbf{u}_2$, $\sigma = \tau_2$ in (4.16) and using (ii) and (4.8) implies

$$\begin{aligned} \alpha Re\, \|\mathbf{u}_2\|_t^2 \,+\, &\frac{\lambda}{2}\, \|\tau_2\|_t^2 \,+\, \frac{1}{2}\|\tau_2\|^2 + \frac{\alpha(1-\alpha)}{2}\|\nabla \mathbf{u}_2\|^2 \\ &+ \left(\lambda \nu h \,-\, \nu^2 h^2 \left(\frac{1}{2} \,+\, \frac{\alpha}{2(1-\alpha)}\right)\right) \|\mathbf{u}_1 \cdot \nabla \tau_2\|^2 \\ &\le 2\alpha \|\mathbf{f}\|_{-1} \|\mathbf{u}_2\|_1 \,+\, \lambda \|g_a(\tau_1, \nabla \mathbf{u}_1)\| \left(\|\tau_2\| \,+\, \nu h \, \|\mathbf{u}_1 \cdot \nabla \tau_2\|\right) \\ &\le \frac{2(1 + C_{PF}^2)}{(1-\alpha)} \|\mathbf{f}\|_{-1}^2 \,+\, \frac{\alpha^2(1-\alpha)}{2}\|\nabla \mathbf{u}_2\|^2 \,+\, \lambda^2 \|g_a(\tau_1, \nabla \mathbf{u}_1)\|^2 \\ &\quad + \frac{1}{2}\|\tau_2\|^2 \,+\, \frac{\nu^2 h^2}{2}\, \|\mathbf{u}_1 \cdot \nabla \tau_2\|^2. \end{aligned} \tag{4.17}$$

Thus for $c_1 = \min\{\alpha Re,\ \lambda/2\}$ and the restriction $\nu h \le 2\lambda(1-\alpha)/(2-\alpha)$,

$$\frac{d}{dt}\left(\|\mathbf{u}_2\|^2 \,+\, \|\tau_2\|^2\right) \le \frac{2(1 + C_{PF}^2)}{c_1(1-\alpha)}\|\mathbf{f}\|_{-1}^2 \,+\, \frac{\lambda^2}{c_1}\|g_a(\tau_1, \nabla \mathbf{u}_1)\|^2.$$

Hence for $0 \le t \le T$,

$$\begin{aligned} \|\mathbf{u}_2\|^2(t) \,+\, \|\tau_2\|^2(t) \le \|\mathbf{u}_2\|^2(0) \,+\, \|\tau_2\|^2(0) \,+\, &\frac{2(1 + C_{PF}^2)}{c_1(1-\alpha)}\|\mathbf{f}\|_{0,-1}^2 \\ + &\frac{\lambda^2\, \acute{d}^2}{c_1}\|\tau_1\|_{\infty,\infty}^2\, \|\nabla \mathbf{u}_1\|_{0,0}^2. \end{aligned} \tag{4.18}$$

By the equivalence of norm in finite dimensional spaces (and $\mathbf{u}_2(0) = \mathbf{u}_1(0)$, $\tau_2(0) = \tau_1(0)$), we therefore have that $(\mathbf{u}_2,\ \tau_2) \in L^2(0, T; Z_h) \times L^\infty(0, T; S_h)$.

Note that (4.18) also establishes that the mapping ξ is bounded on bounded sets.

Step 3. *Existence of an invariant ball for ξ.* We begin by defining an invariant ball. Let $R = c^* h^{\min\{k,m,q+1\}}$ for $0 < c^* < 1$, and define the ball B_h as

$$\begin{aligned} B_h := \Bigg\{ &(\mathbf{v}, \sigma) \in L^2(0, T; Z_h) \times L^\infty(0, T; S_h) \ : \\ &\int_0^T \|\mathbf{u} - \mathbf{v}\|^2 + \|\nabla(\mathbf{u} - \mathbf{v})\|^2 \, dt \le R^2, \sup_{0 \le t \le T} \|\tau - \sigma\| \le R \Bigg\}. \end{aligned} \tag{4.19}$$

The exact solution $(\mathbf{u}, p, \tau)$ of (2.9)–(2.11) satisfies

$$\begin{aligned} &2\alpha Re\,(\mathbf{u}_t, \mathbf{v}) + 2\alpha Re\, \tilde{c}(\mathbf{u}, \mathbf{u}, \mathbf{v}) \,+\, A(\mathbf{u}_1, (\mathbf{u}, \tau), (\mathbf{v}, \sigma)) \,+\, \lambda\, (\tau_t, \sigma) \,+\, \lambda\, B(\mathbf{u}, \mathbf{u}_1, \tau, \sigma) \\ &\qquad = 2\alpha(p, \nabla \cdot \mathbf{v}) \,+\, 2\alpha(\mathbf{f}, \mathbf{v}) - \lambda(g_a(\tau, \nabla \mathbf{u}), \sigma_{u_1}) \ \ \forall\ (\mathbf{v}, \sigma) \in Z \times S. \end{aligned} \tag{4.20}$$

Subtracting (4.16) from (4.20) implies that

$$\begin{aligned}
2\alpha Re\,((\mathbf{u}-\mathbf{u}_2)_t,\mathbf{v}) + 2\alpha Re\,\tilde{c}(\mathbf{u},\mathbf{u},\mathbf{v}) \,-\, 2\alpha Re\,\tilde{c}(\mathbf{u}_1,\mathbf{u}_2,\mathbf{v})& \\
+ A(\mathbf{u}_1,(\mathbf{u}-\mathbf{u}_2,\tau-\tau_2),(\mathbf{v},\sigma))& \\
+\lambda\,((\tau-\tau_2)_t,\sigma) \,+\, \lambda\,B(\mathbf{u}_1,(\tau-\tau_2),\sigma)& \\
= 2\alpha(p,\nabla\cdot\mathbf{v}) \,-\, \lambda\,((g_a(\tau,\nabla\mathbf{u}),\sigma_{u_1}) - (g_a(\tau_1,\nabla\mathbf{u}_1),\sigma_{u_1}))& \\
-\lambda B(\mathbf{u},\mathbf{u}_1,\tau,\sigma) \,+\, \lambda B(\mathbf{u}_1,\tau,\sigma) \;\; \forall\;(\mathbf{v},\sigma)\in Z_h\times S_h.&
\end{aligned}\tag{4.21}$$

Let

$$\Lambda := \mathbf{u}-\mathcal{U}, \qquad E := \mathcal{U}-\mathbf{u}_2, \tag{4.22}$$

$$\Gamma := \tau-\mathcal{T}, \qquad F := \mathcal{T}-\tau_2, \tag{4.23}$$

$$\text{and}\quad \epsilon_{\mathbf{u}} := \Lambda + E \,=\, \mathbf{u}-\mathbf{u}_2, \qquad \epsilon_\tau := \Gamma + F \,=\, \tau-\tau_2. \tag{4.24}$$

Rewriting (4.21) using these definitions, along with the choice $\sigma = F$, $\mathbf{v} = E$, we obtain

$$\begin{aligned}
2\alpha Re\,(E_t,E) + 2\alpha Re\,\tilde{c}(\mathbf{u},\mathbf{u},E) \,-\, 2\alpha Re\,\tilde{c}(\mathbf{u}_1,\mathbf{u}_2,E) \,+\, A(\mathbf{u}_1,(E,F),(E,F))& \\
+\,\lambda\,(F_t,F) \,+\, \lambda\,B(\mathbf{u}_1,F,F)& \\
= -2\alpha Re\,(\Lambda_t,E) \,-\, A(\mathbf{u}_1,(\Lambda,\Gamma),(E,F)) \,-\, \lambda\,(\Gamma_t,F) \,-\, \lambda\,B(\mathbf{u}_1,\Gamma,F)& \\
+\,2\alpha(p,\nabla\cdot E) \,-\, \lambda\,((g_a(\tau,\nabla\mathbf{u}),F_{u_1}) - (g_a(\tau_1,\nabla\mathbf{u}_1),F_{u_1}))& \\
-\lambda B(\mathbf{u},\mathbf{u}_1,\tau,F) \,+\, \lambda B(\mathbf{u}_1,\tau,F).&
\end{aligned}\tag{4.25}$$

We now proceed to bound E in terms of F, $\mathbf{u}$, and $\mathbf{u}_1$. For the $\tilde{c}$ terms we have

$$\begin{aligned}
\tilde{c}(\mathbf{u},\mathbf{u},E) \,-\, \tilde{c}(\mathbf{u}_1,\mathbf{u}_2,E) &= \tilde{c}(\mathbf{u}-\mathbf{u}_1,\mathbf{u},E) \,+\, \tilde{c}(\mathbf{u}_1,\mathbf{u}-\mathbf{u}_2,E) \\
&= \tilde{c}(\mathbf{u}-\mathbf{u}_1,\mathbf{u},E) \,+\, \tilde{c}(\mathbf{u}_1,E+\Lambda,E) \\
&= \tilde{c}(\mathbf{u}-\mathbf{u}_1,\mathbf{u},E) \,+\, \tilde{c}(\mathbf{u}_1,\Lambda,E) \quad \text{(using (4))}.
\end{aligned}\tag{4.26}$$

We estimate the first term on the right-hand side (rhs) of (4.26) by

$$\begin{aligned}
|\tilde{c}(\mathbf{u}-\mathbf{u}_1,\mathbf{u},E)| &\le C_1\,\|\mathbf{u}-\mathbf{u}_1\|^{1/2}\|\nabla(\mathbf{u}-\mathbf{u}_1)\|^{1/2}\|\nabla\mathbf{u}\|\|\nabla E\| \;\; \text{(using (4.5))} \\
&\le \epsilon_1\|\nabla E\|^2 \,+\, \frac{C_1^2}{4\epsilon_1}\|\mathbf{u}-\mathbf{u}_1\|\|\nabla(\mathbf{u}-\mathbf{u}_1)\|\|\nabla\mathbf{u}\|^2.
\end{aligned}\tag{4.27}$$

For the second term on the rhs of (4.26),

$$\begin{aligned}
|\tilde{c}(\mathbf{u}_1,\Lambda,E)| &\le |-\tilde{c}((\mathbf{u}-\mathbf{u}_1),\Lambda,E)| \,+\, |\tilde{c}(\mathbf{u},\Lambda,E)| \\
&\le C_1\,\|\mathbf{u}-\mathbf{u}_1\|^{1/2}\|\nabla(\mathbf{u}-\mathbf{u}_1)\|^{1/2}\|\nabla\Lambda\|\|\nabla E\| \,+\, C_2\|\mathbf{u}\|_\infty\|\nabla\Lambda\|\|\nabla E\| \\
&\le \epsilon_3\|\nabla E\|^2 \,+\, \frac{C_1^2}{4\epsilon_3}\|\mathbf{u}-\mathbf{u}_1\|_1^2\|\nabla\Lambda\|^2 \,+\, \epsilon_4\|\nabla E\|^2 \,+\, \frac{C_2^2}{4\epsilon_4}\|\mathbf{u}\|_\infty^2\|\nabla\Lambda\|^2.
\end{aligned}\tag{4.28}$$

In view of the estimates (4.8) and (4.6), we proceed next to consider the terms on the rhs of (4.25):

$$(\Lambda_t,E) \le \|\Lambda_t\|\|E\| \;\le\; \epsilon_5\|\nabla E\|^2 \,+\, \frac{C_{PF}^2}{4\epsilon_5}\|\Lambda_t\|^2, \tag{4.29}$$

$$(\Gamma_t,F) \le \|\Gamma_t\|\|F\| \;\le\; \epsilon_6\|F\|^2 \,+\, \frac{1}{4\epsilon_6}\|\Gamma_t\|^2. \tag{4.30}$$

For the pressure term we have

$$\begin{aligned}2\alpha\,|(p,\nabla\cdot E)| \;=\; 2\alpha\,|((p-\mathcal{P}),\nabla\cdot E)| &\leq 2\alpha\,\|p-\mathcal{P}\|\|\nabla\cdot E\| \\ &\leq 2\alpha\acute{d}^{1/2}\,\|p-\mathcal{P}\|\|\nabla E\| \\ &\leq \frac{\alpha^2\acute{d}}{\epsilon_7}\,\|p-\mathcal{P}\|^2 \,+\, \epsilon_7\|\nabla E\|^2.\end{aligned}\tag{4.31}$$

Writing out the A term on the rhs of (4.25), we have the terms

$$A(\mathbf{u}_1,(\Lambda,\Gamma),(E,F)) \;=\; (\Gamma,F_{u_1})-2\alpha(D(\Lambda),F_{u_1})+2\alpha(\Gamma,D(E))+\alpha(1-\alpha)(\nabla\Lambda,\nabla E).\tag{4.32}$$

For the first term in A we obtain

$$\begin{aligned}(\Gamma,F_{u_1}) &= (\Gamma,F) \,+\, (\Gamma,\nu h\,\mathbf{u}_1\cdot\nabla F) \\ &= \|\Gamma\|\,\|F\| \,+\, \|\Gamma\|\,\nu h\,\|\mathbf{u}_1\cdot\nabla F\| \\ &= \epsilon_8\|F\|^2 \,+\, \frac{1}{4\epsilon_8}\|\Gamma\|^2 \,+\, \nu^2h^2\,\|\mathbf{u}_1\cdot\nabla F\|^2 \,+\, \frac{1}{4}\|\Gamma\|^2.\end{aligned}\tag{4.33}$$

Similarly,

$$2\alpha\,(D(\Lambda),F_{u_1}) \leq \epsilon_9\|F\|^2 \,+\, \frac{\alpha^2}{\epsilon_9}\|D(\Lambda)\|^2 \,+\, \nu^2h^2\,\|\mathbf{u}_1\cdot\nabla F\|^2 \,+\, \alpha^2\|D(\Lambda)\|^2,\tag{4.34}$$

$$2\alpha\,(\Gamma,D(E)) \leq \epsilon_{10}\|\nabla E\|^2 \,+\, \frac{\alpha^2}{4\epsilon_{10}}\|\Gamma\|^2,\tag{4.35}$$

$$\alpha(1-\alpha)\,(\nabla\Gamma,\nabla E) \leq \epsilon_{11}\|\nabla E\|^2 \,+\, \frac{\alpha^2(1-\alpha)^2}{4\epsilon_{11}}\|\nabla\Gamma\|^2.\tag{4.36}$$

Bounding the $g_a(\cdot,\cdot)$ terms on the rhs of (4.25) is more involved. We rewrite the difference as the sum of three terms and then bound each of the terms individually.

We have that

$$\begin{aligned}(g_a(\tau,\nabla\mathbf{u})-g_a(\tau_1,\nabla\mathbf{u}_1),F_{u_1}) &= (g_a(\tau-\tau_1,\nabla\mathbf{u}),\,F_{u_1})+(g_a(\tau_1,\nabla(\mathbf{u}-\mathbf{u}_1)),F_{u_1}) \\ &= (g_a(\tau-\tau_1,\nabla\mathbf{u}),F_{u_1})+(g_a(\tau_1-\tau,\nabla(\mathbf{u}-\mathbf{u}_1)),F_{u_1}) \\ &\quad+\,(g_a(\tau,\nabla(\mathbf{u}-\mathbf{u}_1)),F_{u_1})\,.\end{aligned}\tag{4.37}$$

For the first term on the rhs of (4.37) we have

$$\begin{aligned}(g_a(\tau-\tau_1,\nabla\mathbf{u}),F_{u_1}) &\leq 4\|(\tau-\tau_1)\nabla\mathbf{u}\|\|F\|+4\|(\tau-\tau_1)\nabla\mathbf{u}\|\|\nu h\mathbf{u}_1\cdot\nabla F\| \\ &\leq 4\acute{d}\|\nabla\mathbf{u}\|_\infty\|(\tau-\tau_1)\|\|F\|+4\acute{d}\|\nabla\mathbf{u}\|_\infty\|(\tau-\tau_1)\|\|\nu h\mathbf{u}_1\cdot\nabla F\| \\ &\leq \epsilon_{12}\|F\|^2+\frac{4\acute{d}^2}{\epsilon_{12}}\|\nabla\mathbf{u}\|_\infty^2\|(\tau-\tau_1)\|^2+\nu^2h^2\|\nu h\mathbf{u}_1\cdot\nabla F\|^2 \\ &\quad+4\acute{d}^2\|\nabla\mathbf{u}\|_\infty^2\|(\tau-\tau_1)\|^2.\end{aligned}\tag{4.38}$$

For the second term we have

$$\begin{aligned}(g_a(\tau-\tau_1,\nabla(\mathbf{u}\,-\,\mathbf{u}_1)),F_{u_1}) &\leq 4\|(\tau-\tau_1)\nabla(\mathbf{u}-\mathbf{u}_1)\|\|F\| \\ &\quad+4\|(\tau-\tau_1)\nabla(\mathbf{u}-\mathbf{u}_1)\|\|\nu h\mathbf{u}_1\cdot\nabla F\| \\ &\leq \epsilon_{13}\|F\|^2+\frac{4}{\epsilon_{13}}\|(\tau-\tau_1)\nabla(\mathbf{u}-\mathbf{u}_1)\|^2+\nu^2h^2\|\nu h\mathbf{u}_1\cdot\nabla F\|^2 \\ &\quad+4\|(\tau-\tau_1)\nabla(\mathbf{u}-\mathbf{u}_1)\|^2.\end{aligned}\tag{4.39}$$

Note that

$$\|(\tau - \tau_1)\,\nabla(\mathbf{u} - \mathbf{u}_1)\| \le \|(\tau - \tau_1)\|_{L^4}\,\|\nabla(\mathbf{u} - \mathbf{u}_1)\|_{L^4},$$

and, using (3.7),

$$\begin{aligned}\|\tau_1 - \mathcal{T}\|_{L^4} &\le C_I h^{-\acute{d}/4}\|\tau_1 - \mathcal{T}\|\\ &\le C_I h^{-\acute{d}/4}\|\tau_1 - \tau\| \,+\, C_I h^{-\acute{d}/4}\|\tau - \mathcal{T}\|.\end{aligned}$$

Thus,

$$\begin{aligned}\|\tau - \tau_1\|_{L^4} &\le \|\tau - \mathcal{T}\|_{L^4} \,+\, \|\mathcal{T} - \tau_1\|_{L^4}\\ &\le \|\tau - \mathcal{T}\|_{L^4} \,+\, Ch^{-\acute{d}/4}\|\tau_1 - \tau\| \,+\, Ch^{-\acute{d}/4}\|\tau - \mathcal{T}\|\\ &\le 2C_I h^{m+1-\acute{d}/4}\|\tau\|_{m+1} \,+\, C_I h^{-\acute{d}/4}\|\tau_1 - \tau\|. \end{aligned}\tag{4.40}$$

Similarly,

$$\begin{aligned}\|\nabla(\mathbf{u} - \mathbf{u}_1)\|_{L^4} &\le \|\nabla(\mathbf{u} - \mathcal{U})\|_{L^4} \,+\, Ch^{-\acute{d}/4}\|\mathbf{u} - \mathbf{u}_1\|_1 \,+\, Ch^{-\acute{d}/4}\|\mathbf{u} - \mathcal{U}\|_1\\ &\le 2C_I h^{k-\acute{d}/4}\|\mathbf{u}\|_{k+1} \,+\, C_I h^{-\acute{d}/4}\|\mathbf{u} - \mathbf{u}_1\|_1. \end{aligned}\tag{4.41}$$

Combining (4.40), (4.41) with (4.39) yields

$$\begin{aligned}|(g_a(\tau - \tau_1, \nabla(\mathbf{u} - \mathbf{u}_1)), F_{u_1})| &\le \epsilon_{13}\|F\|^2 + \nu^2 h^2\|\mathbf{u}_1 \cdot \nabla F\|^2\\ &\quad + \left(\frac{4}{\epsilon_{13}} + 4\right)\left(2C_I h^{m+1-\acute{d}/4}\|\tau\|_{m+1} + C_I h^{-\acute{d}/4}\|\tau_1 - \tau\|\right)^2\\ &\quad \times \left(2C_I h^{k-\acute{d}/4}\|\mathbf{u}\|_{k+1} + C_I h^{-\acute{d}/4}\|\mathbf{u} - \mathbf{u}_1\|_1\right)^2. \end{aligned}\tag{4.42}$$

For the third $g_a(\cdot,\cdot)$ terms on the rhs of (4.37) we have

$$\begin{aligned}|(g_a(\tau, \nabla(\mathbf{u} - \mathbf{u}_1)),\, F_{u_1})| &\le 4\|\tau\,\nabla(\mathbf{u} - \mathbf{u}_1)\|\,\|F\| \;+\; 4\|\tau\,\nabla(\mathbf{u} - \mathbf{u}_1)\|\;\|\nu h\,\mathbf{u}_1 \cdot \nabla F\|\\ &\le 4\acute{d}\,\|\tau\|_\infty\|\nabla(\mathbf{u} - \mathbf{u}_1)\|\,\|F\| \,+\, 4\acute{d}\,\|\tau\|_\infty\|\nabla(\mathbf{u} - \mathbf{u}_1)\|\,\|\nu h\,\mathbf{u}_1 \cdot \nabla F\|\\ &\le \epsilon_{14}\|F\| \,+\, \frac{4\acute{d}^2}{\epsilon_{14}}\|\tau\|_\infty^2\|\nabla(\mathbf{u} - \mathbf{u}_1)\|^2 \,+\, \nu^2 h^2\,\|\mathbf{u}_1 \cdot \nabla F\|^2\\ &\qquad + 4\acute{d}^2\|\tau\|_\infty^2\|\nabla(\mathbf{u} - \mathbf{u}_1)\|^2. \end{aligned}\tag{4.43}$$

What remains is to estimate the three B terms on the rhs of (4.25). We begin by rewriting the terms in a more convenient form:

$$\begin{aligned}-B(\mathbf{u}_1, \Gamma, F\,) \,-\, B(\mathbf{u}, \mathbf{u}_1, \tau, F) \,+\, B(\mathbf{u}_1, \tau, F) &= B(\mathbf{u}_1, \mathcal{T}, F) \,-\, B(\mathbf{u}, \mathbf{u}_1, \tau, F)\\ &= -B(\mathbf{u} - \mathbf{u}_1, \mathbf{u}_1, \mathcal{T}, F) \,-\, B(\mathbf{u}, \mathbf{u}_1, \Gamma, F)\\ &= B(\mathbf{u} - \mathbf{u}_1, \mathbf{u}_1, \Gamma, F) \,-\, B(\mathbf{u} - \mathbf{u}_1, \mathbf{u}_1, \tau, F)\\ &\qquad - B(\mathbf{u}, \mathbf{u}_1, \Gamma, F). \end{aligned}\tag{4.44}$$

For the first B term in (4.44) we have

$$\begin{aligned}B(\mathbf{u} - \mathbf{u}_1, \mathbf{u}_1, \Gamma, F) &= ((\mathbf{u} - \mathbf{u}_1) \cdot \nabla\Gamma, F) \,+\, ((\mathbf{u} - \mathbf{u}_1) \cdot \nabla\Gamma, \nu h\,\mathbf{u}_1 \cdot \nabla F)\\ &\quad + \frac{1}{2}\,(\nabla \cdot (\mathbf{u} - \mathbf{u}_1)\,\Gamma, F)\end{aligned}$$

$$\begin{aligned}
&\le \|(\mathbf{u}-\mathbf{u}_1)\cdot\nabla\Gamma\|\|F\| + \|(\mathbf{u}-\mathbf{u}_1)\cdot\nabla\Gamma\|\|\nu h\,\mathbf{u}_1\cdot\nabla F\| \\
&\quad + \frac{1}{2}\|\nabla\cdot(\mathbf{u}-\mathbf{u}_1)\,\Gamma\|\|F\| \\
&\le \epsilon_{15}\|F\|^2 + \left(\frac{1}{4\epsilon_{15}} + \frac{1}{4}\right)\|(\mathbf{u}-\mathbf{u}_1)\cdot\nabla\Gamma\|^2 + \nu^2h^2\|\mathbf{u}_1\cdot\nabla F\|^2 \\
&\quad + \epsilon_{16}\|F\|^2 + \frac{1}{16\epsilon_{16}}\|\nabla\cdot(\mathbf{u}-\mathbf{u}_1)\Gamma\|^2.
\end{aligned}\tag{4.45}$$

For I_u the interpolant of $\mathbf{u}$ we have, using (3.7), (3.8),

$$\begin{aligned}
\|\mathbf{u}-\mathbf{u}_1\|_\infty &\le \|\mathbf{u}-I_u\|_\infty + \|I_u-\mathbf{u}_1\|_\infty \\
&\le C_n h^{k+1-\acute{d}/2}\|\mathbf{u}\|_{k+1} + C_v h^{-\acute{d}/2}\|I_u-\mathbf{u}_1\| \\
&\le C_n h^{k+1-\acute{d}/2}\|\mathbf{u}\|_{k+1} + C_v h^{-\acute{d}/2}\|I_u-\mathbf{u}\| + C_v h^{-\acute{d}/2}\|\mathbf{u}-\mathbf{u}_1\| \\
&\le C_{nv} h^{k+1-\acute{d}/2}\|\mathbf{u}\|_{k+1} + C_v h^{-\acute{d}/2}\|\mathbf{u}-\mathbf{u}_1\|.
\end{aligned}\tag{4.46}$$

Using this estimate, we obtain that

$$\begin{aligned}
\|(\mathbf{u}-\mathbf{u}_1)\cdot\nabla\Gamma\| &\le \acute{d}\|\mathbf{u}-\mathbf{u}_1\|_\infty\|\nabla\Gamma\| \\
&\le \acute{d}\,(C_{nv}h^{k+1-\acute{d}/2}\|\mathbf{u}\|_{k+1} + C_v h^{-\acute{d}/2}\|\mathbf{u}-\mathbf{u}_1\|)\|\nabla\Gamma\|.
\end{aligned}\tag{4.47}$$

Also,

$$\begin{aligned}
\|\nabla\cdot(\mathbf{u}-\mathbf{u}_1)\,\Gamma\| &\le \acute{d}^{3/2}\|\nabla(\mathbf{u}-\mathbf{u}_1)\|\|\Gamma\|_\infty \\
&\le C_{vi}\acute{d}^{3/2}h^{m+1-\acute{d}/2}\,\|\mathbf{u}-\mathbf{u}_1\|_1\|\tau\|_{m+1}.
\end{aligned}\tag{4.48}$$

Combining (4.45), (4.47), and (4.48), we have

$$\begin{aligned}
B(\mathbf{u}-\mathbf{u}_1,\mathbf{u}_1,\Gamma,F) &\le (\epsilon_{15}+\epsilon_{16})\|F\|^2 + \nu^2h^2\|\mathbf{u}_1\cdot\nabla F\|^2 \\
&\quad + \left(\frac{1}{4\epsilon_{15}}+\frac{1}{4}\right)\acute{d}^2\,(C_{nv}h^{k+1-\acute{d}/2}\|\mathbf{u}\|_{k+1} + C_v h^{-\acute{d}/2}\|\mathbf{u}-\mathbf{u}_1\|)^2\|\nabla\Gamma\|^2 \\
&\quad + \frac{1}{16\epsilon_{16}}(C_{vi}\acute{d}^{3/2}h^{m+1-\acute{d}/2}\,\|\mathbf{u}-\mathbf{u}_1\|_1\|\tau\|_{m+1})^2.
\end{aligned}\tag{4.49}$$

For the second B term on the rhs of (4.44) we have

$$\begin{aligned}
B(\mathbf{u}-\mathbf{u}_1,\mathbf{u}_1,\tau,F) &= ((\mathbf{u}-\mathbf{u}_1)\cdot\nabla\tau,F) + ((\mathbf{u}-\mathbf{u}_1)\cdot\nabla\tau,\nu h\,\mathbf{u}_1\cdot\nabla F) \\
&\quad + \frac{1}{2}\,(\nabla\cdot(\mathbf{u}-\mathbf{u}_1)\,\tau,F) \\
&\le \|(\mathbf{u}-\mathbf{u}_1)\cdot\nabla\tau\|\,\|F\| + \|(\mathbf{u}-\mathbf{u}_1)\cdot\nabla\tau\|\,\|\nu h\,\mathbf{u}_1\cdot\nabla F\| \\
&\quad + \frac{1}{2}\|\nabla\cdot(\mathbf{u}-\mathbf{u}_1)\,\tau\|\,\|F\| \\
&\le \epsilon_{17}\|F\|^2 + \frac{1}{4\epsilon_{17}}\|(\mathbf{u}-\mathbf{u}_1)\cdot\nabla\tau\|^2 + \nu^2h^2\,\|\mathbf{u}_1\cdot\nabla F\|^2 \\
&\quad + \frac{1}{4}\|(\mathbf{u}-\mathbf{u}_1)\cdot\nabla\tau\|^2 \\
&\quad + \epsilon_{18}\|F\|^2 + \frac{1}{16\epsilon_{18}}\|\nabla\cdot(\mathbf{u}-\mathbf{u}_1)\,\tau\|^2 \\
&\le (\epsilon_{17}+\epsilon_{18})\|F\|^2 + \nu^2h^2\,\|\mathbf{u}_1\cdot\nabla F\|^2 \\
&\quad + \acute{d}^3\left(\frac{1}{4\epsilon_{17}}+\frac{1}{4}\right)\|\nabla\tau\|_\infty^2\|\mathbf{u}-\mathbf{u}_1\|^2 + \frac{\acute{d}^3}{16\epsilon_{18}}\|\tau\|_\infty^2\|\nabla(\mathbf{u}-\mathbf{u}_1)\|^2.
\end{aligned}\tag{4.50}$$

For the third B term on the rhs of (4.44) we have

$$\begin{aligned}
B(\mathbf{u}, \mathbf{u}_1, \Gamma, F) &= (\mathbf{u} \cdot \nabla\Gamma, F) \,+\, (\mathbf{u} \cdot \nabla\Gamma, \nu h\, \mathbf{u}_1 \cdot \nabla F) \,+\, \frac{1}{2}\,(\nabla \cdot \mathbf{u}\,\Gamma, F) \\
&\le \|\mathbf{u} \cdot \nabla\Gamma\|\,\|F\| \,+\, \|\mathbf{u} \cdot \nabla\Gamma\|\,\nu h\,\|\mathbf{u}_1 \cdot \nabla F\| \,+\, \frac{1}{2}\|\nabla \cdot \mathbf{u}\,\Gamma\|\,\|F\| \\
&\le (\epsilon_{19} \,+\, \epsilon_{20})\|F\|^2 \,+\, \nu^2 h^2\,\|\mathbf{u}_1 \cdot \nabla F\|^2 \\
&\quad + \acute{d}\left(\frac{1}{4\epsilon_{19}} + \frac{1}{4}\right)\|\mathbf{u}\|_\infty^2\|\nabla\Gamma\|^2 \,+\, \frac{\acute{d}^2}{16\epsilon_{20}}\|\nabla\mathbf{u}\|_\infty^2\|\Gamma\|^2. \qquad (4.51)
\end{aligned}$$

Returning to (4.25) and putting everything back together, we obtain

$$\begin{aligned}
&\alpha Re\frac{d}{dt}\|E\|^2 + \frac{\lambda}{2}\frac{d}{dt}\|F\|^2 + \frac{1}{2}\|F\|^2 + \frac{\alpha(1-\alpha)}{2}\|\nabla E\|^2 \\
&+ \left(\lambda\nu h - \frac{1}{2}\nu^2h^2 - \frac{\alpha\nu^2h^2}{2(1-\alpha)}\right)\|\mathbf{u}_1 \cdot \nabla F\|^2 - 2\alpha(\epsilon_1 + \epsilon_3 + \epsilon_4)\|\nabla E\|^2 \\
&- 2\alpha\frac{C_1^2}{4\epsilon_1}\|\nabla\mathbf{u}\|^2\|\mathbf{u} - \mathbf{u}_1\|_1^2 - 2\alpha\frac{C_1^2}{4\epsilon_3}\|\mathbf{u} - \mathbf{u}_1\|_1^2\|\nabla\Lambda\|^2 - 2\alpha\frac{C_2^2}{4\epsilon_4}\|\mathbf{u}\|_\infty^2\|\nabla\Lambda\|^2 \\
&\le 2\alpha Re\,\frac{C_{PF}^2}{4\epsilon_5}\|\Lambda_t\|^2 + 2\alpha\epsilon_5\|\nabla E\|^2 + \frac{1}{4\epsilon_6}\|\Gamma\|^2 + \epsilon_6\|F\|^2 \\
&+ \epsilon_8\|F\|^2 \,+\, \frac{1}{4\epsilon_8}\|\Gamma\|^2 \,+\, \frac{1}{4}\|\Gamma\|^2 \,+\, \nu^2h^2\|\mathbf{u}_1 \cdot \nabla F\|^2 \\
&+ \frac{1}{4}\|\Gamma\|^2 \,+\, \nu^2h^2\|\mathbf{u}_1 \cdot \nabla F\|^2 \,+\, \frac{\alpha^2}{2\epsilon_9}\|\nabla\Lambda\|^2 \,+\, \epsilon_9\|F\|^2 \,+\, \frac{\alpha^2}{2}\|\nabla\Lambda\|^2 \\
&+ \nu^2h^2\|\mathbf{u}_1 \cdot \nabla F\|^2 \,+\, \frac{\alpha^2}{4\epsilon_{10}}\|\Gamma\|^2 \,+\, \epsilon_{10}\|\nabla E\|^2 \,+\, \frac{\alpha^2(1-\alpha)^2}{4\epsilon_{11}}\|\nabla\Gamma\|^2 \\
&+ \epsilon_{11}\|\nabla E\|^2 + \frac{\lambda}{4\epsilon_6}\|\Gamma_t\|^2 \,+\, \lambda\epsilon_6\|F\|^2 \,+ +\, \frac{\alpha^2}{\epsilon_7}\acute{d}\|p - \mathcal{P}\|^2 \,+\, \epsilon_7\|\nabla E\|^2 \\
&+ \lambda\frac{4\acute{d}^2}{\epsilon_{12}}\|\nabla\mathbf{u}\|_\infty^2\|\tau - \tau_1\|^2 \,+\, \lambda 4\acute{d}^2\,\|\nabla\mathbf{u}\|_\infty^2\|\tau - \tau_1\|^2 \,+\, \lambda\epsilon_{12}\|F\|^2 \\
&+ \lambda\nu^2h^2\|\mathbf{u}_1 \cdot \nabla F\|^2 \\
&+ \lambda\left(\frac{4}{\epsilon_{13}} + 4\right)\left(2C_I h^{m+1-\acute{d}/4}\|\tau\|_{m+1} \,+\, C_I h^{-\acute{d}/4}\|\tau_1 - \tau\|\right)^2 \\
&\quad \times\left(2C_I h^{k-\acute{d}/4}\|\mathbf{u}\|_{k+1} \,+\, C_I h^{-\acute{d}/4}\|\mathbf{u} - \mathbf{u}_1\|_1\right)^2 \\
&+ \lambda\epsilon_{13}\|F\|^2 \,+\, \lambda\nu^2h^2\|\mathbf{u}_1 \cdot \nabla F\|^2 \,+\, \lambda 4\acute{d}^2\left(\frac{1}{\epsilon_{14}} + 1\right)\|\tau\|_\infty^2\|\mathbf{u} - \mathbf{u}_1\|^2 \\
&+ \lambda\epsilon_{14}\|F\|^2 + \lambda\nu^2h^2\|\mathbf{u}_1 \cdot \nabla F\|^2 \\
&+ \lambda\left(\frac{1}{4\epsilon_{15}} + \frac{1}{4}\right)\acute{d}^2\,\left(C_{nv}h^{k+1-\acute{d}/2}\|\mathbf{u}\|_{k+1} \,+\, C_v h^{-\acute{d}/2}\|\mathbf{u} - \mathbf{u}_1\|\right)^2\|\nabla\Gamma\|^2 \\
&+ \lambda\frac{1}{16\,\epsilon_{16}}\left(C_{vi}\acute{d}^{3/2}h^{m+1-\acute{d}/2}\,\|\mathbf{u} - \mathbf{u}_1\|_1\|\tau\|_{m+1}\right)^2 + \lambda\,(\epsilon_{15} \,+\, \epsilon_{16})\|F\|^2 \\
&+ \lambda\,\nu^2h^2\|\mathbf{u}_1 \cdot \nabla F\|^2 + \lambda\,(\epsilon_{17} \,+\, \epsilon_{18})\|F\|^2 \,+\, \lambda\,\nu^2h^2\,\|\mathbf{u}_1 \cdot \nabla F\|^2 \\
&+ \lambda\,\acute{d}^3\left(\frac{1}{4\epsilon_{17}} + \frac{1}{4}\right)\|\nabla\tau\|_\infty\,\|\mathbf{u} - \mathbf{u}_1\| \,+\, \lambda\,\frac{\acute{d}^3}{16\epsilon_{18}}\|\tau\|_\infty\|\nabla(\mathbf{u} - \mathbf{u}_1)\|^2 \\
&+ \lambda\,(\epsilon_{19} \,+\, \epsilon_{20})\|F\|^2 \,+\, \lambda\,\nu^2h^2\,\|\mathbf{u}_1 \cdot \nabla F\|^2
\end{aligned}$$

$$+\ \lambda\,\acute{d}\left(\frac{1}{4\epsilon_{19}}+\frac{1}{4}\right)\|\mathbf{u}\|_\infty^2\|\nabla\Gamma\|^2\ +\ \lambda\,\frac{\acute{d}^2}{16\epsilon_{20}}\|\nabla\mathbf{u}\|_\infty^2\|\Gamma\|^2. \tag{4.52}$$

We rewrite (4.52), collecting the terms involving E, ∇E, F, and ∇F on the lhs. The remaining terms are collected on the rhs and grouped as terms "controlled" by the *ball*, terms controlled by *interpolation approximation*, and terms controlled by both the ball and interpolation approximation. The resulting inequality is

$$\begin{aligned}
\alpha Re\,\frac{d}{dt}\|E\|^2 + \frac{\lambda}{2}\frac{d}{dt}\|F\|^2 + \left(\frac{\alpha(1-\alpha)}{2} - 2\alpha(\epsilon_1+\epsilon_3+\epsilon_4+\epsilon_5) - (\epsilon_7+\epsilon_{10}+\epsilon_{11})\right)\|\nabla E\|^2 \\
+ \left(\frac{1}{2} - (\epsilon_6+\epsilon_8+\epsilon_9) - \lambda(\epsilon_6+\epsilon_{12}+\epsilon_{13}+\epsilon_{14}+\epsilon_{15}+\epsilon_{16}+\epsilon_{17}\right. \\
\left. + \epsilon_{18}+\epsilon_{19}+\epsilon_{20})\right)\|F\|^2 \\
+ \left(\lambda\nu h - \nu^2h^2\left(\frac{7}{2} - \frac{\alpha}{2(1-\alpha)} + 6\lambda\right)\right)\|\mathbf{u}_1\cdot\nabla F\|^2 \\
\le \|\mathbf{u}-\mathbf{u}_1\|^2\left\{\lambda 4\acute{d}^2\left(\frac{1}{\epsilon_{14}}+1\right)\|\tau\|_\infty^2 + \lambda\acute{d}^3\left(\frac{1}{4\epsilon_{17}}+\frac{1}{4}\right)\|\nabla\tau\|_\infty^2\right\} \\
+ \|\tau-\tau_1\|^2\left\{\lambda\frac{4\acute{d}^2}{\epsilon_{12}}\|\nabla\mathbf{u}\|_\infty^2\ +\ 4\lambda\acute{d}^2\|\nabla\mathbf{u}\|_\infty^2\right\} \\
+ \|\mathbf{u}-\mathbf{u}_1\|_1^2\left\{2\alpha\frac{C_1^2}{4\epsilon_1}\|\nabla\mathbf{u}\|^2 + \lambda\frac{\acute{d}^3}{16\epsilon_{18}}\|\tau\|_\infty^2\right\} \\
+ \|\mathbf{u}-\mathbf{u}_1\|_1^2\|\tau-\tau_1\|^2\left\{\lambda\left(\frac{4}{\epsilon_{13}}+4\right)C_I^2 4h^{-\acute{d}}\right\} \\
+ \|\Gamma\|^2\left\{\frac{1}{4}+\frac{1}{4\epsilon_6}+\frac{\alpha^2}{4\epsilon_{10}}+\lambda\frac{\acute{d}^2}{16\epsilon_{20}}\|\nabla\mathbf{u}\|_\infty^2\right\}\ +\ \|\Gamma_t\|^2\left\{\frac{\lambda}{4\epsilon_6}\right\} \\
+ \|\nabla\Gamma\|^2\left\{\frac{\alpha^2(1-\alpha)^2}{4\epsilon_{11}}\ +\ 2C_{nv}^2h^{2k+2-\acute{d}}\|\mathbf{u}\|_{k+1}^2\ +\ \lambda\acute{d}\left(\frac{1}{4\epsilon_{19}}+\frac{1}{4}\right)\|\mathbf{u}\|_\infty^2\ \right\} \\
+ \|\Lambda_t\|^2\left\{2\alpha Re\frac{C_{PF}^2}{4\epsilon_5}\right\}\ +\ \|\nabla\Lambda\|^2\left\{2\alpha\frac{C_2^2}{4\epsilon_4}\|\mathbf{u}\|_\infty^2\ +\ \frac{\alpha^2}{2\epsilon_9}\ +\ \frac{\alpha^2}{2}\right\} \\
+ \|p-\mathcal{P}\|^2\left\{\acute{d}\frac{\alpha^2}{\epsilon_7}\right\}\ +\ \lambda\left(\frac{4}{\epsilon_{13}}+4\right)C_I^2 64h^{2m+2k+2-\acute{d}}\|\tau\|_{m+1}^2\|\mathbf{u}\|_{k+1}^2 \\
+ \|\mathbf{u}-\mathbf{u}_1\|^2\left\{2\lambda\left(\frac{1}{4\epsilon_{15}}+\frac{1}{4}\right)\acute{d}^2C_v^2h^{-\acute{d}}\|\nabla\Gamma\|^2\right\} \\
+ \|\mathbf{u}-\mathbf{u}_1\|_1^2\left\{2\alpha\frac{C_1^2}{4\epsilon_3}\|\nabla\Lambda\|^2\right\} \\
+ \|\tau-\tau_1\|^2\left\{\lambda\left(\frac{4}{\epsilon_{13}}+4\right)C_I^2 16h^{2k-\acute{d}}\|\mathbf{u}\|_{k+1}^2\right\} \\
+ \|\mathbf{u}-\mathbf{u}_1\|_1^2\left\{\lambda\left(\frac{4}{\epsilon_{13}}+4\right)C_I^2 16h^{2m+2-\acute{d}}\|\tau\|_{m+1}^2\right. \\
\left. + \frac{1}{16\epsilon_{16}}C_{vi}^2\acute{d}^3h^{2m+2-\acute{d}}\|\tau\|_{m+1}^2\right\}.
\end{aligned} \tag{4.53}$$

With our assumptions that $0 < \alpha < 1$ and $\lambda > 0$, we can choose values for the ϵ_i's, and νh sufficiently small, such that the lhs of (4.53) is bounded below by

$$\alpha Re \frac{d}{dt}\|E\|^2 + \frac{\lambda}{2}\frac{d}{dt}\|F\|^2 + \frac{1}{4}\|F\|^2 + \frac{\alpha(1-\alpha)}{4}\|\nabla E\|^2 + \frac{\lambda \nu h}{2}\|\mathbf{u}_1 \cdot \nabla F\|^2. \tag{4.54}$$

Let D_i, $i = 1, \dots, 6$, denote constants dependent upon $\mathbf{u}$, p, τ, their derivatives, and T. (Recall the definition of c^*, R in (4.19), and D_0 in Theorem 4.4.) As usual, C_j, $j = 4, \dots, 10$, denote constants independent of the solution $\mathbf{u}$, p, τ and the mesh parameter h.

Using (4.54) and integrating (4.53), we obtain

$$\begin{aligned} \|E\|^2(t) \ + \ \|F\|^2(t) \ + \ \int_0^t \|\nabla E\|^2(s)\, ds \le {} & R^2 C_4 D_0 \\ & + R^4 C_5 h^{-\acute{d}} \\ & + D_1 h^{2m+2} \ + \ D_2 h^{2m+2} \\ & + C_6 D_0 h^{2m} \ + \ D_3 h^{2k+2m+2-\acute{d}} \\ & + D_4 h^{2k+2} \ + \ C_7 D_0 h^{2k} \\ & + C_8 D_0 h^{2q+2} \ + \ D_5 h^{2k+2m+2-\acute{d}} \\ & + R^2 C_9 D_0 h^{2m-\acute{d}} \\ & + R^2 D_6 h^{2k} \\ & + R^2 C_6 D_0 h^{2k-\acute{d}} \\ & + R^2 C_7 D_0 h^{2m+2-\acute{d}}. \end{aligned} \tag{4.55}$$

Now, in view of (4.24), we have that for h, D_0, and c^* sufficiently small

$$\begin{aligned} \|\tau - \tau_2\|^2(t) &\le 2\|F\|^2(t) \ + \ 2\|\Gamma\|^2(t) \\ &\le cR^2 + C_{10} D_0 \left(h^{2m} + h^{2k}\right) + 2D_0 h^{2m+2} \\ &\le \tilde{c} R^2, \end{aligned} \tag{4.56}$$

where $0 < \tilde{c} < 1$. Similarly, for h sufficiently small

$$\begin{aligned} \|\mathbf{u} - \mathbf{u}_2\|^2(t) &\le 2\|E\|^2(t) \ + \ 2\|\Lambda\|^2(t) \\ &\le cR^2 + C_{10} D_0 \left(h^{2m} + h^{2k}\right) + 2D_0 h^{2k+2}; \end{aligned} \tag{4.57}$$

hence

$$\int_0^T \|\mathbf{u} - \mathbf{u}_2\|^2(t)\, dt \le \frac{\tilde{c}}{2} R^2. \tag{4.58}$$

Also, for h sufficiently small

$$\begin{aligned} \int_0^T \|\nabla(\mathbf{u} - \mathbf{u}_2)\|^2(t)\, dt &\le 2\int_0^T \|\nabla E\|^2(t)\, dt \ + \ 2\int_0^T \|\nabla \Lambda\|^2(t)\, dt \\ &\le c1 R^2 \ + \ 2D_0 T h^{2k} \\ &\le \frac{\tilde{c}}{2} R^2. \end{aligned} \tag{4.59}$$

Combining (4.56)–(4.59), we have for h sufficiently small that ξ is a strict contraction on the *ball* defined in (4.19).

Step 4. A direct application of Schauder's fixed point theorem now establishes the uniqueness of the approximation and the stated error estimates. □

5. Fully discrete approximation. In this section we analyze a fully discrete approximation to (2.14), (2.15).

We assume that the fluid flow satisfies the following properties:

$$\|\mathbf{u}\|_\infty,\ \|\tau\|_\infty,\ \|\nabla\mathbf{u}\|_\infty,\ \|\nabla\tau\|_\infty\ \le M \tag{5.1}$$

for all $t \in [0, T]$.

Note that it follows from (5.1) and inverse estimates that

$$\|\mathcal{U}^n\|_\infty, \|\nabla\mathcal{U}^n\|_\infty \le \tilde{M} \approx M. \tag{5.2}$$

Below, for simplicity, we take $\tilde{M} = M$.

To simplify the notation, the following definition is used in the analysis.

Definition 5.1.

$$b(\mathbf{u}, \tau, \psi)\ :=\ (\mathbf{u}\cdot\nabla\tau, \psi). \tag{5.3}$$

To obtain the fully discretized approximation, the time derivatives are replaced by backward differences, and the nonlinear terms are lagged. As we are assuming "slow flow," i.e., $Re \equiv O(1)$, we use a conforming FE method to discretize the momentum equation. For the constitutive equation for stress, we use a SUPG discretization to control the production of spurious oscillations in the approximation. The discrete approximating system of equations is then the following.

Approximating system. For $n = 1, 2, \ldots, N$, find $\mathbf{u}_h^n \in Z_h, \tau_\mathbf{h}^\mathbf{n} \in S_h$ such that

$$Re\,(d_t\mathbf{u}_h^n, \mathbf{v}) + Re\ c\left(\mathbf{u}_h^{n-1}, \mathbf{u}_h^n, \mathbf{v}\right) + (1-\alpha)\,(\nabla\mathbf{u}_h^n, \nabla\mathbf{v}) + (\tau_h^n, D(\mathbf{v})) = (\mathbf{f}^n, \mathbf{v})\,, \qquad \mathbf{v} \in Z_h, \tag{5.4}$$

$$\frac{1}{\lambda}\,(\tau_h^n, \tilde{\sigma}) + (d_t\tau_h^n, \sigma) + b\left(\mathbf{u}_h^{n-1}, \tau_h^n, \tilde{\sigma}\right) - \overline{\lambda}\,(D(\mathbf{u}_h^n), \tilde{\sigma}) = -\left(g_a(\tau_h^{n-1}, \nabla\mathbf{u}_h^{n-1}), \tilde{\sigma}\right), \qquad \sigma \in S_h, \tag{5.5}$$

where $\tilde{\sigma} := \sigma + \nu\sigma_u^n$, $\sigma_u^n := \mathbf{u}_h^{n-1}\cdot\nabla\sigma$, ν is a small positive constant, and $\overline{\lambda} := (2\alpha)/\lambda$.

The parameter $\nu > 0$ is used to suppress the production of spurious oscillations in the approximation. Note that for $\nu = 0$ the discretization of the constitutive equation is a conforming Galerkin method. The goal in choosing ν is to keep it as small as possible, but large enough to control the generation of catastrophic spurious oscillations in the approximate stress.

To ensure computability of the algorithm, we begin by showing that (5.4)–(5.5) is uniquely solvable for $\mathbf{u}_h$ and τ_h at each time step n. We use the following induction hypothesis:

$$\left\|\mathbf{u}_h^{n-1}\right\|_\infty, \left\|\tau_h^{n-1}\right\|_\infty \le K. \tag{IH1}$$

Lemma 5.2. *Assume* (IH1) *is true. For sufficiently small step size Δt, there exists a unique solution $(\mathbf{u}_h^n, \tau_h^n) \in Z_h \times S_h$ satisfying* (5.4)–(5.5).

Proof. For notational simplicity, in this proof we drop the subscript h from the variables. Choosing $\mathbf{v} = \mathbf{u}_h^n,\ \sigma = \tau_h^n$, multiplying (5.4) by $\overline{\lambda}$, and adding to (5.5), we obtain

$$\begin{aligned} &a(\mathbf{u}^n, \tau^n; \mathbf{u}^n, \tau^n) \\ &\quad = \overline{\lambda}\,(\mathbf{f}^n, \mathbf{u}^n) + \overline{\lambda}\frac{Re}{\Delta t}\left(\mathbf{u}^{n-1}, \mathbf{u}^n\right) - \left(g_a\left(\tau^{n-1}, \nabla\mathbf{u}^{n-1}\right), \tilde{\tau}^n\right) + \frac{1}{\Delta t}\left(\tau^{n-1}, \tau^n\right), \end{aligned} \tag{5.6}$$

where the bilinear form $a(\mathbf{u},\tau;\mathbf{v},\sigma)$ is defined as

$$\begin{aligned}
&a(\mathbf{u},\tau;\mathbf{v},\sigma)\\
&\quad := \overline{\lambda}\frac{Re}{\Delta t}\,(\mathbf{u},\mathbf{v}) + \overline{\lambda}\;Re\;c(\mathbf{u}^{n-1},\mathbf{u},\mathbf{v}) + \overline{\lambda}(1-\alpha)\,(\nabla\mathbf{u},\nabla\mathbf{v}) + \frac{1}{\lambda}\,(\tau,\tilde{\sigma}) + \frac{1}{\Delta t}\,(\tau,\sigma)\\
&\qquad + b\left(\mathbf{u}^{n-1},\tau,\sigma\right) + b\left(\mathbf{u}^{n-1},\tau,\nu\mathbf{u}^{n-1}\cdot\nabla\sigma\right) - \overline{\lambda}\left(D(\mathbf{u}),\nu\mathbf{u}^{n-1}\cdot\nabla\sigma\right).
\end{aligned}$$

We now estimate the terms in $a(\mathbf{u}^n,\tau^n;\mathbf{u}^n,\tau^n)$. We have

$$\begin{aligned}
\left|c(\mathbf{u}^{n-1},\mathbf{u},\mathbf{u})\right| = \left|\left(\mathbf{u}^{n-1}\cdot\nabla\mathbf{u},\mathbf{u}\right)\right| &\le \acute{d}^{\frac{1}{2}}\left\|\mathbf{u}^{n-1}\right\|_\infty\|\nabla\mathbf{u}\|\,\|\mathbf{u}\|\\
&\le \epsilon_1\|\nabla\mathbf{u}\|^2 + \frac{\acute{d}K^2}{4\epsilon_1}\|\mathbf{u}\|^2,\\
\left|b(\mathbf{u}^{n-1},\tau,\tau)\right| = \left|\left(\mathbf{u}^{n-1}\cdot\nabla\tau,\tau\right)\right| &\le \left\|\mathbf{u}^{n-1}\cdot\nabla\tau\right\|\|\tau\|\\
&\le \epsilon_2\left\|\mathbf{u}^{n-1}\cdot\nabla\tau\right\|^2 + \frac{1}{4\epsilon_2}\|\tau\|^2,\\
b(\mathbf{u}^{n-1},\tau,\nu\mathbf{u}^{n-1}\cdot\nabla\tau) &= \nu\left\|\mathbf{u}^{n-1}\cdot\nabla\tau\right\|^2,\\
\left|\left(D(\mathbf{u}),\nu\mathbf{u}^{n-1}\cdot\nabla\tau\right)\right| &\le \|D(\mathbf{u})\|\left\|\nu\mathbf{u}^{n-1}\cdot\nabla\tau\right\|\\
&\le \epsilon_3\|D(\mathbf{u})\|^2 + \frac{\nu^2}{4\epsilon_3}\left\|\mathbf{u}^{n-1}\cdot\nabla\tau\right\|^2\\
&\le \epsilon_3\|\nabla\mathbf{u}\|^2 + \frac{\nu^2}{4\epsilon_3}\left\|\mathbf{u}^{n-1}\cdot\nabla\tau\right\|^2.
\end{aligned}$$

Applying these inequalities to the bilinear form $a(\cdot,\cdot;\cdot,\cdot)$ yields

$$\begin{aligned}
a(\mathbf{u}^n,\tau^n;\mathbf{u}^n,\tau^n) \ge{}& \overline{\lambda}\;Re\left(\frac{1}{\Delta t} - \frac{\acute{d}K^2}{4\epsilon_1}\right)\|\mathbf{u}^n\|^2 + \overline{\lambda}\left((1-\alpha) - Re\;\epsilon_1 - \epsilon_3\right)\|\nabla\mathbf{u}\|^2\\
&+\left(\frac{1}{\lambda} + \frac{1}{\Delta t} - \frac{1}{4\epsilon_2}\right)\|\tau^n\|^2 + \left(\nu - \epsilon_2 - \frac{\nu^2}{4\epsilon_3}\right)\left\|\mathbf{u}^{n-1}\cdot\nabla\tau^n\right\|^2.
\end{aligned}$$

Choosing $\epsilon_1 = \frac{(1-\alpha)}{4\,Re}, \epsilon_2 = \frac{\nu}{3}, \epsilon_3 = \frac{(1-\alpha)}{4}, \nu \le \frac{2(1-\alpha)}{3}$, and $\Delta t \le \min\{\frac{1-\alpha}{Re\;\acute{d}K^2},\nu\}$, it follows that the bilinear form $a(\cdot,\cdot;\cdot,\cdot)$ is positive. Hence, (5.6) has at most one solution. Since (5.6) is a finite dimensional linear system, the uniqueness of the solution implies the existence of the solution. □

The discrete Gronwall's lemma plays an important role in the following analysis.

LEMMA 5.3 (discrete Gronwall's lemma; see [9]). *Let Δt, H, and a_n, b_n, c_n, γ_n (for integers $n \ge 0$) be nonnegative numbers such that*

$$a_l \;+\; \Delta t\sum_{n=0}^{l} b_n \;\le\; \Delta t\sum_{n=0}^{l}\gamma_n\,a_n \;+\; \Delta t\sum_{n=0}^{l} c_n \;+\; H \quad \textit{for } l \ge 0.$$

Suppose that $\Delta t\,\gamma_n < 1\,\forall\,n$, and set $\sigma_n = (1-\Delta t\,\gamma_n)^{-1}$. Then

$$a_l \;+\; \Delta t\sum_{n=0}^{l} b_n \;\le\; \exp\left(\Delta t\sum_{n=0}^{l}\sigma_n\,\gamma_n\right)\left\{\Delta t\sum_{n=0}^{l} c_n \;+\; H\right\} \quad \textit{for } l \ge 0. \tag{5.7}$$

□

5.1. Analysis of the fully discrete approximation. In this section we analyze the error between the finite element approximation given by (5.4), (5.5) and the true solution. A priori error estimates for the approximation are given in Theorem 5.4.

THEOREM 5.4. *Assume that the system* (2.3)–(2.8) *(and thus,* (2.14)–(2.15)*) has a solution* $(\mathbf{u}, \tau, \mathbf{p}) \in C^2(0,T;H^{k+1}) \times C^2(0,T;H^{m+1}) \times C(0,T;H^{q+1})$*. In addition, assume that* $\Delta t,\ \nu \leq c\,h^{d/2}$*, and*

$$\|\mathbf{u}\|_\infty, \|\nabla \mathbf{u}\|_\infty, \|\tau\|_\infty, \|\nabla\tau\|_\infty \leq M \ \forall t \in [0,T]. \tag{5.8}$$

Then, the finite element approximation (5.4)–(5.5) *is convergent to the solution of* (2.14)–(2.15) *on the interval* $(0,T)$ *as* $\Delta t,\ h \to 0$*. In addition, the approximation* $(\mathbf{u}_h, \tau_h)$ *satisfies the following error estimates:*

$$\left\|\left|\mathbf{u}_h - \mathbf{u}\right|\right\|_{\infty,0} + \left\|\left|\tau_h - \tau\right|\right\|_{\infty,0} \leq \mathbf{F}(\Delta t, \nu, h), \tag{5.9}$$

$$\left\|\left|\mathbf{u}_h - \mathbf{u}\right|\right\|_{0,1} + \left\|\left|\tau_h - \tau\right|\right\|_{0,0} \leq \mathbf{F}(\Delta t, \nu, h), \tag{5.10}$$

where

$$\begin{aligned}
\mathbf{F}(\Delta t, \nu, h) = C\ &\left(h^k \left\|\left|\mathbf{u}\right|\right\|_{0,k+1} + h^{k+1} \left\|\left|\mathbf{u}_t\right|\right\|_{0,k+1}\right) \\
&+ C\ \left(h^m \left\|\left|\tau\right|\right\|_{0,m+1} + h^{m+1} \left\|\left|\tau_t\right|\right\|_{0,m+1}\right) \\
&+ C\ h^{q+1} \left\|\left|p\right|\right\|_{0,q+1} + C\left(h^{k+1} \left\|\left|\mathbf{u}\right|\right\|_{\infty,k+1} + h^{m+1} \left\|\left|\tau\right|\right\|_{\infty,m+1}\right) \\
&+ C\,|\Delta t| \left(\|\mathbf{u}_t\|_{0,1} + \|\mathbf{u}_{tt}\|_{0,0} + \|\tau_t\|_{0,1} + \|\tau_{tt}\|_{0,0}\right) \\
&+ C\ \nu \left(\left\|\left|\tau_t\right|\right\|_{0,1} + \left\|\left|\tau_t\right|\right\|_{\infty,0}\right).
\end{aligned}$$

In order to establish the estimates (5.9)–(5.10), we begin by introducing the following notation. Let $\mathbf{u}^n = \mathbf{u}(t_n), \tau^n = \tau(t_n)$ represent the solution of (2.14)–(2.15), and $\mathbf{u}^n_h, \tau^n_h$ denote the solution of (5.4)–(5.5).

Define $\mathbf{\Lambda}^n, \mathbf{E}^n, \mathbf{\Gamma}^n, \mathbf{F}^n, \epsilon_u, \epsilon_\tau$ as

$$\begin{aligned}
\mathbf{\Lambda}^n &= \mathbf{u}^n - \mathcal{U}^n, & \mathbf{E}^n &= \mathcal{U}^n - \mathbf{u}^n_h, \\
\mathbf{\Gamma}^n &= \tau^n - \mathcal{T}^n, & \mathbf{F}^n &= \mathcal{T}^n - \tau^n_h, \\
\epsilon_u &= \mathbf{u} - \mathbf{u}^n_h, & \epsilon_\tau &= \tau - \tau^n_h.
\end{aligned}$$

The proof of Theorem 5.4 is established in three steps:

1. Prove a lemma, assuming two induction hypotheses.
2. Show that the induction hypotheses are true.
3. Prove the error estimates given in (5.9), (5.10).

Step 1. We prove the following lemma.

LEMMA 5.5. *Under the induction hypothesis* (IH1) *and the additional assumption*

$$\sum_{n=1}^{l-1} \Delta t \left\|\nabla E^n\right\|_\infty \leq 1, \tag{IH2}$$

we have that

$$\left\|\mathbf{E}^l\right\|^2 + \left\|\mathbf{F}^l\right\|^2 \leq G(\Delta t, h, \nu), \tag{5.11}$$

where

$$\begin{aligned}
G(\Delta t, h, \nu) = C\ & \left(h^{2k}\, |||\mathbf{u}|||^2_{0,k+1} + h^{2k+2}\, |||\mathbf{u}_t|||^2_{0,k+1}\right) \\
& + C\ \left(h^{2m}\, |||\tau|||^2_{0,m+1} + h^{2m+2}\, |||\tau_t|||^2_{0,m+1}\right) \\
& + C\ h^{2q+2}\, |||p|||^2_{0,q+1} + C\, |\Delta t|^2 \left(|||\mathbf{u}_t|||^2_{0,1} + |||\mathbf{u}_{tt}|||^2_{0,0} + |||\tau_t|||^2_{0,1} + |||\tau_{tt}|||^2_{0,0}\right) \\
& + C\ \nu^2 \left(|||\tau_t|||^2_{0,1} + |||\tau_t|||^2_{\infty,0}\right).
\end{aligned}$$

Proof of Lemma 5.5. From (2.14)–(2.15), it is clear that the true solution $(\mathbf{u}, \tau)$ satisfies

$$\begin{aligned}
& Re\ \left(d_t\mathbf{u}^n, \mathbf{v}\right) + Re\ c\left(\mathbf{u}_h^{n-1}, \mathbf{u}^n, \mathbf{v}\right) + (1-\alpha)\left(\nabla\mathbf{u}^n, \nabla\mathbf{v}\right) + \left(\tau^n, D(\mathbf{v})\right) \\
& \qquad = \left(\mathbf{f}^n, \mathbf{v}\right) + \left(p^n, \nabla\cdot\mathbf{v}\right) + R_1(\mathbf{v}) \qquad \forall\ \mathbf{v} \in Z_h, \qquad (5.12)
\end{aligned}$$

$$\begin{aligned}
& \left(d_t\tau^n, \sigma\right) + b\left(\mathbf{u}_h^{n-1}, \tau^n, \tilde{\sigma}\right) - \hat{\lambda}\left(D(\mathbf{u}^n), \tilde{\sigma}\right) + \frac{1}{\lambda}\left(\tau^n, \tilde{\sigma}\right) \\
& \qquad = -\left(g_a\left(\tau_h^{n-1}, \nabla\mathbf{u}_h^{n-1}\right), \tilde{\sigma}\right) + R_2(\sigma) \qquad \forall \sigma \in S_h, \qquad (5.13)
\end{aligned}$$

where $\hat{\lambda} := (2\alpha)/\lambda$,

$$R_1(\mathbf{v}) := Re\ \left(d_t\mathbf{u}^n, \mathbf{v}\right) - Re\ \left(\mathbf{u}_t^n, \mathbf{v}\right) + Re\ c(\mathbf{u}_h^{n-1}, \mathbf{u}^n, \mathbf{v}) - Re\ c(\mathbf{u}^n, \mathbf{u}^n, \mathbf{v}),$$

and

$$\begin{aligned}
R_2(\sigma) := {} & \left(d_t\tau^n, \sigma\right) - \left(\tau_t^n, \sigma\right) - \nu\left(\tau_t^n, \mathbf{u}_h^{n-1}\cdot\nabla\sigma\right) + b(\mathbf{u}_h^{n-1}, \tau^n, \tilde{\sigma}) \\
& - b(\mathbf{u}^n, \tau^n, \tilde{\sigma}) + \left(g_a\left(\tau_h^{n-1}, \nabla\mathbf{u}_h^{n-1}\right), \tilde{\sigma}\right) - \left(g_a\left(\tau^n, \nabla\mathbf{u}^n\right), \tilde{\sigma}\right).
\end{aligned}$$

Subtracting (5.4)–(5.5) from (5.12)–(5.13), we obtain the following equations for ϵ_u and ϵ_τ:

$$\begin{aligned}
& Re\ \left(d_t\epsilon_u, \mathbf{v}\right) + Re\ c(\mathbf{u}_h^{n-1}, \epsilon_u, \mathbf{v}) + (1-\alpha)\left(\nabla\epsilon_u, \nabla\mathbf{v}\right) + \left(\epsilon_\tau, D(\mathbf{v})\right) \\
& \qquad = \left(p^n, \nabla\cdot\mathbf{v}\right) + R_1(\mathbf{v})\ \ \forall\ \mathbf{v}\ \in\ Z_h, \qquad (5.14)
\end{aligned}$$

$$(5.15) \quad \left(d_t\epsilon_\tau, \sigma\right) + b(\mathbf{u}_h^{n-1}, \epsilon_\tau, \tilde{\sigma}) - \hat{\lambda}\left(D(\epsilon_u), \tilde{\sigma}\right) + \frac{1}{\lambda}\left(\epsilon_\tau, \tilde{\sigma}\right) = R_2(\sigma)\quad \forall\ \sigma\ \in\ S_h.$$

Substituting $\epsilon_u = \mathbf{E}^n + \mathbf{\Lambda}^n,\ \epsilon_\tau = \mathbf{F}^n + \mathbf{\Gamma}^n,\ \mathbf{v} = \mathbf{E}^n,\ \sigma = \mathbf{F}^n$ into (5.14)–(5.15), we obtain

(5.16)

$$Re\left(d_t\mathbf{E}^n, \mathbf{E}^n\right) + Re\, c(\mathbf{u}_h^{n-1}, \mathbf{E}^n, \mathbf{E}^n) + (1-\alpha)\left(\nabla\mathbf{E}^n, \nabla\mathbf{E}^n\right) + \left(\mathbf{F}^n, D(\mathbf{E}^n)\right) = \mathcal{F}_1(\mathbf{E}^n),$$

$$(5.17) \qquad \left(d_t\mathbf{F}^n, \mathbf{F}^n\right) + b(\mathbf{u}_h^{n-1}, \mathbf{F}^n, \tilde{\mathbf{F}}^n) - \hat{\lambda}(D(\mathbf{E}^n), \tilde{\mathbf{F}}^n) + \frac{1}{\lambda}(\mathbf{F}^n, \tilde{\mathbf{F}}^n) = \mathcal{F}_2(\mathbf{F}^n),$$

where

$$\begin{aligned}
\mathcal{F}_1(\mathbf{E}^n) = {} & \left(p^n, \nabla\cdot\mathbf{E}^n\right) + R_1(\mathbf{E}^n) - Re\ \left(d_t\mathbf{\Lambda}^n, \mathbf{E}^n\right) - Re\ c(\mathbf{u}_h^{n-1}, \mathbf{\Lambda}^n, \mathbf{E}^n) \\
& - (1-\alpha)\left(\nabla\mathbf{\Lambda}^n, \nabla\mathbf{E}^n\right) - \left(\mathbf{\Gamma}^n, D(\mathbf{E}^n)\right), \\
\mathcal{F}_2(\mathbf{F}^n) = {} & R_2(\mathbf{F}^n) - \left(d_t\mathbf{\Gamma}^n, \mathbf{F}^n\right) - b(\mathbf{u}_h^{n-1}, \mathbf{\Gamma}^n, \tilde{\mathbf{F}}^n) + \hat{\lambda}(D(\mathbf{\Lambda}^n), \tilde{\mathbf{F}}^n) - \frac{1}{\lambda}(\mathbf{\Gamma}^n, \tilde{\mathbf{F}}^n).
\end{aligned}$$

Multiplying (5.16) by $\hat{\lambda}$ and adding to (5.17), we obtain the single equation

$$\begin{aligned} &\mathit{Re}\,\hat{\lambda}\,(d_t\mathbf{E}^n,\mathbf{E}^n) + \mathit{Re}\,\hat{\lambda}\,c(\mathbf{u}_h^{n-1},\mathbf{E}^n,\mathbf{E}^n) + (1-\alpha)\hat{\lambda}\,(\nabla\mathbf{E}^n,\nabla\mathbf{E}^n) + (d_t\mathbf{F}^n,\mathbf{F}^n) \\ &\qquad + b(\mathbf{u}_h^{n-1},\mathbf{F}^n,\tilde{\mathbf{F}}^n) - \hat{\lambda}\left(D(\mathbf{E}^n),\nu\mathbf{u}_h^{n-1}\cdot\nabla\mathbf{F}^n\right) + \frac{1}{\lambda}(\mathbf{F}^n,\tilde{\mathbf{F}}^n) \\ &\qquad = \hat{\lambda}\mathcal{F}_1(\mathbf{E}^n) + \mathcal{F}_2(\mathbf{F}^n). \end{aligned} \tag{5.18}$$

Note that

$$\begin{aligned} (d_t\mathbf{E}^n,\mathbf{E}^n) &= \frac{1}{\Delta t}\left[(\mathbf{E}^n,\mathbf{E}^n) - \left(\mathbf{E}^{n-1},\mathbf{E}^n\right)\right] \\ &\geq \frac{1}{\Delta t}\left[\|\mathbf{E}^n\|^2 - \|\mathbf{E}^n\|\,\left\|\mathbf{E}^{n-1}\right\|\right] \\ &\geq \frac{1}{2\Delta t}\left[\|\mathbf{E}^n\|^2 - \left\|\mathbf{E}^{n-1}\right\|^2\right], \end{aligned}$$

and similarly, $(d_t\mathbf{F}^n,\mathbf{F}^n) \geq \frac{1}{2\Delta t}[\|\mathbf{F}^n\|^2 - \|\mathbf{F}^{n-1}\|^2]$. Thus, we have

$$\begin{aligned} &\frac{\mathit{Re}\,\hat{\lambda}}{2\Delta t}\left[\|\mathbf{E}^n\|^2 - \left\|\mathbf{E}^{n-1}\right\|^2\right] + \frac{1}{2\Delta t}\left[\|\mathbf{F}^n\|^2 - \left\|\mathbf{F}^{n-1}\right\|^2\right] + (1-\alpha)\hat{\lambda}\,\|\nabla\mathbf{E}^n\|^2 \\ &\qquad + \nu\left\|\mathbf{u}_h^{n-1}\cdot\nabla\mathbf{F}^n\right\|^2 + \frac{1}{\lambda}\,\|\mathbf{F}^n\|^2 \\ &\leq -\mathit{Re}\,\hat{\lambda}\,c(\mathbf{u}_h^{n-1},\mathbf{E}^n,\mathbf{E}^n) - b(\mathbf{u}_h^{n-1},\mathbf{F}^n,\mathbf{F}^n) + \hat{\lambda}\left(D(\mathbf{E}^n),\nu\mathbf{u}_h^{n-1}\cdot\nabla\mathbf{F}^n\right) \\ &\quad - \frac{1}{\lambda}\left(\mathbf{F}^n,\nu\mathbf{u}_h^{n-1}\cdot\nabla\mathbf{F}^n\right) + \hat{\lambda}\mathcal{F}_1(\mathbf{E}^n) + \mathcal{F}_2(\mathbf{F}^n). \end{aligned} \tag{5.19}$$

Multiplying (5.19) by Δt and summing from $n = 1$ to l yields

$$\begin{aligned} &\frac{\mathit{Re}\,\hat{\lambda}}{2}\left[\left\|\mathbf{E}^l\right\|^2 - \left\|\mathbf{E}^0\right\|^2\right] + \frac{1}{2}\left[\left\|\mathbf{F}^l\right\|^2 - \left\|\mathbf{F}^0\right\|^2\right] + (1-\alpha)\hat{\lambda}\sum_{n=1}^{l}\Delta t\,\|\nabla\mathbf{E}^n\|^2 \\ &\qquad + \nu\sum_{n=1}^{l}\Delta t\left\|\mathbf{u}_h^{n-1}\cdot\nabla\mathbf{F}^n\right\|^2 + \frac{1}{\lambda}\sum_{n=1}^{l}\Delta t\,\|\mathbf{F}^n\|^2 \\ &\quad \leq \Delta t\sum_{n=1}^{l}\Bigg[-\mathit{Re}\,\hat{\lambda}\,c(\mathbf{u}_h^{n-1},\mathbf{E}^n,\mathbf{E}^n) - b(\mathbf{u}_h^{n-1},\mathbf{F}^n,\mathbf{F}^n) \\ &\qquad\qquad + \hat{\lambda}\left(D(\mathbf{E}^n),\nu\mathbf{u}_h^{n-1}\cdot\nabla\mathbf{F}^n\right) - \frac{1}{\lambda}\left(\mathbf{F}^n,\nu\mathbf{u}_h^{n-1}\cdot\nabla\mathbf{F}^n\right)\Bigg] \\ &\qquad + \hat{\lambda}\Delta t\sum_{n=1}^{l}\mathcal{F}_1(\mathbf{E}^n) + \Delta t\sum_{n=1}^{l}\mathcal{F}_2(\mathbf{F}^n). \end{aligned} \tag{5.20}$$

We now estimate each term on the rhs of (5.20). For $c(\mathbf{u}_h^{n-1},\mathbf{E}^n,\mathbf{E}^n)$ we have that

$$\begin{aligned} \left|c(\mathbf{u}_h^{n-1},\mathbf{E}^n,\mathbf{E}^n)\right| &\leq \left|\left(\mathbf{u}_h^{n-1}\cdot\nabla\mathbf{E}^n,\mathbf{E}^n\right)\right| \\ &\leq \left\|\mathbf{u}_h^{n-1}\cdot\nabla\mathbf{E}^n\right\|\,\|\mathbf{E}^n\| \\ &\leq \left\|\mathbf{u}_h^{n-1}\right\|_\infty \acute{d}^{\frac{1}{2}}\,\|\nabla\mathbf{E}^n\|\,\|\mathbf{E}^n\| \\ &\leq \epsilon_1\,\|\nabla\mathbf{E}^n\|^2 + \frac{\acute{d}K^2}{4\epsilon_1}\,\|\mathbf{E}^n\|^2, \qquad \text{using (IH1).} \end{aligned} \tag{5.21}$$

Note that for $\mathbf{v} = 0$ on $\partial\Omega$, applying Green's theorem, we have

$$b(\mathbf{v}, \tau, \sigma) = -b(\mathbf{v}, \sigma, \tau) - (\nabla \cdot \mathbf{v}\, \tau, \sigma)\,, \tag{5.22}$$

which implies

$$b(\mathbf{v}, \tau, \tau) = -\frac{1}{2}\,(\nabla \cdot \mathbf{v}\, \tau, \tau)\,. \tag{5.23}$$

Using (5.23),

$$\begin{aligned}
\left|b(\mathbf{u}_h^{n-1}, \mathbf{F}^n, \mathbf{F}^n)\right| &= \frac{1}{2}\left|\left(\nabla \cdot \mathbf{u}_h^{n-1}\, \mathbf{F}^n, \mathbf{F}^n\right)\right| \\
&= \frac{1}{2}\left|\left(\nabla \cdot (\mathbf{u}_h^{n-1} - \mathcal{U}^{n-1})\, \mathbf{F}^n, \mathbf{F}^n\right) + \left(\nabla \cdot \mathcal{U}^{n-1}\, \mathbf{F}^n, \mathbf{F}^n\right)\right| \\
&\le \frac{1}{2}\left\|\nabla \cdot \mathbf{E}^{n-1}\right\|_\infty \|\mathbf{F}^n\|^2 + \frac{1}{2}\left\|\nabla \cdot \mathcal{U}^{n-1}\right\|_\infty \|\mathbf{F}^n\|^2 \\
&\le \frac{1}{2}\left\|\nabla \cdot \mathbf{E}^{n-1}\right\|_\infty \|\mathbf{F}^n\|^2 + \frac{1}{2}M\,\|\mathbf{F}^n\|^2\,, \qquad \text{using (5.2).}
\end{aligned}$$

Next,

$$\begin{aligned}
\left|\left(D(\mathbf{E}^n), \nu\mathbf{u}_h^{n-1} \cdot \nabla\mathbf{F}^n\right)\right| &\le \|D(\mathbf{E}^n)\|\,\left\|\nu\mathbf{u}_h^{n-1} \cdot \nabla\mathbf{F}^n\right\| \\
&\le \|\nabla\mathbf{E}^n\|\,\left\|\nu\mathbf{u}_h^{n-1} \cdot \nabla\mathbf{F}^n\right\| \\
&\le \epsilon_2\,\|\nabla\mathbf{E}^n\|^2 + \frac{\nu^2}{4\epsilon_2}\left\|\mathbf{u}_h^{n-1} \cdot \nabla\mathbf{F}^n\right\|^2.
\end{aligned}$$

Also,

$$\begin{aligned}
\left|\left(\mathbf{F}^n, \nu\mathbf{u}_h^{n-1} \cdot \nabla\mathbf{F}^n\right)\right| &= \nu\left|\left(\mathbf{F}^n, \mathbf{u}_h^{n-1} \cdot \nabla\mathbf{F}^n\right)\right| \\
&\le \nu\,\|\mathbf{F}^n\|\,\left\|\mathbf{u}_h^{n-1} \cdot \nabla\mathbf{F}^n\right\| \\
&\le \|\mathbf{F}^n\|^2 + \frac{\nu^2}{4}\left\|\mathbf{u}_h^{n-1} \cdot \nabla\mathbf{F}^n\right\|^2.
\end{aligned}$$

Thus, for the first summation on the rhs of (5.20), we have

$$\begin{aligned}
\Delta t \sum_{n=1}^{l} \Bigg[&- Re\;\hat{\lambda}\, c(\mathbf{u}_h^{n-1}, \mathbf{E}^n, \mathbf{E}^n) - b(\mathbf{u}_h^{n-1}, \mathbf{F}^n, \mathbf{F}^n) + \hat{\lambda}\left(D(\mathbf{E}^n), \nu\mathbf{u}_h^{n-1} \cdot \nabla\mathbf{F}^n\right) \\
&\qquad - \frac{1}{\lambda}\left(\mathbf{F}^n, \nu\mathbf{u}_h^{n-1} \cdot \nabla\mathbf{F}^n\right)\Bigg] \\
&\le \Delta t \sum_{n=1}^{l}(Re\;\hat{\lambda}\epsilon_1 + \hat{\lambda}\epsilon_2)\,\|\nabla\mathbf{E}^n\|^2 + \Delta t \sum_{n=1}^{l} \frac{Re\;\hat{\lambda}\acute{d}K^2}{4\epsilon_1}\,\|\mathbf{E}^n\|^2 \\
&\qquad + \Delta t \sum_{n=1}^{l}\left(\frac{\hat{\lambda}\nu^2}{4\epsilon_2} + \frac{\nu^2}{\lambda 4\epsilon_3}\right)\left\|\mathbf{u}_h^{n-1} \cdot \nabla\mathbf{F}^n\right\|^2 \\
&\qquad + \Delta t \sum_{n=1}^{l}\left(\frac{1}{2}M + \ \frac{1}{2}\left\|\nabla \cdot \mathbf{E}^{n-1}\right\|_\infty + \frac{\epsilon_3}{\lambda}\right)\|\mathbf{F}^n\|^2\,.
\end{aligned} \tag{5.24}$$

Next we consider $\mathcal{F}_1(\mathbf{E}^n)$:

$$
\begin{aligned}
|(p^n, \nabla \cdot \mathbf{E}^n)| &= |(p^n - \mathcal{P}^n, \nabla \cdot \mathbf{E}^n)| \\
&\le \|p^n - \mathcal{P}^n\| \, \acute{d}^{\frac{1}{2}} \, \|\nabla \mathbf{E}^n\| \\
&\le \epsilon_4 \, \|\nabla \mathbf{E}^n\|^2 + \frac{\acute{d}}{4\epsilon_4} \, \|p^n - \mathcal{P}^n\|^2 \, , \qquad (5.25) \\
|(d_t \mathbf{\Lambda}^n, \mathbf{E}^n)| &\le \|\mathbf{E}^n\| \, \|d_t \mathbf{\Lambda}^n\| \\
&\le \|\mathbf{E}^n\|^2 + \frac{1}{4} \, \|d_t \mathbf{\Lambda}^n\|^2 \, , \\
\left| c(\mathbf{u}_h^{n-1}, \mathbf{\Lambda}^n, \mathbf{E}^n) \right| &\le \|\mathbf{E}^n\| \, \left\| \mathbf{u}_h^{n-1} \cdot \nabla \mathbf{\Lambda}^n \right\| \\
&\le \|\mathbf{E}^n\| \, \left\| \mathbf{u}_h^{n-1} \right\|_\infty \acute{d}^{\frac{1}{2}} \, \|\nabla \mathbf{\Lambda}^n\| \\
&\le \|\mathbf{E}^n\|^2 + \frac{K^2 \acute{d}}{4} \, \|\nabla \mathbf{\Lambda}^n\|^2 \, , \qquad \text{using (IH1)}, \qquad (5.26) \\
|(\nabla \mathbf{\Lambda}^n, \nabla \mathbf{E}^n)| &\le \|\nabla \mathbf{E}^n\| \, \|\nabla \mathbf{\Lambda}^n\| \\
&\le \epsilon_5 \, \|\nabla \mathbf{E}^n\|^2 + \frac{1}{4\epsilon_5} \, \|\nabla \mathbf{\Lambda}^n\|^2 \, , \qquad (5.27) \\
|(\mathbf{\Gamma}^n, D(\mathbf{E}^n))| &\le \|D(\mathbf{E}^n)\| \, \|\mathbf{\Gamma}^n\| \\
&\le \|\nabla \mathbf{E}^n\| \, \|\mathbf{\Gamma}^n\| \\
&\le \epsilon_6 \, \|\nabla \mathbf{E}^n\|^2 + \frac{1}{4\epsilon_6} \, \|\mathbf{\Gamma}^n\|^2 \, . \qquad (5.28)
\end{aligned}
$$

For the $R_1(\mathbf{E}^n)$ terms we have

$$
|(d_t \mathbf{u}^n, \mathbf{E}^n) - (\mathbf{u}_t^n, \mathbf{E}^n)| \le \|\mathbf{E}^n\|^2 + \frac{1}{4} \, \|d_t \mathbf{u}^n - \mathbf{u}_t^n\|^2 \, , \qquad (5.29)
$$

$$
\begin{aligned}
&\left| c(\mathbf{u}_h^{n-1}, \mathbf{u}^n, \mathbf{E}^n) - c(\mathbf{u}^n, \mathbf{u}^n, \mathbf{E}^n) \right| \\
&\quad = |c(\mathbf{u}_h^{n-1} - \mathcal{U}^{n-1}, \mathbf{u}^n, \mathbf{E}^n) + c(\mathcal{U}^{n-1} - \mathbf{u}^{n-1}, \mathbf{u}^n, \mathbf{E}^n) \\
&\qquad + c(\mathbf{u}^{n-1} - \mathbf{u}^n, \mathbf{u}^n, \mathbf{E}^n)| \\
&\quad \le \|\mathbf{E}^{n-1} \cdot \nabla \mathbf{u}^n\| \, \|\mathbf{E}^n\| \\
&\qquad + \|\mathbf{\Lambda}^{n-1} \cdot \nabla \mathbf{u}^n\| \, \|\mathbf{E}^n\| + \|(\mathbf{u}^n - \mathbf{u}^{n-1}) \cdot \nabla \mathbf{u}^n\| \, \|\mathbf{E}^n\| \\
&\quad \le \acute{d} M \|\mathbf{E}^{n-1}\| \, \|\mathbf{E}^n\| + \acute{d} M \|\mathbf{\Lambda}^{n-1}\| \, \|\mathbf{E}^n\| + \acute{d} M \|(\mathbf{u}^n - \mathbf{u}^{n-1})\| \, \|\mathbf{E}^n\| \\
&\quad \le \frac{\acute{d} M}{2} \left\| \mathbf{E}^{n-1} \right\|^2 + \left(\frac{\acute{d} M}{2} + 2 \right) \|\mathbf{E}^n\|^2 + \frac{\acute{d}^2 M^2}{4} \left\| \mathbf{\Lambda}^{n-1} \right\|^2 \\
&\qquad + \frac{\acute{d}^2 M^2}{4} \Delta t \int_{t_{n-1}}^{t_n} \|\mathbf{u}_t\|^2 \, dt. \qquad (5.30)
\end{aligned}
$$

Combining (5.25)–(5.30), we have the following estimate for $\mathcal{F}_1(\mathbf{E}^n)$:

$$
\begin{aligned}
|\hat{\lambda} \mathcal{F}_1(\mathbf{E}^n)| &\le \hat{\lambda} (\epsilon_4 + \epsilon_5 + \epsilon_6) \, \|\nabla \mathbf{E}^n\|^2 + \hat{\lambda} \; Re \left(\frac{\acute{d} M}{2} + 5 \right) \|\mathbf{E}^n\|^2 \\
&\quad + \hat{\lambda} \; Re \frac{\acute{d} M}{2} \left\| \mathbf{E}^{n-1} \right\|^2 + \hat{\lambda} \frac{\acute{d}}{4\epsilon_4} \, \|(p^n - \mathcal{P}^n)\|^2 + \hat{\lambda} \; Re \frac{\acute{d}^2 M^2}{4} \left\| \mathbf{\Lambda}^{n-1} \right\|^2 \\
&\quad + \hat{\lambda} \left(\frac{Re \; K^2 \acute{d}}{4} + \frac{(1-\alpha)}{4\epsilon_5} \right) \|\nabla \mathbf{\Lambda}^n\|^2 + Re \frac{1}{4} \, \|d_t \mathbf{\Lambda}^n\|^2 + \hat{\lambda} \frac{1}{4\epsilon_6} \, \|\mathbf{\Gamma}^n\|^2 \\
&\quad + \hat{\lambda} \; Re \frac{1}{4} \, \|d_t \mathbf{u}^n - \mathbf{u}_t^n\|^2 + \hat{\lambda} \; Re \frac{\acute{d}^2 M^2}{4} \Delta t \int_{t_{n-1}}^{t_n} \|\mathbf{u}_t\|^2 \, dt. \qquad (5.31)
\end{aligned}
$$

Next we consider the terms in $\mathcal{F}_2(\mathbf{F}^n)$:

$$|(d_t\mathbf{\Gamma}^n,\mathbf{F}^n)| \le \|\mathbf{F}^n\|^2 + \frac{1}{4}\|d_t\mathbf{\Gamma}^n\|^2 , \tag{5.32}$$

$$\begin{aligned}
|b(\mathbf{u}_h^{n-1},\mathbf{\Gamma}^n,\tilde{\mathbf{F}}^n)| &= \left|b(\mathbf{u}_h^{n-1},\mathbf{\Gamma}^n,\mathbf{F}^n) + b(\mathbf{u}_h^{n-1},\mathbf{\Gamma}^n,\nu\mathbf{F}_u^n)\right| \\
&\le \left\|\mathbf{u}_h^{n-1}\cdot\nabla\mathbf{\Gamma}^n\right\|\|\mathbf{F}^n\| + \left\|\mathbf{u}_h^{n-1}\cdot\nabla\mathbf{\Gamma}^n\right\|\|\nu\mathbf{F}_u^n\| \\
&\le \acute{d}^{\frac{1}{2}}\left\|\mathbf{u}_h^{n-1}\right\|_\infty\|\nabla\mathbf{\Gamma}^n\|\|\mathbf{F}^n\| + \acute{d}^{\frac{1}{2}}\left\|\mathbf{u}_h^{n-1}\right\|_\infty\|\nabla\mathbf{\Gamma}^n\|\|\nu\mathbf{F}_u^n\| \\
&\le \|\mathbf{F}^n\|^2 + \nu^2\|\mathbf{F}_u^n\|^2 + \frac{\acute{d}K^2}{2}\|\nabla\mathbf{\Gamma}^n\|^2 , \qquad (5.33)
\end{aligned}$$

$$\begin{aligned}
|(D(\mathbf{\Lambda}^n)\tilde{\mathbf{F}}^n)| &= |(D(\mathbf{\Lambda}^n),\mathbf{F}^n) + (D(\mathbf{\Lambda}^n),\nu\mathbf{F}_u^n)| \\
&\le \|\mathbf{F}^n\|^2 + \nu^2\|\mathbf{F}_u^n\|^2 + \frac{1}{2}\|\nabla\mathbf{\Lambda}^n\|^2 , \qquad (5.34)
\end{aligned}$$

$$\begin{aligned}
|(\mathbf{\Gamma}^n\tilde{\mathbf{F}}^n)| &= |(\mathbf{\Gamma}^n,\mathbf{F}^n) + \nu\,(\mathbf{\Gamma}^n,\nu\mathbf{F}_u^n)| \\
&\le \|\mathbf{F}^n\|^2 + \nu^2\|\mathbf{F}_u^n\|^2 + \frac{1}{2}\|\mathbf{\Gamma}^n\|^2 . \qquad (5.35)
\end{aligned}$$

For the terms making up $R_2(\mathbf{F}^n)$ we have

$$|(d_t\tau^n,\mathbf{F}^n) - (\tau_t^n,\mathbf{F}^n)| \le \|\mathbf{F}^n\|^2 + \frac{1}{4}\|d_t\tau^n - \tau_t^n\|^2 , \tag{5.36}$$

$$\begin{aligned}
|(\tau_t^n,\nu\mathbf{F}_u^n)| &= \left|\left(\tau_t^n,\nu\mathbf{u}_h^{n-1}\cdot\nabla\mathbf{F}^n\right)\right| \\
&= \left|b(\nu\mathbf{u}_h^{n-1},\mathbf{F}^n,\tau_t^n)\right| \\
&\le \left|b(\nu\mathbf{u}_h^{n-1},\tau_t^n,\mathbf{F}^n)\right| + \left|\left(\nabla\cdot\mathbf{u}_h^{n-1}\,\nu\mathbf{F}^n,\tau_t^n\right)\right| \quad \text{(using (5.22))} \\
&\le \nu\left\|\mathbf{u}_h^{n-1}\cdot\nabla\tau_t^n\right\|\|\mathbf{F}^n\| + \left|\left(\nabla\cdot(\mathbf{u}_h^{n-1} - \mathcal{U}^{n-1})\,\nu\mathbf{F}^n,\tau_t^n\right)\right| \\
&\quad + \left|\left(\nabla\cdot\mathcal{U}^{n-1}\,\nu\mathbf{F}^n,\tau_t^n\right)\right| \\
&\le \nu\left\|\mathbf{u}_h^{n-1}\right\|_\infty\acute{d}^{\frac{1}{2}}\|\nabla\tau_t^n\|\|\mathbf{F}^n\| + \nu\left\|\nabla\cdot(\mathbf{u}_h^{n-1} - \mathcal{U}^{n-1})\right\|_\infty\|\mathbf{F}^n\|\|\tau_t^n\| \\
&\quad + \left\|\nabla\cdot\mathcal{U}^{n-1}\right\|_\infty\nu\|\mathbf{F}^n\|\|\tau_t^n\| \\
&\le \left(2 + \left\|\nabla\mathbf{E}^{n-1}\right\|_\infty\right)\|\mathbf{F}^n\|^2 + \frac{\nu^2}{4}\acute{d}^2\left(M^2 + \left\|\nabla\mathbf{E}^{n-1}\right\|_\infty\right)\|\tau_t^n\|^2 \\
&\quad + \frac{\nu^2}{4}K^2\acute{d}\,\|\nabla\tau_t^n\|^2 \quad \text{(using (5.2) and (IH1))}, \qquad (5.37)
\end{aligned}$$

$$\begin{aligned}
|b(\mathbf{u}_h^{n-1},\tau^n,\tilde{\mathbf{F}}^n) - b(\mathbf{u}^n,\tau^n,\tilde{\mathbf{F}}^n)| &= |((\mathbf{u}_h^{n-1} - \mathbf{u}^n)\cdot\nabla\tau^n\tilde{\mathbf{F}}^n)| \\
&\le \left\|(\mathbf{u}_h^{n-1} - \mathbf{u}^n)\cdot\nabla\tau^n\right\|\|\tilde{\mathbf{F}}^n\| \\
&\le \frac{1}{2}\|\tilde{\mathbf{F}}^n\|^2 + \frac{1}{2}\acute{d}^3\|\nabla\tau^n\|_\infty^2\left\|\mathbf{u}_h^{n-1} - \mathbf{u}^n\right\|^2 \\
&\le \|\mathbf{F}^n\|^2 + \nu^2\|\mathbf{F}_u^n\|^2 + \frac{1}{2}\acute{d}^3M^2\left\|-\mathbf{E}^{n-1} - \mathbf{\Lambda}^{n-1} + \mathbf{u}^{n-1} - \mathbf{u}^n\right\|^2 \\
&\le \|\mathbf{F}^n\|^2 + \nu^2\|\mathbf{F}_u^n\|^2 + \frac{3}{2}\acute{d}^3M^2\left\|\mathbf{E}^{n-1}\right\|^2 + \frac{3}{2}\acute{d}^3M^2\left\|\mathbf{\Lambda}^{n-1}\right\|^2 \\
&\quad + \frac{3}{2}\acute{d}^3M^2\Delta t\int_{t_{n-1}}^{t_n}\|\mathbf{u}_t\|^2\,dt. \qquad (5.38)
\end{aligned}$$

In order to estimate the g_a terms in $\mathcal{F}_2(\cdot)$, note that

$$\begin{aligned} g_a\left(\tau_h^{n-1}, \nabla \mathbf{u}_h^{n-1}\right) - g_a\left(\tau^n, \nabla \mathbf{u}^n\right) \\ &= g_a\left(\tau_h^{n-1}, \nabla(\mathbf{u}_h^{n-1} - \mathcal{U}^{n-1})\right) + g_a\left(\tau_h^{n-1}, \nabla(\mathcal{U}^{n-1} - \mathbf{u}^{n-1})\right) \\ &\quad + g_a\left(\tau_h^{n-1}, \nabla(\mathbf{u}^{n-1} - \mathbf{u}^n)\right) + g_a\left(\tau_h^{n-1} - \mathcal{T}^{n-1}, \nabla \mathbf{u}^n\right) \\ &\quad + g_a\left(\mathcal{T}^{n-1} - \tau^{n-1}, \nabla \mathbf{u}^n\right) + g_a\left(\tau^{n-1} - \tau^n, \nabla \mathbf{u}^n\right) \\ &= -g_a\left(\tau_h^{n-1}, \nabla \mathbf{E}^{n-1}\right) - g_a\left(\tau_h^{n-1}, \nabla \mathbf{\Lambda}^{n-1}\right) - g_a\left(\tau_h^{n-1}, \nabla(\mathbf{u}^n - \mathbf{u}^{n-1})\right) \\ &\quad - g_a\left(\mathbf{F}^{n-1}, \nabla \mathbf{u}^n\right) - g_a\left(\mathbf{\Gamma}^{n-1}, \nabla \mathbf{u}^n\right) - g_a\left(\tau^n - \tau^{n-1}, \nabla \mathbf{u}^n\right). \end{aligned} \tag{5.39}$$

Bounding each of the terms on the rhs of (5.39), we obtain

$$\begin{aligned} |(g_a\left(\tau_h^{n-1}, \nabla \mathbf{E}^{n-1}\right)\tilde{\mathbf{F}}^n)| &\le \left\|g_a\left(\tau_h^{n-1}, \nabla \mathbf{E}^{n-1}\right)\right\| \|\tilde{\mathbf{F}}^n\| \\ &\le 4\acute{d}\left\|\tau_h^{n-1}\right\|_\infty \left\|\nabla \mathbf{E}^{n-1}\right\| \|\tilde{\mathbf{F}}^n\| \\ &\le \epsilon_7 \left\|\nabla \mathbf{E}^{n-1}\right\|^2 + \frac{8\acute{d}^2K^2}{\epsilon_7}\|\mathbf{F}^n\|^2 + \frac{8\acute{d}^2K^2}{\epsilon_7}\nu^2\|\mathbf{F}_u^n\|^2, \end{aligned} \tag{5.40}$$

$$|(g_a\left(\tau_h^{n-1}, \nabla \mathbf{\Lambda}^{n-1}\right)\tilde{\mathbf{F}}^n)| \le 8\acute{d}^2K^2\left\|\nabla \mathbf{\Lambda}^{n-1}\right\|^2 + \|\mathbf{F}^n\|^2 + \nu^2\|\mathbf{F}_u^n\|^2, \tag{5.41}$$

$$|(g_a\left(\tau_h^{n-1}, \nabla(\mathbf{u}^n - \mathbf{u}^{n-1})\right)\tilde{\mathbf{F}}^n)| \le 8\acute{d}^2K^2\Delta t\int_{t_{n-1}}^{t_n}\|\nabla \mathbf{u}_t\|^2\, dt + \|\mathbf{F}^n\|^2 + \nu^2\|\mathbf{F}_u^n\|^2, \tag{5.42}$$

$$|(g_a\left(\mathbf{F}^{n-1}, \nabla \mathbf{u}^n\right)\tilde{\mathbf{F}}^n)| \le 8\acute{d}^2M^2\left\|\mathbf{F}^{n-1}\right\|^2 + \|\mathbf{F}^n\|^2 + \nu^2\|\mathbf{F}_u^n\|^2, \tag{5.43}$$

$$|(g_a\left(\mathbf{\Gamma}^{n-1}, \nabla \mathbf{u}^n\right)\tilde{\mathbf{F}}^n)| \le 8\acute{d}^2M^2\left\|\mathbf{\Gamma}^{n-1}\right\|^2 + \|\mathbf{F}^n\|^2 + \nu^2\|\mathbf{F}_u^n\|^2, \tag{5.44}$$

$$|(g_a\left(\tau^n - \tau^{n-1}, \nabla \mathbf{u}^n\right)\tilde{\mathbf{F}}^n)| \le 8\acute{d}^2M^2\Delta t\int_{t_{n-1}}^{t_n}\|\tau_t\|^2\, dt + \|\mathbf{F}^n\|^2 + \nu^2\|\mathbf{F}_u^n\|^2. \tag{5.45}$$

Combining the estimates in (5.32)–(5.38), (5.40)–(5.45), we obtain the following estimate for $\mathcal{F}_2(\mathbf{F}^n)$:

$$\begin{aligned} |\mathcal{F}_2(\mathbf{F}^n)| \le \epsilon_7\left\|\nabla \mathbf{E}^{n-1}\right\|^2 + \nu^2\|\mathbf{F}_u^n\|^2\left(7 + \frac{8\acute{d}^2K^2}{\epsilon_7} + \hat{\lambda} + \frac{1}{\lambda}\right) \\ + \|\mathbf{F}^n\|^2\left(11 + \frac{8\acute{d}^2K^2}{\epsilon_7} + \left\|\nabla \mathbf{E}^{n-1}\right\|_\infty + \hat{\lambda} + \frac{1}{\lambda}\right) \\ + \left\|\mathbf{E}^{n-1}\right\|^2\left(\frac{3}{2}\acute{d}^3M^2\right) + \left\|\mathbf{F}^{n-1}\right\|^2\left(8\acute{d}^2M^2\right) \\ + \|\nabla \mathbf{\Lambda}^n\|^2\left(\frac{\hat{\lambda}}{2}\right) + \|\nabla \mathbf{\Gamma}^n\|^2\left(\frac{\acute{d}K^2}{2}\right) + \|\mathbf{\Gamma}^n\|^2\left(\frac{1}{2\lambda}\right) + \|d_t\mathbf{\Gamma}^n\|^2\left(\frac{1}{4}\right) \\ + \left\|\nabla \mathbf{\Lambda}^{n-1}\right\|^2(8\acute{d}^2K^2) + \left\|\mathbf{\Lambda}^{n-1}\right\|^2\left(\frac{3}{2}\acute{d}^3M^2\right) + \left\|\mathbf{\Gamma}^{n-1}\right\|^2(8\acute{d}^2M^2) \end{aligned}$$

$$
\begin{aligned}
&+\ \frac{1}{4}\left\|d_t\tau^n-\tau_t^n\right\|^2+\frac{\nu^2}{4}\acute{d}^2\left(M^2+\left\|\nabla\mathbf{E}^{n-1}\right\|_\infty\right)\left\|\tau_t^n\right\|^2\\
&+\ \frac{\nu^2}{4}K^2\acute{d}\left\|\nabla\tau_t^n\right\|^2\\
&+\ \frac{3}{2}\acute{d}^3M^2\Delta t\int_{t_{n-1}}^{t_n}\left\|\mathbf{u}_t\right\|^2\,dt+8\acute{d}^2M^2\Delta t\int_{t_{n-1}}^{t_n}\left\|\tau_t\right\|^2\,dt\\
&+\ 8\acute{d}^2K^2\Delta t\int_{t_{n-1}}^{t_n}\left\|\nabla\mathbf{u}_t\right\|^2\,dt. \qquad (5.46)
\end{aligned}
$$

With the choices $\epsilon_1=\frac{(1-\alpha)}{12\ Re\ \hat{\lambda}}, \epsilon_2=\epsilon_4=\epsilon_5=\epsilon_6=\epsilon_7=\frac{(1-\alpha)}{12\hat{\lambda}}, \mathbf{u}_h^0=\mathcal{U}^0(\Rightarrow\mathbf{E}^0=0), \tau_h^0=\mathcal{T}^0(\Rightarrow\mathbf{F}^0=0)$, substituting (5.24), (5.31), (5.46) into (5.20) yields

$$
\begin{aligned}
&\frac{Re\ \hat{\lambda}}{2}\left\|\mathbf{E}^l\right\|^2+\frac{1}{2}\left\|\mathbf{F}^l\right\|^2+\frac{(1-\alpha)}{2}\hat{\lambda}\sum_{n=1}^{l}\Delta t\left\|\nabla\mathbf{E}^n\right\|^2\\
&\qquad+\ \left[\nu-\nu^2\left(\frac{3\hat{\lambda}^2+96\acute{d}^2K^2\hat{\lambda}}{(1-\alpha)}+7+\hat{\lambda}+\frac{5}{4\lambda}\right)\right]\sum_{n=1}^{l}\Delta t\left\|\mathbf{F}_u^n\right\|^2\\
&\quad\le C_1\sum_{n=1}^{l}\Delta t\left\|\mathbf{E}^n\right\|^2+C_2\sum_{n=1}^{l}\Delta t\left\|\mathbf{F}^n\right\|^2+C_3\sum_{n=1}^{l}\Delta t\left\|\nabla\mathbf{E}^{n-1}\right\|_\infty\left\|\mathbf{F}^n\right\|^2\\
&\qquad+\ C_4\sum_{n=1}^{l}\Delta t\left\|\mathbf{\Lambda}^n\right\|^2+C_5\sum_{n=1}^{l}\Delta t\left\|\nabla\mathbf{\Lambda}^n\right\|^2\\
&\qquad+\ \frac{1}{4}\sum_{n=1}^{l}\Delta t\left\|d_t\mathbf{\Lambda}^n\right\|^2+C_6\sum_{n=1}^{l}\Delta t\left\|\mathbf{\Gamma}^n\right\|^2+Re\frac{\hat{\lambda}}{4}\sum_{n=1}^{l}\Delta t\left\|d_t\mathbf{u}^n-\mathbf{u}_t^n\right\|^2\\
&\qquad+\ \left(\frac{\acute{d}K^2}{2}\right)\sum_{n=1}^{l}\Delta t\left\|\nabla\mathbf{\Gamma}^n\right\|^2+\frac{1}{4}\sum_{n=1}^{l}\Delta t\left\|d_t\mathbf{\Gamma}^n\right\|^2+\frac{1}{4}\sum_{n=1}^{l}\Delta t\left\|d_t\tau^n-\tau_t^n\right\|^2\\
&\qquad+\ \frac{\nu^2}{4}\sum_{n=1}^{l}\Delta t\acute{d}^2\left(M^2+\left\|\nabla\mathbf{E}^{n-1}\right\|_\infty\right)\left\|\tau_t^n\right\|^2+\sum_{n=1}^{l}\Delta t\left\|p^n-\mathcal{P}^n\right\|^2\\
&\qquad+\ |\Delta t|^2\,\acute{d}\left(Re\ \acute{d}M^2\frac{\hat{\lambda}}{4}\left\|\mathbf{u}_t\right\|_{0,0}^2+\frac{3}{2}\acute{d}^2M^2\left\|\mathbf{u}_t\right\|_{0,0}^2\right.\\
&\qquad\qquad\left.+\ 8\acute{d}M^2\left\|\tau_t\right\|_{0,0}^2+8\acute{d}K^2\left\|\mathbf{u}_t\right\|_{0,1}^2\right)+\ \frac{\nu^2}{4}K^2\acute{d}\ \vert\vert\vert\nabla\tau_t\vert\vert\vert_{0,0}^2\,. \qquad (5.47)
\end{aligned}
$$

We now apply the interpolation properties of the approximating spaces to estimate the terms on the rhs of (5.47). Using elements of order k for velocity, elements of order m for stress, and elements of order q for pressure, we have

$$
\begin{aligned}
&\sum_{n=1}^{l}\Delta t\left\|\nabla\mathbf{\Lambda}^n\right\|^2+\sum_{n=1}^{l}\Delta t\left\|\nabla\mathbf{\Gamma}^n\right\|^2\\
&\qquad\le C\left(h^{2k}\sum_{n=1}^{l}\Delta t\left\|\mathbf{u}^n\right\|_{k+1}^2+h^{2m}\sum_{n=1}^{l}\Delta t\left\|\tau^n\right\|_{m+1}^2\right)\\
&\qquad\le C\left(h^{2k}\,\vert\vert\vert\mathbf{u}\vert\vert\vert_{0,k+1}^2+h^{2m}\,\vert\vert\vert\tau\vert\vert\vert_{0,m+1}^2\right), \qquad (5.48)
\end{aligned}
$$

$$
\begin{aligned}
\sum_{n=1}^{l} \Delta t \left\| \mathbf{\Lambda}^n \right\|^2 &+ \sum_{n=1}^{l} \Delta t \left\| \mathbf{\Gamma}^n \right\|^2 + \sum_{n=1}^{l} \Delta t \left\| p - \mathcal{P}^n \right\|^2 \\
&\leq C \left(h^{2k+2} \sum_{n=1}^{l} \Delta t \left\| \mathbf{u}^n \right\|_{k+1}^2 + h^{2m+2} \sum_{n=1}^{l} \Delta t \left\| \tau^n \right\|_{m+1}^2 + h^{2q+2} \sum_{n=1}^{l} \Delta t \left\| p^n \right\|_{q+1}^2 \right) \\
&\leq C \left(h^{2k+2} \left\| \mathbf{u} \right\|_{0,k+1}^2 + h^{2m+2} \left\| \tau \right\|_{0,m+1}^2 + h^{2q+2} \left\| p \right\|_{0,q+1}^2 \right),
\end{aligned}
$$

(5.49)

$$
\begin{aligned}
\sum_{n=1}^{l} \Delta t \left\| d_t \mathbf{\Lambda}^n \right\|^2 &= \sum_{n=1}^{l} \Delta t \left\| \frac{1}{\Delta t} \int_{t_{n-1}}^{t_n} 1 \frac{\partial \Lambda}{\partial t} dt \right\|^2 \\
&\leq \sum_{n=1}^{l} \Delta t \left(\frac{1}{\Delta t} \right)^2 \int_{\Omega} \left(\int_{t_{n-1}}^{t_n} 1 \, dt \right) \left(\int_{t_{n-1}}^{t_n} \left(\frac{\partial \Lambda}{\partial t} \right)^2 \, dt \right) d\mathbf{x} \\
&\leq C h^{2k+2} \left\| \mathbf{u}_t \right\|_{0,k+1}^2, \qquad (5.50)
\end{aligned}
$$

and similarly,

$$
\sum_{n=1}^{l} \Delta t \left\| d_t \mathbf{\Gamma}^n \right\|^2 \leq C h^{2m+2} \left\| \tau_t \right\|_{0,m+1}^2. \qquad (5.51)
$$

Note that $d_t \mathbf{u}^n - \mathbf{u}_t^n$ may be expressed as

$$
d_t \mathbf{u}^n - \mathbf{u}_t^n = \frac{1}{2\Delta t} \int_{t_{n-1}}^{t_n} \mathbf{u}_{tt}(\cdot, t)(t_{n-1} - t) \, dt.
$$

Also,

$$
\begin{aligned}
\left(\frac{1}{2\Delta t} \int_{t_{n-1}}^{t_n} \mathbf{u}_{tt}(\cdot, t)(t_{n-1} - t) \, dt \right)^2 &\leq \frac{1}{4 \left| \Delta t \right|^2} \int_{t_{n-1}}^{t_n} \mathbf{u}_{tt}^2(\cdot, t) \, dt \int_{t_{n-1}}^{t_n} (t_{n-1} - t)^2 \, dt \\
&= \frac{1}{12} \Delta t \int_{t_{n-1}}^{t_n} \mathbf{u}_{tt}^2(\cdot, t) \, dt.
\end{aligned}
$$

Therefore it follows that

$$
\begin{aligned}
\sum_{n=1}^{l} \Delta t \left\| d_t \mathbf{u}^n - \mathbf{u}_t^n \right\|^2 &\leq \sum_{n=1}^{l} \Delta t \int_{\Omega} \frac{1}{12} \Delta t \int_{t_{n-1}}^{t_n} \mathbf{u}_{tt}^2(\cdot, t) \, dt \, d\mathbf{x} \\
&= \frac{1}{12} \left| \Delta t \right|^2 \left\| \mathbf{u}_{tt} \right\|_{0,0}^2. \qquad (5.52)
\end{aligned}
$$

Similarly, for $d_t \tau^n - \tau_t^n$ we have

$$
\sum_{n=1}^{l} \Delta t \left\| d_t \tau^n - \tau_t^n \right\|^2 \leq \frac{1}{12} \left| \Delta t \right|^2 \left\| \tau_{tt} \right\|_{0,0}^2. \qquad (5.53)
$$

In view of (5.48)–(5.53), our induction hypotheses (IH1), (IH2), and with ν chosen such that

$$
\nu \leq \frac{1}{2} \left(\frac{3\hat{\lambda}^2 + 96 d'^2 K^2 \hat{\lambda}}{(1 - \alpha)} + 7 + \hat{\lambda} + \frac{5}{4\lambda} \right)^{-1}, \qquad (5.54)
$$

from (5.47) we obtain

$$
\begin{aligned}
&\frac{Re\ \hat{\lambda}}{2}\left\|\mathbf{E}^l\right\|^2+\frac{1}{2}\left\|\mathbf{F}^l\right\|^2+\frac{(1-\alpha)}{2}\hat{\lambda}\sum_{n=1}^{l}\Delta t\left\|\nabla\mathbf{E}^n\right\|^2+\frac{\nu}{2}\sum_{n=1}^{l}\Delta t\left\|\mathbf{F}_u^n\right\|^2\\
&\qquad\le C\sum_{n=1}^{l}\Delta t\left(\left\|\mathbf{E}^n\right\|^2+\left\|\mathbf{F}^n\right\|^2\right)+C\sum_{n=1}^{l}\Delta t\left\|\nabla\mathbf{E}^{n-1}\right\|_\infty\left\|\mathbf{F}^n\right\|^2\\
&\qquad\quad+C\nu^2\left(|||\tau_t|||_{0,1}^2+|||\tau_t|||_{\infty,0}^2\right)\\
&\qquad\quad+C\left|\Delta t\right|^2\left(\left\|\mathbf{u}_t\right\|_{0,1}^2+\left\|\mathbf{u}_{tt}\right\|_{0,0}^2+\left\|\tau_t\right\|_{0,0}^2+\left\|\tau_{tt}\right\|_{0,0}^2\right)+Ch^{2k+2}|||\mathbf{u}|||_{0,k+1}^2\\
&\qquad\quad+Ch^{2m+2}|||\tau|||_{0,m+1}^2+Ch^{2q+2}|||p|||_{0,q+1}^2+Ch^{2k}\ |||\mathbf{u}|||_{0,k+1}^2\\
&\qquad\quad+Ch^{2k+2}\left\|\mathbf{u}_t\right\|_{0,k+1}^2+Ch^{2m}|||\tau|||_{0,m+1}^2+Ch^{2m+2}\left\|\tau_t\right\|_{0,m+1}^2\,,
\end{aligned}
\tag{5.55}
$$

where the C's denote constants independent of $l, \Delta t, h, \nu$. Applying Gronwall's lemma and (IH2) to (5.55), the estimate given in (5.11) follows. □

Step 2. We show that the induction hypotheses (IH1) and (IH2) are true.

Verification of (IH1). Assume that (IH1) holds true for $n = 1, 2, \ldots, l-1$. By interpolation properties, inverse estimates, and (5.11), we have that

$$
\begin{aligned}
\left\|\mathbf{u}_h^l\right\|_\infty &\le \left\|\mathbf{u}_h^l-\mathbf{u}^l\right\|_\infty+\left\|\mathbf{u}^l\right\|_\infty\\
&\le \left\|\mathbf{E}^l\right\|_\infty+\left\|\Lambda^l\right\|_\infty+M\\
&\le Ch^{-\frac{\acute{d}}{2}}\left\|\mathbf{E}^l\right\|_0+Ch^{-\frac{\acute{d}}{2}}\left\|\Lambda^l\right\|_0+M\\
&\le C\left(\left|\Delta t\right|h^{-\frac{\acute{d}}{2}}+\nu h^{-\frac{\acute{d}}{2}}+h^{k-\frac{\acute{d}}{2}}+h^{m-\frac{\acute{d}}{2}}+h^{q+1-\frac{\acute{d}}{2}}+h^{k+1-\frac{\acute{d}}{2}}\right)+M.
\end{aligned}
\tag{5.56}
$$

Note that the expression $C(\left|\Delta t\right|h^{-\frac{\acute{d}}{2}}+\nu h^{-\frac{\acute{d}}{2}}+h^{k-\frac{\acute{d}}{2}}+h^{m-\frac{\acute{d}}{2}}+h^{q+1-\frac{\acute{d}}{2}}+h^{k+1-\frac{\acute{d}}{2}})$ is independent of l. Hence, if we set $k, m \ge \frac{\acute{d}}{2}, q \ge \frac{\acute{d}}{2}-1$, and choose $h, \Delta t, \nu$ such that

$$
h^{k-\frac{\acute{d}}{2}}, h^{m-\frac{\acute{d}}{2}}, h^{q+1-\frac{\acute{d}}{2}} \le \frac{1}{C}, \qquad \Delta t, \nu \le \frac{h^{\frac{\acute{d}}{2}}}{C},
\tag{5.57}
$$

then from (5.56)

$$
\left\|\mathbf{u}_h^l\right\|_\infty \le M+6.
$$

Similarly it follows that $\left\|\tau_h^l\right\|_\infty \le M+6$. □

Verification of (IH2). Assume that (IH2) is true for $n = 1, 2, \ldots, l-1$. Equations (5.11) and (5.55) imply

$$
\sum_{n=1}^{l}\Delta t\left\|\nabla\mathbf{E}^n\right\|_0^2 \le C\left(h^{2k}+h^{2m}+h^{2q+2}+\left|\Delta t\right|^2+\nu^2\right).
\tag{5.58}
$$

Applying the inverse estimate and using the inequality

$$
\sum_{n=1}^{l}a_n \le \sqrt{l}\left(\sum_{n=1}^{l}a_n^2\right)^{\frac{1}{2}},
$$

from (5.58) we obtain

$$\begin{aligned}\sum_{n=1}^{l} \Delta t \, \|\nabla \mathbf{E}^n\|_\infty &\le C h^{-\frac{d}{2}} \sum_{n=1}^{l} \Delta t \, \|\nabla \mathbf{E}^n\| \\ &\le C h^{-\frac{d}{2}} \sqrt{\Delta t} \, \sqrt{l} \left(\sum_{n=1}^{l} \Delta t \, \|\nabla \mathbf{E}^n\|^2 \right)^{\frac{1}{2}} \\ &\le \tilde{C} \left(\Delta t \, h^{-\frac{d}{2}} + \nu h^{-\frac{d}{2}} + h^{k-\frac{d}{2}} + h^{m-\frac{d}{2}} + h^{q+1-\frac{d}{2}} \right),\end{aligned}$$

where $\tilde{C} = C\sqrt{T}$ is a constant independent of $l, h, \Delta t$, and ν. Hence when

$$\nu, \Delta t \le \frac{h^{\frac{d}{2}}}{5\tilde{C}} \tag{5.59}$$

and

$$h^{k-\frac{d}{2}}, h^{m-\frac{d}{2}}, h^{q+1-\frac{d}{2}} \le \frac{1}{5\tilde{C}},$$

(IH2) holds. □

Step 3. We derive the error estimates in (5.9) and (5.10).

Proof of Theorem 5.4. Using estimates (5.11) and (approximation properties), we have

$$\begin{aligned}|||\mathbf{u} - \mathbf{u}_h|||_{\infty,0}^2 + |||\tau - \tau_h|||_{\infty,0}^2 &\le |||\mathbf{E}|||_{\infty,0}^2 + |||\Lambda|||_{\infty,0}^2 + |||\mathbf{F}|||_{\infty,0}^2 + |||\Gamma|||_{\infty,0}^2 \\ &\le G(\Delta t, h, \nu) + C \left(h^{2k+2} \, |||\mathbf{u}|||_{\infty,k+1}^2 + h^{2m+2} \, |||\tau|||_{\infty,m+1}^2 \right).\end{aligned}$$

Note the restrictions on ν from (5.54), (5.57), (5.59), and on Δt from (3.1), (5.57), (5.59). Hence, we obtain the stated estimate (5.9).

To establish (5.10), from (5.11), (5.55) we have

$$|||\nabla \mathbf{E}|||_{0,0}^2 + \Delta t \, |||\mathbf{F}_u|||_{0,0}^2 \le C(T+1) G(\Delta t, h, \nu) \tag{5.60}$$

and

$$|||\mathbf{E}|||_{0,0}^2 + |||\mathbf{F}|||_{0,0}^2 \le T \, G(\Delta t, h, \nu). \tag{5.61}$$

Hence

$$|||\mathbf{E}|||_{1,0}^2 + |||\mathbf{F}|||_{0,0}^2 \le \tilde{C} G(\Delta t, h, \nu). \qquad \square \tag{5.62}$$

We conclude this analysis with some comments on the sensitivity of the error bounds to the physical parameters in the modeling equations. From (5.47) we note that the constants C_1, C_2, C_3 involve the terms $K^2, M^2, Re, \bar{\lambda}(= \lambda/2\alpha), \lambda^{-1}$. Thus, in view of the exponential multiplicative factor in the discrete Gronwall's lemma, we have that the generic constants C in (5.9), (5.10), (5.11) depend exponentially on these terms.

REFERENCES

[1] J. Baranger and A. Machmoum, *Existence of approximate solutions and error bounds for viscoelastic fluid flow: Characteristics method*, Comput. Methods Appl. Mech. Engrg., 148 (1997), pp. 39–52.

[2] J. Baranger and D. Sandri, *Finite element approximation of viscoelastic fluid flow: Existence of approximate solutions and error bounds.* I. *Discontinuous constraints*, Numer. Math., 63 (1992), pp. 13–27.

[3] J. Baranger and S. Wardi, *Numerical Analysis of an FEM for a transient viscoelastic flow*, Comput. Methods Appl. Mech. Engrg., 125 (1995), pp. 171–185.

[4] R.B. Bird, R.C. Armstrong, and O. Hassager, *Dynamics of Polymeric Liquids*, John Wiley & Sons, New York, 1987.

[5] S.C. Brenner and L.R. Scott, *The Mathematical Theory of Finite Element Methods*, Springer-Verlag, New York, 1994.

[6] C. Corduneanu, *Principals of Differential and Integral Equations*, Chelsea, New York, 1977.

[7] V. Girault and P.A. Raviart, *Finite element methods for Navier–Stokes equations*, Springer-Verlag, Berlin, Heidelberg, 1986.

[8] C. Guillope and J.C. Saut, *Existence results for the flow of viscoelastic fluids with a differential constitutive law*, Nonlinear Anal., 15 (1990), pp. 849–869.

[9] J.G. Heywood and R. Rannacher, *Finite-element approximations of the nonstationary Navier–Stokes problem. Part* IV: *Error analysis for second-order time discretization*, SIAM J. Numer. Anal., 27 (1990), pp. 353–384.

[10] W. Layton and L. Tobiska, *A two-level method with backtracking for the Navier–Stokes equations*, SIAM J. Numer. Anal., 35 (1998), pp. 2035–2054.

[11] B. Liu, *The analysis of a finite element method with streamline diffusion for the compressible Navier–Stokes equations*, SIAM J. Numer. Anal., 38 (2000), pp. 1–16.

[12] K. Najib and D. Sandri, *On a decoupled algorithm for solving a finite element problem for the approximation of viscoelastic fluid flow*, Numer. Math., 72 (1995), pp. 223–238.

[13] M. Renardy, *Mathematical Analysis of Viscoelastic Flows*, CBMS-NSF Reg. Conf. Ser. Appl. Math. 73, SIAM, Philadelphia, 2000.

[14] D. Sandri, *Finite element approximation of viscoelastic fluid flow: Existence of approximate solutions and error bounds. Continuous approximation of the stress*, SIAM J. Numer. Anal., 31 (1994), pp. 362–377.

SIAM J. NUMER. ANAL.
Vol. 41, No. 2, pp. 487–515

ADAPTIVE MESH METHODS FOR ONE- AND TWO-DIMENSIONAL HYPERBOLIC CONSERVATION LAWS*

HUAZHONG TANG† AND TAO TANG‡

Abstract. We develop efficient moving mesh algorithms for one- and two-dimensional hyperbolic systems of conservation laws. The algorithms are formed by two independent parts: PDE evolution and mesh-redistribution. The first part can be any appropriate high-resolution scheme, and the second part is based on an iterative procedure. In each iteration, meshes are first redistributed by an equidistribution principle, and then on the resulting new grids the underlying numerical solutions are updated by a *conservative-interpolation* formula proposed in this work. The iteration for the mesh-redistribution at a given time step is complete when the meshes governed by a nonlinear equation reach the equilibrium state. The main idea of the proposed method is to keep the mass-conservation of the underlying numerical solution at each redistribution step. In one dimension, we can show that the underlying numerical approximation obtained in the mesh-redistribution part satisfies the desired TVD property, which guarantees that the numerical solution at any time level is TVD, provided that the PDE solver in the first part satisfies such a property. Several test problems in one and two dimensions are computed using the proposed moving mesh algorithm. The computations demonstrate that our methods are efficient for solving problems with shock discontinuities, obtaining the same resolution with a much smaller number of grid points than the uniform mesh approach.

Key words. adaptive mesh method, hyperbolic conservation laws, finite volume method

AMS subject classifications. 65M93, 35L64, 76N10

PII. S003614290138437X

1. Introduction.

Adaptive mesh methods have important applications for a variety of physical and engineering areas such as solid and fluid dynamics, combustion, heat transfer, material science, etc. The physical phenomena in these areas develop dynamically singular or nearly singular solutions in fairly localized regions, such as shock waves, boundary layers, detonation waves, etc. The numerical investigation of these physical problems may require extremely fine meshes over a small portion of the physical domain to resolve the large solution variations. In multidimensions, developing effective and robust adaptive grid methods for these problems becomes necessary. Successful implementation of the adaptive strategy can increase the accuracy of the numerical approximations and also decrease the computational cost. In the past two decades, there has been important progress in developing mesh methods for PDEs, including the variational approach of Winslow [36], Brackbill [5], and Brackbill and Saltzman [6]; finite element methods by Miller and Miller [25] and Davis and Flaherty [11]; the moving mesh PDEs of Cao, Huang, and Russell [7], Stockie, Mackenzie, and Russell [33], Li and Petzold [23], and Ceniceros and Hou [8]; and moving mesh methods based on harmonic mapping of Dvinsky [12] and Li, Tang, and Zhang [21, 22].

*Received by the editors February 1, 2001; accepted for publication (in revised form) September 26, 2002; published electronically April 23, 2003. This research was supported by Hong Kong Baptist University and Hong Kong Research Grants Council.

http://www.siam.org/journals/sinum/41-2/38437.html

†Institute of Computational Mathematics, Academy of Mathematics and System Sciences, Chinese Academy of Sciences, P.O. Box 2719, Beijing 100080, People's Republic of China. Current address: School of Mathematical Sciences, Peking University, Beijing 100871, People's Republic of China (hztang@lsec.cc.ac.cn, hztang@math.pku.edu.cn). The research of this author was partially supported by the National Natural Science Foundation of China (19901031) and the Special Funds for Major State Basic Research Projects of China.

‡Department of Mathematics, Hong Kong Baptist University, Kowloon Tong, Hong Kong, People's Republic of China (ttang@math.hkbu.edu.hk).

Harten and Hyman [14] began the earliest study in this direction, by moving the grid at an adaptive speed in each time step to improve the resolution of shocks and contact discontinuities. After their work, many other moving mesh methods for hyperbolic problems have been proposed in the literature, including those of Azarenok and collaborators [1, 2, 3], Fazio and LeVeque [13], Liu, Ji, and Liao [24], Saleri and Steinberg [29], and Stockie, Mackenzie, and Russell [33]. However, many existing moving mesh methods for hyperbolic problems are designed for one space dimension. In one dimension, it is generally possible to compute on a very fine grid, and so the need for moving mesh methods may not be clear. Multidimensional moving mesh methods are often difficult to use in fluid dynamics problems, since the grid will typically suffer large distortions and possible tangling. It is therefore useful to design a simple and robust moving mesh algorithm for hyperbolic problems in multidimensions.

The main objective of this paper is to develop one- and two-dimensional (1D and 2D) moving mesh methods for hyperbolic systems of conservation laws. Following Li, Tang, and Zhang [21] we propose a moving mesh method containing two separate parts: PDE time-evolution and mesh-redistribution. The first part can be any suitable high-resolution method such as the wave-propagation algorithm, central schemes, and ENO methods. Once numerical solutions are obtained at the given time level, the mesh will be redistributed using an iteration procedure. At each iteration, the grid is moved according to a variational principle, and the underlying numerical solution on the new grid will be updated using some simple methods (such as conventional interpolation). It is noted that the direct use of conventional interpolation is unsatisfactory for hyperbolic problems, since many physical properties such as mass-conservation and TVD (in one dimension) may be destroyed. In order to preserve these physical properties, we propose to use conservative-interpolation in the solution-updating step. The idea of using conservative-interpolation is new and is shown to work successfully for hyperbolic problems. This approach also preserves the total mass of the numerical solutions, and by the well-known Lax–Wendroff theory the numerical solutions converge to the weak solution of the underlying hyperbolic system.

The paper is organized as follows. In section 2, we briefly review some theory of the variational approach for moving mesh methods, which is relevant to the mesh-redistribution part of our algorithm. In section 3, we propose a 1D moving mesh algorithm for solving hyperbolic systems of conservation laws, which will be extended to a 2D algorithm in section 4. Numerical experiments are carried out in sections 5 and 6, where several 1D and 2D examples are considered.

2. Mesh generation based on the variational approach. Let $\vec{x} = (x_1, x_2, \dots, x_d)$ and $\vec{\xi} = (\xi_1, \xi_2, \dots, \xi_d)$ denote the physical and computational coordinates, respectively. Here $d \geq 1$ denotes the number of spatial dimensions. A one-to-one coordinate transformation from the computational (or logical) domain Ω_c to the physical domain Ω_p is denoted by

$$\vec{x} = \vec{x}(\vec{\xi}), \qquad \vec{\xi} \in \Omega_c. \tag{2.1}$$

Its inversion is denoted by

$$\vec{\xi} = \vec{\xi}(\vec{x}), \qquad \vec{x} \in \Omega_p. \tag{2.2}$$

In the variational approach, the mesh map is provided by the minimizer of a functional of the following form:

$$E(\vec{\xi}) = \frac{1}{2}\sum_k \int_{\Omega_p} \nabla\xi_k^T G_k^{-1} \nabla\xi_k d\vec{x}, \tag{2.3}$$

where $\nabla := (\partial_{x_1}, \partial_{x_2}, \ldots, \partial_{x_d})^T$ and G_k are given symmetric positive definite matrices called *monitor functions*. In general, monitor functions depend on the underlying solution to be adapted. More terms can be added to the functional (2.3) to control other aspects of the mesh such as orthogonality and mesh alignment with a given vector field [5, 6].

The variational mesh is determined by the Euler–Lagrange equation of the above functional:

$$\nabla \cdot \left(G_k^{-1}\nabla\xi_k\right) = 0, \qquad 1 \le k \le d. \tag{2.4}$$

One of the simplest choices of monitor functions is $G_k = \omega I$, $1 \le k \le d$, where I is the identity matrix and ω is a positive weight function, e.g., $\omega = \sqrt{1 + |\nabla u|^2}$. Here u is the solution of the underlying PDE. In this case, we obtain Winslow's variable diffusion method [36]:

$$\nabla \cdot \left(\frac{1}{\omega}\nabla\xi_k\right) = 0, \qquad 1 \le k \le d. \tag{2.5}$$

By using the above equations, a map between the physical domain Ω_p and the logical domain Ω_c can be computed. Typically, the map transforms a uniform mesh in the logical domain, clustering grid points in those regions of the physical domain where the solution has the largest gradients.

2.1. 1D case. Although the main objective of this work is to provide an effective moving mesh algorithm for 2D conservation laws, it is easier to illustrate the basic moving mesh ideas by starting with some 1D discussions. Let x and ξ denote the physical and computational coordinates, respectively, which are (without loss of generality) assumed to be in $[a, b]$ and $[0, 1]$, respectively. A one-to-one coordinate transformation between these domains is denoted by

$$\begin{aligned} &x = x(\xi), \qquad && \xi \in [0, 1], \\ &x(0) = a, && x(1) = b. \end{aligned} \tag{2.6}$$

The 1D Euler–Lagrange equation has the form

$$(\omega^{-1}\xi_x)_x = 0. \tag{2.7}$$

Using the above equation, we can obtain the conventional 1D *equidistribution principle*: $\omega x_\xi =$ constant, or equivalently,

$$(\omega x_\xi)_\xi = 0. \tag{2.8}$$

Both (2.7) and (2.8) have the same form, and therefore solving either of them will yield the desired mesh map $x = x(\xi)$. However, the situation is different in the 2D case, where we will choose to solve equations of the form (2.8), as will be described in the next subsection.

2.2. 2D case. We will consider the Winslow's variable diffusion method (2.5). The extension to the general Euler–Lagrange equation (2.4) is straightforward. Let $(x, y) = (x(\xi, \eta), y(\xi, \eta))$ be the mesh map in two dimensions. Then (2.5) becomes

$$\begin{aligned} &(\omega^{-1}\xi_x)_x + (\omega^{-1}\xi_y)_y = 0, \\ &(\omega^{-1}\eta_x)_x + (\omega^{-1}\eta_y)_y = 0. \end{aligned} \tag{2.9}$$

In practice, the physical domain Ω_p may have a very complex geometry, and as a result, solving the elliptic system (2.9) directly on structured grids is unrealistic. Therefore we usually solve the corresponding mesh generation equations on the computational domain Ω_c by interchanging the dependent and independent variables in (2.9):

$$\begin{aligned} &\frac{x_\xi}{J}\left[\left(x_\eta \frac{1}{J\omega} x_\eta + y_\eta \frac{1}{J\omega} y_\eta\right)_\xi - \left(x_\xi \frac{1}{J\omega} x_\eta + y_\xi \frac{1}{J\omega} y_\eta\right)_\eta\right] \\ &\qquad + \frac{x_\eta}{J}\left[-\left(x_\eta \frac{1}{J\omega} x_\xi + y_\eta \frac{1}{J\omega} y_\xi\right)_\xi + \left(x_\xi \frac{1}{J\omega} x_\xi + y_\xi \frac{1}{J\omega} y_\xi\right)_\eta\right] = 0, \\ &\frac{y_\xi}{J}\left[\left(x_\eta \frac{1}{J\omega} x_\eta + y_\eta \frac{1}{J\omega} y_\eta\right)_\xi - \left(x_\xi \frac{1}{J\omega} x_\eta + y_\xi \frac{1}{J\omega} y_\eta\right)_\eta\right] \\ &\qquad + \frac{y_\eta}{J}\left[-\left(x_\eta \frac{1}{J\omega} x_\xi + y_\eta \frac{1}{J\omega} y_\xi\right)_\xi + \left(x_\xi \frac{1}{J\omega} x_\xi + y_\xi \frac{1}{J\omega} y_\xi\right)_\eta\right] = 0. \end{aligned} \tag{2.10}$$

Note that system (2.10) is more complicated than the Euler–Lagrange equation (2.9), which requires more computational effort in obtaining numerical approximations. An alternative approach, as observed by Ceniceros and Hou [8], is to consider a functional defined in the computational domain,

$$\tilde{E}[x, y] = \frac{1}{2}\int_{\Omega_c} \left(\widetilde{\nabla}^T x G_1 \widetilde{\nabla} x + \widetilde{\nabla}^T y G_2 \widetilde{\nabla} y\right)\, d\xi d\eta, \tag{2.11}$$

to replace the conventional functional (2.3), where G_k are monitor functions, and $\widetilde{\nabla} = (\partial_\xi, \partial_\eta)^T$. The corresponding Euler–Lagrange equation is

$$\begin{aligned} &\partial_\xi(G_1\partial_\xi x) + \partial_\eta(G_1\partial_\eta x) = 0, \\ &\partial_\xi(G_2\partial_\xi y) + \partial_\eta(G_2\partial_\eta y) = 0. \end{aligned} \tag{2.12}$$

In particular, with the choice $G = \omega I$ we have

$$\widetilde{\nabla} \cdot (\omega\widetilde{\nabla} x) = 0, \qquad \widetilde{\nabla} \cdot (\omega\widetilde{\nabla} y) = 0. \tag{2.13}$$

The monitor functions will be chosen based on the properties of the physical solutions. A typical choice used in [8] is $\omega = \sqrt{1 + \alpha_1|u|^2 + \alpha_2|\widetilde{\nabla} u|^2}$ or $\omega = \sqrt{1 + \alpha_1|u|^2 + \alpha_2|\nabla u|^2}$, where α_1, α_2 are some nonnegative constants.

3. 1D algorithm. For convenience, we assume that a fixed uniform mesh on the computational domain is given by $\xi_j = j/(J+1), 0 \le j \le J+1$. We denote the cell average of the solution $u(x)$ over the interval $[x_j, x_{j+1}]$ as

$$u_{j+\frac{1}{2}} = \frac{1}{\Delta x_{j+\frac{1}{2}}}\int_{x_j}^{x_{j+1}} u(x)\, dx,$$

where $\Delta x_{j+\frac{1}{2}} = x_{j+1} - x_j$. In practice, the monitor function ω is always associated with the underlying solution u or/and its derivatives, but without loss of generality we assume that $\omega = \omega(u)$. For monitor functions involving first or second derivatives, central differencing will be used to approximate these derivatives.

3.1. Mesh-redistribution. In order to solve the mesh-redistribution equation (2.8), we introduce an artificial time τ and solve

$$x_\tau = (\omega x_\xi)_\xi, \qquad 0 < \xi < 1, \tag{3.1}$$

subject to boundary conditions $x(0,\tau) = a$ and $x(1,\tau) = b$. We discretize (3.1) on the uniform mesh in Ω_c:

$$\tilde{x}_j = x_j + \frac{\Delta\tau}{\Delta\xi^2}\left[\omega(u_{j+\frac{1}{2}})(x_{j+1} - x_j) - \omega(u_{j-\frac{1}{2}})(x_j - x_{j-1})\right], \quad 1 \le j \le J, \tag{3.2}$$

where $\Delta\xi = 1/(J+1)$ is the step size in Ω_c. Solving (3.2) with boundary conditions $x_0 = a$ and $x_{J+1} = b$ leads to a new grid in the physical domain Ω_p. Some advantages of using the approach (3.1) and (3.2) to solve the mesh redistribution equation (2.8) will be seen from Lemma 3.1 and Theorem 3.1.

3.2. Solution-updating on new grids. After obtaining the new grid $\{\widetilde{x}_j\}$, we need to update u at the grid point $\widetilde{x}_{j+\frac{1}{2}} = (\widetilde{x}_j + \widetilde{x}_{j+1})/2$ based on the knowledge of $\{x_{j+\frac{1}{2}}, \widetilde{x}_{j+\frac{1}{2}}, u_{j+\frac{1}{2}}\}$. The traditional way to do this is using the conventional interpolation

$$\widetilde{u}_{j+\frac{1}{2}} = u_{k+\frac{1}{2}} + \frac{u_{k+\frac{1}{2}} - u_{k-\frac{1}{2}}}{x_{k+\frac{1}{2}} - x_{k-\frac{1}{2}}}(\widetilde{x}_{j+\frac{1}{2}} - x_{k+\frac{1}{2}}) \qquad \text{if } \widetilde{x}_{j+\frac{1}{2}} \in [x_{k-\frac{1}{2}}, x_{k+\frac{1}{2}}]. \tag{3.3}$$

Since the monitor function ω is dependent on the underlying solution u, the grid redistribution equations (3.2)–(3.3) form a nonlinear system. It is therefore natural to make several iterations to solve (3.2)–(3.3) in order to gain better control of the grid distribution near those regions where the solution u has a large gradient. In solving hyperbolic conservation laws with strong discontinuities (e.g., shocks), iteration techniques based on (3.2)–(3.3) have been employed, and it is found that the results for the solution and the mesh are not satisfactory. The main problem is that the linear interpolation (3.3) cannot preserve conservation of mass, which, by the Lax–Wendroff theory, is an essential requirement for a good numerical scheme for hyperbolic conservation laws.

In the following we will introduce a new method to update u, noting that mass-conservation is an essential requirement for hyperbolic conservation laws. To begin with, assume that the difference between $\tilde{x}_{j+\frac{1}{2}}$ and $x_{j+\frac{1}{2}}$ is small. Let $\widetilde{u}_{j+\frac{1}{2}}$ and $u_{j+\frac{1}{2}}$ be cell averages of the solution $u(x)$ over the intervals $[\widetilde{x}_j, \widetilde{x}_{j+1}]$ and $[x_j, x_{j+1}]$, respectively. We will derive a formula for $\widetilde{u}_{j+\frac{1}{2}}$ using the perturbation method. If

$\widetilde{x} = x - c(x)$ with a small displacement $c(x)$, i.e., $|c(x)| \ll 1$, then we have

$$\begin{aligned}\int_{\widetilde{x}_j}^{\widetilde{x}_{j+1}} \tilde{u}(\widetilde{x})\, d\widetilde{x} &= \int_{x_j}^{x_{j+1}} u(x - c(x))(1 - c'(x))\, dx \\ &\approx \int_{x_j}^{x_{j+1}} (u(x) - c(x)u_x(x))(1 - c'(x))\, dx \\ &\approx \int_{x_j}^{x_{j+1}} (u(x) - (cu)_x)\, dx \\ &= \int_{x_j}^{x_{j+1}} u(x)\, dx - ((cu)_{j+1} - (cu)_j), \end{aligned} \tag{3.4}$$

where we have neglected higher-order terms, and $(cu)_j$ denotes the value of cu at the jth cell interface. The following almost conservative-interpolation formula follows from (3.4):

$$\Delta\widetilde{x}_{j+\frac{1}{2}}\widetilde{u}_{j+\frac{1}{2}} = \Delta x_{j+\frac{1}{2}} u_{j+\frac{1}{2}} - \big((cu)_{j+1} - (cu)_j\big), \tag{3.5}$$

where $\Delta\widetilde{x}_{j+\frac{1}{2}} = \widetilde{x}_{j+1} - \widetilde{x}_j$ and $c_j = x_j - \widetilde{x}_j$. Note that the above solution-updating method guarantees the conservation of mass in the following sense:

$$\sum_j \Delta\widetilde{x}_{j+\frac{1}{2}}\widetilde{u}_{j+\frac{1}{2}} = \sum_j \Delta x_{j+\frac{1}{2}} u_{j+\frac{1}{2}}. \tag{3.6}$$

The linear flux cu in (3.5) will be approximated by some upwinding numerical flux; see (3.11) below.

If the function u is suitably smooth, then it can be shown that the size of the moving speed $c(x)$ is small. It is known that the first and second derivatives of the parabolic-type equation (3.1) are bounded, provided that the initial data and the monitor function satisfy some regularity requirements. By the definition of $c(x)$, we have

$$\begin{aligned} c(x) &= x - \widetilde{x} = -(x_\tau)\Delta\tau = \mathcal{O}(\Delta\tau), \\ c'(x) &= 1 - \tilde{x}_x = 1 - \frac{\widetilde{x}_\xi}{x_\xi} = -\frac{x_{\xi\tau}}{x_\xi}\Delta\tau = \mathcal{O}(\Delta\tau), \end{aligned}$$

which indicate that the moving speed in each cell is indeed very small.

3.3. Solution procedure. Our solution procedure is based on two independent parts: a mesh-redistribution algorithm and a solution algorithm. The first part will be based on an iteration procedure using (3.2) and (3.5). The second part will be independent of the first, and it can be any of the standard codes for solving the given PDEs, such as ENO schemes [31, 39], central schemes [17, 27], relaxation schemes [16, 26, 34], BGK schemes [38, 35], and several other types of high-resolution methods (see, e.g., [20, 15, 18]). The solution procedure can be illustrated by the following flowchart.

ALGORITHM 0.

Step 1. Given a uniform (fixed) partition of the logical domain Ω_c, use the equidistribution principle (2.8) to generate an initial partition $x_j^{[0]} := x_j$ of the physical domain Ω_p. Then compute the grid values $u_{j+\frac{1}{2}}^{[0]}$ based on the cell average for the initial data $u(x, 0)$.

Step 2. Move grid $\{x_j^{[\nu]}\}$ to $\{x_j^{[\nu+1]}\}$ based on scheme (3.2), and compute $\{u_{j+\frac{1}{2}}^{[\nu+1]}\}$ on the new grid based on scheme (3.5) for $\nu \geq 0$. Repeat the updating procedure for a fixed number of iterations or until $\|x^{[\nu+1]} - x^{[\nu]}\| \leq \epsilon$. The mesh-redistribution scheme (3.2) can be also replaced by the Gauss–Seidel iteration procedure (3.32), as discussed at the end of this section.

Step 3. Evolve the underlying PDEs using a high-resolution finite volume method on the mesh $\{x_j^{[\nu+1]}\}$ to obtain the numerical approximations $u_{j+\frac{1}{2}}^{n+1}$ at the time level t_{n+1}.

Step 4. If $t_{n+1} \leq T$, then let $u_{j+\frac{1}{2}}^{[0]} := u_{j+\frac{1}{2}}^{n+1}$ and $x_j^{[0]} := x_j^{[\nu+1]}$ and go to **Step 2**.

3.3.1. Some discussions on Step 2. A new mesh $x_j^{[\nu+1]}$ is obtained using (3.2):

$$x_j^{[\nu+1]} = \alpha_{j+\frac{1}{2}} x_{j+1}^{[\nu]} + (1 - \alpha_{j+\frac{1}{2}} - \alpha_{j-\frac{1}{2}}) x_j^{[\nu]} + \alpha_{j-\frac{1}{2}} x_{j-1}^{[\nu]} \tag{3.7}$$

where

$$\alpha_{j+\frac{1}{2}} = \frac{\Delta\tau}{\Delta\xi^2} \omega(u_{j+\frac{1}{2}}^{[\nu]}).$$

The above equation is solved subject to the following stability condition:

$$\max_j \alpha_{j+\frac{1}{2}} \leq \frac{1}{2}. \tag{3.8}$$

Next, numerical solutions are updated on the new grids $\{x_j^{[\nu+1]}\}$ (at the same time level) using (3.5),

$$u_{j+\frac{1}{2}}^{[\nu+1]} = \beta_j^{[\nu]} u_{j+\frac{1}{2}}^{[\nu]} - \gamma_j^{[\nu]} ((\widehat{cu})_{j+1}^{[\nu]} - (\widehat{cu})_j^{[\nu]}), \tag{3.9}$$

where

$$\gamma_j^{[\nu]} = (x_{j+1}^{[\nu+1]} - x_j^{[\nu+1]})^{-1}, \qquad \beta_j^{[\nu]} = \gamma_j^{[\nu]} \cdot (x_{j+1}^{[\nu]} - x_j^{[\nu]}), \tag{3.10}$$

and the numerical flux $\widehat{cu}_j$ is defined by

$$(\widehat{cu})_j = \frac{c_j}{2}(u_{j+\frac{1}{2}} + u_{j-\frac{1}{2}}) - \frac{|c_j|}{2}(u_{j+\frac{1}{2}} - u_{j-\frac{1}{2}}). \tag{3.11}$$

The wave speed c_j above is defined by $c_j^{[\nu]} = x_j^{[\nu]} - x_j^{[\nu+1]}$.

Remark 3.1. In our numerical computation, the first-order numerical flux $(\widehat{cu})_j$ defined by (3.11) will be replaced by a second-order one as the following:

$$(\widehat{cu})_j = \frac{c_j}{2}(u_j^+ + u_j^-) - \frac{|c_j|}{2}(u_j^+ - u_j^-), \tag{3.12}$$

where u_j^+ and u_j^- will be defined by (3.19) below.

Remark 3.2. In practice, it is common to use some temporal or spatial *smoothing* on the monitor function to obtain smoother meshes. One of the reasons for using smoothing is to avoid very singular mesh and/or large approximation errors near those regions where the solution has a large gradient. In this work, we apply the following low pass filter to smooth the monitor:

$$\omega_{j+\frac{1}{2}} \leftarrow \frac{1}{4}(\omega_{j+\frac{3}{2}} + 2\omega_{j+\frac{1}{2}} + \omega_{j-\frac{1}{2}}), \tag{3.13}$$

where $\omega_{j+\frac{1}{2}} = \omega(u_{j+\frac{1}{2}})$.

3.3.2. Some discussions on Step 3. This step is independent of Step 2, and, as a result, it can be done using any efficient modern numerical technique for hyperbolic conservation laws. As an example, we consider a second-order finite volume approach to solving the 1D scalar hyperbolic conservation laws

$$u_t + f(u)_x = 0, \qquad t > 0, \tag{3.14}$$

with compactly supported initial condition

$$u(x,0) = u_0(x), \qquad u_0 \in L^\infty \cap BV. \tag{3.15}$$

Integrating (3.14) over the control volume $[t_n, t_{n+1}) \times [x_j, x_{j+1}]$ leads to the following (explicit) finite volume method:

$$u^{n+1}_{j+\frac{1}{2}} = u^n_{j+\frac{1}{2}} - \frac{t_{n+1} - t_n}{x_{j+1} - x_j}\left(\widehat{f}^n_{j+1} - \widehat{f}^n_j\right), \tag{3.16}$$

where $\widehat{f}^n_j$ is some appropriate numerical flux satisfying

$$\widehat{f}^n_j = \widehat{f}(u^{n,-}_j, u^{n,+}_j), \qquad \widehat{f}(u,u) = f(u). \tag{3.17}$$

An example of such a numerical flux is the Lax–Friedrichs flux:

$$\widehat{f}(a,b) = \frac{1}{2}\left[f(a) + f(b) - \max_u \{|f_u|\}\,(b-a)\right]. \tag{3.18}$$

In (3.17), $u^{n,\pm}_j$ are defined by

$$u^{n,\pm}_j = u^n_{j\pm\frac{1}{2}} + \frac{1}{2}(x_j - x_{j\pm1})\tilde{S}_{j\pm\frac{1}{2}}, \tag{3.19}$$

where $\tilde{S}_{j+\frac{1}{2}}$ is an approximation of the slope u_x at $x_{j+\frac{1}{2}}$, defined by

$$\tilde{S}_{j+\frac{1}{2}} = \left(\operatorname{sign}(\tilde{S}^+_{j+\frac{1}{2}}) + \operatorname{sign}(\tilde{S}^-_{j+\frac{1}{2}})\right)\frac{|\tilde{S}^+_{j+\frac{1}{2}}\tilde{S}^-_{j+\frac{1}{2}}|}{|\tilde{S}^+_{j+\frac{1}{2}}| + |\tilde{S}^-_{j+\frac{1}{2}}|}, \tag{3.20}$$

with

$$\tilde{S}^+_{j+\frac{1}{2}} = \frac{u^n_{j+\frac{3}{2}} - u^n_{j+\frac{1}{2}}}{x_{j+\frac{3}{2}} - x_{j+\frac{1}{2}}}, \qquad \tilde{S}^-_{j+\frac{1}{2}} = \frac{u^n_{j+\frac{1}{2}} - u^n_{j-\frac{1}{2}}}{x_{j+\frac{1}{2}} - x_{j-\frac{1}{2}}}.$$

The MUSCL (monotone upstream-centered scheme for conservation laws)-type finite volume method (3.16)–(3.18), which is of second-order accuracy in smooth regions, will be applied in the 1D numerical experiments.

3.4. Some theoretical results on the adaptive mesh solutions. In one dimension, some good theoretical guarantees for the numerical grids can be obtained. In the following, we prove some theoretical results for the mesh-redistribution equation (3.7) and the solution-updating equation (3.9). We first demonstrate that the new mesh $x^{[\nu+1]}$ generated by (3.7) keeps the monotonic order of $x^{[\nu]}$.

Lemma 3.1. *Assume $x^{[\nu]}_{j+1} > x^{[\nu]}_j$ for $0 \le j \le J$. If the new mesh $x^{[\nu+1]}$ is obtained using* (3.7)*, with $\alpha_{j+\frac{1}{2}}$ satisfying the stability condition* (3.8)*, then $x^{[\nu+1]}_{j-1} < x^{[\nu]}_j < x^{[\nu+1]}_{j+1}$ for $1 \le j \le J$, and $x^{[\nu+1]}_{j+1} > x^{[\nu+1]}_j$ for $0 \le j \le J$.*

Proof. Using the stability condition (3.8) gives $1-\alpha_{j+\frac{1}{2}}-\alpha_{j-\frac{1}{2}}\geq 0$. Moreover, $\alpha_{j\pm\frac{1}{2}}$ are all positive. Therefore, it follows from (3.7) and the assumption $x_{j+1}^{[\nu]}>x_j^{[\nu]}$ that $x_{j-1}^{[\nu+1]}<x_j^{[\nu]}<x_{j+1}^{[\nu+1]}$. We now rewrite (3.7) into the following form:

$$x_j^{[\nu+1]}=\alpha_{j+\frac{1}{2}}\Delta x_{j+\frac{1}{2}}^{[\nu]}+x_j^{[\nu]}-\alpha_{j-\frac{1}{2}}\Delta x_{j-\frac{1}{2}}^{[\nu]}, \tag{3.21}$$

where $\Delta x_{j+\frac{1}{2}}=x_{j+1}-x_j$. It follows from the above equation that

$$\Delta x_{j-\frac{1}{2}}^{[\nu+1]}=\alpha_{j+\frac{1}{2}}\Delta x_{j+\frac{1}{2}}^{[\nu]}+(1-2\alpha_{j-\frac{1}{2}})\Delta x_{j-\frac{1}{2}}^{[\nu]}+\alpha_{j-\frac{3}{2}}\Delta x_{j-\frac{3}{2}}^{[\nu]}.$$

Since the first and last coefficients of the right-hand side are positive and the second one is nonnegative (due to the stability condition (3.8)), the assumption $x_{j+1}^{[\nu]}>x_j^{[\nu]}$ yields $\Delta x_{j-\frac{1}{2}}^{[\nu+1]}>0$. This shows that $x_{j+1}^{[\nu+1]}>x_j^{[\nu+1]}$ for $0\leq j\leq J$. □

Remark 3.3. A consequence of Lemma 3.1 is that $x_j^{[\nu+1]}\in(x_{j-1}^{[\nu]},x_{j+1}^{[\nu]})$, which implies that the speed of mesh moving is finite. This is important in better controlling grid distribution near the regions of large gradients in the solution.

Next, we provide a necessary condition under which the updated solution $u_{j+\frac{1}{2}}^{[\nu+1]}$ satisfies the TVD property.

LEMMA 3.2. *Assume that the initial data* $u^{[0]}$ *is compactly supported and that the stability condition* (3.8) *is satisfied. If* $x_{j-1}^{[\nu+1]}\leq x_j^{[\nu]}\leq x_{j+1}^{[\nu+1]}$ *and* $x_{j+1}^{[\nu+1]}>x_j^{[\nu+1]}$, *then the solution-updating scheme* (3.9)–(3.11) *satisfies*

$$\mathsf{TV}(u^{[\nu+1]})\leq\mathsf{TV}(u^{[\nu]}),$$

where the total variation is defined by

$$\mathsf{TV}(u):=\sum_j\left|u_{j+\frac{1}{2}}-u_{j-\frac{1}{2}}\right|.$$

Proof. For ease of notation we denote $\tilde{x}=x^{[\nu+1]},x=x^{[\nu]},\tilde{u}=u^{[\nu+1]},u=u^{[\nu]}$. Note that $c_{j+1}-c_j=\Delta x_{j+\frac{1}{2}}-\Delta\tilde{x}_{j+\frac{1}{2}}$. This fact, together with the scheme (3.9) and the numerical flux (3.11), gives

$$\begin{aligned}\Delta\tilde{x}_{j+\frac{1}{2}}\tilde{u}_{j+\frac{1}{2}}&=\left(c_{j+1}-c_j+\Delta\tilde{x}_{j+\frac{1}{2}}\right)u_{j+\frac{1}{2}}+\frac{1}{2}\left(|c_{j+1}|-c_{j+1}\right)u_{j+\frac{3}{2}}\\&\quad+\frac{1}{2}\left(c_j-|c_j|-c_{j+1}-|c_{j+1}|\right)u_{j+\frac{1}{2}}+\frac{1}{2}\left(|c_j|+c_j\right)u_{j-\frac{1}{2}}\\&=\Delta\tilde{x}_{j+\frac{1}{2}}u_{j+\frac{1}{2}}+m_{j+1}u_{j+\frac{3}{2}}-m_{j+1}u_{j+\frac{1}{2}}-M_ju_{j+\frac{1}{2}}+M_ju_{j-\frac{1}{2}},\end{aligned}$$

where $M_j=\max(0,c_j)$ and $m_j=-\min(0,c_j)$. Note that both M_j and m_j are nonnegative. It follows from the above result that

$$\tilde{u}_{j+\frac{1}{2}}=u_{j+\frac{1}{2}}+\frac{m_{j+1}}{\Delta\tilde{x}_{j+\frac{1}{2}}}\Delta u_{j+1}-\frac{M_j}{\Delta\tilde{x}_{j+\frac{1}{2}}}\Delta u_j, \tag{3.22}$$

where $\Delta u_j=u_{j+\frac{1}{2}}-u_{j-\frac{1}{2}}$. It follows from (3.22) that

$$\Delta\tilde{u}_j=\Delta u_j+\frac{m_{j+1}}{\Delta\tilde{x}_{j+\frac{1}{2}}}\Delta u_{j+1}-\left(\frac{M_j}{\Delta\tilde{x}_{j+\frac{1}{2}}}+\frac{m_j}{\Delta\tilde{x}_{j-\frac{1}{2}}}\right)\Delta u_j+\frac{M_{j-1}}{\Delta\tilde{x}_{j-\frac{1}{2}}}\Delta u_{j-1},$$

which gives

(3.23)

$$\sum_j |\Delta \tilde{u}_j| \le \sum_j \frac{m_j}{\Delta \tilde{x}_{j-\frac{1}{2}}} |\Delta u_j| + \sum_j \left| 1 - \frac{M_j}{\Delta \tilde{x}_{j+\frac{1}{2}}} - \frac{m_j}{\Delta \tilde{x}_{j-\frac{1}{2}}} \right| |\Delta u_j| \sum_j \frac{M_j}{\Delta \tilde{x}_{j+\frac{1}{2}}} |\Delta u_j|.$$

It can be verified using the definition of c_j that the condition $\tilde{x}_{j-1} \le x_j \le \tilde{x}_{j+1}$ is equivalent to $-\Delta \tilde{x}_{j-\frac{1}{2}} \le c_j \le \Delta \tilde{x}_{j+\frac{1}{2}}$. This fact, together with the observation that $M_j = 0$ when $c_j \le 0$ and $m_j = 0$ when $c_j \ge 0$, yields

$$1 - \frac{M_j}{\Delta \tilde{x}_{j+\frac{1}{2}}} - \frac{m_j}{\Delta \tilde{x}_{j-\frac{1}{2}}} \ge 0. \tag{3.24}$$

It follows from (3.23) and (3.24) that $\mathsf{TV}(\tilde{u}) \le \mathsf{TV}(u)$. □

With the two ingredients above, the following TVD property for Step 2 of Algorithm 0 is established.

THEOREM 3.1. *Assume that the initial data* $u^{[0]}$ *is compactly supported and that the stability condition* (3.8) *is satisfied. Then the iterated mesh and solution* $\{x^{[\nu+1]}, u^{[\nu+1]}\}$ *generated by* (3.7)–(3.11) *satisfies* $\mathsf{TV}(u^{[\nu+1]}) \le \mathsf{TV}(u^{[\nu]})$.

Remark 3.4. If the PDE solver in Step 3 of Algorithm 0 is TVB (i.e., TV-bounded) (or TVD), then the above theorem guarantees the TVB (or TVD) property of the moving mesh solution at any time level. It can be also proved similarly that the l^∞- and the l^1-stabilities are also preserved.

THEOREM 3.2. *Assume that the initial function* u_0 *in* (3.15) *is compactly supported and that the stability condition* (3.8) *is satisfied. Then the moving mesh solution generated by Algorithm* 0, *with* (3.7)–(3.11) *for Step* 2 *and* (3.16) *for Step* 3, *is a weak solution of the conservation law* (3.14).

Proof. Without loss of generality we assume that only one iteration is used in Step 2 of Algorithm 0. Then, given $\{x_j^n, u_{j+\frac{1}{2}}^n\}$, the solution $\{x_j^{n+1}, u_{j+\frac{1}{2}}^{n+1}\}$ is computed by the following operator-splitting-type algorithm:

$$U_{j+\frac{1}{2}}^n = u_{j+\frac{1}{2}}^n - \frac{\Delta t_n}{\Delta x_{j+\frac{1}{2}}^n} \left(f_{j+1}^n - f_j^n \right), \tag{3.25}$$

$$x_j^{n+1} = x_j^n + \frac{\Delta \tau}{\Delta \xi^2} \left[\omega(U_{j+\frac{1}{2}}^n)(x_{j+1}^n - x_j^n) - \omega(U_{j-\frac{1}{2}}^n)(x_j^n - x_{j-1}^n) \right], \tag{3.26}$$

$$\Delta x_{j+\frac{1}{2}}^{n+1} u_{j+\frac{1}{2}}^{n+1} = \Delta x_{j+\frac{1}{2}}^n U_{j+\frac{1}{2}}^n - ((cU^n)_{j+1} - (cU^n)_j), \tag{3.27}$$

where $c_j = x_j^n - x_j^{n+1}$ and the numerical flux f_j^n satisfies the consistency requirement. Multiplying the first equation above by a test function $\phi \in C_0^\infty(\mathbf{R} \times (0, T])$ gives

$$\Delta x_{j+\frac{1}{2}}^n U_{j+\frac{1}{2}}^n \phi(x_j^n, t_n) = \Delta x_{j+\frac{1}{2}}^n u_{j+\frac{1}{2}}^n \phi(x_j^n, t_n) - \Delta t_n \left(f_{j+1}^n - f_j^n \right) \phi(x_j^n, t_n),$$

which, together with the interpolation step (3.27), gives

$$\begin{aligned} &\Delta x_{j+\frac{1}{2}}^{n+1} u_{j+\frac{1}{2}}^{n+1} \phi(x_j^n, t_n) + ((cU^n)_{j+1} - (cU^n)_j) \phi(x_j^n, t_n) \\ &\quad = \Delta x_{j+\frac{1}{2}}^n u_{j+\frac{1}{2}}^n \phi(x_j^n, t_n) - \Delta t_n \left(f_{j+1}^n - f_j^n \right) \phi(x_j^n, t_n). \end{aligned} \tag{3.28}$$

Standard summation by parts yields

$$\sum_j \sum_{n=0}^{N} \left[\Delta x_{j+\frac{1}{2}}^{n+1} u_{j+\frac{1}{2}}^{n+1} - \Delta x_{j+\frac{1}{2}}^{n} u_{j+\frac{1}{2}}^{n}\right] \phi(x_j^n, t_n)$$

$$= -\sum_j \sum_{n=0}^{N} ((cU^n)_{j+1} - (cU^n)_j)\phi(x_j^n, t_n) - \sum_{n=0}^{N} \sum_j \Delta t_n \left(f_{j+1}^n - f_j^n\right) \phi(x_j^n, t_n)$$

and then

$$\begin{aligned} -\sum_j \Delta x_{j+\frac{1}{2}}^0 u_{j+\frac{1}{2}}^0 \phi(x_j^0, 0) &- \sum_j \sum_{n=1}^{N} [\phi(x_j^n, t_n) - \phi(x_j^{n-1}, t_{n-1})] \Delta x_{j+\frac{1}{2}}^n u_{j+\frac{1}{2}}^n \\ &= \sum_j \sum_{n=0}^{N} [\phi(x_j^n, t_n) - \phi(x_{j-1}^n, t_n)](x_j^n - x_j^{n+1}) U_j^n \\ &\quad + \sum_{n=0}^{N} \sum_j \Delta t_n \left[\phi(x_j^n, t_n) - \phi(x_{j-1}^n, t_n)\right] f_j^n, \end{aligned} \tag{3.29}$$

where we have used the fact $\phi(x, t_N) = 0$ with $t_N = T$. We can show that $\Delta x_{j+\frac{1}{2}}^0 \to 0$ as $J \to \infty$. Without loss of generality, assume that the monitor function is the one associated with the equidistribution principle, i.e., $\omega(u) = \sqrt{1 + u_x^2}$. Then

$$L = \sqrt{1 + u_x^2(x_j^0, 0)}\Delta x_{j+\frac{1}{2}}^0, \qquad 0 \le j \le J - 1, \tag{3.30}$$

is a constant independent of j. It follows from the definition of L that

$$L \le \left(1 + |u_x(x_j^0, 0)|\right) \Delta x_{j+\frac{1}{2}}^0, \qquad 0 \le j \le J - 1,$$

which gives

$$JL \le \text{the size of } u_0\text{'s support} + \mathsf{TV}(u^0).$$

This, together with the definition (3.30), leads to

$$\Delta x_{j+\frac{1}{2}}^0 \le \text{const. } J^{-1} \to 0 \qquad \text{as} \quad J \to \infty. \tag{3.31}$$

Moreover, since $\{x_j^n\}$, with $n \ge 1$, are obtained by solving a parabolic equation, we have $\Delta x_{j+\frac{1}{2}}^n \sim \mathcal{O}(\Delta\xi) \to 0$ for $n \ge 1$. Taking limits on both sides of (3.29) leads to

$$-\int u(x, 0)\phi(x, 0)dx - \int\int (\phi_x x_t + \phi_t) u dx dt = -\int\int \phi_x x_t u dx dt + \int\int \phi_x f(u) dx dt,$$

provided that the numerical solution is convergent to u almost everywhere, where for the second term on the LHS (left-hand side) of (3.29) we have used the fact $\phi(x, t)_t = \phi_t x_t + \phi_t$, and for the first term on the RHS (right-hand side) we have used the fact that $c_j = x_j^n - x_j^n \sim x_t dt$. The above result leads to

$$\int\int (\phi_t u + \phi_x f(u)) dx dt + \int u(x, 0)\phi(x, 0) dx = 0,$$

which indicates that the moving mesh solution is indeed a weak solution of the underlying conservation law. □

3.5. Grid-motion with Gauss–Seidel iteration. In practice, we also use the following Gauss–Seidel-type iteration to solve the mesh moving equation (2.8):

$$\omega(u^{[\nu]}_{j+\frac{1}{2}})(x^{[\nu]}_{j+1} - x^{[\nu+1]}_j) - \omega(u^{[\nu]}_{j-\frac{1}{2}})(x^{[\nu+1]}_j - x^{[\nu+1]}_{j-1}) = 0. \tag{3.32}$$

It can also be demonstrated that the new mesh $x^{[\nu+1]}$ generated by (3.32) keeps the monotonic order of $x^{[\nu]}$.

LEMMA 3.3. *Assume $x^{[\nu]}_{j+1} > x^{[\nu]}_j$ for $0 \leq j \leq J$. If the new mesh $x^{[\nu+1]}$ is obtained by using the Gauss–Seidel iterative scheme* (3.32)*, with positive monitor function ω, then $x^{[\nu+1]}_j > x^{[\nu+1]}_{j-1}$ for $1 \leq j \leq J+1$. Moreover, $x^{[\nu]}_j > x^{[\nu+1]}_{j-1}$ for $1 \leq j \leq J+1$.*

Proof. We again denote $\tilde{x} = x^{[\nu+1]}, x = x^{[\nu]}$. It follows from (3.32) that

$$-\alpha_j x_{j+1} + \tilde{x}_j - \beta_j \tilde{x}_{j-1} = 0, \tag{3.33}$$

where $\alpha_j > 0, \beta_j > 0$ (due to the positivity assumption of the monitor function), and $\alpha_j + \beta_j = 1$. It follows from the above equation that

$$\tilde{x}_j - x_{j+1} - \beta_j(\tilde{x}_{j-1} - x_j) = \beta_j(x_j - x_{j+1}) \leq 0,$$

which gives that

$$\tilde{x}_j - x_{j+1} \leq \left(\prod_{k=1}^{j} \beta_k\right)(\tilde{x}_0 - x_1) = \left(\prod_{k=1}^{j} \beta_k\right)(x_0 - x_1) < 0.$$

The above result yields $\tilde{x}_j < x_{j+1}$, which, together with (3.32), also leads to $\tilde{x}_j > \tilde{x}_{j-1}$. □

If we can further show that $x^{[\nu]}_j \leq x^{[\nu+1]}_{j+1}$ for the Gauss–Seidel iteration (3.32), then based on Lemma 3.2 the solution-updating scheme (3.9) together with the mesh-redistribution scheme (3.32) will also satisfy the TVD property, i.e., $\mathsf{TV}(u^{[\nu+1]}) \leq \mathsf{TV}(u^{[\nu]})$. However, it seems unlikely that $x^{[\nu]}_j \leq x^{[\nu+1]}_{j+1}$ holds for (3.32) in general situations. On the other hand, the combination of (3.9) and (3.32) has been employed in our numerical experiments, and the numerical results are quite satisfactory. Therefore, both (3.7) and (3.32) can be used in Step 2 of Algorithm 0 to redistribute grid points. In most test problems considered in this work, the grid updating procedure at each time step takes five full Gauss–Seidel iterations, although the difference between solutions with three and five iterations is very small.

4. 2D algorithm.

One of the advantages of the adaptive mesh methods described in the last section is that they can be naturally extended to two dimensions. In the following, we briefly discuss this extension, together with the boundary point redistribution technique which is necessary for 2D mesh redistribution.

4.1. A conservative solution-updating method. In two dimensions, the logical domain $\bar{\Omega}_c = \{(\xi, \eta) | 0 \leq \xi \leq 1, 0 \leq \eta \leq 1\}$ is covered by the square mesh:

$$\left\{(\xi_j, \eta_k) \,\middle|\, \xi_j = \frac{j}{(J_x+1)}, \eta_k = \frac{k}{(J_y+1)};\ 0 \leq j \leq J_x+1, 0 \leq k \leq J_y+1\right\}.$$

Correspondingly, the numerical approximations to $x = x(\xi, \eta)$ and $y = y(\xi, \eta)$ are denoted by $x_{j,k} = x(\xi_j, \eta_k)$ and $y_{j,k} = y(\xi_j, \eta_k)$. As in the 1D case, we will derive a

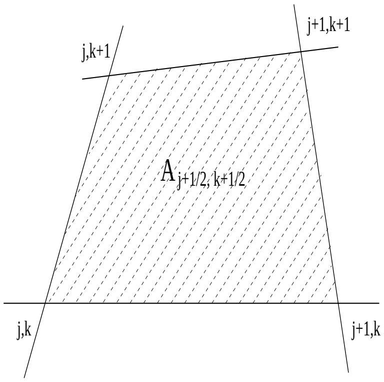


FIG. 4.1. *A control volume.*

conservative scheme to evaluate approximate values at new grid points. Let $A_{j+\frac{1}{2},k+\frac{1}{2}}$ be a control volume as shown in Figure 4.1. Let $\widetilde{A}_{j+\frac{1}{2},k+\frac{1}{2}}$ denote the quadrangle of the finite control volume with four vertices $(\widetilde{x}_{j+p,k+q}, \widetilde{y}_{j+p,k+q})$, $0 \leq p, q \leq 1$, which is of setup similar to Figure 4.1.

Assume that $\widetilde{u}_{j+\frac{1}{2},k+\frac{1}{2}}$ and $u_{j+\frac{1}{2},k+\frac{1}{2}}$ are cell averages of $u(x,,y)$ over $\widetilde{A}_{j+\frac{1}{2},k+\frac{1}{2}}$ and $A_{j+\frac{1}{2},k+\frac{1}{2}}$, respectively. As in the 1D case, we use the perturbation method to evaluate the numerical approximation on the resulting new grids $(\widetilde{x}_{j,k}, \widetilde{y}_{j,k})$. If $(\widetilde{x}, \widetilde{y}) = (x - c^x(x,y),\, y - c^y(x,y))$, where we assume that the speeds (c^x, c^y) have small amplitude, then we have

$$\begin{aligned}
&\int_{\widetilde{A}_{j+\frac{1}{2},k+\frac{1}{2}}} \widetilde{u}(\widetilde{x},\widetilde{y})\ d\widetilde{x}d\widetilde{y} \\
&= \int_{A_{j+\frac{1}{2},k+\frac{1}{2}}} u(x-c^x,\, y-c^y) \det\left(\frac{\partial(\tilde{x},\tilde{y})}{\partial(x,y)}\right) dxdy \\
&\approx \int_{A_{j+\frac{1}{2},k+\frac{1}{2}}} (u(x,y) - c^x u_x - c^y u_y)(1 - c^x_x - c^y_y)\ dxdy \\
&\approx \int_{A_{j+\frac{1}{2},k+\frac{1}{2}}} [u(x,y) - c^x u_x - c^y u_y - c^x_x u - c^y_y u] dxdy \\
&= \int_{A_{j+\frac{1}{2},k+\frac{1}{2}}} [u(x,y) - (c^x u)_x - (c^y u)_y] dxdy \\
&= \int_{A_{j+\frac{1}{2},k+\frac{1}{2}}} u(x,y)\ dxdy - \left[(c_n u)_{j+1,k+\frac{1}{2}} + (c_n u)_{j,k+\frac{1}{2}}\right] \\
&\qquad - \left[(c_n u)_{j+\frac{1}{2},k+1} + (c_n u)_{j+\frac{1}{2},k}\right],
\end{aligned} \tag{4.1}$$

where we have neglected higher-order terms, $c_n := c^x n_x + c^y n_y$ with (n_x, n_y) the unit normal, and $(c_n u)_{j,k+\frac{1}{2}}$ and $(c_n u)_{j+\frac{1}{2},k}$ denote the values of $c_n u$ at the corresponding surfaces of the control volume $A_{j+\frac{1}{2},k+\frac{1}{2}}$. From (4.1), we obtain a conservative-interpolation:

$$\begin{aligned}
&|\widetilde{A}_{j+\frac{1}{2},k+\frac{1}{2}}| \widetilde{u}_{j+\frac{1}{2},k+\frac{1}{2}} = |A_{j+\frac{1}{2},k+\frac{1}{2}}| u_{j+\frac{1}{2},k+\frac{1}{2}} \\
&\quad - \left[(c_n u)_{j+1,k+\frac{1}{2}} + (c_n u)_{j,k+\frac{1}{2}}\right] - \left[(c_n u)_{j+\frac{1}{2},k+1} + (c_n u)_{j+\frac{1}{2},k}\right],
\end{aligned} \tag{4.2}$$

where $|\widetilde{A}|$ and $|A|$ denote the areas of the control volumes $\widetilde{A}$ and A, respectively. It can be verified that the above solution-updating scheme satisfies mass-conservation:

$$\sum_{j,k} |\widetilde{A}_{j+\frac{1}{2},k+\frac{1}{2}}|\widetilde{u}_{j+\frac{1}{2},k+\frac{1}{2}} = \sum_{j,k} |A_{j+\frac{1}{2},k+\frac{1}{2}}|u_{j+\frac{1}{2},k+\frac{1}{2}}. \tag{4.3}$$

4.2. Solution procedure. The solution procedure of our adaptive mesh strategy for two-dimensional hyperbolic problems is almost the same as that of Algorithm 0 provided in section 3.3. Some details of the steps used for our 2D algorithm are given below.

Step i. Give an initial partition $\vec{z}^{[0]}_{j,k} = \big(x^{[0]}_{j,k}, y^{[0]}_{j,k}\big) := (x_{j,k}, y_{j,k})$ of the physical domain Ω_p and a uniform (fixed) partition of the logical domain Ω_c, and compute grid values $u^{[0]}_{j+\frac{1}{2},k+\frac{1}{2}}$ by cell averaging the initial data $u(x,y,0)$ over the control volume $A_{j+\frac{1}{2},k+\frac{1}{2}}$.

Step ii. For $\nu = 0, 1, 2, \ldots,$ do the following:

(a) Move grid $\vec{z}^{[\nu]}_{j,k} = \{(x^{[\nu]}_{j,k}, y^{[\nu]}_{j,k})\}$ to $\vec{z}^{[\nu+1]}_{j,k} = \{(x^{[\nu+1]}_{j,k}, y^{[\nu+1]}_{j,k})\}$ by solving $\vec{z}_\tau = (\omega \vec{z}_\xi)_\xi + (\omega \vec{z}_\eta)_\eta$ with the conventional explicit scheme. This step can be also done by solving $(\omega \vec{z}_\xi)_\xi + (\omega \vec{z}_\eta)_\eta = 0$ with the following Gauss–Seidel iteration:

$$\begin{aligned} &\alpha_{j+\frac{1}{2},k}\big(\vec{z}^{[\nu]}_{j+1,k} - \vec{z}^{[\nu+1]}_{j,k}\big) - \alpha_{j-\frac{1}{2},k}\big(\vec{z}^{[\nu+1]}_{j,k} - \vec{z}^{[\nu+1]}_{j-1,k}\big) \\ &\quad + \beta_{j,k+\frac{1}{2}}\big(\vec{z}^{[\nu]}_{j,k+1} - \vec{z}^{[\nu+1]}_{j,k}\big) - \beta_{j,k-\frac{1}{2}}\big(\vec{z}^{[\nu+1]}_{j,k} - \vec{z}^{[\nu+1]}_{j,k-1}\big) = 0 \end{aligned} \tag{4.4}$$

for $1 \le j \le J_x$ and $1 \le k \le J_y$, where

$$\begin{aligned} \alpha_{j\pm\frac{1}{2},k} &= \omega\big(u^{[\nu]}_{j\pm\frac{1}{2},k}\big) = \omega\big(\tfrac{1}{2}(u^{[\nu]}_{j\pm\frac{1}{2},k+\frac{1}{2}} + u^{[\nu]}_{j\pm\frac{1}{2},k-\frac{1}{2}})\big), \\ \beta_{j,k\pm\frac{1}{2}} &= \omega\big(u^{[\nu]}_{j,k\pm\frac{1}{2}}\big) = \omega\big(\tfrac{1}{2}(u^{[\nu]}_{j+\frac{1}{2},k\pm\frac{1}{2}} + u^{[\nu]}_{j-\frac{1}{2},k\pm\frac{1}{2}})\big). \end{aligned}$$

(b) Compute $\{u^{[\nu+1]}_{j+\frac{1}{2},k+\frac{1}{2}}\}$ on the new grid using the conservative-interpolation (4.2). The approximations for c^x, c^y, etc. are direct extensions of those defined for the 1D case.

(c) Repeat the updating procedure (a) and (b) for a fixed number of iterations (say, three or five) or until $\|\vec{z}^{[\nu+1]} - \vec{z}^{[\nu]}\| \le \epsilon$.

Step iii. Evolve the underlying PDEs using 2D high-resolution finite volume methods on the mesh $\{(x^{[\nu+1]}_{j,k}, y^{[\nu+1]}_{j,k})\}$ to obtain the numerical approximations $u^{n+1}_{j+\frac{1}{2},k+\frac{1}{2}}$ at the time level t_{n+1}.

Step iv. If $t_{n+1} \le T$, then let $u^{[0]}_{j+\frac{1}{2},k+\frac{1}{2}} := u^{n+1}_{j+\frac{1}{2},k+\frac{1}{2}}$ and $(x^{[0]}_{j,k}, y^{[0]}_{j,k}) := (x^{[\nu+1]}_{j,k}, y^{[\nu+1]}_{j,k})$, and go to Step ii.

4.3. Boundary redistribution. In many flow situations, discontinuities may initially exist in boundaries or move to boundaries at a later time. As a consequence, boundary point redistribution should be made in order to improve the quality of the solution near boundaries. A simple redistribution strategy is proposed as follows. (For convenience, our attention is restricted to the case in which the physical domain Ω_p is rectangular.) Assume that a new set of grid points $\{\tilde{x}_{j,k}, \tilde{y}_{j,k}\}$ is obtained in Ω_p by solving the moving mesh equation (4.4). Then the speeds of the internal grid point $(x_{j,k}, y_{j,k})$ are given by

$$(c^1, c^2)_{j,k} := (\widetilde{x} - x, \widetilde{y} - y)_{j,k} \qquad \text{for} \quad 1 \le j \le J_x,\ 1 \le j \le J_y.$$

We assume that the points of the left and bottom boundaries are moving with the same speed as the tangential component of the speed for the internal points adjacent to those boundary points, namely,

$$\begin{aligned}(c^1, c^2)_{0,k} &= (0, c^2_{1,k}), \qquad 1 \le j \le J_y,\\ (c^1, c^2)_{j,0} &= (c^1_{j,1}, 0), \qquad 1 \le j \le J_x.\end{aligned}$$

Thus new boundary points $(\widetilde{x}_{0,k}, \widetilde{y}_{0,k})$ and $(\widetilde{x}_{j,0}, \widetilde{y}_{j,0})$ are defined by

$$\begin{aligned}(\widetilde{x}, \widetilde{y})_{0,k} &= (x, y)_{0,k} + (c^1, c^2)_{0,k}, \qquad 1 \le k \le J_y,\\ (\widetilde{x}, \widetilde{y})_{j,0} &= (x, y)_{j,0} + (c^1, c^2)_{j,0}, \qquad 1 \le j \le J_x.\end{aligned}$$

The redistribution for other boundaries can be carried out in a similar way. Numerical experiments show that the above procedure for moving the boundary points is useful in improving the solution resolution.

5. Numerical experiments for 1D problems. In this section, we first implement our adaptive mesh methods presented in the last section for several 1D model problems. One of the main advantages of Algorithm 0 is that the solution algorithm (i.e., PDE solver) and the mesh redistribution algorithm are independent of each other. Several solution schemes, such as the MUSCL-type finite volume method (3.16)–(3.18), the second-order MUSCL-type gas-kinetic approach [38], and the second-order central scheme [27], have been employed to evolve the underlying PDEs in Step 3 of Algorithm 0. The results obtained by the three methods are in good agreement.

5.1. 1D example. Three examples will be considered in this subsection. All of them have been used by several authors to test various numerical schemes.

Example 5.1. *Burgers' equation.* This example is the inviscid Burgers' equation

$$u_t + \left(\frac{u^2}{2}\right)_x = 0, \qquad 0 \le x \le 2\pi, \tag{5.1}$$

subject to the 2π-periodic initial data

$$u(x, 0) = 0.5 + \sin(x), \quad x \in [0, 2\pi).$$

The solution propagates to the right, steepening until the critical time $t_c = 1$, at which a shock forms. Figure 5.1 shows the solutions at $t = 2$, when the shock is well developed. Also shown in Figure 5.1 is the trajectory of the grid points up to $t = 2$, obtained with $J = 30$ and $J = 50$. The ability of the adaptive mesh method to capture and follow the moving shock is clearly demonstrated in this figure. Some details in this example are the following: the monitor function used in the computation is $\omega = \sqrt{1 + 0.2|u_\xi|^2}$; the number of Gauss–Seidel iterations used is 5; the scheme for evolving Burgers' equation is a (formally) second-order MUSCL finite volume scheme (with the Lax–Friedrichs flux) together with a second-order Runge–Kutta discretization; the CFL number used is 0.3.

In Table 5.1, L^1-error and convergence rate are listed for $t = 0.9$ and $t = 0.999$. It is observed that a second-order rate of convergence can be obtained for the adaptive mesh method.

Example 5.2. *Nonconvex conservation laws.* Here we apply the adaptive mesh algorithm to the Riemann problem of a scalar hyperbolic conservation law with a nonlinear nonconvex flux:

$$u_t + f(u)_x = 0, \qquad f(u) = \frac{1}{4}(u^2 - 1)(u^2 - 4). \tag{5.2}$$

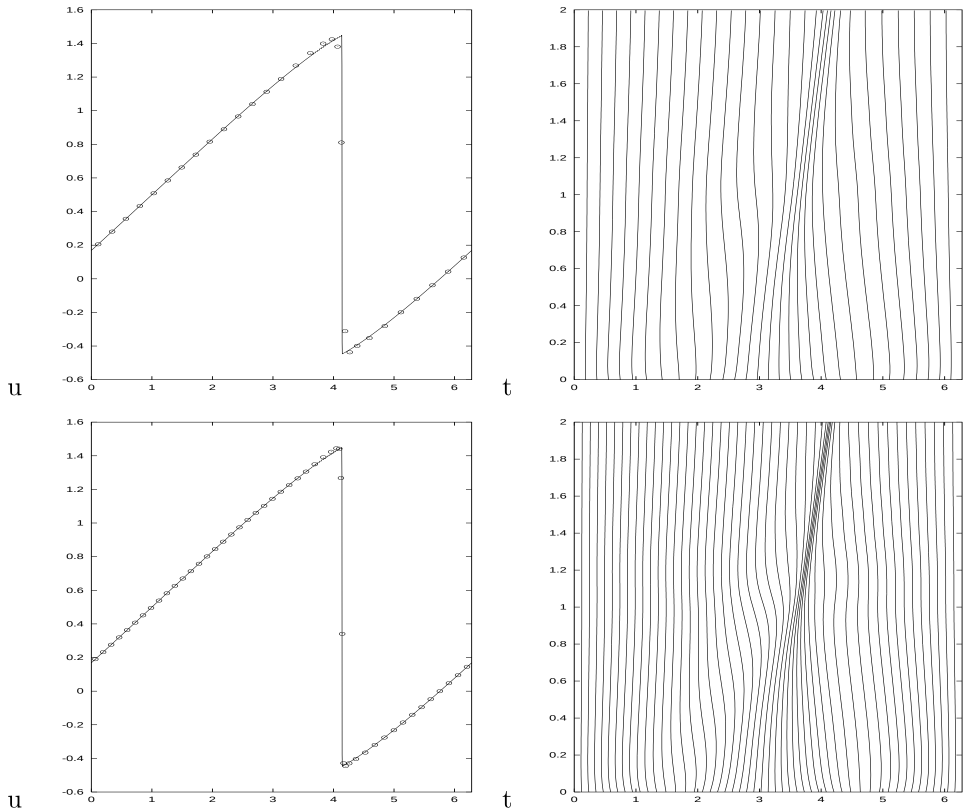


FIG. 5.1. *Example* 5.1. *Left: numerical ("o") and exact solutions (solid line) at* $t = 2$. *Right: trajectory of the mesh for* $0 \leq t \leq 2$. *Top:* $J = 30$; *and bottom:* $J = 50$.

TABLE 5.1
Example 5.1: L^1*-error and convergence order at* $t = 0.9$ *and* $t = 0.999$.

J	40	80	160	320
$t = 0.9$	4.73e-2 (–)	1.48e-2 (1.68)	3.76e-3 (1.98)	7.90e-4 (2.25)
$t = 0.999$	5.84e-2 (–)	1.85e-2 (1.67)	5.23e-3 (1.82)	1.33e-3 (1.98)

The initial data are $u(x, 0) = -2\text{sign}(x)$.

The problem was also considered in [17]. In contrast with Burgers' equation, the flux function for this problem is nonconvex, which leads to difficulties with some numerical schemes and so serves as a good test problem. The numerical solution at $t = 1.2$ is shown for an adaptive mesh in Figure 5.2, with $J = 30$ and 50. Some details in this example are the following: the monitor function used in the computation is $\omega = \sqrt{1 + |u_\xi|^2}$; the number of Gauss–Seidel iterations is 5; the scheme for evolving (5.2) is a (formally) second-order MUSCL finite volume scheme (with the Lax–Friedrichs flux) together with a second-order Runge–Kutta discretization. It is seen that the numerical solution gives sharp shock profiles.

Example 5.3. *Euler equations of gas dynamics.* In this example, we test our

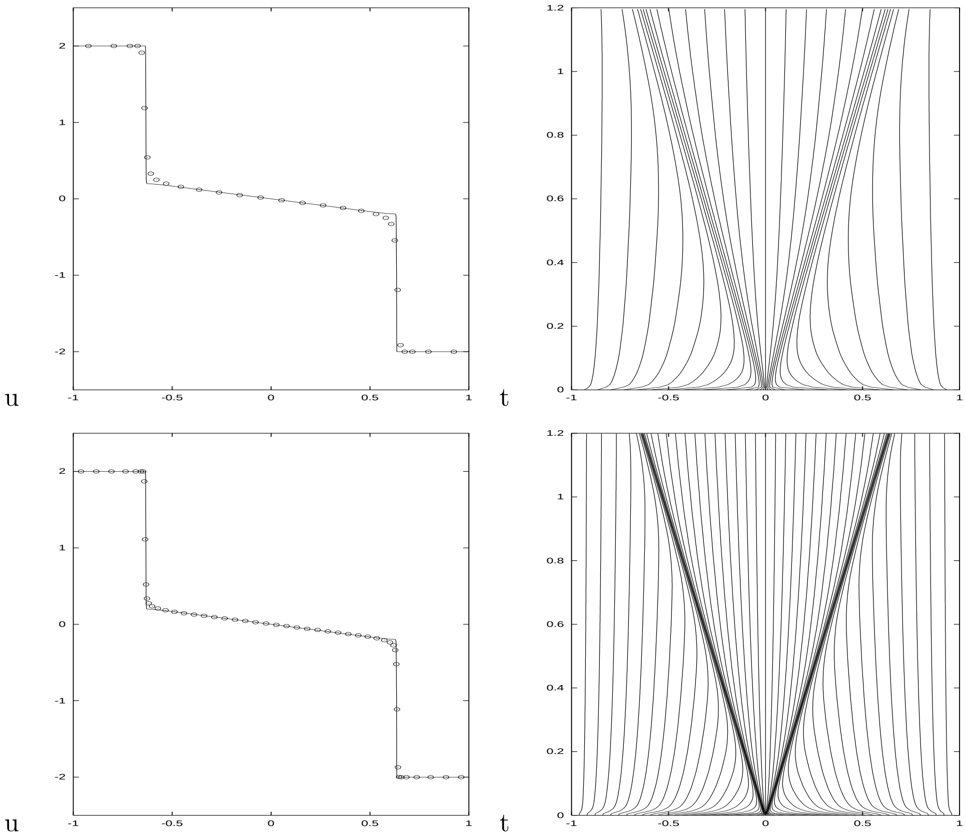


FIG. 5.2. *Example* 5.2. *Left: numerical ("o") and exact solutions (solid line) at* $t = 1.2$*. Right: trajectory of the mesh for* $0 \le t \le 1.2$*. Top:* $J = 30$*; and bottom:* $J = 50$*.*

adaptive mesh algorithm with the one-dimensional Euler equations of gas dynamics,

$$\begin{bmatrix} \rho \\ \rho u \\ E \end{bmatrix}_t + \begin{bmatrix} \rho u \\ \rho u^2 + p \\ u(E+p) \end{bmatrix}_x = 0, \tag{5.3}$$

where ρ, u, p, and E are density, velocity, pressure, and total energy, respectively. The above system is closed by the equation of state, $p = (\gamma - 1)(E - \rho u^2/2)$. The initial data are chosen as

$$(\rho, \rho u, E) = \begin{cases} (1,\ 0,\ 2.5) & \text{if } x < 0.5, \\ (0.125,\ 0,\ 0.25) & \text{if } x > 0.5. \end{cases}$$

This is a well known test problem proposed by Sod [32]. The monitor function employed for this computation is $G = \omega I$ with

$$\omega = \sqrt{1 + \alpha_1 \left(\frac{u_\xi}{\max_\xi |u_\xi|} \right)^2 + \alpha_2 \left(\frac{s_\xi}{\max_\xi |s_\xi|} \right)^2}, \tag{5.4}$$

where $s = p/\rho^\gamma$, and the parameters α_i $(i = 1, 2)$ are some nonnegative constants. The above monitor function was suggested by Stockie, Mackenzie, and Russell [33], who also discussed several other choices for the monitor function. The numerical results are obtained with $J = 100$, $\alpha_1 = 20, \alpha_2 = 100$ and are presented in Figure 5.3.

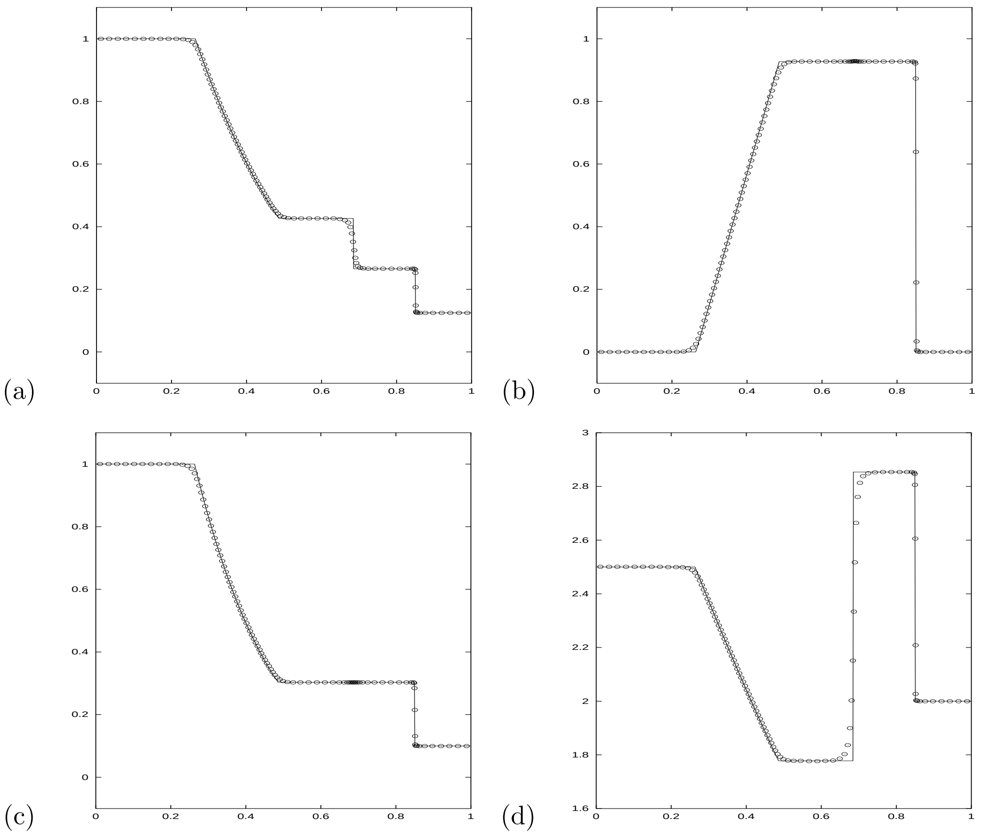


FIG. 5.3. *Example* 5.3: *adaptive mesh solution at* $t = 0.2$. (a) *density,* (b) *velocity,* (c) *pressure, and* (d) *internal energy. "o" and solid lines denote numerical and exact solutions, respectively.*

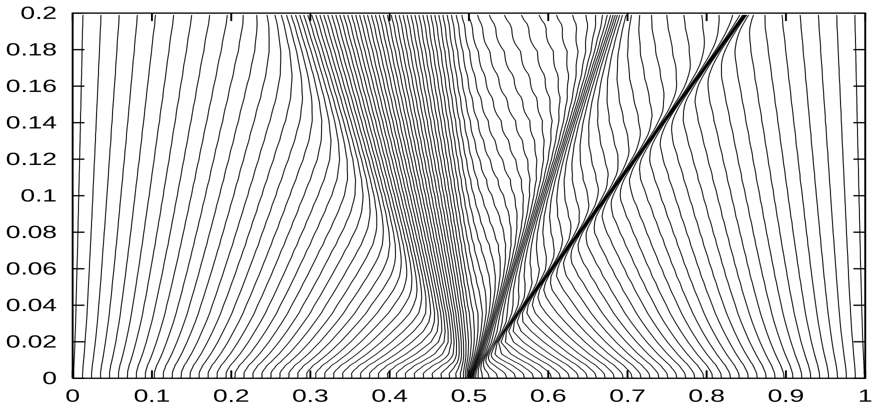


FIG. 5.4. *Example* 5.3: *trajectory of the grid points.*

It is found that the contact and shock discontinuities are well resolved, although quite a number of grid points are also moved to the rarefaction wave region. This can also be observed from the mesh contour plotted in Figure 5.4.

6. Numerical experiment for 2D problems. In this section, we will test our adaptive mesh algorithm presented in section 3.3 for some 2D problems, including 2D Riemann problems, a double-Mach reflection problem, and flow past a circular cylinder.

6.1. 2D grid generation. We begin by testing Step 2 of Algorithm 0, i.e., testing the mesh distribution part with given functions.

Example 6.1. *2D grid generation.* We consider grid generation in the physical domain $[-1,1] \times [-1,1]$ for the following functions:

$$u(x,y) = \exp(-8(4x^2 + 9y^2 - 1)^2), \tag{6.1}$$

$$u(x,y) = \exp(-100(y - x^2 + 0.5)^2), \tag{6.2}$$

$$u(x,y) = 50\exp(-2500(x^2 + y^2)), \tag{6.3}$$

$$u(x,y) = \begin{cases} 1 & \text{if } |x| \leq |y|, \\ 0 & \text{otherwise.} \end{cases} \tag{6.4}$$

The monitor function is taken as $G = \omega I$ with $\omega = \sqrt{1 + \alpha u^2}$, with $\alpha = 100$. Grid generations based on the above functions have been investigated by many authors; see e.g., [7, 21, 28]. Our results plotted in Figure 6.1 can be favorably compared with published results. Our results indicate that Step 2 of Algorithm 0 for two dimensions performs well for functions with large gradients or singularities.

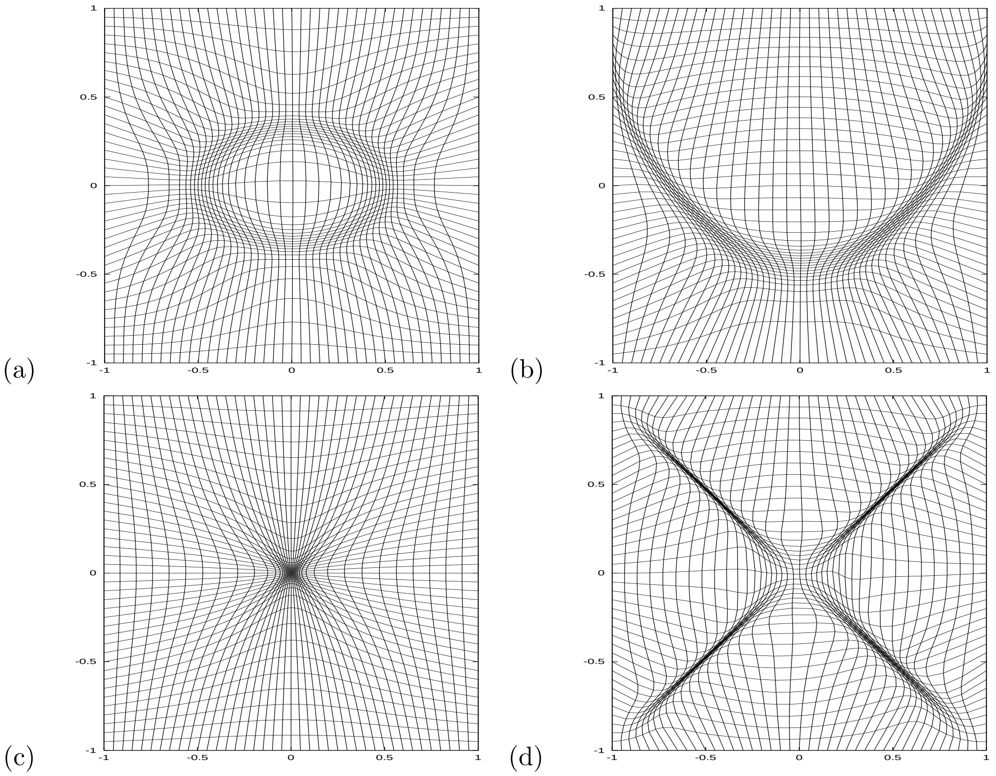


Fig. 6.1. *Example* 6.1: *the adaptive meshes for* (a) *function* (6.1), (b) *function* (6.2), (c) *function* (6.3), *and* (d) *function* (6.4).

6.2. 2D examples for Euler equations of gas dynamics. In this subsection we consider some well known test examples in two dimensions, including three Riemann problems and a double-Mach reflection problem.

Example 6.2. 2*D Riemann problem* I: *Shock waves.* Two-dimensional Euler equations of gas dynamics can be written as

$$\begin{bmatrix} \rho \\ \rho u \\ \rho v \\ E \end{bmatrix}_t + \begin{bmatrix} \rho u \\ \rho u^2 + p \\ \rho uv \\ u(E+p) \end{bmatrix}_x + \begin{bmatrix} \rho v \\ \rho uv \\ \rho v^2 + p \\ v(E+p) \end{bmatrix}_y = 0, \tag{6.5}$$

where ρ, (u, v), p, and E are the density, velocity, pressure, and total energy, respectively. For an ideal gas, the equation of state, $p = (\gamma - 1)(E - \rho(u^2 + v^2)/2)$, is provided. The initial data are chosen as

$$(\rho, u, v, p) = \begin{cases} (1.1,\ 0.0,\ 0.0,\ 1.1) & \text{if } x > 0.5, \quad y > 0.5, \\ (0.5065,\ 0.8939,\ 0.0,\ 0.35) & \text{if } x < 0.5, \quad y > 0.5, \\ (1.1,\ 0.8939,\ 0.8939,\ 1.1) & \text{if } x < 0.5, \quad y < 0.5, \\ (0.5065,\ 0.0,\ 0.8939,\ 0.35) & \text{if } x > 0.5, \quad y < 0.5, \end{cases}$$

which corresponds to the case of left forward shock, right backward shock, upper backward shock, and lower forward shock. We refer the readers to [19, 30] for details.

In [19], Lax and Liu computed 2D Riemann problems with various initial data using positive schemes. The problem considered here corresponds to Configuration 4 discussed in their paper. We use our adaptive mesh algorithm with $(J_x, J_y) = (50, 50)$ and $(J_x, J_y) = (100, 100)$ to compute this Riemann problem and display the mesh and density at $t = 0.25$ in Figure 6.2. It is found that our results with $J_x = J_y = 100$ give sharper shock resolution than that of the positive schemes with $(J_x, J_y) = (400, 400)$ (see [19, p. 333]). The monitor function used in this computation is $G = \omega I$, with $\omega = \sqrt{1 + 2(\rho_\xi^2 + \rho_\eta^2)x}$.

Example 6.3. 2*D Riemann problem* II: *Contact discontinuities.* We reconsider Configurations 6 and 7 in Lax and Liu's paper [19], whose solutions contain contact discontinuities. The first configuration has initial data

$$(\rho, u, v, p) = \begin{cases} (1,\ 0.75,\ -0.5,\ 1) & \text{if } x > 0.5, \quad y > 0.5, \\ (2,\ 0.75,\ 0.5,\ 1) & \text{if } x < 0.5, \quad y > 0.5, \\ (1,\ -0.75,\ 0.5,\ 1) & \text{if } x < 0.5, \quad y < 0.5, \\ (3,\ -0.75,\ -0.5,\ 1) & \text{if } x > 0.5, \quad y < 0.5, \end{cases}$$

and the second configuration has initial data

$$(\rho, u, v, p) = \begin{cases} (1,\ 0.1,\ 0.1,\ 1) & \text{if } x > 0.5, \quad y > 0.5, \\ (0.5197,\ -0.6259,\ 0.1,\ 0.4) & \text{if } x < 0.5, \quad y > 0.5, \\ (0.8,\ 0.1,\ 0.1,\ 0.4) & \text{if } x < 0.5, \quad y < 0.5, \\ (0.5197,\ 0.1,\ -0.6259,\ 0.4) & \text{if } x > 0.5, \quad y < 0.5. \end{cases}$$

The adaptive mesh results for Configuration 6 at $t = 0.3$ and for Configuration 7 at $t = 0.25$ are displayed in Figure 6.3. The number of grid points are $(J_x, J_y) = (100, 100)$. It can be observed that the adaptive mesh results with $J_x = J_y = 100$ for Configuration 7 are comparable with those obtained by using the positive schemes with $J_x = J_y = 400$ (see [19, p. 334]). However, this seems not to be the case for the results for Configuration 6. Although the resolution can be improved by using more

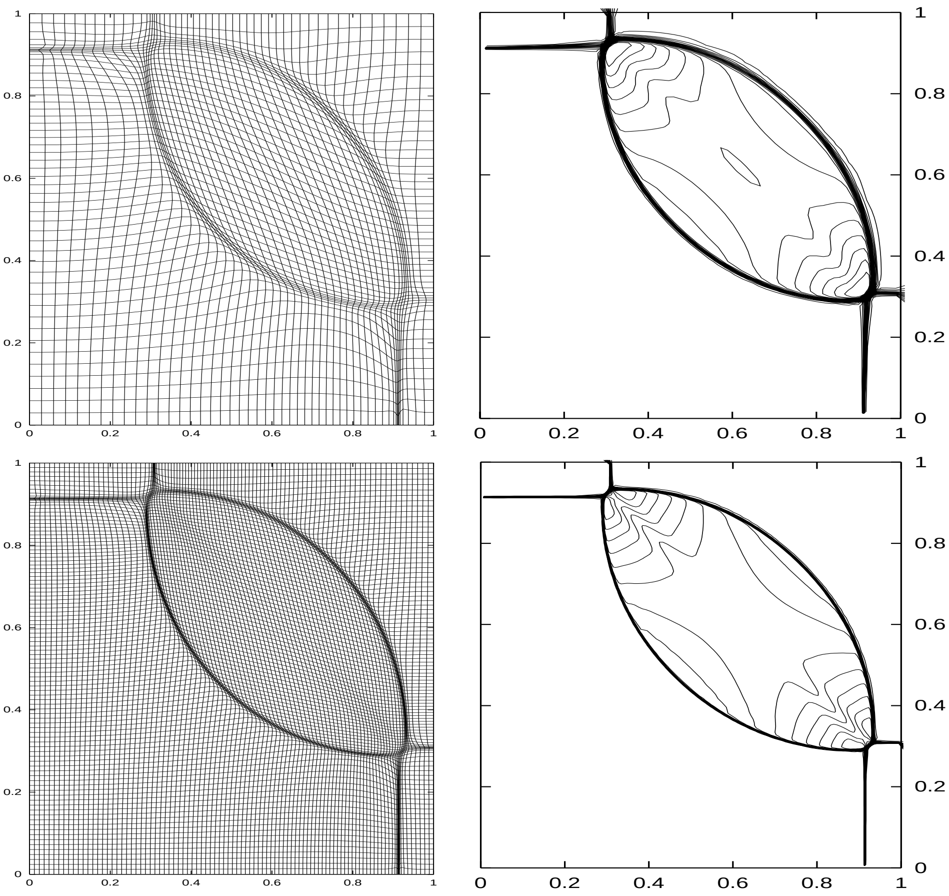


FIG. 6.2. *Example* 6.2. *The contours of the mesh (left) and the density (right). Top:* $J_x = J_y = 50$*; and bottom:* $J_x = J_y = 100$*.* 30 *equally spaced contour lines are used for the density.*

grid points, it is quite clear from this computation and Example 6.2 that the treatment of contact discontinuities is less effective than the treatment of shocks. It is expected that the monitor functions used in this paper are suitable for shock discontinuities but may be less appropriate for contact discontinuities. Therefore, it requires further investigation to obtain more effective monitor functions for contact discontinuities.

For this computation, the monitor function can be chosen as $G = \omega I$, with $\omega = \sqrt{1 + \alpha(\rho_\xi^2 + \rho_\eta^2)}$. It is found that in both cases if the parameter α is chosen in the range $[0.1, 1]$, then the efficiency and effectiveness of the adaptive mesh approach seem satisfactory. The results in Figure 6.3 are obtained using $\alpha = 0.1$ (for Configuration 6) and $\alpha = 0.9$ (for Configuration 7).

Example 6.4. *The double-Mach reflection problem.* This problem was studied extensively in Woodward and Colella [37] and later by many others. We use exactly the same setup as in [37], i.e., the same initial and boundary conditions and same solution domain $\Omega_p = [0, 4] \times [0, 1]$. Initially a right-moving Mach 10 shock is positioned at $x = \frac{1}{6}$, $y = 0$ and makes a 60° angle with the x-axis. More precisely, the initial data are

$$U = \begin{cases} U_L & \text{for} \quad y \geq h(x, 0), \\ U_R & \text{otherwise}, \end{cases}$$

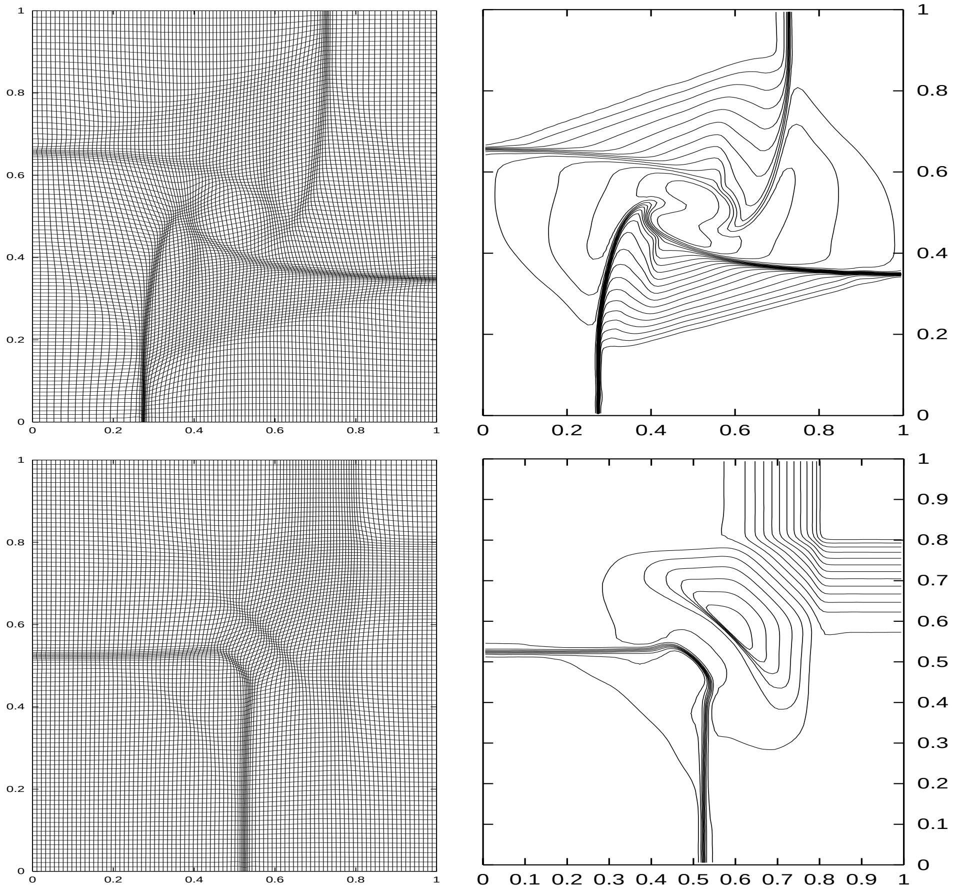


FIG. 6.3. *Adaptive mesh results for Example* 6.3 *with* 100×100 *grid points. Top: Configuration* 6; *bottom: Configuration* 7. *Left: the adaptive mesh; Right: density.* 19 *equally spaced contour lines are used for the density.*

where the state on the left, the state on the right, and the shock strength are, respectively,

$$U_L = (8,\ 57.1597,\ -33.0012,\ 563.544)^T,$$
$$U_R = (1.4,\ 0.0\ 0.0,\ 2.5)^T, \qquad h(x,t) = \sqrt{3}(x - 1/6) - 20t.$$

As in [37], only the results in $[0,3] \times [0,1]$ are displayed. In Figure 6.4, the adaptive meshes with $(J_x, J_y) = (80, 20)$, $(160, 40)$, and $(320, 80)$ are displayed, while the corresponding contours of density are displayed in Figure 6.5. By comparing the density plots, it is found that the adaptive computation results with $(J_x, J_y) = (320, 80)$ have similar resolution to the results obtained by the second-order discontinuous Galerkin method with $(J_x, J_y) = (960, 240)$ (see [10, p. 214]) and by the second-order central scheme with $(J_x, J_y) = (960, 240)$ (see [10, p. 67]). Moreover, the adaptive results with $(J_x, J_y) = (160, 40)$ have slightly better resolution than the results of fifth-order weighted ENO and the fourth-order ENO with 480×119 grids [10, p. 406]. Of course, this is not too surprising, since these published results are computed using uniform meshes.

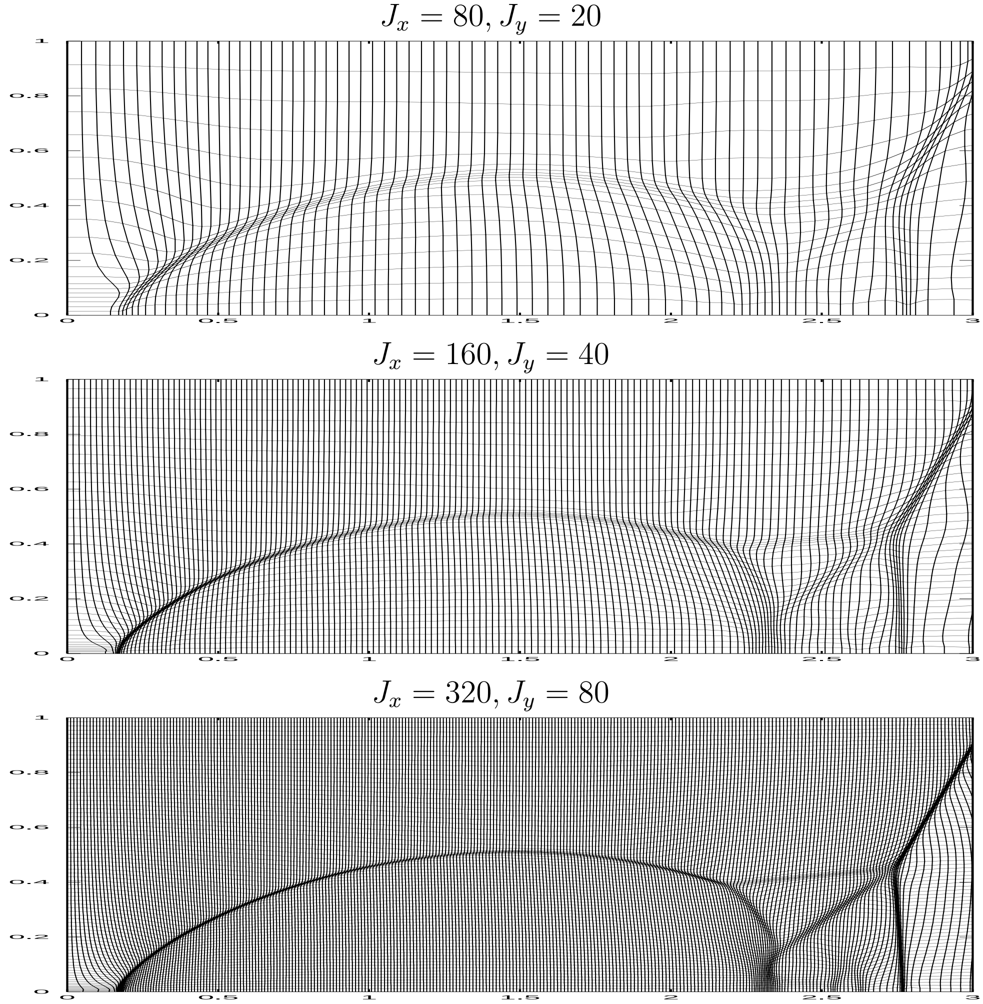
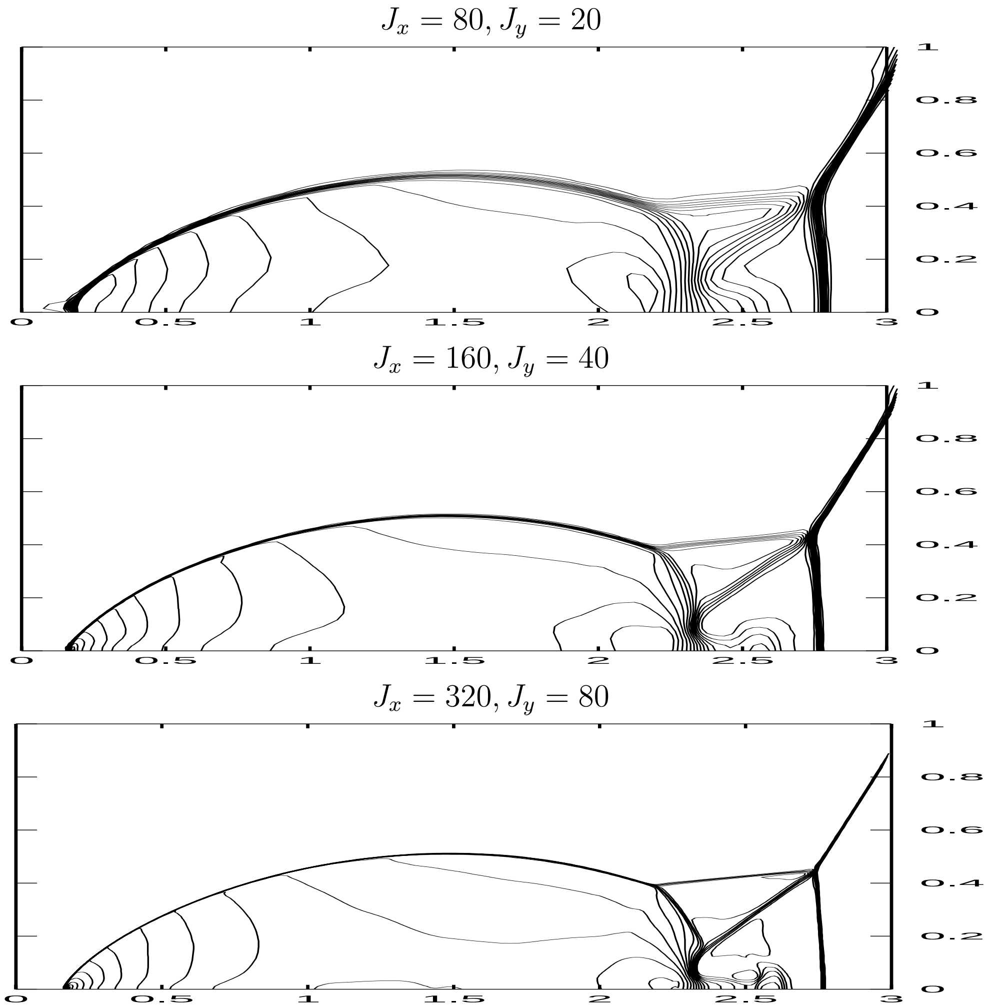


FIG. 6.4. *2D double-Mach reflection at* $t = 0.2$*: the contours of meshes. From top to bottom:* $(J_x, J_y) = (80, 20)$, $(160, 40)$, *and* $(320, 80)$.

We also show a *blow up* portion around the double-Mach region in Figure 6.6. In our computations, we used 640×160 and 960×240 grid points. The corresponding mesh contours in the blow up region are shown in Figure 6.7. The smallest Δx and Δy in these runs are listed in Tables 6.1 and 6.2. It is seen that ratios between the largest and smallest mesh sizes in the adaptive grids are quite large (≥ 20), which is a desired feature of the adaptive grid methods. The fine details of the complicated structure in this region were previously obtained by Cockburn and Shu [9], who used high-order discontinuous Galerkin (RKDG) methods with 960×240 and 1920×480 grid points. Although the moving mesh algorithm gives a good resolution in this blow up portion, it is observed that, even with approximately the same number of grid points (960×240), the third-order RKDG results [10, p. 216] have a slightly better resolution of the complex structure. The monitor function used for this example is taken as $G = \omega I$, with $\omega = \sqrt{1 + 0.125(\rho_\xi^2 + \rho_\eta^2)}$.



$J_x = 80, J_y = 20$

$J_x = 160, J_y = 40$

$J_x = 320, J_y = 80$

FIG. 6.5. $2D$ *double-Mach reflection at* $t = 0.2$: *the contours of density. From top to bottom:* $(J_x, J_y) = (80, 20)$, $(160, 40)$, *and* $(320, 80)$. 30 *equally spaced contour lines are used.*

6.3. Example of a nonconvex physical domain. So far, the numerical examples in two dimensions have been restricted to rectangular domains. In the final example, we consider a test problem whose domain is not even convex. In this case, as long as the domain can be smoothly transformed to a rectangle, the adaptive mesh algorithm can be handily applied.

Example 6.5. *Flow past a cylinder.* This example is concerned with the supersonic flow past a cylinder with unit radius, which is positioned at the origin on an x-y plane. The problem is initialized by a Mach 3 *free-stream* moving toward the cylinder from the left. Since the physical domain Ω_p is nonconvex, we first transform Ω_p to a square domain $\widehat{\Omega}_c = [0, 1] \times [0, 1]$ by using the following mapping:

$$\begin{aligned} x &= -(R_x - (R_x - 1)\widehat{x}) \cos(\theta(2\widehat{y} - 1)), \\ y &= (R_y - (R_y - 1)\widehat{x}) \sin(\theta(2\widehat{y} - 1)), \end{aligned} \tag{6.6}$$

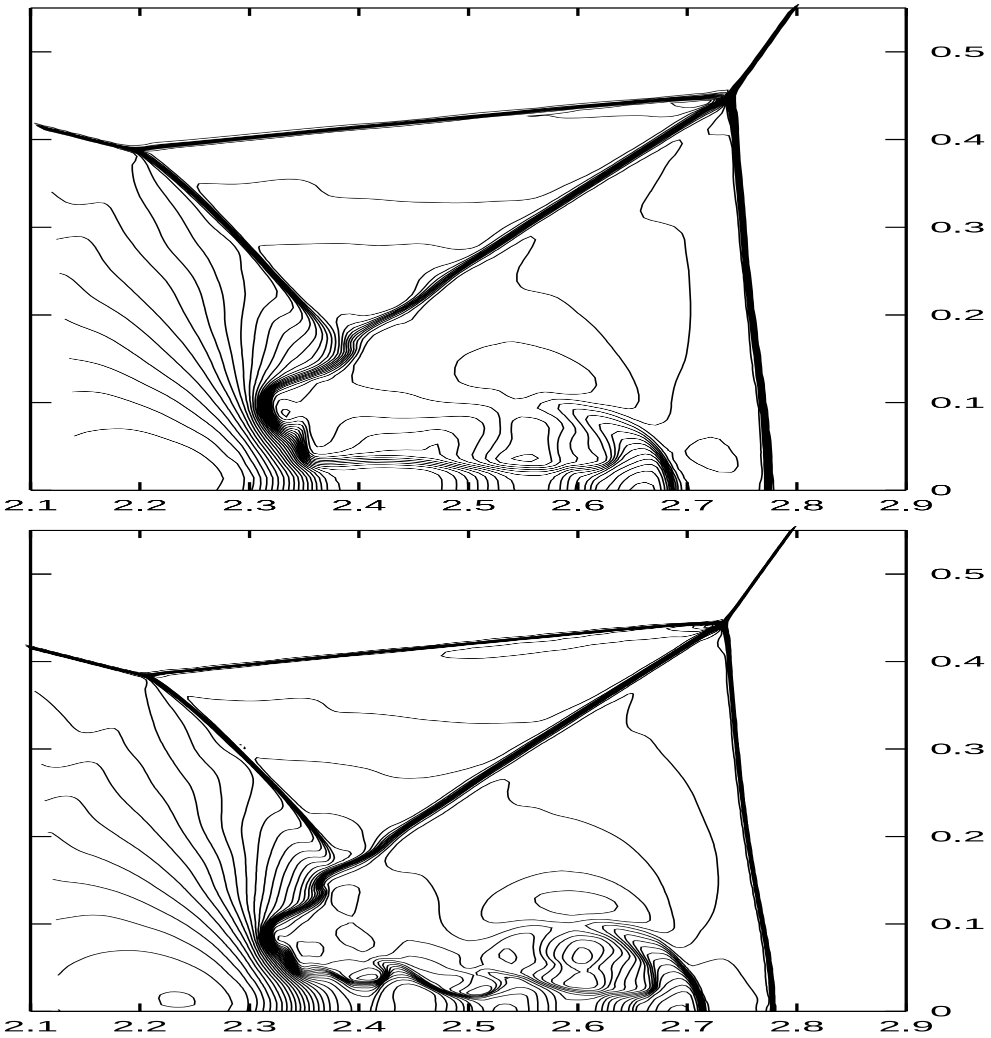


FIG. 6.6. *Double-Mach reflection problem: density ρ in blowup region around the double-Mach stems. Top: $(J_x, J_y) = (640, 160)$; bottom: $(J_x, J_y) = (960, 240)$. 45 equally spaced contour lines are used.*

with $R_x = 3$, $R_y = 5$, and $\theta = 5\pi/12$. A reflective boundary condition is imposed at the surface of the cylinder, i.e., $\hat{x} = 1$; inflow boundary condition is applied at $\hat{x} = 0$; and outflow boundary conditions are applied at $\hat{y} = 0$ and 1. We then solve the problem in $\widehat{\Omega}_c$ using the adaptive mesh algorithm, with a logical domain Ω_c as before. This procedure will lead to numerical solution in $\widehat{\Omega}_c$, and the mapping (6.6) finally gives the numerical approximation in the physical domain Ω_p.

We present an illustration of the mesh in the physical space and the pressure contour in Figure 6.8, by using 30×40, 60×80, and 120×160 grid points. The monitor function used for this example is taken as $G = \omega I$ with $\omega = \sqrt{1 + 0.125(\rho_\xi^2 + \rho_\eta^2)}$. As can be seen from these figures, the advantages of the adaptive mesh methods are quite obvious. The shock location computed by our adaptive mesh algorithm is approximately 0.703 (the distance between the shock curve and the surface of the cylinder), which is in good agreement with the experimental results reported in [4].

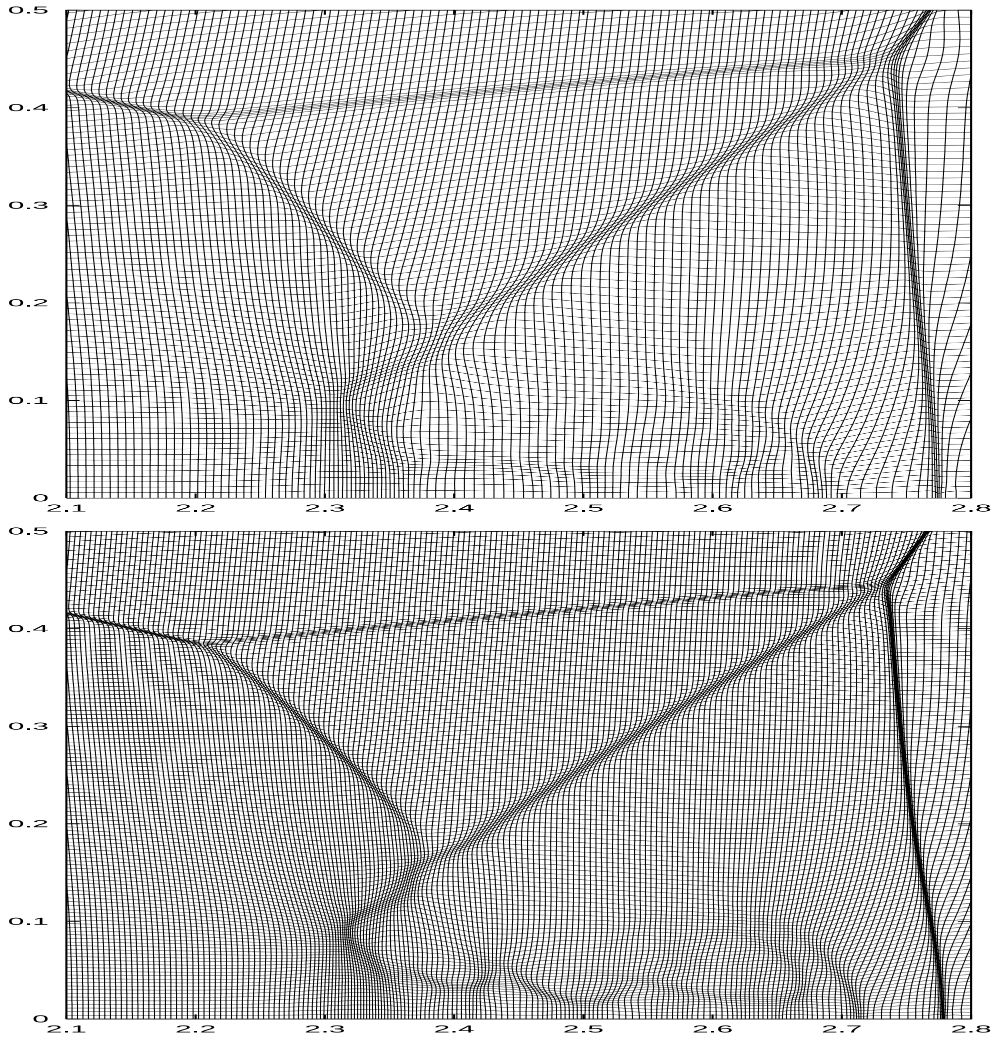


FIG. 6.7. *Double-Mach reflection problem: the adaptive mesh in blowup region around the double-Mach stems. Top:* $(J_x, J_y) = (640, 160)$*; bottom:* $(J_x, J_y) = (960, 240)$.

TABLE 6.1
The smallest mesh size for the double-Mach reflection problem with 640×160 *grid points.*

	$\min\{\Delta x\}$	$\max\{\Delta x\}$	$\max\{\Delta x\}/\min\{\Delta x\}$
Δx	6.5e-04	2.0e-02	30.8
Δy	4.8e-04	1.0e-02	20.8
$\sqrt{\Delta x^2 + \Delta y^2}$	8.2e-04	2.0e-02	24.4

Acknowledgments. The authors thank Professors Chi-Wang Shu and Eitan Tadmor for numerous discussions during the preparation of this work. We also thank the referees for many helpful suggestions.

Table 6.2

The smallest mesh sizes for the double-Mach reflection problem with 960×240 *grid points.*

	$\min\{\Delta x\}$	$\max\{\Delta x\}$	$\max\{\Delta x\}/\min\{\Delta x\}$
Δx	4.3e-04	1.1e-02	25.6
Δy	3.1e-04	5.9e-03	19.0
$\sqrt{\Delta x^2 + \Delta y^2}$	5.3e-04	1.2e-02	22.6

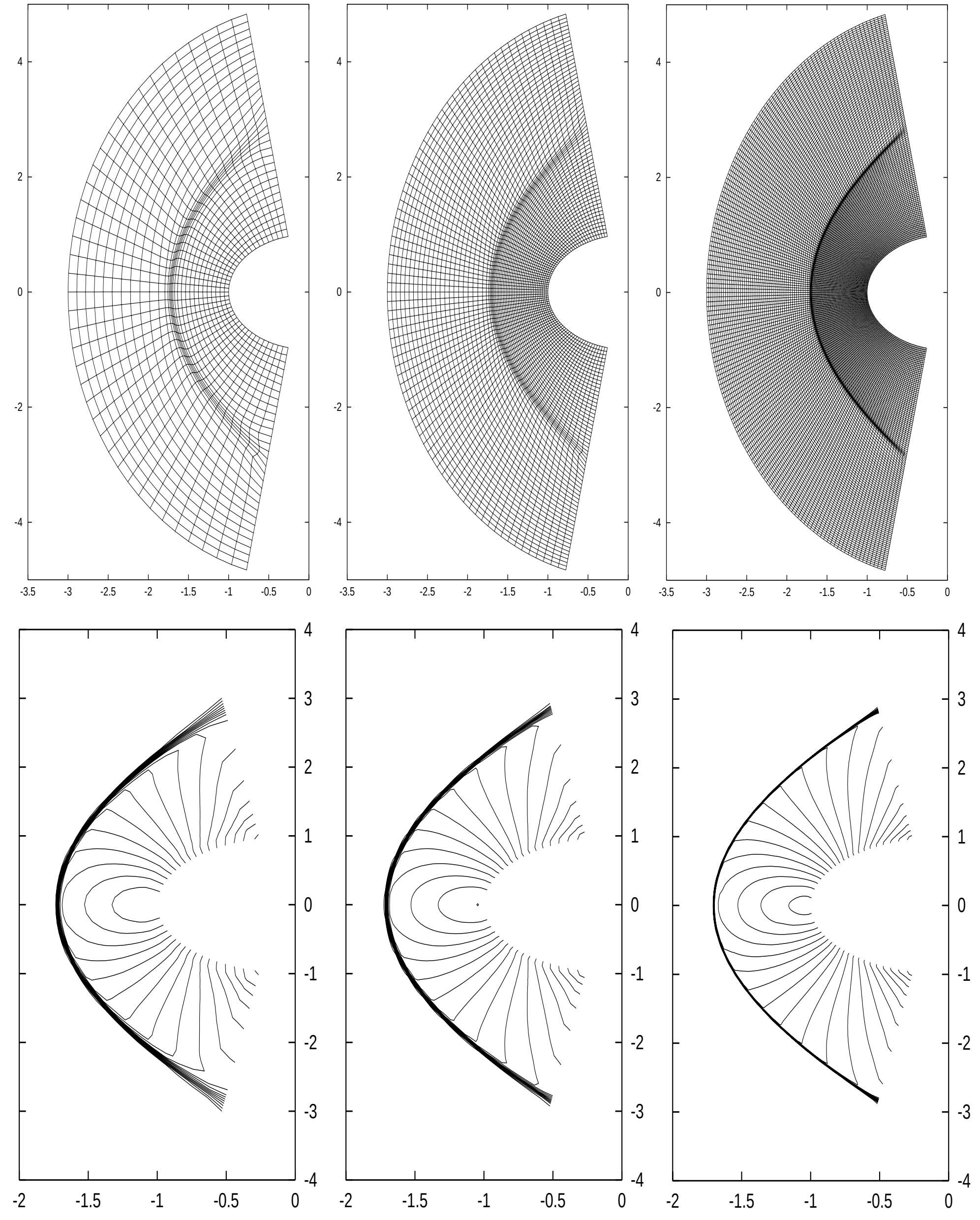


Fig. 6.8. *Example 6.5. Top: adaptive mesh; bottom: pressure. From left to right:* 30×40, 60×80, *and* 120×160 *grid points.* 20 *equally spaced contour lines are used for the pressure.*

REFERENCES

[1] B. N. Azarenok, *Variational barrier method of adaptive grid generation in hyperbolic problems of gas dynamics*, SIAM J. Numer. Anal., 40 (2002), pp. 651–682.

[2] B. N. Azarenok and S. A. Ivanenko, *Application of adaptive grids in numerical analysis of time-dependent problems in gas dynamics*, Comput. Math. Math. Phys., 40 (2000), pp. 1330–1349.

[3] B. N. Azarenok, S. A. Ivanenko, and T. Tang, *Adaptive mesh redistribution method based on Godunov's scheme*, Comm. Math. Sci., 1 (2003), pp. 152–179.

[4] O. M. Belotserkovskii, *Computation of flows around the circular cylinder with detached shock waves*, Comput. Math., 3 (1958), pp. 149–185 (in Russian).

[5] J. U. Brackbill, *An adaptive grid with directional control*, J. Comput. Phys., 108 (1993), pp. 38–50.

[6] J. U. Brackbill and J. S. Saltzman, *Adaptive zoning for singular problems in two dimensions*, J. Comput. Phys., 46 (1982), pp. 342–368.

[7] W. M. Cao, W. Z. Huang, and R. D. Russell, *An r-adaptive finite element method based upon moving mesh PDEs*, J. Comput. Phys., 149 (1999), pp. 221–244.

[8] H. D. Ceniceros and T. Y. Hou, *An efficient dynamically adaptive mesh for potentially singular solutions*, J. Comput. Phys., 172 (2001), pp. 609–639.

[9] B. Cockburn and C.-W. Shu, *The Runge-Kutta discontinuous Galerkin method for conservation laws* V: *Multidimensional systems*, J. Comput. Phys., 141 (1998), pp. 199–224.

[10] B. Cockburn, C. Johnson, C.-W. Shu, and E. Tadmor, *Advanced Numerical Approximation of Nonlinear Hyperbolic Equations*, papers from the C.I.M.E. Summer School in Centraro, Italy, 1997, A. Quarteroni, ed., Lecture Notes in Math. 1697, Springer-Verlag, Berlin, 1998.

[11] S. F. Davis and J. E. Flaherty, *An adaptive finite element method for initial-boundary value problems for partial differential equations*, SIAM J. Sci. Statist. Comput., 3 (1982), pp. 6–27.

[12] A. S. Dvinsky, *Adaptive grid generation from harmonic maps on Riemannian manifolds*, J. Comput. Phys., 95 (1991), pp. 450–476.

[13] R. Fazio and R. LeVeque, *Moving-mesh methods for one-dimensional hyperbolic problems using CLAWPACK*, Comp. Math. Appl., to appear.

[14] A. Harten and J. M. Hyman, *Self-adjusting grid methods for one-dimensional hyperbolic conservation laws*, J. Comput. Phys., 50 (1983), pp. 235–269.

[15] K. H. Karlsen, K.-A. Lie, and N. H. Risebro, *A front tracking method for conservation laws with boundary conditions*, in Hyperboilic Problems: Theory, Numerics, Applications, M. Fey and R. Jeltsch, eds., Internat. Ser. Numer. Math. 129, Birkhäuser Verlag, 1999, pp. 493–502.

[16] S. Jin and Z. P. Xin, *The relaxation schemes for systems of conservation laws in arbitrary space dimensions*, Comm. Pure Appl. Math., 48 (1995), pp. 235–276.

[17] A. Kurganov and E. Tadmor, *New high-resolution central schemes for nonlinear conservation laws and convective-diffusion equations*, J. Comput. Phys., 160 (2000), pp. 214–282.

[18] P. D. Lax and X. D. Liu, *Positive schemes for solving multi-dimensional hyperbolic systems of conservation laws*, J. Comput. Fluid Dynam., 5 (1996), pp. 133–156.

[19] P. D. Lax and X.-D. Liu, *Solutions of two-dimensional Riemann problems of gas dynamics by positive schemes*, SIAM J. Sci. Comput., 19 (1998), pp. 319–340.

[20] R. J. LeVeque, *High-resolution finite volume methods on arbitrary grids via wave propagation*, J. Comput. Phys., 78 (1988), pp. 36–63.

[21] R. Li, T. Tang, and P. W. Zhang, *Moving mesh methods in multiple dimensions based on harmonic maps*, J. Comput. Phys., 170 (2001), pp. 562–588.

[22] R. Li, T. Tang, and P. W. Zhang, *A moving mesh finite element algorithm for singular problems in two and three space dimensions*, J. Comput. Phys., 177 (2002), pp. 365–393.

[23] S. Li and L. Petzold, *Moving mesh methods with upwinding schemes for time-dependent PDEs*, J. Comput. Phys., 131 (1997), pp. 368–377.

[24] F. Liu, S. Ji, and G. Liao, *An adaptive grid method and its application to steady Euler flow calculations*, SIAM J. Sci. Comput., 20 (1998), pp. 811–825.

[25] K. Miller and R. N. Miller, *Moving finite elements.* I, SIAM J. Numer Anal., 18 (1981), pp. 1019–1032.

[26] R. Natanlini, *Convergence to equilibrium for the relaxation approximations of conservation laws*, Comm. Pure Appl. Math., 49 (1996), pp. 795–823.

[27] H. Nessyahu and E. Tadmor, *Non-oscillatory central differencing for hyperbolic conservation laws*, J. Comput. Phys., 87 (1990), pp. 408–463.

[28] W. Q. Ren and X. P. Wang, *An iterative grid redistribution method for singular problems in*

multiple dimensions, J. Comput. Phys., 159 (2000), pp. 246–273.

[29] K. Saleri and S. Steinberg, *Flux-corrected transport in a moving grid*, J. Comput. Phys., 111 (1994), pp. 24–32.

[30] C. W. Schulz-Rinne, J. P. Collins, and H. M. Glaz, *Numerical solution of the Riemann problem for two-dimensional gas dynamics*, SIAM J. Sci. Comput., 14 (1993), pp. 1394–1414.

[31] C.-W. Shu and S. Osher, *Efficient implementation of essentially non-oscillatory shock-capturing schemes* II, J. Comput. Phys., 83 (1989), pp. 32–78.

[32] G. A. Sod, *A survey of finite difference methods for systems of nonlinear hyperbolic conservation laws*, J. Comput. Phys., 27 (1978), pp. 1–31.

[33] J. M. Stockie, J. A. Mackenzie, and R. D. Russell, *A moving mesh method for one-dimensional hyperbolic conservation laws*, SIAM J. Sci. Comput., 22 (2001), pp. 1791–1813.

[34] E. Tadmor and T. Tang, *Pointwise error estimates for relaxation approximations to conservation laws*, SIAM J. Math. Anal., 32 (2000), pp. 870–886.

[35] H. Z. Tang and H.-M. Wu, *Kinetic flux vector splitting for radiation hydrodynamical equations*, Comput. & Fluids, 29 (2000), pp. 917–933.

[36] A. Winslow, *Numerical solution of the quasi-linear Poisson equation*, J. Comput. Phys., 1 (1967), pp. 149–172.

[37] P. Woodward and P. Colella, *The numerical simulation of two dimensional fluid flow with strong shocks*, J. Comput. Phys., 54 (1984), pp. 115–173.

[38] K. Xu, *Gas-Kinetic Schemes for Unsteady Compressible Flow Simulations*, lecture notes from the von Karman Institute for Fluid Dynamics Lecture Series 1998-03, Rhode-Saint-Genèse, Belgium, 1998.

[39] H. Yang, *A local extrapolation method for hyperbolic conservation laws* I: *The ENO underlying schemes*, J. Sci. Comput., 15 (2000), pp. 231–264.

SIAM J. NUMER. ANAL.
Vol. 41, No. 2, pp. 516–538

ON THE CONVERGENCE OF DIFFERENCE SCHEMES FOR HYPERBOLIC PROBLEMS WITH CONCENTRATED DATA*

BOSKO S. JOVANOVIĆ[†] AND LUBIN G. VULKOV[‡]

Abstract. Hyperbolic equations with unbounded coefficients and even generalized functions (in particular, Dirac-delta functions) occur both naturally and artificially and must be treated in numerical schemes. An abstract operator method is proposed for studying these equations. For finite difference schemes approximating several one-dimensional initial-boundary value problems convergence rate estimates in special discrete energetic Sobolev's norms, compatible with the smoothness of the solutions, are obtained.

Key words. concentrated mass, difference scheme, Sobolev spaces, generalized solution, rate of convergence

AMS subject classification. 65M15

PII. S0036142900382983

1. Introduction.

The study of properties of numerical schemes for discretizing of hyperbolic equations is of great interest in applied mathematics. Numerous works have been concerned with classical schemes for these equations in homogeneous whole space (especially studies of stability and dispersion relations). Other works have studied the approximations of waves at plane boundaries or interfaces. Here we analyze the convergence of difference schemes for hyperbolic equations in the case when the coefficients are discontinuous or change sharply; an important application is in the situation where these are unbounded functions and even generalized functions (in particular, Dirac-delta functions).

The solutions of hyperbolic problems with nonsmooth or discontinuous data (coefficients, initial and boundary functions) are weak solutions, i.e., functions from Sobolev space [20, 21]. Since such solutions do not possess continuous partial derivatives, one cannot use the classical Taylor technique to establish the convergence of discrete approximations. The role of Taylor's formula is often taken by the Bramble–Hilbert lemma and its generalizations [4, 12, 17].

The theory of convergence rate estimates *compatible* with the smoothness of the differential problem solutions was formulated in the 80s by Samarskii, Lazarov, and Makarov (see, e.g., [12]). This concept has been systematically developed in the monographs [4] and [17], where a key role in the analysis is played by the Bramble–Hilbert lemma. Further, results for hyperbolic problems have been obtained in [4, 5, 6]. In all these works, problems with variable coefficients including discontinuous coefficients have been considered.

*Received by the editors December 27, 2000; accepted for publication (in revised form) September 13, 2002; published electronically April 23, 2003. The U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government purposes. Copyright is owned by SIAM to the extent not limited by these rights.

http://www.siam.org/journals/sinum/41-2/38298.html

[†] University of Belgrade, Faculty of Mathematics, 11000 Belgrade, Yugoslavia (bosko@matf.bg.ac.yu). The research of this author was supported by the MSTD of the Republic of Serbia, grant 1840.

[‡] Department of Mathematics, University of Rousse, 7017 Rousse, Bulgaria (vulkov@ami.ru.acad.bg).

The basic mechanical system corresponding to the hyperbolic problem considered in the present paper is that of forced oscillations of a string with concentrated mass at the ends or in interior points of the string. Our aim is to treat these problems as a second order abstract evolution equation (2.1) with self-adjoint positive linear operators A, B defined in a Hilbert space H, and then to use energy methods from the theory of operators on a Hilbert space. Discrete analysis of appropriate subspaces of the Sobolev spaces are used with alternative inner products that are equivalent to the Sobolev inner product and yet that allow the discrete operators to be self-adjoint on the space involved. The second important idea of our method consists of constructing special integral representations of the truncation error of the difference schemes. The use of the Bramble–Hilbert lemma for truncation error estimates involves all partial derivatives of the solution up to the corresponding order, although only some of them could have discontinuity on the interface. Therefore, the present approach gives more precise results than does the Bramble–Hilbert lemma.

The remainder of this paper is organized as follows. Energy estimates for the solutions of an abstract Cauchy problem for a second order evolution equation and for an operator-difference scheme can be found in the next section. Section 3 is devoted to the derivation of convergence rate estimates in special discrete Sobolev norms of difference scheme approximations to wave equations with discontinuous coefficients and dynamical conditions of conjugation, i.e., in which the time derivative of the solution is involved. In sections 4 and 5 we treat hyperbolic second order equations with dynamical boundary conditions and elliptic equations with dynamical conditions of conjugation. Analogous results for parabolic equations are obtained in [9, 10]. Results concerning finite difference schemes on uniform meshes for the equation of a vibrating string with concentrated mass are reported in [11].

In this article we consider one-dimensional problems. The approach presented in this paper will be applied to similar two-dimensional problems in a forthcoming paper. The method proposed here can be developed for evolution problems modeling vibrations of beam-mass systems, i.e., problems of the type (2.1) in which A is a fourth order elliptic operator. Also, the method works for interface problems in which the Dirac-delta functions appear in the lowest coefficients [7], or on the right-hand side of the equation.

Convergence to classical solutions for parabolic and hyperbolic equations with dynamical boundary conditions or dynamical conditions of conjugations are studied in [1, 2, 3, 21].

2. Preliminary results.

Let H be a real separable Hilbert space with scalar product $(\cdot,\cdot)$ and norm $\|\cdot\|$, and let S be an unbounded self-adjoint positively defined linear operator on a domain $D(S)$ which is dense in H. The bilinear form $(u,v)_S = (Su,v)$, $u,v \in D(S)$, satisfies the axioms of the scalar product. Let the Hilbert space $H_S \subset H$ be the completion of $D(S)$ in the norm $\|u\|_S = (u,u)_S^{1/2}$. Then the scalar product (u,v) is continuously extended on $H_S^* \times H_S$, where H_S^* is a space which is dual to H_S, and the operator S is extended to the map $S : H_S \to H_S^*$. There exists the unbounded self-adjoint positive linear operator $S^{1/2}$ (see [13, 15]) and $D(S^{1/2}) = H_S$, $(u,v)_S = (Su,v) = (S^{1/2}u, S^{1/2}v)$. We also introduce the space $L_2(a,b;H)$ and the function $u = u(t)$ mapping the segment $(a,b) \subset R$ in H with the scalar product $(u,v)_{L_2(a,b;H)} = \int_a^b (u(t),v(t))dt$ (see [13]).

Let A and B be unbounded self-adjoint positive linear operators that do not depend on t and are defined in the Hilbert space H. We suppose that $D(A)$ is dense in H and $H_A \subset H_B$. In the general case, A and B are noncommutative. Let us consider the abstract Cauchy problem (see [22])

$$Bu'' + Au = f(t), \quad 0 < t < T, \qquad u(0) = u_0, \quad u'(0) = u_1, \tag{2.1}$$

where $u_0 \in H_0$, $u_1 \in H_B$, $f(t) \in L_2(0,T;H_{B^{-1}})$ are given and $u(t) \in H_A$ is the unknown function. Letting $f(t) = g'(t)$ in (2.1), we get the Cauchy problem

$$Bu'' + Au = g'(t), \quad 0 < t < T, \qquad u(0) = u_0, \quad u'(0) = u_1. \tag{2.2}$$

The following lemma holds.

LEMMA 2.1. *The a priori estimate for the solution of the problem* (2.1) *is valid*

$$\max_{t\in[0,T]} \left[\|u(t)\|_A^2 + \|u'(t)\|_B^2\right] \le C\left[\|u_0\|_A^2 + \|u_1\|_B^2 + \int_0^T \|f(t)\|_{B^{-1}}^2 dt\right].$$

At the less strong assumptions, $u_0 \in H_B$, $Bu_1 \in H_{A^{-1}}$, and $f \in L_2(0,T;H_{A^{-1}})$, the estimate

$$\max_{t\in[0,T]} \|u(t)\|_B^2 \le C\left[\|u_0\|_B^2 + \|Bu_1\|_{A^{-1}}^2 + \int_0^T \|f(t)\|_{A^{-1}}^2 dt\right]$$

holds true. For the solution of the problem (2.2), *if $u_0 \in H_B$, $Bu_1 \in H_{A^{-1}}$, and $g \in L_2(0,T;H_{B^{-1}})$, the estimate*

$$\max_{t\in[0,T]} \|u(t)\|_B^2 \le C\left[\|u_0\|_B^2 + \|Bu_1 - g(0)\|_{A^{-1}}^2 + \int_0^T \|g(t)\|_{B^{-1}}^2 dt\right]$$

holds true.

Proof. The proof can be found by using the energy method and Grönwall's lemma.

Similar results are true for operator-difference schemes. Let H_h be a finite-dimensional real Hilbert space with scalar product $(\cdot,\cdot)_h$ and norm $\|\cdot\|_h$. Also, let A_h and B_h be constant self-adjoint positively defined in H_h linear operators, i.e.,

$$A_h \ne A_h(t), \quad A_h = A_h^* \ge d_1 E_h, \qquad B_h \ne B_h(t), \quad B_h = B_h^* \ge d_2 E_h,$$

where $d_1, d_2 = \text{const} > 0$ and E_h is the identity operator in H_h. As in the previous case, we assume that the operators A_h and B_h in the general case are noncommutative. By H_{S_h}, where $S_h = S_h^* > 0$, we denote the space $H_{S_h} = H_h$ with scalar product and norm

$$(v,w)_{S_h} = (S_h v, w)_h, \qquad \|v\|_{S_h} = (S_h v, v)_h^{1/2}.$$

Let ω_τ be a uniform mesh on $(0,T)$ with step $\tau = T/m$, $\omega_\tau^- = \omega_\tau \cup \{0\}$, and $\bar\omega_\tau = \omega_\tau \cup \{0,T\}$. Further we shall make use of standard notations of the difference schemes [17]: $v = v(t)$, $\hat v = v(t+\tau)$, $\check v = v(t-\tau)$, $v_t = (\hat v - v)/\tau$, $v_{\bar t} = (v - \check v)/\tau$.

Let us consider the simplest three-layer operator-difference scheme with weights

$$B_h v_{t\bar t} + A_h v^{(\sigma)} = \phi(t), \quad t \in \omega_\tau, \qquad v(0) = v_0, \quad v_t(0) = v_1, \tag{2.3}$$

where $\sigma \ge 1/4$ is the weight parameter; $v^{(\sigma)} = \sigma\hat v + (1-2\sigma)v + \sigma\check v$; v_0, v_1 are given elements of H_h; and $\phi(t)$ and $v(t)$ are given unknown mesh functions with values in H_h. We also study the scheme

$$B_h v_{t\bar t} + A_h v^{(\sigma)} = \psi_{\bar t}, \quad t \in \omega_\tau, \qquad v(0) = v_0, \quad v_t(0) = v_1, \tag{2.4}$$

where $\psi(t)$ is a given mesh function with values in H_h. The following analogue of Lemma 2.1 holds.

LEMMA 2.2. *For the solution of the problem* (2.3) *the a priori estimates are valid:*

$$\max_{t\in\omega_\tau^-}\left[\left\|\frac{(\hat{v}+v)}{2}\right\|^2_{A_h}+\|v_t\|^2_{B_h}\right]\le C\left[\|v_0\|^2_{A_h}+\|v_1\|^2_{B_h}+\tau^2\|v_1\|^2_{A_h}+\tau\sum_{t\in\omega_\tau}\|\phi\|^2_{B_h^{-1}}\right],$$

$$\max_{t\in\omega_\tau^-}\left\|\frac{(\hat{v}+v)}{2}\right\|^2_{B_h}\le C\left[\|v_0\|^2_{B_h}+\|B_hv_1\|^2_{A_h^{-1}}+\tau^2\|v_1\|^2_{B_h}+\tau\sum_{t\in\omega_\tau}\|\phi\|^2_{A_h^{-1}}\right].$$

For the solution of problem (4) *the a priori estimate holds:*

$$\max_{t\in\omega_\tau^-}\left\|\frac{(\hat{v}+v)}{2}\right\|^2_{B_h}\le C\left[\|v_0\|^2_{B_h}+\|B_hv_1-\psi(0)\|^2_{A_h^{-1}}+\tau^2\|v_1\|^2_{B_h}+\tau\sum_{t\in\omega_\tau^-}\|\psi\|^2_{B_h^{-1}}\right].$$

Proof. The proof is analogous to the proof of Lemma 2.1.

3. Equation of string vibrations with concentrated mass. Let us consider the first initial-boundary value problem (IBVP) for the equation of vibrating string with concentrated mass at the interior point $x=\xi$ (see [19]):

(3.1)
$$\big[c(x)+K\delta(x-\xi)\big]\frac{\partial^2u}{\partial t^2}-\frac{\partial}{\partial x}\Big(a(x)\frac{\partial u}{\partial x}\Big)=f(x,t),\quad (x,t)\in Q=(0,1)\times(0,T),$$

$$u(0,t)=0,\quad u(1,t)=0,\quad 0<t<T, \tag{3.2}$$

$$u(x,0)=u_0(x),\quad \frac{\partial u(x,0)}{\partial t}=u_1(x),\quad x\in(0,1), \tag{3.3}$$

where $K>0$, $0<c_1\le a(x)\le c_2$, $0<c_3\le c(x)\le c_4$, and $\delta(x)$ is the Dirac distribution [22]. It follows from (3.1) that the solution for this problem satisfies for $(x,t)\in Q_1=(0,\xi)\times(0,T)$ and $(x,t)\in Q_2=(\xi,1)\times(0,T)$ the differential equation

$$c(x)\frac{\partial^2u}{\partial t^2}-\frac{\partial}{\partial x}\left(a(x)\frac{\partial u}{\partial x}\right)=f(x,t),$$

and for $x=\xi$ the conditions of conjugation

$$[u]_{x=\xi}\equiv u(\xi+0,t)-u(\xi-0,t)=0,\qquad \left[a\frac{\partial u}{\partial x}\right]_{x=\xi}=K\frac{\partial^2u(\xi,t)}{\partial t^2}.$$

It is easy to see that the IBVP (3.1)–(3.3) can be reduced to the form (2.1), letting $H=L_2(0,1)$, $Au=-\frac{\partial}{\partial x}(a(x)\frac{\partial u}{\partial x})$, and $Bu=\big[c(x)+K\delta(x-\xi)\big]u(x,t)$. Then

$$\|w\|^2_A=\int_0^1a(x)[w'(x)]^2dx\asymp\|w\|^2_{W_2^1(0,1)},\qquad w\in\overset{\circ}{W}{}_2^1(0,1),$$

$$\|w\|_B^2 = \int_0^1 c(x)w^2(x)dx + Kw^2(\xi) \asymp \|w\|_{L_2(0,1)}^2 + w^2(\xi).$$

In such a way $H_A = \overset{\circ}{W}{}_2^1(0,1)$, $H_{A^{-1}} = W_2^{-1}(0,1)$, H_B is the space of all $w \in L_2(0,1)$ with finite norm $\|w\|_B$, and $H_{B^{-1}}$ is the space of all Schwarz distributions on $(0,1)$ with finite norm

$$\|w\|_{B^{-1}} = \sup_{u \in H_B} \frac{|(u,w)|}{\|u\|_B}.$$

Further, we assume that the function $c(x)$ is continuous on $[0,1]$ and $a(x)$ has finite jump in the point $x = \xi$.

3.1. The functional spaces $\boldsymbol{\widetilde{W}_2^k(0,1)}$ and $\boldsymbol{\widetilde{W}_2^k(Q)}$. By $\widetilde{L}_2(0,1) = \widetilde{W}_2^0(0,1)$ we denote the subspace of functions $w(x) \in L_2(0,1)$ with scalar product and norm

$$(u,w)_{\widetilde{L}_2(0,1)} = \int_0^1 u(x)w(x)dx + u(\xi)w(\xi), \qquad \|w\|_{\widetilde{L}_2(0,1)} = (u,w)_{\widetilde{L}_2(0,1)}^{1/2}.$$

Further, we let $\widetilde{W}_2^1(0,1) = \overset{\circ}{W}{}_2^1(0,1)$ and $\widetilde{W}_2^k(0,1) = \overset{\circ}{W}{}_2^1(0,1) \cap W_2^k(0,\xi) \cap W_2^k(\xi,1)$, $k = 2,3,\dots$.

We define also the spaces $\widetilde{W}_2^k(Q)$ $(k = 0,1,2,\dots)$ as subsets of functions $w \in L_2(Q)$ for which

$$\begin{aligned}
&w(0,t) = w(1,t) = 0,\\
&\frac{\partial^i w}{\partial t^i} \in L_2(0,T;\widetilde{L}_2(0,1)), \quad i = 0,1,\dots,k,\\
&\frac{\partial^i w}{\partial x \partial t^{i-1}} \in L_2(Q), \quad i = 1,2,\dots,k,\\
&\frac{\partial^i w}{\partial x^j \partial t^{i-j}} \in L_2(Q_1) \cap L_2(Q_2), \quad 2 \le j \le k,\ i = j,\dots,k,
\end{aligned}$$

and norms defined as usual

$$\|w\|_{\widetilde{W}_2^k(Q)}^2 = \sum_{i=0}^k \left(\left\| \frac{\partial^i w(\xi,\cdot)}{\partial t^i} \right\|_{L_2(0,T)}^2 + \left\| \frac{\partial^i w}{\partial t^i} \right\|_{L_2(Q)}^2 \right)$$

$$+ \sum_{i=1}^k \left\| \frac{\partial^i w}{\partial x \partial t^{i-1}} \right\|_{L_2(Q)}^2 + \sum_{j=2}^k \sum_{i=j}^k \left(\left\| \frac{\partial^i w}{\partial x^j \partial t^{i-j}} \right\|_{L_2(Q_1)}^2 + \left\| \frac{\partial^i w}{\partial x^j \partial t^{i-j}} \right\|_{L_2(Q_2)}^2 \right).$$

Differentiating (3.1) with respect to x and t and applying Lemma 2.1, we easily obtain the following assertion.

LEMMA 3.1. (i) *Let* $a \in W_2^2(0,\xi) \cap W_2^2(\xi,1)$, $c \in W_2^2(0,1)$, $f \in W_2^2(Q_1) \cap W_2^2(Q_2)$, $f(0,t) = f(1,t) = 0$, $[f]_{x=\xi} = 0$, $u_0 \in \widetilde{W}_2^3(0,1)$, $u_1 \in \widetilde{W}_2^2(0,1)$, *and the compatibility conditions hold*

$$U_{tt}(0) = U_{tt}(1) = 0, \quad [U_{tt}]_{x=\xi} = 0, \quad [au_0']_{x=\xi} = K \lim_{x \to \xi} U_{tt}(x),$$

where

$$U_{tt}(x) = \frac{a(x)u_0''(x) + a'(x)u_0'(x) + f(x,0)}{c(x)}, \qquad ' \equiv \frac{d}{dx}.$$

Then the problem (3.1)–(3.3) *has unique solution* $u \in \widetilde{W}_2^3(Q)$.

(ii) *Let* $a \in W_2^3(0,\xi) \cap W_2^3(\xi,1)$, $c \in W_2^3(0,1)$, $f \in W_2^3(Q_1) \cap W_2^3(Q_2)$, $f(0,t) = f(1,t) = 0$, $[f]_{x=\xi} = 0$, $u_0 \in \widetilde{W}_2^4(0,1)$, $u_1 \in \widetilde{W}_2^3(0,1)$, *and in addition to the last compatibility conditions, the following hold:*

$$U_{ttt}(0) = U_{ttt}(1) = 0, \quad [U_{ttt}]_{x=\xi} = 0, \quad [au_1']_{x=\xi} = K \lim_{x\to\xi} U_{ttt}(x),$$

where

$$U_{ttt}(x) = \left(a(x)u_1''(x) + a'(x)u_1'(x) + \frac{\partial f(x,0)}{\partial t} \right) / c(x).$$

Then problem (3.1)–(3.3) *has unique solution* $u \in \widetilde{W}_2^4(Q)$.

3.2. The difference scheme. Let $\omega_h = \{x_1, x_2, \ldots, x_{n-1}\}$ be a nonuniform mesh in $(0,1)$, chosen so that ξ is a grid point. Define $\omega_h^- = \omega_h \cup \{x_0\}$, $\omega_h^+ = \omega_h \cup \{x_n\}$, $\bar\omega_h = \omega_h \cup \{x_0, x_n\}$, $x_0 = 0$, $x_n = 1$, and $h_i = x_i - x_{i-1}$. Also, let

$$v_x = \frac{(v_+ - v)}{h_+}, \quad v_{\bar x} = \frac{(v - v_-)}{h}, \quad v_{\hat x} = \frac{(v_+ - v)}{\hbar},$$

$$v = v(x), \quad v_\pm = v(x_\pm), \quad x = x_i, \quad x_\pm = x_{i\pm1}, \quad \hbar = \frac{(h + h_+)}{2}.$$

We assume that the following condition is fulfilled:

$$\frac{1}{c_0} \le \frac{h_+}{h} \le c_0, \quad c_0 = \text{const} \ge 1.$$

We approximate the problem (3.1)–(3.3) on the mesh $\bar\omega_h \times \bar\omega_\tau$ by the weighted difference scheme with averaged right-hand side

$$(c + K\delta_h)v_{t\bar t} - (\tilde a v_{\bar x}^{(\sigma)})_{\hat x} = T_x^2 T_t^2 f, \quad (x,t) \in \omega_h \times \omega_\tau, \tag{3.4}$$

$$v(0,t) = 0, \quad v(1,t) = 0, \quad t \in \bar\omega_\tau, \quad v(x,0) = u_0(x), \quad x \in \omega_h, \tag{3.5}$$

(3.6)
$$(c + K\delta_h)v_t(x,0) = T_x^2(cu_1) + K\delta_h u_1 + 0.5\tau T_x^2 \left[\widetilde{T}_t^2 f(x,0) + (au_0'(x))' \right], \quad x \in \omega_h,$$

where $\sigma \ge 1/4$, $\tilde a(x) = \big[a(x) + a(x-h)\big]/2$ if $x \ne \xi, \xi_+$, $\tilde a(\xi) = \big[a(\xi - 0) + a(\xi_-)\big]/2$, $\tilde a(\xi_+) = \big[a(\xi_+) + a(\xi + 0)\big]/2$,

$$\delta_h = \delta_h(x - \xi) = \begin{cases} 0, & x \in \omega_h \setminus \{\xi\}, \\ 1/\hbar, & x = \xi, \end{cases}$$

is the discrete Dirac-function, and

$$T_t f(x,t) = T_t^- f\left(x, t+\frac{\tau}{2}\right) = T_t^+ f\left(x, t-\frac{\tau}{2}\right) = \frac{1}{\tau}\int_{t-\tau/2}^{t+\tau/2} f(x,t')dt',$$

$$\widetilde{T}_t^2 f(x,0) = \frac{2}{\tau}\int_0^\tau \left(1-\frac{t'}{\tau}\right) f(x,t')dt',$$

$$T_x^- f(x,t) = \frac{1}{h}\int_{x_-}^{x} f(x',t)dx', \qquad T_x^+ f(x,t) = \frac{1}{h_+}\int_{x}^{x_+} f(x',t)dx',$$

$$T_x^2 f(x,t) = \frac{1}{\hbar}\int_{x_-}^{x_+} \kappa(x,x')f(x',t)dx', \quad \kappa(x,x') = \begin{cases} 1+(x'-x)/h, & x_- < x' < x, \\ 1-(x'-x)/h_+, & x < x' < x_+, \end{cases}$$

are Steklov averaged operators [4, 10, 12, 17]. Note that these operators are commutative and map derivatives into finite differences, for example,

$$T_x^2 \frac{\partial^2 u}{\partial x^2} = u_{\bar{x}\hat{x}}, \qquad T_t^- \frac{\partial u}{\partial t} = u_{\bar{t}}.$$

Let H_h be the set of mesh functions defined on the mesh $\bar{\omega}_h$ and zero at $x=0$ and $x=1$. We define the scalar products

$$(v,w)_h = \sum_{x\in\omega_h} v(x)w(x)\hbar, \qquad (v,w]_{h\star} = \sum_{x\in\omega_h^+} v(x)w(x)h,$$

and the corresponding norms

$$\|w\|_h = \|w\|_{L_{2,h}} = (w,w)_h^{1/2}, \qquad \|w]|_{h\star} = (w,w]_{h\star}^{1/2}.$$

The difference scheme (3.4)–(3.6) can be written in the form (2.3), setting $A_h v = -(\tilde{a}v_{\bar{x}})_{\hat{x}}$ and $B_h v = (c + K\delta_h)v$. For $w \in H_h$ we have

$$\|w\|_{A_h}^2 = (A_h w, w)_h = \sum_{x\in\omega_h^+} \tilde{a}(x)w_{\bar{x}}^2(x)h = \|w_{\bar{x}}]|_{h\star}^2,$$

$$\|w\|_{B_h}^2 = (B_h w, w)_h = \sum_{x\in\omega_h} c(x)w^2(x)\hbar + Kw^2(\xi) \asymp \|w\|_{B_{0h}}^2,$$

and

$$\|w\|_{B_h^{-1}}^2 = (B_h^{-1}w, w)_h = \sum_{x\in\omega_h\setminus\{\xi\}} \frac{w^2(x)}{c(x)}\hbar + \frac{\hbar^2(\xi)}{K+\hbar c(\xi)}w^2(\xi) = \|w\|_{B_{0h}^{-1}}^2,$$

where $B_{0h}w = (1+\delta_h)w$.

We define the following norms of Sobolev type:

$$\|w\|_{\widetilde{L}_{2,h}}^2 = \|w\|_{B_{0h}}^2 = \|w\|_{L_{2,h}}^2 + w^2(\xi), \qquad \|w\|_{\widetilde{W}_{2,h}^1}^2 = \|w_{\bar{x}}]|_{h\star}^2 + \|w\|_h^2,$$

$$\|v\|_{h\tau}^{(0)} = \max_{t\in\omega_\tau^-} \left\| \frac{(v(\cdot,t+\tau)+v(\cdot,t))}{2} \right\|_{\widetilde{L}_{2,h}},$$

$$\|v\|_{h\tau}^{(1)} = \max_{t\in\omega_\tau^-} \left[\left\| \frac{(v(\cdot,t+\tau)+v(\cdot,t))}{2} \right\|_{\widetilde{W}_{2,h}^1}^2 + \|v_t(\cdot,t)\|_{\widetilde{L}_{2,h}}^2 \right]^{1/2}.$$

3.3. Convergence of the difference scheme in the norm $\|\cdot\|^{(0)}_{h\tau}$. Let u be the solution of the problem (3.1)–(3.3) and v the solution of (3.4)–(3.6). The error $z = u - v$ satisfies the problem

$$(c + K\delta_h)z_{t\bar t} - (\tilde a z^{(\sigma)}_{\bar x})_{\hat x} = \varphi_{\hat x} + \eta_{t\bar t}, \quad (x,t) \in \omega_h \times \omega_\tau, \tag{3.7}$$

$$z(0,t) = 0, \quad z(1,t) = 0, \quad t \in \bar\omega_\tau, \quad z(x,0) = 0, \quad x \in \omega_h, \tag{3.8}$$

$$(c + K\delta_h)z_t(x,0) = \eta_t(x,0) + \chi_{\hat x}(x), \quad x \in \omega_h, \tag{3.9}$$

where it is denoted

$$\varphi = T_x^- T_t^2\Big(a\frac{\partial u}{\partial x}\Big) - \tilde a u^{(\sigma)}_{\bar x} - \frac{h^2}{6}(cu)_{\bar x t\bar t}, \qquad \eta = cu - T_x^2(cu) + \Big(\frac{h^2}{6}(cu)_{\bar x}\Big)_{\hat x},$$

$$\chi = \Big\{\frac{\tau}{2}T_x^-\Big[a\Big(\widetilde{T}_t^2\frac{\partial u}{\partial x} - \frac{du_0}{dx}\Big)\Big] - \frac{h^2}{6}(cu)_{\bar x t}\Big\}\Big|_{t=0}.$$

From Lemma 2.2, using the inequality

$$\|\varphi_{\hat x}\|_{A_h^{-1}} = \max_{w\in H_h}\frac{|(\varphi_{\hat x}, w)_h|}{\|w\|_{A_h}} = \max_{w\in H_h}\frac{|-(\varphi, w_{\bar x}]_{h\star}|}{\|w\|_{A_h}} \le \max_{w\in H_h}\frac{\|\varphi]|_{h\star}\|w_{\bar x}]|_{h\star}}{\|w\|_{A_h}} \le \frac{1}{c_1}\|\varphi]|_{h\star}, \tag{3.10}$$

we immediately get the following a priori estimate for the solution of problem (3.7)–(3.9):

$$\|z\|^{(0)}_{h\tau} \le C\left[\|\chi]|^2_{h\star} + \tau^2\|\chi_{\hat x}\|^2_{B_{0h}^{-1}} + \tau\sum_{t\in\omega_\tau}\|\varphi(\cdot,t)]|^2_{h\star} + \tau\sum_{t\in\bar\omega_\tau^-}\|\eta_t(\cdot,t)\|^2_{B_{0h}^{-1}}\right]^{1/2}. \tag{3.11}$$

Therefore, in order to estimate the rate of convergence of the difference scheme (3.4)–(3.6) in the norm $\|\cdot\|^{(0)}_{h\tau}$, it is sufficient to estimate the right-hand side of (3.11).

We let

$$\varphi = \varphi_1 + \varphi_2 + \varphi_3 + \varphi_4 + \varphi_5, \quad \text{where} \quad \varphi_1 = T_x^- T_t^2\Big(a\frac{\partial u}{\partial x}\Big) - (T_x^- a)\Big(T_x^- T_t^2\frac{\partial u}{\partial x}\Big),$$

$$\varphi_2 = [(T_x^- a) - \tilde a]\Big(T_x^- T_t^2\frac{\partial u}{\partial x}\Big), \quad \varphi_3 = \tilde a\Big[\Big(T_x^- T_t^2\frac{\partial u}{\partial x}\Big) - \Big(T_x^-\frac{\partial u}{\partial x}\Big)\Big],$$

$$\varphi_4 = -\sigma\tau^2\tilde a\Big(T_x^- T_t^2\frac{\partial^3 u}{\partial x\partial t^2}\Big), \quad \varphi_5 = -\frac{h^2}{6}\Big(T_x^- T_t^2\frac{\partial^3(cu)}{\partial x\partial t^2}\Big).$$

Using the integral representation

$$\varphi_1(x,t) = \frac{1}{2h^2\tau}\int_{t-\tau}^{t+\tau}\int_{x_-}^{x}\int_{x_-}^{x}\int_{x''}^{x'}\int_{x'}^{x''}\Big(1-\frac{|t'-t|}{\tau}\Big)a'(y')\frac{\partial^2 u(y'',t')}{\partial x^2}dy''dy'dx''dx'dt',$$

$$\varphi_2(x,t) = \left(\frac{-1}{2h}\int_{x_-}^{x}\int_{x'}^{x}\int_{x'}^{x''} a''(x''')dx'''dx''dx'\right) \cdot \left[\frac{1}{h\tau}\int_{x_-}^{x}\int_{t-\tau}^{t+\tau}\Big(1-\frac{|t'-t|}{\tau}\Big)\frac{\partial u(x',t')}{\partial x}dx'dt'\right],$$

$$\varphi_3(x,t) = \frac{a(x)+a(x_-)}{2h\tau}\int_{x_-}^{x}\int_{t-\tau}^{t+\tau}\int_{t}^{t'}\int_{t}^{t''}\Big(1-\frac{|t'-t|}{\tau}\Big)\frac{\partial^3 u(x',t''')}{\partial x\partial t^2}dt'''dt''dt'dx',$$

$$\varphi_4(x,t) = -\sigma\tau^2\frac{a(x)+a(x_-)}{2h\tau}\int_{x_-}^{x}\int_{t-\tau}^{t+\tau}\Big(1-\frac{|t'-t|}{\tau}\Big)\frac{\partial^3 u(x',t')}{\partial x\partial t^2}dt'dx',$$

$$\varphi_5(x,t) = -\frac{h^2}{6h\tau}\int_{x_-}^{x}\int_{t-\tau}^{t+\tau}\Big(1-\frac{|t'-t|}{\tau}\Big)\frac{\partial^3 (cu)(x',t')}{\partial x\partial t^2}dt'dx',$$

we immediately get by summing over the grid,

$$\Big\{\tau\sum_{t\in\omega_\tau}\|\varphi(\cdot,t)]|_{h\star}^2\Big\}^{1/2} \le C(h_{max}^2+\tau^2)\Big(\|a\|_{W_2^2(0,\xi)}+\|a\|_{W_2^2(\xi,1)}+\|c\|_{W_2^2(0,1)}\Big)$$

$$\times\Big(\Big\|\frac{\partial^3 u}{\partial x\partial t^2}\Big\|_{L_2(Q)}+\Big\|\frac{\partial^2 u}{\partial x^2}\Big\|_{L_2(Q_1)}+\Big\|\frac{\partial^2 u}{\partial x^2}\Big\|_{L_2(Q_2)}+\Big\|\frac{\partial u}{\partial x}\Big\|_{L_2(Q)}\Big). \tag{3.12}$$

The integral formula

$$\begin{aligned}\eta_t(x,t) = {} & \frac{1}{\hbar\tau}\int_{t-\tau}^{t}\int_{x_-}^{x}\int_{x'}^{x}\int_{x}^{x''}\Big(1+\frac{x'-x}{h}\Big)\frac{\partial^3 (cu)(x''',t')}{\partial x^2\partial t}dx'''dx''dx'dt' \\ & + \frac{1}{\hbar\tau}\int_{t-\tau}^{t}\int_{x}^{x_+}\int_{x'}^{x}\int_{x}^{x''}\Big(1-\frac{x'-x}{h_+}\Big)\frac{\partial^3 (cu)(x''',t')}{\partial x^2\partial t}dx'''dx''dx'dt' \\ & + \frac{h}{6\hbar\tau}\int_{t-\tau}^{t}\int_{x_-}^{x}\int_{x'}^{x}\frac{\partial^3 (cu)(x'',t')}{\partial x^2\partial t}dx''dx'dt' \\ & + \frac{h_+}{6\hbar\tau}\int_{t-\tau}^{t}\int_{x}^{x_+}\int_{x}^{x'}\frac{\partial^3 (cu)(x'',t')}{\partial x^2\partial t}dx''dx'dt'\end{aligned} \tag{3.13}$$

implies

$$\left\{\tau\sum_{t\in\omega_\tau^-}\|\eta_t(\cdot,t)\|_{B_{0h}^{-1}}^2\right\}^{1/2} \le Ch_{max}^2\|c\|_{W_2^2(0,1)}$$

$$\times\Big(\Big\|\frac{\partial^3 u}{\partial x^2\partial t}\Big\|_{L_2(Q_1)}+\Big\|\frac{\partial^3 u}{\partial x^2\partial t}\Big\|_{L_2(Q_2)}+\Big\|\frac{\partial^2 u}{\partial x\partial t}\Big\|_{L_2(Q)}\Big), \tag{3.14}$$

and the representation

$$\chi(x) = \frac{1}{h}\int_{x_-}^{x}\int_{0}^{\tau}\int_{0}^{t'}\Big(1-\frac{t'}{\tau}\Big)a(x')\frac{\partial^2 u(x',t'')}{\partial x\partial t}dt''dt'dx' \tag{3.15}$$

$$- \frac{h}{6\tau}\int_{x_-}^{x}\int_{0}^{\tau}\frac{\partial^2 (cu)(x',t')}{\partial x\partial t}dt'dx'$$

implies

$$\|\chi]|_{h\star} \le C\tau^{3/2}\Big(\|a\|_{W_2^1(0,\xi)} + \|a\|_{W_2^1(\xi,1)}\Big)\Big\|\frac{\partial^2 u}{\partial x\partial t}\Big\|_{L_2(Q_\tau)}$$

$$+ Ch_{max}^2\tau^{-1/2}\|c\|_{W_2^1(0,1)}\Big\|\frac{\partial^2 u}{\partial x\partial t}\Big\|_{L_2(Q_\tau)},$$

where $Q_\tau = (0,1)\times(0,\tau)$. Hence, using the inequality (see [14])

$$\|g\|_{L_2(0,\varepsilon)} \le C\sqrt{\varepsilon}\|g\|_{W_2^1(0,1)}, \qquad 0<\varepsilon<1, \tag{3.16}$$

we get

$$\|\chi]|_{h\star} \le C(h_{max}^2+\tau^2)(\|a\|_{W_2^1(0,\xi)} + \|a\|_{W_2^1(\xi,1)} + \|c\|_{W_2^1(0,1)})$$

$$\times\Big(\Big\|\frac{\partial^3 u}{\partial x\partial t^2}\Big\|_{L_2(Q)} + \Big\|\frac{\partial^2 u}{\partial x\partial t}\Big\|_{L_2(Q)}\Big). \tag{3.17}$$

Exploiting (3.15) again, we obtain

$$\tau\|\chi\|_{B_{0h}^{-1}} \le C(h_{max}^2+\tau^2)(\|a\|_{W_2^2(0,\xi)} + \|a\|_{W_2^2(\xi,1)} + \|c\|_{W_2^2(0,1)})$$

$$\times\Big(\Big\|\frac{\partial^3 u}{\partial x^2\partial t}\Big\|_{L_2(Q_1)} + \Big\|\frac{\partial^3 u}{\partial x^2\partial t}\Big\|_{L_2(Q_2)} + \Big\|\frac{\partial^3 u}{\partial x\partial t^2}\Big\|_{L_2(Q)} + \Big\|\frac{\partial^2 u}{\partial x\partial t}\Big\|_{L_2(Q)}\Big). \tag{3.18}$$

Now from (3.11), (3.12), (3.14), (3.17), and (3.18) we get the desired convergence rate estimate of the difference scheme (3.4)–(3.6) as follows.

THEOREM 3.2. *Let the assumptions of part* (i) *of Lemma* 3.1 *hold. Then*

$$\|z\|_{h\tau}^{(0)} \le C(h_{max}^2+\tau^2)(\|a\|_{W_2^2(0,\xi)} + \|a\|_{W_2^2(\xi,1)} + \|c\|_{W_2^2(0,1)})\|u\|_{\widetilde{W}_2^3(Q)}. \tag{3.19}$$

Therefore, although the difference scheme on the nonuniform mesh has first order of approximation in the space, the rate of convergence in the "weak" norm $\|\cdot\|_{h\tau}^{(0)}$ is second order with respect to h_{max}.

3.4. Approximation and convergence in the norm $\|\cdot\|_{h\tau}^{(1)}$. Following [18], we approximate (3.1) as follows:

$$(c+K\delta_h)v_{t\bar t} + \Big(\frac{h^2}{6}(cv)_{\bar x t\bar t}\Big)_{\hat x} - (\tilde a v_{\bar x}^{(\sigma)})_{\hat x} - \frac{h_+ - h}{6}\big(a_x v_{\bar x\hat x}^{(\sigma)} - a_{\bar x\hat x}v_{\bar x}^{(\sigma)}\big) = T_x^2T_t^- f, \tag{3.20}$$

$$(x,t) \in \omega_h \times \omega_\tau .$$

In the expressions $a_x(\xi_-)$ and $a_{\bar x\hat x}(\xi_-)$, the value $a(\xi)$ must be changed by $a(\xi-0)$, and in the expressions $a_x(\xi)$ and $a_{\bar x\hat x}(\xi_+)$, the value $a(\xi)$ must be replaced by $a(\xi+0)$. We approximate the boundary and the first initial condition as above, by (3.5). The second initial conditions we approximate letting

(3.21)
$$(c+K\delta_h)v_t(x,0) = T_x^2(cu_1) - \Big(\frac{h^2}{6}(cu_1)_{\bar x}\Big)_{\hat x} + K\delta_h u_1 + 0.5\tau T_x^2\big[f(x,0) + (au_0')'\big],$$

$$x \in \omega_h .$$

As compared with (3.4)–(3.6), the scheme (3.20), (3.5), (3.21) has second order local approximation in the space variable. It can be written in the form

$$(B_h + B_{h1})v_{t\bar t} + (A_h + A_{h1})v^{(\sigma)} = \tilde f(t),$$

where $B_{h1}v = (\frac{h^2}{6}(cv)_{\bar x})_{\hat x}$ and $A_{h1}v = -\frac{h_+-h}{6}\big(a_x v_{\bar x\hat x} - a_{\bar x\hat x}v_{\bar x}\big)$ are "small" nonselfadjoint operators in H_h.

For the grid $\bar\omega_h$ we additionally suppose $h_+ = h$ at $x = \xi$.

The error $z = u - v$, where u is the solution of the problem (3.1)–(3.3) and v is the solution of the difference problem (3.20), (3.5), (3.21), satisfies the difference scheme

$$(3.22)\qquad (c+K\delta_h)z_{t\bar t} + \Big(\frac{h^2}{6}(cz)_{\bar x t\bar t}\Big)_{\hat x} - (\tilde a z_{\bar x}^{(\sigma)})_{\hat x} - \frac{h_+-h}{6}\big(a_x z_{\bar x\hat x}^{(\sigma)} - a_{\bar x\hat x} z_{\bar x}^{(\sigma)}\big) = \phi,$$

$$(x,t) \in \omega_h \times \omega_\tau ,$$

with homogeneous boundary and first initial condition (3.8). The second initial condition takes the form

$$(3.23)\qquad (c+K\delta_h)z_t(x,0) = \zeta, \qquad x \in \omega_h .$$

In (3.22) and (3.23)

$$\phi = \phi_1 + \phi_2 = T_t^2\bigg[c\frac{\partial^2 u}{\partial t^2} - T_x^2\Big(c\frac{\partial^2 u}{\partial t^2}\Big) + \Big(\frac{h^2}{6}\Big(c\frac{\partial^2 u}{\partial t^2}\Big)_{\bar x}\Big)_{\hat x}\bigg]$$

$$- \bigg[(\tilde a u_{\bar x}^{(\sigma)})_{\hat x} + \frac{h_+-h}{6}\big(a_x u_{\bar x\hat x}^{(\sigma)} - a_{\bar x\hat x}u_{\bar x}^{(\sigma)}\big) - T_x^2 T_t^2\frac{\partial}{\partial x}\Big(a\frac{\partial u}{\partial x}\Big)\bigg],$$

$$\zeta = \zeta_1 + \zeta_2 + \zeta_3 = \bigg[cu_1 - T_x^2(cu_1) + \Big(\frac{h^2}{6}(cu_1)_{\bar x}\Big)_{\hat x}\bigg]$$

$$+ \frac{\tau}{2}(c+K\delta_h)\Big(\widetilde{T}_t^2\frac{\partial^2 u}{\partial t^2} - \frac{\partial^2 u}{\partial t^2}\Big)\bigg|_{t=0} + \frac{\tau}{2}\bigg[c\frac{\partial^2 u}{\partial t^2} - T_x^2\Big(c\frac{\partial^2 u}{\partial t^2}\Big)\bigg]\bigg|_{t=0}.$$

Using the energy method and Grönwall's lemma, one easily obtains the following result.

LEMMA 3.3. *If* $c \in C^1[0,1]$ *and* $a \in C^1[0,\xi] \cap C^1[\xi,1]$, *then the difference scheme* (3.22), (3.8), (3.23) *is stable in the norm* $\|\cdot\|_{h\tau}^{(1)}$ *and the following a priori estimate holds:*

$$\|z\|_{h\tau}^{(1)} \le C\Big\{\|\zeta\|^2_{B_{0h}^{-1}} + \tau^2\|(B_{0h}^{-1}\zeta)_{\bar{x}}]|^2_{h\star} + \sum_{t\in\omega_\tau}\|\phi(\cdot,t)\|^2_{B_{0h}^{-1}}\Big\}^{1/2}. \tag{3.24}$$

Therefore, in order to estimate the rate of convergence of the difference scheme (3.20), (3.5), (3.21) in the norm $\|\cdot\|_{h\tau}^{(1)}$, it is sufficient to estimate the right-hand side of (3.24).

Using the identity

$$\phi_1 = \eta_{t\bar{t}}$$

and the integral representation (3.13), we get the estimate

$$\Big\{\tau\sum_{t\in\omega_\tau}\|\phi_1(\cdot,t)\|^2_{B_{0h}^{-1}}\Big\}^{1/2} \le Ch^2_{max}\|c\|_{W_2^2(0,1)}\bigg(\Big\|\frac{\partial^4 u}{\partial x^2\partial t^2}\Big\|_{L_2(Q_1)}$$

$$+\Big\|\frac{\partial^4 u}{\partial x^2\partial t^2}\Big\|_{L_2(Q_2)} + \Big\|\frac{\partial^3 u}{\partial x\partial t^2}\Big\|_{L_2(Q)}\bigg). \tag{3.25}$$

For $x \ne \xi$ the addendum ϕ_2 can be expanded as follows:

$$\phi_2 = \phi_{20} + \phi_{21} + \phi_{22} + \phi_{23} + \phi_{24} + \phi_{25} + \phi_{26} + \phi_{27} + \phi_{28} + \phi_{29}$$

$$= -\frac{1}{2}\Big(T_x^2\frac{\partial^2(au)}{\partial x^2} - T_x^2T_t^2\frac{\partial^2(au)}{\partial x^2}\Big) - \frac{1}{2}\Big(a + \frac{h_+ - h}{3}a_x\Big)\Big(T_x^2\frac{\partial^2 u}{\partial x^2} - T_x^2T_t^2\frac{\partial^2 u}{\partial x^2}\Big)$$

$$-\frac{1}{2}\Big(a - T_x^2 a + \frac{h_+ - h}{3}a_x\Big)T_x^2T_t^2\frac{\partial^2 u}{\partial x^2} - \frac{1}{2}\Big[\big(T_x^2 a\big)\Big(T_x^2T_t^2\frac{\partial^2 u}{\partial x^2}\Big) - T_x^2T_t^2\Big(a\frac{\partial^2 u}{\partial x^2}\Big)\Big]$$

$$+\frac{1}{2}\big(T_x^2 a''\big)\Big(u + \frac{h_+ - h}{3}u_{\bar{x}} - T_t^2 u - \frac{h_+ - h}{3}T_t^2 u_{\bar{x}}\Big) + \frac{1}{2}\big(T_x^2 a''\big)T_t^2\Big(u - T_x^2 u + \frac{h_+ - h}{3}u_{\bar{x}}\Big)$$

$$+\frac{1}{2}\Big[\big(T_x^2 a''\big)\big(T_x^2T_t^2 u\big) - T_x^2T_t^2\big(a''u\big)\Big] - \frac{\sigma\tau^2}{2}T_x^2T_t^2\frac{\partial^4(au)}{\partial x^2\partial t^2}$$

$$-\frac{\sigma\tau^2}{2}\Big(a + \frac{h_+ - h}{3}a_x\Big)T_x^2T_t^2\frac{\partial^4 u}{\partial x^2\partial t^2} + \frac{\sigma\tau^2}{2}\Big(T_x^2 a\Big)\Big[\Big(T_t^2\frac{\partial^2 u}{\partial t^2}\Big) + \frac{h_+ - h}{3}\Big(T_t^2\frac{\partial^2 u}{\partial t^2}\Big)_{\bar{x}}\Big].$$

The integral representation

$$\phi_{20}(x,t) = -\frac{1}{2\hbar\tau}\int_{x_-}^{x_+}\int_{t-\tau}^{t}\int_{t'}^{t}\int_{t}^{t''}\kappa(x,x')\Big(1 - \frac{|t'-t|}{\tau}\Big)\frac{\partial^4(au)(x',t''')}{\partial x^2\partial t^2}dt'''dt''dt'dx',$$

$$x \neq \xi,$$

implies the estimate

(3.26)
$$\Big\{\tau \sum_{t\in\omega_\tau} \sum_{x\in\omega_h,\, x<\xi} \phi_{20}^2(x,t)\hbar\Big\}^{1/2} \le C\tau^2 \|a\|_{W_2^2(0,\xi)} \Big(\Big\|\frac{\partial^4 u}{\partial x^2\partial t^2}\Big\|_{L_2(Q_1)} + \Big\|\frac{\partial^3 u}{\partial x\partial t^2}\Big\|_{L_2(Q_1)}\Big).$$

In a similar way, we obtain

$$\Big\{\tau \sum_{t\in\omega_\tau} \sum_{x\in\omega_h,\, x<\xi} \phi_{21}^2(x,t)\hbar\Big\}^{1/2} \le C\tau^2 \|a\|_{W_2^1(0,\xi)} \Big\|\frac{\partial^4 u}{\partial x^2\partial t^2}\Big\|_{L_2(Q_1)}. \tag{3.27}$$

Using a known estimate for an expression of the form $a - T_x^2 a + \frac{h_+ - h}{3} a_x$ (see [8]), we find

$$\Big\{\tau \sum_{t\in\omega_\tau} \sum_{x\in\omega_h,\, x<\xi} \phi_{22}^2(x,t)\hbar\Big\}^{1/2} \le Ch_{max}^2 \|a\|_{W_2^3(0,\xi)} \Big\|\frac{\partial^2 u}{\partial x^2}\Big\|_{L_2(Q_1)}. \tag{3.28}$$

The addendum ϕ_{25} can be estimated in a similar way:

$$\Big\{\tau \sum_{t\in\omega_\tau} \sum_{x\in\omega_h,\, x<\xi} \phi_{25}^2(x,t)\hbar\Big\}^{1/2} \le Ch_{max}^2 \|a\|_{W_2^3(0,\xi)} \Big\|\frac{\partial^2 u}{\partial x^2}\Big\|_{L_2(Q_1)}. \tag{3.29}$$

Using the integral representation

$$\phi_{23}(x,t) = -\frac{1}{4\hbar^2\tau} \int_{x_-}^{x_+} \int_{x_-}^{x_+} \int_{t-\tau}^{t+\tau} \kappa(x,x')\kappa(x,x'')\Big(1 - \frac{|t'-t|}{\tau}\Big)$$

$$\times \Big(\int_{x''}^{x'} a'(x''')dx'''\Big)\Big(\int_{x'}^{x''} \frac{\partial^3 u(x''',t')}{\partial x^3} dx'''\Big) dt' dx'' dx', \quad x \neq \xi,$$

and the embedding $W_2^2(0,\xi) \subset C^1[0,\xi]$, we get

$$\Big\{\tau \sum_{t\in\omega_\tau} \sum_{x\in\omega_h,\, x<\xi} \phi_{23}^2(x,t)\hbar\Big\}^{1/2} \le Ch_{max}^2 \|a\|_{W_2^2(0,\xi)} \Big\|\frac{\partial^3 u}{\partial x^3}\Big\|_{L_2(Q_1)}. \tag{3.30}$$

In a similar way we obtain

(3.31)
$$\Big\{\tau \sum_{t\in\omega_\tau} \sum_{x\in\omega_h,\, x<\xi} \phi_{26}^2(x,t)\hbar\Big\}^{1/2} \le Ch_{max}^2 \|a\|_{W_2^3(0,\xi)} \Big(\Big\|\frac{\partial^2 u}{\partial x^2}\Big\|_{L_2(Q_1)} + \Big\|\frac{\partial u}{\partial x}\Big\|_{L_2(Q_1)}\Big).$$

The obvious inequality

$$|\phi_{24}| \le C\|a''\|_{C[0,\xi]} \max_{x\in[0,\xi]} |u - T_t^2 u|$$

and the embedding $W_2^3(0,\xi) \subset C^2[0,\xi]$ imply

$$\Big\{\tau \sum_{t\in\omega_\tau} \sum_{x\in\omega_h,\ x<\xi} \phi_{24}^2(x,t)\hbar\Big\}^{1/2} \le C\tau^2 \|a\|_{W_2^3(0,\xi)} \Big\|\frac{\partial^3 u}{\partial x \partial t^2}\Big\|_{L_2(Q_1)}. \tag{3.32}$$

In a similar way we get

$$\Big\{\tau \sum_{t\in\omega_\tau} \sum_{x\in\omega_h,\ x<\xi} \phi_{29}^2(x,t)\hbar\Big\}^{1/2} \le C\tau^2 \|a\|_{W_2^2(0,\xi)} \bigg(\Big\|\frac{\partial^3 u}{\partial x \partial t^2}\Big\|_{L_2(Q_1)} + \Big\|\frac{\partial^2 u}{\partial t^2}\Big\|_{L_2(Q_1)}\bigg). \tag{3.33}$$

From the integral representation

$$\phi_{27}(x,t) = -\frac{\sigma\tau}{2\hbar} \int_{x_-}^{x_+} \int_{t-\tau}^{t+\tau} \kappa(x,x')\Big(1 - \frac{|t'-t|}{\tau}\Big) \frac{\partial^4 (au)(x',t')}{\partial x^2 \partial t^2} dt' dx'$$

follows the estimate

$$\Big\{\tau \sum_{t\in\omega_\tau} \sum_{x\in\omega_h,\ x<\xi} \phi_{27}^2(x,t)\hbar\Big\}^{1/2} \le C\tau^2 \|a\|_{W_2^2(0,\xi)} \bigg(\Big\|\frac{\partial^4 u}{\partial x^2 \partial t^2}\Big\|_{L_2(Q_1)} + \Big\|\frac{\partial^3 u}{\partial x \partial t^2}\Big\|_{L_2(Q_1)}\bigg). \tag{3.34}$$

In a similar way we estimate ϕ_{28}:

$$\Big\{\tau \sum_{t\in\omega_\tau} \sum_{x\in\omega_h,\ x<\xi} \phi_{28}^2(x,t)\hbar\Big\}^{1/2} \le C\tau^2 \|a\|_{W_2^1(0,\xi)} \Big\|\frac{\partial^4 u}{\partial x^2 \partial t^2}\Big\|_{L_2(Q_1)}. \tag{3.35}$$

For $x = \xi$ we set

$$\phi_2 = \phi_{2,10} + \phi_{2,11} + \phi_{2,12} = -[(\tilde a u_{\bar x})_{\hat x} - T_t^2(\tilde a u_{\bar x})_{\hat x}]$$

$$- \bigg[T_t^2(\tilde a u_{\bar x})_{\hat x} - T_x^2 T_t 2 - \frac{\partial}{\partial x}\Big(a\frac{\partial u}{\partial x}\Big)\bigg] - \sigma\tau^2(\tilde a u_{\bar x})_{\hat x t\bar t}.$$

The integral representations

$$\hbar\phi_{2,10}(\xi,t) = \frac{a(\xi-0)+a(\xi-h)}{2h\tau} \int_{\xi_-}^{\xi} \int_{t-\tau}^{t+\tau} \int_{t'}^{t} \int_{t}^{t''} \Big(1 - \frac{|t'-t|}{\tau}\Big) \frac{\partial^3 u(x',t''')}{\partial x \partial t^2} dt''' dt'' dt' dx'$$

$$-\frac{a(\xi+0)+a(\xi+h)}{2h\tau} \int_{\xi}^{\xi_+} \int_{t-\tau}^{t+\tau} \int_{t'}^{t} \int_{t}^{t''} \Big(1 - \frac{|t'-t|}{\tau}\Big) \frac{\partial^3 u(x',t''')}{\partial x \partial t^2} dt''' dt'' dt' dx',$$

$$\hbar\phi_{2,11}(\xi,t) = \frac{1}{2h\tau} \int_{\xi_-}^{\xi} \int_{x'}^{\xi} \int_{x'}^{x''} \int_{t-\tau}^{t+\tau} \Big(1 - \frac{|t'-t|}{\tau}\Big) \bigg(a''(x''') \frac{\partial u(x'',t')}{\partial x} - a'(x'') \frac{\partial^2 u(x''',t')}{\partial x^2}\bigg) dx''' dx'' dx' dt'$$

$$-\frac{1}{2h\tau}\int\limits_{\xi}^{\xi_+}\int\limits_{x'}^{\xi_+}\int\limits_{x'}^{x''}\int\limits_{t-\tau}^{t+\tau}\Big(1-\frac{|t'-t|}{\tau}\Big)\Big(a''(x''')\frac{\partial u(x'',t')}{\partial x}-a'(x'')\frac{\partial^2 u(x''',t')}{\partial x^2}\Big)dx'''dx''dx'dt',$$

and

$$\hbar\phi_{2,12}(\xi,t)=\sigma\tau\frac{a(\xi-0)+a(\xi-h)}{2h}\int_{\xi_-}^{\xi}\int_{t-\tau}^{t+\tau}\Big(1-\frac{|t'-t|}{\tau}\Big)\frac{\partial^3 u(x',t')}{\partial x\partial t^2}dt'dx'$$

$$-\sigma\tau\frac{a(\xi+0)+a(\xi+h)}{2h}\int_{\xi}^{\xi_+}\int_{t-\tau}^{t+\tau}\Big(1-\frac{|t'-t|}{\tau}\Big)\frac{\partial^3 u(x',t')}{\partial x\partial t^2}dt'dx'$$

imply

$$\Big\{\tau\sum_{t\in\omega_\tau}\phi_2^2(\xi,t)\hbar^2\Big\}^{1/2}\le C(h_{max}^2+\tau^2)\Big[\|a\|_{W_2^3(0,\xi)}\Big(\Big\|\frac{\partial^4 u}{\partial x^2\partial t^2}\Big\|_{L_2(Q_1)}+\Big\|\frac{\partial^3 u}{\partial x\partial t^2}\Big\|_{L_2(Q_1)}$$

$$+\Big\|\frac{\partial^3 u}{\partial x^3}\Big\|_{L_2(Q_1)}+\Big\|\frac{\partial^2 u}{\partial x^2}\Big\|_{L_2(Q_1)}+\Big\|\frac{\partial u}{\partial x}\Big\|_{L_2(Q_1)}\Big)+\|a\|_{W_2^1(\xi,1)}\Big(\Big\|\frac{\partial^4 u}{\partial^2 x\partial t^2}\Big\|_{L_2(Q_2)}$$

$$+\Big\|\frac{\partial^3 u}{\partial x\partial t^2}\Big\|_{L_2(Q_2)}+\Big\|\frac{\partial^3 u}{\partial x^3}\Big\|_{L_2(Q_1)}+\Big\|\frac{\partial^2 u}{\partial x^2}\Big\|_{L_2(Q_1)}+\Big\|\frac{\partial u}{\partial x}\Big\|_{L_2(Q_1)}\Big)\Big]. \tag{3.36}$$

From (3.26)–(3.35), similar estimates for $x>\xi$, and (3.36), we finally obtain an estimate for ϕ_2:

$$\Big\{\tau\sum_{t\in\omega_\tau}\|\phi_2(\cdot,t)\|_{B_{0h}^{-1}}^2\Big\}^{1/2}\le C(h_{max}^2+\tau)\big(\|a\|_{W_2^3(0,\xi)}+\|a\|_{W_2^3(\xi,1)}\big)\|u\|_{\widetilde{W}_2^4(Q)}. \tag{3.37}$$

It remains to estimate the expression ζ. The addendum ζ_1 has a form similar to η, η_t, and $\eta_{t\bar t}=\phi_1$. Therefore,

$$\|\zeta_1\|_{B_{0h}^{-1}}\le Ch_{max}^2\|c\|_{W_2^2(0,1)}(\|u_1''\|_{L_2(0,\xi)}+\|u_1''\|_{L_2(\xi,1)}+\|u_1'\|_{L_2(0,1)}),$$

from which, applying the trace theorem, we get

$$\|\zeta_1\|_{B_{0h}^{-1}}\le Ch_{max}^2\|c\|_{W_2^2(0,1)}\Big(\Big\|\frac{\partial^4 u}{\partial x^2\partial t^2}\Big\|_{L_2(Q_1)}+\Big\|\frac{\partial^3 u}{\partial x^2\partial t}\Big\|_{L_2(Q_1)}$$

$$+\Big\|\frac{\partial^4 u}{\partial x^2\partial t^2}\Big\|_{L_2(Q_2)}+\Big\|\frac{\partial^3 u}{\partial x^2\partial t}\Big\|_{L_2(Q_2)}+\Big\|\frac{\partial^3 u}{\partial x\partial t^2}\Big\|_{L_2(Q)}+\Big\|\frac{\partial^2 u}{\partial x\partial t}\Big\|_{L_2(Q)}\Big). \tag{3.38}$$

Also, we have

$$\tau\|(B_{0h}^{-1}\zeta_1)_{\bar x}]|\le Ch_{max}\tau\|c\|_{W_2^2(0,1)}\Big(\Big\|\frac{\partial^4 u}{\partial x^2\partial t^2}\Big\|_{L_2(Q_1)}+\Big\|\frac{\partial^3 u}{\partial x^2\partial t}\Big\|_{L_2(Q_1)}$$

$$+\Big\|\frac{\partial^4 u}{\partial x^2\partial t^2}\Big\|_{L_2(Q_2)}+\Big\|\frac{\partial^3 u}{\partial x^2\partial t}\Big\|_{L_2(Q_2)}+\Big\|\frac{\partial^3 u}{\partial x\partial t^2}\Big\|_{L_2(Q)}+\Big\|\frac{\partial^2 u}{\partial x\partial t}\Big\|_{L_2(Q)}\Big). \tag{3.39}$$

From the integral representation

$$B_h^{-1}\zeta_2(x) = \frac{1}{\hbar}\int_{x_-}^{x_+}\int_0^{\tau}\int_0^{t'} \kappa(x,x')\Big(1-\frac{t'}{\tau}\Big)\frac{\partial^3 u(x',t'')}{\partial t^3}dt''dt'dx'$$

$$+\frac{1}{\hbar}\int_{x_-}^{x_+}\int_{x'}^{x}\int_0^{\tau}\int_0^{t'} \kappa(x,x')\Big(1-\frac{t'}{\tau}\Big)\frac{\partial^4 u(x'',t'')}{\partial x\partial t^3}dt''dt'dx''dx', \quad x \neq \xi,$$

$$B_h^{-1}\zeta_2(\xi) = \int_0^{\tau}\int_0^{t'} \Big(1-\frac{t'}{\tau}\Big)\frac{\partial^3 u(\xi,t'')}{\partial t^3}dt''dt',$$

and

$$(B_h^{-1}\zeta_2)_x = \frac{1}{h}\int_{x_-}^{x}\int_0^{\tau}\int_0^{t'} \Big(1-\frac{t'}{\tau}\Big)\frac{\partial^4 u(x',t'')}{\partial x\partial t^3}dt''dt'dx',$$

applying (3.16), we find

$$\|\zeta_2\|_{B_{0h}^{-1}} \leq C(h_{max}^2+\tau^2)\Big(\Big\|\frac{\partial^4 u}{\partial t^4}\Big\|_{L_2(Q)} + \Big\|\frac{\partial^3 u}{\partial t^3}\Big\|_{L_2(Q)}$$

$$+\Big\|\frac{\partial^4 u}{\partial x\partial t^3}\Big\|_{L_2(Q)} + \Big\|\frac{\partial^4 u(\xi,\cdot)}{\partial t^4}\Big\|_{L_2(0,T)} + \Big\|\frac{\partial^3 u(\xi,\cdot)}{\partial t^3}\Big\|_{L_2(0,T)}\Big) \tag{3.40}$$

and

$$\tau\|(B_{0h}^{-1}\zeta_2)_{\bar{x}}]| \leq C\tau^{5/2}\Big\|\frac{\partial^4 u}{\partial x\partial t^3}\Big\|_{L_2(Q)}. \tag{3.41}$$

Next, using the integral formula

$$\zeta_3(x) = \frac{\tau}{2\hbar}\int_{x_-}^{x_+}\int_{x'}^{x}\kappa(x,x')\frac{\partial}{\partial x}\Big(c\frac{\partial^2 u}{\partial t^2}\Big)\Big|_{(x'',0)}dx''dx'$$

and the trace theorem, we obtain the estimate

$$\|\zeta_3\|_{B_{0h}^{-1}} \leq C(h_{max}^2+\tau^2)\|c\|_{W_2^2(0,1)}\Big(\Big\|\frac{\partial^4 u}{\partial x\partial t^3}\Big\|_{L_2(Q)}$$

$$+\Big\|\frac{\partial^3 u}{\partial x\partial t^2}\Big\|_{L_2(Q)} + \Big\|\frac{\partial^3 u}{\partial t^3}\Big\|_{L_2(Q)} + \Big\|\frac{\partial^2 u}{\partial t^2}\Big\|_{L_2(Q)}\Big) \tag{3.42}$$

and

$$\tau\|(B_{0h}^{-1}\zeta_3)_{\bar{x}}]| \leq C\tau^2\|c\|_{W_2^2(0,1)}\Big(\Big\|\frac{\partial^4 u}{\partial x\partial t^3}\Big\|_{L_2(Q)}$$

$$+\Big\|\frac{\partial^3 u}{\partial x\partial t^2}\Big\|_{L_2(Q)} + \Big\|\frac{\partial^3 u}{\partial t^3}\Big\|_{L_2(Q)} + \Big\|\frac{\partial^2 u}{\partial t^2}\Big\|_{L_2(Q)}\Big). \tag{3.43}$$

Finally, from (3.25), (3.37), and (3.38)–(3.43) we get the required convergence rate estimate of the scheme (3.20), (3.5), (3.21).

THEOREM 3.4. *Let the assumptions of part* (ii) *of Lemma* 3.1 *hold. Then*

$$\|z\|_{h\tau}^{(1)} \le C(h_{max}^2 + \tau)(\|a\|_{W_2^3(0,\xi)} + \|a\|_{W_2^3(\xi,1)} + \|c\|_{W_2^2(0,1)} + 1)\|u\|_{\widetilde{W}_2^4(Q)}. \tag{3.44}$$

Remark. Analogous results hold for the IBVP where the mass is concentrated at several points $\xi_i \in (0,1)$, $i = 1, 2, \ldots, L$. In such cases $Bu = [c(x) + \sum_{i=1}^{L} K_i \delta(x - \xi_i)]u(x,t)$ and $\|w\|_B^2 \asymp \|w\|_{L_2(0,1)}^2 + \sum_{i=1}^{L} w^2(\xi_i)$.

4. Problem with dynamical boundary condition. Let us consider the initial-boundary value problem of string vibrations with dynamical boundary condition at $x = 0$ (see [19]):

$$c(x)\frac{\partial^2 u}{\partial t^2} - \frac{\partial}{\partial x}\Big(a(x)\frac{\partial u}{\partial x}\Big) = f(x,t), \quad x \in (0,1), \quad 0 < t < T, \tag{4.1}$$

$$K\frac{\partial^2 u(0,t)}{\partial t^2} = a(0)\frac{\partial u(0,t)}{\partial x}, \quad u(1,t) = 0, \quad 0 < t < T, \tag{4.2}$$

$$u(x,0) = u_0(x), \quad \frac{\partial u(x,0)}{\partial t} = u_1(x), \quad x \in (0,1), \tag{4.3}$$

where, as in section 3, $K > 0$, $0 < c_1 \le a(x) \le c_2$, and $0 < c_3 \le c(x) \le c_4$.

The problem (4.1)–(4.3) can be reduced to a problem of the form (3.1)–(3.3) using even extension of the input data: $c(x) = c(-x)$, $a(x) = a(-x)$, $u_0(x) = u_0(-x)$, $u_1(x) = u_1(-x)$, and $f(x,t) = f(-x,t)$ for $x \in (-1,0)$. It easily follows that the solution $u(x,t)$ can also be extended by even fashion on $(-1,0) \times (0,T)$ and satisfies the problem

$$\big[c(x) + 2K\delta(x)\big]\frac{\partial^2 u}{\partial t^2} - \frac{\partial}{\partial x}\Big(a(x)\frac{\partial u}{\partial x}\Big) = f(x,t), \quad x \in (-1,1), \quad 0 < t < T, \tag{4.4}$$

$$u(-1,t) = 0, \quad u(1,t) = 0, \quad 0 < t < T, \tag{4.5}$$

$$u(x,0) = u_0(x), \quad \frac{\partial u(x,0)}{\partial t} = u_1(x), \quad x \in (-1,1). \tag{4.6}$$

The problem (4.4)–(4.6) can be written in the form (2.1) if one lets $H = L_2(-1,1)$,

$$Au = -\frac{\partial}{\partial x}\Big(a(x)\frac{\partial u}{\partial x}\Big), \qquad \text{and} \qquad Bu = \big[c(x) + 2K\delta(x)\big]u(x,t).$$

If $w(x)$ is an even function on the segment $(-1,1)$, then

$$\|w\|_A^2 = \int_{-1}^{1} a(x)[w'(x)]^2 dx = 2\int_0^1 a(x)[w'(x)]^2 dx,$$

$$\|w\|_B^2 = \int_{-1}^{1} c(x)w^2(x)dx + 2Kw^2(0) = 2\int_0^1 c(x)w^2(x)dx + 2Kw^2(0).$$

Further, we assume that the functions $c(x)$ and $a(x)$ are continuous on $[0,1]$.

By $\widehat{L}_2(0,1) = \widehat{W}_2^0(0,1)$ we denote the subspace of functions $w(x) \in L_2(0,1)$ equipped with scalar product and norm

$$(u,w)_{\widehat{L}_2(0,1)} = \int_0^1 u(x)w(x)dx + u(0)w(0), \qquad \|w\|_{\widehat{L}_2(0,1)} = (u,w)^{1/2}_{\widehat{L}_2(0,1)}.$$

We let $\widehat{W}_2^1(0,1) = \{w \in W_2^1(0,1) : w(1) = 0\}$ and $\widehat{W}_2^k(0,1) = \widehat{W}_2^1(0,1) \cap W_2^k(0,1)$, $k = 2,3,\dots$.

We also define the space $\widehat{W}_2^k(Q)$ $(k = 0,1,2,\dots)$ as the space of functions $w \in W_2^k(Q)$ for which

$$\frac{\partial^i w}{\partial t^i} \in L_2(0,T;\widehat{L}_2(0,1)), \quad i = 0,1,\dots,k,$$

and the norm is defined as follows:

$$\|w\|^2_{\widehat{W}_2^k(Q)} = \|w\|^2_{W_2^k(Q)} + \sum_{i=0}^k \left\| \frac{\partial^i w(\xi,\cdot)}{\partial t^i} \right\|^2_{L_2(0,T)}.$$

The following analogue of Lemma 3.1 holds true.

Lemma 4.1. (i) *Let $a,c \in W_2^2(0,1)$, $f \in W_2^2(Q)$, $f(1,t) = 0$, $u_0 \in \widehat{W}_2^3(0,1)$, $u_1 \in \widehat{W}_2^2(0,1)$, and the compatibility conditions*

$$U_{tt}(1) = 0, \quad a(0)u_0'(0) = KU_{tt}(0)$$

hold, where $U_{tt}(x)$ is defined as in Lemma 3.1. *Then the problem* (4.1)–(4.3) *has a unique solution $u \in \widehat{W}_2^3(Q)$.*

(ii) *Let $a,c \in W_2^3(0,1)$, $f \in W_2^3(Q)$, $f(1,t) = 0$, $u_0 \in \widehat{W}_2^4(0,1)$, $u_1 \in \widehat{W}_2^3(0,1)$, and the compatibility conditions*

$$U_{ttt}(1) = 0, \quad a(0)u_0'(0) = KU_{ttt}(0)$$

hold, where $U_{ttt}(x)$ is defined as in Lemma 3.1. *Then the problem* (4.1)–(4.3) *has a unique solution $u \in \widehat{W}_2^4(Q)$.*

On the segment $[0,1]$ we introduce the nonuniform mesh $\bar{\omega}_h$. Let $\widehat{H}_h$ be the space of mesh functions, equal to zero at $x = 1$. We will use the following scalar product

$$[v,w)_h = \frac{h_1}{2}v(0)w(0) + \sum_{x\in\omega_h} v(x)w(x)\hbar,$$

and the corresponding norm $|[w\|_h = |[w\|_{L_{2,h}} = [w,w)_h^{1/2}$. We also define the mesh norms

$$|[w\|^2_{\widehat{L}_{2,h}} = |[w\|^2_{L_{2,h}} + w^2(0), \qquad |[w\|^2_{\widehat{W}^1_{2,h}} = \|w_{\bar{x}}]|^2_{h\star} + |[w\|^2_h,$$

$$|[v\|^{(0)}_{h\tau} = \max_{t\in\omega_\tau^-} \left| \left[\frac{(v(\cdot,t+\tau) + v(\cdot,t))}{2} \right\|_{\widehat{L}_{2,h}} \right.,$$

$$|[v\|_{h\tau}^{(1)} = \max_{t\in\omega_\tau^-}\left[\left|\left[\frac{(v(\cdot,t+\tau)+v(\cdot,t))}{2}\right\|^2_{\widehat{W}^1_{2,h}} + |[v_t(\cdot,t)\|^2_{\widehat{L}_{2,h}}\right]^{1/2}.$$

We approximate the problem (4.1)–(4.3) by the difference scheme

$$(c+2K\delta_h)v_{t\bar t} - (\tilde a v_{\bar x}^{(\sigma)})_{\hat x} = T_x^2T_t^2 f, \quad (x,t)\in\omega_h^-\times\omega_\tau, \tag{4.7}$$

$$v(1,t)=0, \quad t\in\bar\omega_\tau, \quad v(x,0)=u_0(x), \quad x\in\omega_h^-, \tag{4.8}$$

(4.9)
$$(c+2K\delta_h)v_t(x,0) = T_x^2(cu_1) + \delta_h\big(2Ku_1+\tau a u_0'\big) + \frac{\tau}{2}T_x^2\left[\widetilde{T}_t^2 f(x,0) + (au_0')'\right],$$

$$x\in\omega_h^-,$$

where $\sigma\ge 1/4$, $\delta_h(0)=1/h_1$,

$$(\tilde a v_{\bar x})_{\hat x}|_{x=0} = \frac{2}{h_1}(\tilde a v_{\bar x})\Big|_{x=x_1} \quad \text{and} \quad T_x^2 f(0,t) = \frac{2}{h_1}\int_0^{x_1}\Big(1-\frac{x'}{h_1}\Big)f(x',t)dx'.$$

We also consider the higher order difference scheme

(4.10)
$$(c+2K\delta_h)v_{t\bar t} + \Big(\frac{h^2}{6}(cv)_{\bar x t\bar t}\Big)_{\hat x} - (\tilde a v_{\bar x}^{(\sigma)})_{\hat x} - \theta\frac{h_+-h}{6}\big(a_x v_{\bar x\hat x}^{(\sigma)} - a_{\bar x\hat x}v_{\bar x}^{(\sigma)}\big) = T_x^2T_t^2 f,$$

$$(x,t)\in\omega_h^-\times\omega_\tau, \quad \theta(0)=0, \quad \theta(x)=1 \text{ for } x\in\omega_h,$$

(4.11)
$$(c+2K\delta_h)v_t(x,0) = T_x^2(cu_1) - \Big(\frac{h^2}{6}(cu_1)_{\bar x}\Big)_{\hat x} + \delta_h\big(2Ku_1+\tau a u_0'\big) + \frac{\tau}{2}T_x^2\big[f(x,0)+(au_0')'\big],$$

$$x\in\omega_h^-,$$

where the boundary condition at $x=1$ and the first initial condition are approximated by (4.8).

Using results obtained in section 3, we immediately obtain the following result.

THEOREM 4.2. *If the assumptions of part* (i) *of Lemma* 4.1 *hold, then the difference scheme* (4.7)–(4.9) *converges and the following error bound holds:*

$$|[u-v\|_{h\tau}^{(0)} \le C(h_{max}^2+\tau^2)(\|a\|_{W_2^2(0,1)} + \|c\|_{W_2^2(0,1)})\|u\|_{\widehat{W}_2^3(Q)}.$$

If the assumptions of part (ii) *of Lemma* 4.1 *hold, then the difference scheme* (4.10), (4.8), (4.11) *converges and the following error bound holds:*

$$|[u-v\|_{h\tau}^{(1)} \le C(h_{max}^2+\tau^2)\big(\|a\|_{W_2^3(0,1)} + \|c\|_{W_2^2(0,1)} + 1\big)\|u\|_{\widehat{W}_2^4(Q)}.$$

5. Weakly hyperbolic equation. We consider the initial-boundary value problem

$$\delta(x-\xi)\frac{\partial^2 u}{\partial t^2} - \frac{\partial}{\partial x}\Big(a(x)\frac{\partial u}{\partial x}\Big) = f(x,t), \quad x \in (0,1), \quad 0 < t < T, \tag{5.1}$$

$$u(0,t) = 0, \quad u(1,t) = 0, \quad 0 < t < T, \tag{5.2}$$

$$u(\xi,0) = u_0 = \text{const}, \quad \frac{\partial u(\xi,0)}{\partial t} = u_1 = \text{const}, \tag{5.3}$$

where $0 < c_1 \le a(x) \le c_2$ and $\delta(x)$ is the Dirac-distribution. From (5.1) it follows that the solution at $(x,t) \in Q_1$ and $(x,t) \in Q_2$ satisfies the equation

$$-\frac{\partial}{\partial x}\Big(a(x)\frac{\partial u}{\partial x}\Big) = f(x,t),$$

and at $x = \xi$ the conjugation conditions

$$[u]_{x=\xi} \equiv u(\xi+0,t) - u(\xi-0,t) = 0, \qquad \Big[a\frac{\partial u}{\partial x}\Big]_{x=\xi} = \frac{\partial^2 u(\xi,t)}{\partial t^2}.$$

Therefore, at fixed t, the equation is elliptic on $(0,\xi)$ and $(\xi,1)$, and its hyperbolic character is exhibited only in the point $x = \xi$.

The problem (5.1)–(5.3) also has the form (2.1), where $Au = -\frac{\partial}{\partial x}(a(x)\frac{\partial u}{\partial x})$ and $Bu = \delta(x-\xi)u(x,t)$. The operator A is positively definite in the space $H_A = \overset{\circ}{W}{}_2^1(0,1)$. The operator B is nonnegative in H_A, and

$$\|w\|_B = |w(\xi)|.$$

It is easy to see that in this case the second estimate of Lemma 2.1 in which the operator B^{-1} doesn't participate is valid.

Retaining the notations from section 3, we approximate the problem (5.1)–(5.3) by the weighted difference scheme with averaged right-hand side

$$\delta_h v_{t\bar{t}} - (\tilde{a} v_{\bar{x}}^{(\sigma)})_{\hat{x}} = T_x^2 T_t^2 f, \quad (x,t) \in \omega_h \times \omega_\tau, \tag{5.4}$$

$$v(0,t) = 0, \quad v(1,t) = 0, \quad t \in \bar{\omega}_\tau, \tag{5.5}$$

$$v(\xi,0) = u_0, \quad v_t(\xi,0) = u_1 + \frac{\tau}{2}\Big[a\frac{\partial u}{\partial x}\Big]_{(\xi,0)}. \tag{5.6}$$

At $t = 0$ the problem (5.1)–(5.3) disintegrates in two second order ordinary differential equations

$$-(aU')' = f(x,0), \quad 0 < x < \xi, \qquad U(0) = 0, \quad U(\xi) = u_0$$

and

$$-(aU')' = f(x,0), \quad \xi < x < 1, \qquad U(\xi) = u_0, \quad U(1) = 0,$$

where $U(x) = u(x, 0)$, the solution of which can be derived explicitly. Therefore, we have an effective way to calculate the approximation of the initial boundary condition (5.6):

$$\Big[a\frac{\partial u}{\partial x}\Big]_{(\xi,0)} = [aU']|_{x=\xi} = \int_0^1 f(x,0)dx - \Big(\int_0^\xi \frac{dx}{a(x)}\Big)^{-1}\Big(u_0 + \int_0^\xi \int_0^x \frac{f(x',0)}{a(x)}dx'dx\Big)$$

$$- \Big(\int_\xi^1 \frac{dx}{a(x)}\Big)^{-1}\Big(u_0 + \int_\xi^1 \int_x^1 \frac{f(x',0)}{a(x)}dx'dx\Big).$$

The error $z = u - v$ satisfies the condition

$$\delta_h z_{t\bar{t}} - (\tilde{a} z_{\bar{x}}^{(\sigma)})_{\hat{x}} = \overline{\varphi}_{\hat{x}}, \quad (x,t) \in \omega_h \times \omega_\tau, \tag{5.7}$$

$$z(0,t) = 0, \quad z(1,t) = 0, \quad t \in \bar{\omega}_\tau, \tag{5.8}$$

$$z(\xi, 0) = 0, \quad z_t(\xi, 0) = \mu, \tag{5.9}$$

where (see section 3.3)

$$\overline{\varphi} = \varphi_1 + \varphi_2 + \varphi_3 + \varphi_4 = \tilde{a}u_{\bar{x}} - T_x^- T_t^2\Big(a\frac{\partial u}{\partial x}\Big) \quad \text{and} \quad \mu = \frac{\tau}{2}\Big(\widetilde{T}_t^2 \frac{\partial^2 u}{\partial t^2} - \frac{\partial^2 u}{\partial t^2}\Big)\Big|_{(\xi,0)}.$$

The difference scheme (5.7)–(5.9) takes the form (2.3), where $A_h v = -(\tilde{a}v_{\bar{x}})_{\hat{x}}$ is a positive linear operator in H_h and $B_h v = \delta_h v$ is a nonnegative linear operator in H_h. Also

$$\|w\|_{A_h} = \Big\{\sum_{x\in\omega_h^+} \tilde{a} w_{\bar{x}}^2 \hbar\Big\}^{1/2} \asymp \|w_{\bar{x}}]|_{h\star}, \qquad \|w\|_{B_h} = |w(\xi)|.$$

We also need the norm

$$|v|_{h\tau}^{(0)} = \max_{t\in\omega_\tau^-}\Big|\frac{v(\xi,t) + v(\xi,t+\tau)}{2}\Big|.$$

From Lemma 2.2, using (3.10), we get the a priori estimate

$$|z|_{h\tau}^{(0)} \le C\Big[\|\mu\|_{B_h}^2 + \tau \sum_{t\in\omega_\tau} \|\overline{\varphi}(\cdot,t)]|_{h\star}^2\Big]^{1/2}. \tag{5.10}$$

Since $\varphi = \overline{\varphi}$ at $c(x,t) = 0$, we obtain the estimate of $\overline{\varphi}$ immediately from (3.12). Using the integral representation

$$\mu = \int_0^\tau \int_0^{t'} \Big(1 - \frac{t'}{\tau}\Big)\frac{\partial^3 u(\xi, t'')}{\partial t^3} dt'' dt'$$

and (3.16), we get

$$\|\mu\|_{B_h} = |\mu| \le C\tau^2\Big(\Big\|\frac{\partial^4 u(\xi,\cdot)}{\partial t^4}\Big\|_{L_2(0,T)} + \Big\|\frac{\partial^3 u(\xi,\cdot)}{\partial t^3}\Big\|_{L_2(0,T)}\Big). \tag{5.11}$$

Thus from (5.10), (3.12), and (5.11) we get the following convergence rate estimate for the difference scheme (5.4)–(5.6):

$$|z|_{h\tau}^{(0)} \leq C(h_{max}^2 + \tau^2)(\|a\|_{W_2^2(0,\xi)} + \|a\|_{W_2^2(\xi,1)} + 1)\left(\left\|\frac{\partial^4 u(\xi,\cdot)}{\partial t^4}\right\|_{L_2(0,T)}\right.$$

(5.12)

$$\left. + \left\|\frac{\partial^3 u(\xi,\cdot)}{\partial t^3}\right\|_{L_2(0,T)} + \left\|\frac{\partial^3 u}{\partial x \partial t^2}\right\|_{L_2(Q)} + \left\|\frac{\partial^2 u}{\partial x^2}\right\|_{L_2(Q_1)} + \left\|\frac{\partial^2 u}{\partial x^2}\right\|_{L_2(Q_2)} + \left\|\frac{\partial u}{\partial x}\right\|_{L_2(Q)}\right).$$

The estimate (5.12) guarantees the convergence only at $x = \xi$. However, with its help the error can be estimated in all nodes of the mesh $\omega_h \times \omega_\tau$. In fact, $z^{(\sigma)}$ at $0 \leq x \leq \xi$ satisfies the conditions

$$-(z_{\bar{x}}^{(\sigma)})_{\hat{x}} = \overline{\varphi}_{\hat{x}}, \quad z^{(\sigma)}(0,t) = 0, \quad z^{(\sigma)}(\xi,t) \neq 0.$$

Applying the maximum principle [16], we get

$$\max_{x \in \omega_h \cap [0,\xi]} |z^{(\sigma)}(x,t)| \leq C\Big(\|\overline{\varphi}(\cdot,t)]|_{h\star} + |z^{(\sigma)}(\xi,t)|\Big). \tag{5.13}$$

A similar a priori estimate also holds for $\xi \leq x \leq 1$. At $\sigma = 1/4$ from (3.12), (3.16), (5.12), and (5.13) follows the convergence rate estimate

$$\max_{(x,t) \in \omega_h \times \omega_\tau} |z^{(1/4)}(x,t)| \leq C(h_{max}^2 + \tau^2)(\|a\|_{W_2^2(0,\xi)} + \|a\|_{W_2^2(\xi,1)} + 1)\|u\|_{\widetilde{W}_2^4(Q)}.$$

REFERENCES

[1] I. Braianov, *Convergence of a Crank–Nicolson difference scheme for heat equation with interface in the heat flow and concentrated heat capacity,* Lecture Notes in Comput. Sci., 1196 (1997), pp. 58–65.

[2] I. Braianov and L. Vulkov, *Finite difference schemes with variable weights for parabolic equations with concentrated capacity,* Notes Numer. Fluid Dynam. 62, Vieweg, Braunschweig, Germany, 1998, pp. 208–216.

[3] I. A. Braianov and L. G. Vulkov, *Homogeneous difference schemes for the heat equation with concentrated capacity,* Comput. Math. Math. Phys., 39 (1999), pp. 254–261 (in Russian).

[4] B. S. Jovanović, *Finite difference method for boundary value problems with weak solutions,* Technical report, Posebna izdanja Mat. Instituta 16, Belgrade, Yugoslavia, 1993.

[5] B. S. Jovanović, *Convergence of a finite-difference scheme for hyperbolic equations with variable coefficients,* Z. Angew. Math. Mech., 72 (1992), pp. 493–496.

[6] B. S. Jovanović, L. D. Ivanović, and E. E. Süli, *Convergence of a finite difference scheme for second-order hyperbolic equations with variable coefficients,* IMA J. Numer. Anal., 7 (1987), pp. 39–45.

[7] B. S. Jovanović, J. D. Kandilarov, and L. G. Vulkov, *Construction and convergence of a difference scheme for a model elliptic equation with Dirac-delta function coefficient,* Lecture Notes in Comput. Sci., 1988 (2001), pp. 431–438.

[8] B. S. Jovanović, P. P. Matus, and V. S. Shcheglik, *Convergence rate estimates for parabolic problems with variable coefficients and generalized solutions,* Sib. Zh. Vychisl. Mat., 2 (1999), pp. 123–136 (in Russian).

[9] B. Jovanović and L. Vulkov, *On the convergence of finite difference schemes for the heat equation with concentrated capacity,* Numer. Math., 89 (2001), pp. 715–734.

[10] B. S. Jovanović and L. G. Vulkov, *Operator's approach to the problems with concentrated factors,* Lecture Notes in Comput. Sci., 1988 (2001), pp. 439–450.

[11] B. S. Jovanović and L. G. Vulkov, *On the convergence of difference schemes for the string equation with concentrated mass,* in Proceedings of the 4th International Conference on Finite-Difference Schemes: Theory and Applications, R. Ciegis, A. Samarskiĭ, and M. Sapagovas, eds., IMI, Vilnius, Lithuania, 2000, pp. 107–116.

[12] R. D. Lazarov, V. L. Makarov, and A. A. Samarskiĭ, *Applications of exact difference scheme for construction and studies of difference schemes on generalized solutions,* Math. Sbornik, 117 (1982), pp. 469–480 (in Russian).

[13] J. L. Lions and E. Magenes, *Non Homogeneous Boundary Value Problems and Applications,* Springer-Verlag, Berlin, New York, 1972.

[14] L. A. Oganesyan and L. A. Rukhovets, *Variational-Difference Methods for Solution of Elliptic Equations,* AS Arm., Erevan, Armenia, 1979 (in Russian).

[15] F. Riesz and B. Sz.–Nagy, *Leçons d'analyse fonctionelle,* Akadémiai Kiadó, Budapest, 1972.

[16] A. A. Samarskiĭ, *Theory of Difference Schemes,* Nauka, Moscow, 1989 (in Russian).

[17] A. A. Samarskiĭ, R. D. Lazarov, and V. L. Makarov, *Difference Schemes for Differential Equations with Generalized Solutions,* Vyshaya Shkola, Moscow, 1987 (in Russian).

[18] A. A. Samarskiĭ, V. I. Mazhukin, D. A. Malafeĭ, and P. P. Matus, *Difference schemes of high order of approximation on non uniform in space meshes,* Dokl. RAN (Russian Academy of Sciences), 36 (1999), pp. 1–4 (in Russian).

[19] A. N. Tikhonov and A. A. Samarskiĭ, *Equations of Mathematical Physics,* GITTL, Moscow, 1953 (in Russian).

[20] V. S. Vladimirov, *Equations of Mathematical Physics,* Nauka, Moscow, 1988 (in Russian).

[21] L. Vulkov, *Application of Steklov-type eigenvalues problems to convergence of difference schemes for parabolic and hyperbolic equations with dynamical boundary conditions,* Lecture Notes in Comput. Sci., 1196 (1997), pp. 557–564.

[22] J. Wloka, *Partial Differential Equations,* Cambridge University Press, Cambridge, UK, 1987.

SIAM J. Numer. Anal.
Vol. 41, No. 2, pp. 539–562

H-CONVERGENCE AND NUMERICAL SCHEMES FOR ELLIPTIC PROBLEMS*

ROBERT EYMARD[†] AND THIERRY GALLOUËT[‡]

Abstract. We study the convergence of two coupled numerical schemes, which are a discretization of a so-called elliptic-hyperbolic system. Only weak convergence properties are proved on the discrete diffusion of the elliptic problem, and an adaptation of the H-convergence method gives a convergence property of the elliptic part of the scheme. The limit of the approximate solution is then the solution of an elliptic problem, the diffusion of which is not in the general case the H-limit of the discrete diffusion. In a particular case, a kind of weak limit is then obtained for the hyperbolic equation.

Key words. H-convergence, finite volume schemes, two-phase flow, porous media

AMS subject classifications. 35K65, 35K55

PII. S0036142901397083

1. Introduction.

Numerical simulation takes an important place in oil recovery engineering. In many cases, the engineer should represent at the same time the thermodynamical evolution of the hydrocarbon components during the pressure drop due to the extraction of oil and the mass transfers in the oil reservoir. In this paper, we focus on the consequences of a mobility contrast between an injected fluid (generally water) and the oil in place, in a very simple case: oil and water are assumed to be incompressible immiscible fluid phases with a common pressure, and the reservoir is supposed to be a horizontal homogeneous isotropic domain. Following [3], the conservation equations for such a two-phase flow in this particular case, using Darcy's law, can be written as

$$\left.\begin{aligned} &\frac{\partial s}{\partial t} - \operatorname{div}(\gamma(s)\lambda(s)\nabla u) \\ &\quad = (\bar f)^+\gamma(\bar s) - (\bar f)^-\gamma(s), \\ &\frac{\partial(1-s)}{\partial t} - \operatorname{div}((1-\gamma(s))\lambda(s)\nabla u) \\ &\quad = (\bar f)^+(1-\gamma(\bar s)) - (\bar f)^-(1-\gamma(s)) \end{aligned}\right\} \text{ in } \Omega, \tag{1}$$

with the boundary conditions

$$\begin{aligned} &u = 0 \text{ on } \partial\Omega \times \mathbb{R}_+, \\ &s = \hat s \text{ on } \{(x,t) \in \partial\Omega \times \mathbb{R}_+, \nabla u(x,t)\cdot \mathbf{n}_{\partial\Omega}(x) \ge 0\}. \end{aligned} \tag{2}$$

In (1) and (2), the domain Ω represents the porous medium, u represents the common pressure of the two phases, s represents the saturation of the water phase, $\gamma(s)$ is a nondecreasing function which is called the "fractional flow," with $\gamma(0) = 0$ and $\gamma(1) = 1$, the positive function $\lambda(s)$ is the "total mobility" of the two phases (the sum of the mobility of water and the mobility of oil), the function $\bar f(x,t)$ represents the rates at

[*] Received by the editors October 26, 2001; accepted for publication (in revised form) October 9, 2002; published electronically April 23, 2003.
http://www.siam.org/journals/sinum/41-2/39708.html

[†] Université de Marne-la-Vallée, 5 Boulevard Descartes, Champs-sur-Marne, 77454 Marne-la-vallée Cedex 2, France (eymard@math.univ-mlv.fr).

[‡] Université Aix-Marseille 1, 13453 Marseille Cedex 13, France (gallouet@cmi.univ-mrs.fr).

the wells, $\bar{s}(x,t)$ is the saturation of the injected fluids (the injected rate corresponds to the positive part of the function $\bar{f}$, the produced rate corresponds to the negative part, and the repartition of the production between water and oil is determined by the saturation in the reservoir), and the function $\hat{s}(x,t)$ is the saturation of incoming fluids at the boundary. We denote, for all real value z, $z^+ = \max(z,0)$ and $z^- = \max(-z,0)$.

The existence of a solution to (1) is an open problem if the function λ is not reduced to a constant. A number of numerical schemes for this problem have already been discussed in the literature. Nevertheless, their convergence has only recently been studied in the only case of a constant function λ: the convergence of a numerical scheme involving a finite volume method for the computation of the saturation s and a standard finite element for the computation of the pressure u is proved in [7], whereas a convergence proof for a finite volume method for the discretization of both equations is presented in [21] and a convergence proof for a mixed finite element/finite volume scheme is given in [15].

The objective of this paper is the study of the convergence properties of finite volume methods in the case where the function λ is not a constant function. This problem appears to be very close to the study of the convergence, when $\varepsilon \longrightarrow 0$ of the solution of the problem

$$\left.\begin{aligned} &\frac{\partial s_\varepsilon}{\partial t} - \operatorname{div}(\gamma(s_\varepsilon)\lambda(s_\varepsilon)\nabla u_\varepsilon) - \varepsilon\Delta s_\varepsilon \\ &\qquad = (\bar{f})^+\gamma(\bar{s}) - (\bar{f})^-\gamma(s_\varepsilon), \\ &\frac{\partial(1-s_\varepsilon)}{\partial t} - \operatorname{div}((1-\gamma(s_\varepsilon))\lambda(s_\varepsilon)\nabla u_\varepsilon) + \varepsilon\Delta s_\varepsilon \\ &\qquad = (\bar{f})^+(1-\gamma(\bar{s})) - (\bar{f})^-(1-\gamma(s_\varepsilon)) \end{aligned}\right\} \text{ in } \Omega, \tag{3}$$

where the additional term $\varepsilon\Delta s_\varepsilon$ stands for a diffusive term, which is similar to the diffusion added by the upstream weighted numerical schemes. Such a diffusive term is slightly different from that which comes from the introduction of a capillary pressure term, yielding some degeneration similar to that of the porous media equation (see [1], [2], [4], [11], and [5] for the existence of a solution of the continuous problem and see [10] for the proof of the convergence of a finite volume scheme).

In order to make clear the tools that appear, we shall consider a steady-state version of (1) (see (50) below). The main result of this paper is the proof that, using a coupled finite volume scheme for the approximation of this system of equations, the approximate pressure converges in $L^2(\Omega)$ to the solution of an elliptic problem whose coefficients are obtained by the same method as the classical H-convergence proof (following [17], [14], or [19]), whereas the approximate saturation converges only in a weak sense (namely, in $L^\infty(\Omega)$ for the weak $\star$ topology [13]). The use, in the discrete setting, of a notion similar to H-convergence is natural: indeed, the existence of a limit as $\varepsilon \longrightarrow 0$ to the family of pressures $(u_\varepsilon)_{\varepsilon>0}$, the solution to the sum of a steady version of the equations (3), immediately results from H-convergence (see section 2). Note that an extension of the H-convergence background to a discrete setting has been performed; see [16] and mainly [12] for the proof of the existence of an "H-limit" to a subsequence of a sequence of discrete elliptic operators, using regular structured grids and finite differences. The objective here is to study the limit of a sequence of finite volume approximations on general meshes, whereas the discrete diffusion results from the coupling of the two discrete conservation equations. The fact that the two unknowns are computed in the same grids makes different, in the general case, the notion of continuous and discrete H-limits, which suggests to distinguish the vocabulary devoted to both notions.

It is also interesting to notice that the question of the independence of these limit coefficients on the way that some diffusion is added in (1) is not known. It is, however, clear that the limit (u, s) of the numerical scheme or of the parabolic regularization (namely (3)) is a solution of (1) if a strong convergence result can be proved for the saturation. This sufficient condition seems to be necessary for a large class of data: for example, it is already necessary in the case of a constant function λ when the function γ is genuinely nonlinear.

This paper is organized as follows:

- In section 2, a short review of the concept of H-convergence is made, and some examples of application of this notion are given.
- In section 3, results are recalled on finite volume methods for elliptic problems.
- In section 4, an adaptation of H-convergence to the study of the convergence of numerical schemes for elliptic problems is made.
- The convergence study of the coupled scheme for the two-phase flow problem is done in section 5.
- Some concluding remarks give guidelines for further works.

2. Some results of H-convergence.

The notion of H-convergence is used for the physical description of effective properties, at the macroscopic level, of heterogeneous materials in which some diffusive phenomena occur. The assumption which is then made is that the scale of the heterogeneities is small compared to the macroscopic scale. Let us take the example of the Dirichlet problem, which models, for example, the steady flow of a monophasic incompressible fluid in a heterogeneous porous medium, using Darcy's law. We assume that the pressure of the fluid is constant at the boundary of the domain and that some volumic source terms represent the injection and the production of fluid throughout some wells. The question of the existence of an "effective" permeability field, which could allow the computation of an accurate approximate solution using only a coarse discretization (which means a discretization at the macroscopic scale), is of major interest for the industrial applications; this question can, in some cases, be handled using the notion of H-convergence.

2.1. Notations for the Dirichlet problem.

Let Ω be an open bounded subset of $\mathbb{R}^N$, with $N \in \mathbb{N}_*$, and let α and β be two real numbers, with $0 < \alpha \leq \beta$. We denote by $\mathcal{M}(\alpha, \beta, \Omega)$ the set of measurable functions $M \ : \Omega \longrightarrow \mathcal{L}(\mathbb{R}^N, \mathbb{R}^N)$ such that, for a.e. $x \in \Omega$ and for all $(\xi, \chi) \in (\mathbb{R}^N)^2$, $\alpha|\xi|^2 \leq M(x)\xi \cdot \xi \leq \beta|\xi|^2$, and $M(x)\xi \cdot \chi = \xi \cdot M(x)\chi$. In the particular case where there exists a function $\mu \in L^\infty(\Omega)$ such that, for a.e. $x \in \Omega$, $M(x) = \mu(x)I_N$, where I_N denotes the identity application from $\mathbb{R}^N$ to $\mathbb{R}^N$, we then denote $M = \mu$. In this case, we say that M represents an isotropic field; otherwise, we say that the field M is anisotropic.

For a given source term $b \in H^{-1}(\Omega)$ and a given $M \in \mathcal{M}(\alpha, \beta, \Omega)$, we denote by $\mathcal{F}(b, M)$ the unique solution $\overline{u}$ of

$$\overline{u} \in H_0^1(\Omega) \text{ and } \int_\Omega M(x)\nabla\overline{u}(x) \ \cdot \nabla\bar{v}(x)dx = b(\bar{v}) \ \forall \bar{v} \in H_0^1(\Omega).$$

2.2. The H-convergence theorem.

The following result, given in [17] (in which it was called G-convergence, in reference to some works of de Giorgi), has been extended in [19] to some more general configurations.

Theorem 1 (H-convergence). *Let Ω be an open bounded subset of $\mathbb{R}^N$, with $N \in \mathbb{N}_*$. Let two real numbers α and β be such that $0 < \alpha \leq \beta$. Let $(M_n)_{n\in\mathbb{N}}$ be a sequence of elements of $M_n \in \mathcal{M}(\alpha, \beta, \Omega)$.*

Then there exists a subsequence of $(M_n)_{n\in\mathbb{N}}$, again denoted $(M_n)_{n\in\mathbb{N}}$, and a function $M \in \mathcal{M}(\alpha,\beta,\Omega)$ such that

- *for all $b \in H^{-1}(\Omega)$, $\mathcal{F}(b, M_n)$ weakly converges to $\mathcal{F}(b, M)$ in $H^1_0(\Omega)$ as $n \longrightarrow \infty$;*
- *for all $b \in H^{-1}(\Omega)$, $M_n\nabla\mathcal{F}(b, M_n)$ weakly converges to $M\nabla\mathcal{F}(b, M)$ in $(L^2(\Omega))^N$ as $n \longrightarrow \infty$.*

We then say that the sequence $(M_n)_{n\in\mathbb{N}}$ H-converges to M, called the H-limit of the sequence.

We now give some examples of H-convergence results.

2.3. The one-dimensional case. In the case $N = 1$, let us suppose that $\Omega = (0, 1)$. The sequence $(M_n)_{n\in\mathbb{N}}$ such that for all $n \in \mathbb{N}$, $M_n \in \mathcal{M}(\alpha,\beta,\Omega)$ is then a sequence of functions belonging to $L^\infty(\Omega)$ and $1/M_n(x) \in [1/\beta, 1/\alpha]$ for a.e. $x \in \Omega$. For a given $f \in L^2(\Omega)$, we denote by $\hat{f}$ the continuous function defined, for all $x \in (0, 1)$, by $\hat{f}(x) = \int_{(0,x)} f(s)ds$. We then have, for all $x \in \Omega$,

$$\mathcal{F}(f, M_n)(x) = \frac{\int_{(0,x)}(1/M_n(t))dt}{\int_{(0,1)}(1/M_n(t))dt}\int_{(0,1)}\frac{\hat{f}(t)}{M_n(t)}dt - \int_{(0,x)}\frac{\hat{f}(t)}{M_n(t)}dt.$$

Up to a subsequence, we can suppose that the sequence $(1/M_n)_{n\in\mathbb{N}}$ converges to a function $1/M$ for the weak $\star$ topology of $L^\infty(\Omega)$. We then get that, for all $x \in \Omega$,

$$\lim_{n\longrightarrow\infty}\mathcal{F}(f, M_n)(x) = \frac{\int_{(0,x)}(1/M(t))dt}{\int_{(0,1)}(1/M(t))dt}\int_{(0,1)}\frac{\hat{f}(t)}{M(t)}dt - \int_{(0,x)}\frac{\hat{f}(t)}{M(t)}dt,$$

which proves that M is the H-limit of this subsequence. Unfortunately, such a relation between the limit for the weak $\star$ topology of $L^\infty(\Omega)$ and the H-limit cannot be obtained in the general case $N > 1$.

2.4. Two-dimensional examples. Let $\mu_r > 0$ and $\mu_b > 0$ be two real values, respectively, defining the permeability of two materials, respectively called "red" and "black." We first define the so-called checkerboard problem, setting $M_1 : \mathbb{R}^2 \longrightarrow \mathbb{R}$ by $(x_1, x_2) \rightarrow \mu_r$ if $\text{Int}(x_1)+\text{Int}(x_2) \in 2\mathbb{Z}$ (denoting for all $z \in \mathbb{R}$ by $\text{Int}(z)$ the largest relative integer value lower than z), else $(x_1, x_2) \rightarrow \mu_b$ (for example, $\text{Int}(0.5)+\text{Int}(0.5) = 0$ and $M_1(0.5, 0.5) = \mu_r$, $\text{Int}(1.5)+\text{Int}(-1.5) = 1 - 2 = -1$ and $M_1(1.5, -1.5) = \mu_b$; see Figure 1). Then we define, for all $n \in \mathbb{N}_*$, $M_n : \mathbb{R}^2 \longrightarrow \mathbb{R}$ by $M_n(x_1, x_2) = M_1(nx_1, nx_2)$. It can then be shown that, in all open domain Ω of $\mathbb{R}^2$, the sequence $(M_n)_{n\in\mathbb{N}}$ H-converges to the constant function $(x_1, x_2) \rightarrow \sqrt{\mu_r\mu_b}$. In this case, the H-limit of a sequence of isotropic heterogeneous fields is an isotropic homogeneous field. Another example involving two materials is the multilayer case, obtained with defining $M_1 : \mathbb{R}^2 \longrightarrow \mathbb{R}$ by $(x_1, x_2) \rightarrow \mu_r$ if $\text{Int}(x_1) \in 2\mathbb{Z}$, else $(x_1, x_2) \rightarrow \mu_b$ (for example, $\text{Int}(0.5) = 0$ and $M_1(0.5, 10) = \mu_r$, $\text{Int}(1.5) = 1$ and $M_1(1.5, -4) = \mu_b$; see Figure 2). We again define the sequence $(M_n)_{n\in\mathbb{N}}$, by $M_n : \mathbb{R}^2 \longrightarrow \mathbb{R}$, $(x_1, x_2) \mapsto M_1(nx_1, nx_2)$, for all $n \in \mathbb{N}_*$. Then it can be proved that the sequence $(M_n)_{n\in\mathbb{N}}$ H-converges, in all open domain Ω of $\mathbb{R}^2$, to the constant field, the value of which is the linear function defined by $(1, 0) \rightarrow (\frac{2\mu_r\mu_b}{\mu_r+\mu_b}, 0)$ and $(0, 1) \rightarrow (0, \frac{\mu_r+\mu_b}{2})$. We can remark that $\frac{2\mu_r\mu_b}{\mu_r+\mu_b}$ is the harmonic average of μ_r and μ_b, that is, the invert of the average value of the inverts of μ_r and μ_b (this is exactly the value obtained by H-convergence in the one-dimensional case), whereas $\frac{\mu_r+\mu_b}{2}$ is the arithmetic average of μ_r and μ_b. In this two-dimensional case, the H-limit of a sequence of isotropic heterogeneous fields is an anisotropic homogeneous field.

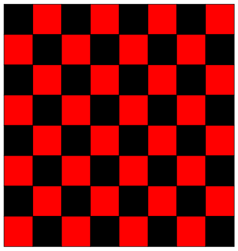


FIG. 1. *The checkerboard case.*

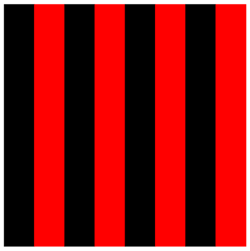


FIG. 2. *The multilayer case.*

Note that in the two above examples the limit of $(M_n)_{n\in\mathbb{N}}$ for the weak $\star$ topology of $L^\infty(\Omega)$ is the constant function $(x_1, x_2) \to \frac{\mu_r+\mu_b}{2}$. Using the notion of nonlinear weak $\star$ convergence (see [9]), the limit of $(M_n)_{n\in\mathbb{N}}$ in terms of Young's measure is the constant field of probability measure $\frac{1}{2}\delta_{\mu_r} + \frac{1}{2}\delta_{\mu_b}$, equivalently given by the function $\mu \in L^\infty(\Omega \times (0,1))$ such that, for a.e. $x \in \Omega$ and $s \in (0, \frac{1}{2})$, $\mu(x,s) = \mu_r$, and for a.e. $x \in \Omega$ and $s \in (\frac{1}{2}, 1)$, $\mu(x,s) = \mu_b$. Thus we see that the notion of nonlinear weak $\star$ convergence does not account for the spatial structure of the heterogeneity and justifies the attempts of finding some more suitable generalized limit (see, for example, [20]).

3. Finite volume meshes and schemes.

3.1. Admissible meshes. We first introduce the notion of admissible discretization [9] which is useful to define a finite volume scheme.

DEFINITION 1 (admissible discretization). *Let Ω be an open bounded polygonal subset of $\mathbb{R}^N$, with $N \in \mathbb{N}_*$ (in general, we have $N = 2$ or $N = 3$). We denote $\partial\Omega = \overline{\Omega} \setminus \Omega$. An admissible finite volume discretization of Ω, denoted by $\mathcal{D}$, is given by $\mathcal{D} = (\mathcal{T}, \mathcal{E}, \mathcal{P})$, where we have the following:*

- *$\mathcal{T}$ is a finite family of nonempty open polygonal convex disjoint subsets of Ω (the "control volumes") such that $\overline{\Omega} = \cup_{K\in\mathcal{T}}\overline{K}$. We then denote, for all $K \in \mathcal{T}$, by $\partial K = \overline{K} \setminus K$ the boundary of K and $m_K > 0$ the N-dimensional Lebesgue measure of K (it is the area of K in the two-dimensional case and the volume in the three-dimensional case).*
- *$\mathcal{E}$ is a finite family of disjoint subsets of $\overline{\Omega}$ (the "edges" of the mesh) such that, for all $\sigma \in \mathcal{E}$, there exists a hyperplane E of $\mathbb{R}^N$ and $K \in \mathcal{T}$ with $\overline{\sigma} = \partial K \cap E$ and σ is a nonempty open subset of E. We then denote $m_\sigma > 0$ the $(N-1)$-dimensional measure of σ. We assume that, for all $K \in \mathcal{T}$, there exists a subset $\mathcal{E}_K$ of $\mathcal{E}$ such that $\partial K = \cup_{\sigma\in\mathcal{E}_K}\overline{\sigma}$. It then results from the previous*

hypotheses that, for all $\sigma \in \mathcal{E}$, either $\sigma \subset \partial\Omega$ or there exists $(K, L) \in \mathcal{T}^2$ with $K \neq L$ such that $\overline{K} \cap \overline{L} = \overline{\sigma}$; we denote in the latter case $\sigma = K|L$.

- *$\mathcal{P}$ is a family of points of Ω indexed by $\mathcal{T}$, denoted by $\mathcal{P} = (x_K)_{K\in\mathcal{T}}$. This family is such that, for all $K \in \mathcal{T}$, $x_K \in K$. For all $\sigma \in \mathcal{E}$ such that there exists $(K, L) \in \mathcal{T}^2$ with $\sigma = K|L$, it is assumed that the straight line (x_K, x_L) going through x_K and x_L is orthogonal to $K|L$. For all $K \in \mathcal{T}$ and all $\sigma \in \mathcal{E}_K$, let y_σ be the orthogonal projection of x_K on σ. We suppose that $y_\sigma \in \sigma$.*

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

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

For all $K \in \mathcal{T}$ and $\sigma \in \mathcal{E}_K$, we denote by $\mathbf{n}_{K,\sigma}$ the unit vector normal to σ outward to K. We define a subset of K associated with the edge σ by

$$D_{K,\sigma} = \{tx_K + (1-t)y,\ t \in (0,1),\ y \in \sigma\}$$

(the letter "D" stands for "diamond") and denote by $d_{K,\sigma}$ the euclidean distance between x_K and σ. We then define

$$\tau_{K,\sigma} = \frac{m_\sigma}{d_{K,\sigma}}.$$

The set of interior (resp., boundary) edges is denoted by $\mathcal{E}_{\text{int}}$ (resp., $\mathcal{E}_{\text{ext}}$), that is, $\mathcal{E}_{\text{int}} = \{\sigma \in \mathcal{E};\ \sigma \not\subset \partial\Omega\}$ (resp., $\mathcal{E}_{\text{ext}} = \{\sigma \in \mathcal{E};\ \sigma \subset \partial\Omega\}$).

3.2. Discrete functional properties.

Definition 2. *Let Ω be an open bounded polygonal subset of $\mathbb{R}^N$, with $N \in \mathbb{N}_*$. Let $\mathcal{D} = (\mathcal{T}, \mathcal{E}, \mathcal{P})$ be an admissible finite volume discretization of Ω in the sense of Definition 1. We denote by $H_{\mathcal{D}}(\Omega) \subset L^2(\Omega)$ the space of functions which admit a constant value in each $K \in \mathcal{T}$. For all $u \in H_{\mathcal{D}}(\Omega)$ and for all $K \in \mathcal{T}$, we denote by u_K the constant value of u in K and we define $(u_\sigma)_{\sigma\in\mathcal{E}}$ by*

$$u_\sigma = 0\ \forall \sigma \in \mathcal{E}_{\text{ext}} \tag{4}$$

and

$$\tau_{K,\sigma}(u_\sigma - u_K) + \tau_{L,\sigma}(u_\sigma - u_L) = 0\ \forall \sigma \in \mathcal{E}_{\text{int}}, \sigma = K|L. \tag{5}$$

We now give a definition of an approximate gradient of the functions of $H_{\mathcal{D}}(\Omega)$. We define a function, denoted $\mathbf{G}_{\mathcal{D}} : H_{\mathcal{D}}(\Omega) \longrightarrow (L^2(\Omega))^N$, $u \longrightarrow \mathbf{G}_{\mathcal{D}}u$ with

$$\mathbf{G}_{\mathcal{D}}u(x) = \frac{N}{d_{K,\sigma}}(u_\sigma - u_K)\mathbf{n}_{K,\sigma}, \quad \textit{for a.e. } x \in D_{K,\sigma}\ \forall K \in \mathcal{T},\ \forall \sigma \in \mathcal{E}_K. \tag{6}$$

Let two real numbers α and β be such that $0 < \alpha \leq \beta$. We denote by $\mathcal{M}_{\mathcal{D}}(\alpha, \beta) \subset L^\infty(\Omega)$ the set of functions μ such that for all $\sigma \in \mathcal{E}$ there exists a constant value, denoted $\mu_\sigma \in [\alpha, \beta]$, such that

$$\mu(x) = \mu_\sigma\ \forall x \in D_{K,\sigma},\ \textit{where } K \textit{ is such that } \sigma \in \mathcal{E}_K.$$

The function which takes the constant value 1 on Ω is denoted by 1. For $(u, v) \in (H_{\mathcal{D}}(\Omega))^2$ and $\varphi \in C^0(\overline{\Omega})$, we denote by

$$[u, v]_{\mathcal{D},\mu,\varphi} = \sum_{K\in\mathcal{T}} \varphi(x_K) \sum_{\sigma\in\mathcal{E}_K} \mu_\sigma \tau_{K,\sigma}(u_\sigma - u_K)(v_\sigma - v_K). \tag{7}$$

We define the following norm in $H_{\mathcal{D}}(\Omega)$ *(see Lemma* 1*) by*

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

Remark 1. *For all edges* σ *such that* $\sigma = K|L$*, the function* $\mathbf{G}_{\mathcal{D}}u$ *is constant on* $D_{K,\sigma} \cup D_{L,\sigma}$.

We have the following properties.

Lemma 1 (discrete Poincaré inequality). *Let* Ω *be an open bounded polygonal subset of* $\mathbb{R}^N$*, with* $N \in \mathbb{N}_*$*. Let* $\mathcal{D} = (\mathcal{T}, \mathcal{E}, \mathcal{P})$ *be an admissible finite volume discretization of* Ω *in the sense of Definition* 1*. Then for all* $u \in H_{\mathcal{D}}(\Omega)$ *(cf. Definition* 2*) one has*

$$\|u\|_{L^2(\Omega)} \le \operatorname{diam}(\Omega)\ |u|_{\mathcal{D}}. \tag{8}$$

The proof of Lemma 1 is given in [9].

Lemma 2 (relative compactness in $L^2(\Omega)$). *Let* Ω *be an open bounded polygonal subset of* $\mathbb{R}^N$*, with* $N \in \mathbb{N}_*$*. We consider a sequence* $(\mathcal{D}_n, u_n)_{n\in\mathbb{N}}$ *such that, for all* $n \in \mathbb{N}$*,* $\mathcal{D}_n$ *is an admissible finite volume discretization of* Ω *in the sense of Definition* 1 *and* $u_n \in H_{\mathcal{D}_n}(\Omega)$ *(cf. Definition* 2*). Let us assume that*

$$\lim_{n\longrightarrow\infty} \operatorname{size}(\mathcal{D}_n) = 0$$

and that there exists $C > 0$ *such that, for all* $n \in \mathbb{N}$*,* $|u_n|_{\mathcal{D}_n} \le C$.

Then there exists a subsequence of $(\mathcal{D}_n, u_n)_{n\in\mathbb{N}}$*, again denoted* $(\mathcal{D}_n, u_n)_{n\in\mathbb{N}}$*, and* $\overline{u} \in H_0^1(\Omega)$ *such that* u_n *tends to* $\overline{u}$ *in* $L^2(\Omega)$ *as* $n \longrightarrow \infty$*,* $\mathbf{G}_{\mathcal{D}_n}u_n$ *weakly tends to* $\nabla\overline{u}$ *in* $(L^2(\Omega))^N$ *as* $n \longrightarrow \infty$*, and*

$$\int_\Omega \varphi(x)(\nabla\overline{u}(x))^2 dx \le \liminf_{n\longrightarrow\infty} [u_n, u_n]_{\mathcal{D}_n,1,\varphi} \quad \forall\varphi \in C^0(\overline{\Omega}, \mathbb{R}_+). \tag{9}$$

Proof. The proof of the existence of a subsequence of $(\mathcal{D}_n, u_n)_{n\in\mathbb{N}}$, again denoted $(\mathcal{D}_n, u_n)_{n\in\mathbb{N}}$, and $\overline{u} \in H_0^1(\Omega)$ such that u_n tends to $\overline{u}$ in $L^2(\Omega)$ as $n \longrightarrow \infty$ is given in [9]. The proof of (9) is given in [10]. Therefore, we have only to prove that, up to a subsequence, $\mathbf{G}_{\mathcal{D}_n}u_n$ weakly tends to $\nabla\overline{u}$ in $(L^2(\Omega))^N$ as $n \longrightarrow \infty$. Since for all $K \in \mathcal{T}$ and $\sigma \in \mathcal{E}_K$ the measure of $D_{K,\sigma}$ is equal to $m_\sigma d_{K,\sigma}/N$, we have

$$\left(\|\mathbf{G}_{\mathcal{D}_n}u_n\|_{(L^2(\Omega))^N}\right)^2 = N\left(|u_n|_{\mathcal{D}_n}\right)^2 \le NC^2. \tag{10}$$

Thus there exists a subsequence of $(\mathcal{D}_n, u_n)_{n\in\mathbb{N}}$, again denoted $(\mathcal{D}_n, u_n)_{n\in\mathbb{N}}$, and $\bar{\mathbf{g}} \in (L^2(\Omega))^N$ such that $\mathbf{G}_{\mathcal{D}_n}u_n$ converges weakly to $\bar{\mathbf{g}}$ in $(L^2(\Omega))^N$ as $n \longrightarrow \infty$. It now remains to prove that $\bar{\mathbf{g}} = \nabla\overline{u}$. Let $\varphi \in (C_c^\infty(\Omega))^N$. Let G_0 be defined by

$$G_0 = \int_\Omega \nabla\overline{u}(x)\cdot\varphi(x)dx = -\int_\Omega \overline{u}(x)\operatorname{div}\varphi(x)dx.$$

For a given $n \in \mathbb{N}$, we consider the expression

$$G_{0,n} = -\int_\Omega u_n \operatorname{div}\varphi(x)dx.$$

We then have $\lim_{n\longrightarrow\infty} G_{0,n} = G_0$. Since we have (omitting indexes n in discrete terms)

$$G_{0,n} = -\sum_{K\in\mathcal{T}} u_K \int_K \operatorname{div}\varphi(x)dx = \sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{E}_K} (u_\sigma - u_K)\int_\sigma \varphi(x)\cdot\mathbf{n}_{K,\sigma}d\gamma(x)$$

(in which $d\gamma(x)$ is the $N-1$-dimensional measure), we get

$$G_{0,n} = \sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}_K} \frac{N}{d_{K,\sigma}}(u_\sigma - u_K) \int_{D_{K,\sigma}} \varphi_\sigma \cdot \mathbf{n}_{K,\sigma} dx,$$

in which we denote

$$\varphi_\sigma = \frac{1}{m_\sigma} \int_\sigma \varphi(x) d\gamma(x) \ \forall \sigma \in \mathcal{E}.$$

If we now set

$$G_{1,n} = \int_\Omega \mathbf{G}_{\mathcal{D}_n} u_n(x) \cdot \varphi(x) dx,$$

we have, on the one hand, $\lim_{n\longrightarrow\infty} G_{1,n} = \int_\Omega \bar{\mathbf{g}}(x) \cdot \varphi(x) dx$ and, on the other hand,

$$G_{1,n} = \sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}_K} \frac{N}{d_{K,\sigma}}(u_\sigma - u_K) \int_{D_{K,\sigma}} \varphi(x) \cdot \mathbf{n}_{K,\sigma} dx.$$

Therefore, denoting $C_{0,\varphi} > 0$ a value such that $|\varphi(x) - \varphi(y)| \le C_{0,\varphi}|x-y|$, we have

$$|G_{1,n} - G_{0,n}| \le C_{0,\varphi} \text{size}(\mathcal{D}_n) \sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}_K} m_\sigma |u_\sigma - u_K|,$$

and thanks to the Cauchy–Schwarz inequality,

$$\begin{aligned} (G_{1,n} - G_{0,n})^2 &\le C_{0,\varphi}^2 \text{size}(\mathcal{D}_n)^2 N m_\Omega \sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}_K} \tau_{K,\sigma}(u_\sigma - u_K)^2 \\ &\le C_{0,\varphi}^2 \text{size}(\mathcal{D}_n)^2 N m_\Omega C, \end{aligned}$$

where m_Ω denotes the measure of Ω which verifies $N m_\Omega = \sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}_K} m_\sigma d_{K,\sigma}$. This proves that $\lim_{n\longrightarrow\infty} G_{0,n} = \lim_{n\longrightarrow\infty} G_{1,n}$, and therefore

$$\int_\Omega \bar{\mathbf{g}}(x) \cdot \varphi(x) dx = \int_\Omega \nabla \overline{u}(x) \cdot \varphi(x) dx.$$

Since the above equation is true for all $\varphi \in (C_c^\infty(\Omega))^N$, we then deduce that $\bar{\mathbf{g}}(x) = \nabla\overline{u}(x)$, for a.e. $x \in \Omega$. Thanks to the uniqueness of this limit, this proves that all of the sequence $(\mathcal{D}_n, u_n)_{n\in\mathbb{N}}$ such that u_n tends to $\overline{u}$ in $L^2(\Omega)$ as $n \longrightarrow \infty$ verifies that $\mathbf{G}_{\mathcal{D}_n} u_n$ weakly tends to $\nabla\overline{u}$ in $(L^2(\Omega))^N$ as $n \longrightarrow \infty$.

Remark 2. *In the preceding proof, the convergence of* $\mathbf{G}_{\mathcal{D}_n} u_n$*, as* $n \longrightarrow \infty$*, cannot be in* $(L^2(\Omega))^N$*, except if it converges to* 0*, since we get* $\liminf_n \|\mathbf{G}_{\mathcal{D}_n} u_n\|_{(L^2(\Omega))^N} \ge \sqrt{N}\|\nabla\overline{u}\|_{(L^2(\Omega))^N}$ *from* (10) *and* (9) *with* $\varphi = 1$ *(see also Remark* 7 *for a more general case).*

3.3. Finite volume scheme. We now give a finite volume scheme for a Dirichlet problem on Ω. Let $\mathcal{D} = (\mathcal{T}, \mathcal{E}, \mathcal{P})$ be an admissible discretization of Ω in the sense of Definition 1. Let two real numbers α and β be such that $0 < \alpha \le \beta$ and let $\mu \in \mathcal{M}_\mathcal{D}(\alpha, \beta)$. For a given $f \in L^2(\Omega)$, let $u \in H_\mathcal{D}(\Omega)$ (cf. Definition 2) be such that

$$-\sum_{\sigma\in\mathcal{E}_K} \mu_\sigma \tau_{K,\sigma}(u_\sigma - u_K) = \int_K f(x) dx \ \forall K \in \mathcal{T} \tag{11}$$

(the existence and uniqueness of a $u \in H_{\mathcal{D}}(\Omega)$ solution of (11) results from the inequality $\mu_\sigma \geq \alpha$ for all $\sigma \in \mathcal{E}$ and from (8); see [9]). Since, for all $v \in H_{\mathcal{D}}(\Omega)$, $\sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}_K} \mu_\sigma \tau_{K,\sigma}(u_\sigma - u_K)v_\sigma = 0$ thanks to (4)–(5), (11) is equivalent to

$$u \in H_{\mathcal{D}}(\Omega) \text{ and } [u,v]_{\mathcal{D},\mu,1} = \int_\Omega f(x)v(x)dx \ \forall v \in H_{\mathcal{D}}(\Omega). \tag{12}$$

Using the results of Lemma 2 for the points concerning the approximate gradient, we then have the following results, given in [9].

LEMMA 3 (finite volume method). *Let Ω be an open bounded polygonal subset of $\mathbb{R}^N$, with $N \in \mathbb{N}_*$. Let $\mathcal{D} = (\mathcal{T}, \mathcal{E}, \mathcal{P})$ be an admissible finite volume discretization of Ω in the sense of Definition* 1. *Let two real numbers α and β be such that $0 < \alpha \leq \beta$ and let $\mu \in \mathcal{M}_{\mathcal{D}}(\alpha, \beta)$. Let $f \in L^2(\Omega)$.*

Then there exists one and only one $u \in H_{\mathcal{D}}(\Omega)$ (cf. Definition 2*) given by* (11). *We then denote $u = F_{\mathcal{D}}(f, \mu)$. Moreover,*

$$\alpha|u|_{\mathcal{D}} \leq \operatorname{diam}(\Omega)\ \|f\|_{L^2(\Omega)}. \tag{13}$$

In the case $\mu = 1$, we have the following convergence results: $F_{\mathcal{D}}(f,1)$ converges to $\mathcal{F}(f,1)$ in $L^2(\Omega)$ as $\text{size}(\mathcal{D}) \longrightarrow 0$, *$\mathbf{G}_{\mathcal{D}}F_{\mathcal{D}}(f,1)$ weakly converges to $\nabla\mathcal{F}(f,1)$ as* $\text{size}(\mathcal{D}) \longrightarrow 0$ *in $(L^2(\Omega))^N$, and*

$$\begin{aligned}&\int_\Omega \varphi(x)(\nabla\mathcal{F}(f,1)(x))^2dx = \lim_{\text{size}(\mathcal{D})\longrightarrow 0} [F_{\mathcal{D}}(f,1), F_{\mathcal{D}}(f,1)]_{\mathcal{D},1,\varphi}\\ &\forall \varphi \in C^0(\overline{\Omega}).\end{aligned} \tag{14}$$

4. Adaptation of H-convergence to numerical schemes.

4.1. The Hd-convergence theorem and relations with H-convergence.

The following theorem (proved in sections 4.2 and 4.3 below) expresses a discrete version of Theorem 1.

THEOREM 2 (Hd-convergence). *Let Ω be an open bounded polygonal subset of $\mathbb{R}^N$, with $N \in \mathbb{N}_*$. Let two real numbers α and β be such that $0 < \alpha \leq \beta$. Let $(\mathcal{D}_n, \mu_n)_{n\in\mathbb{N}}$ be a sequence such that, for all $n \in \mathbb{N}$, $\mathcal{D}_n$ is an admissible discretization of Ω in the sense of Definition* 1, *and $\mu_n \in \mathcal{M}_{\mathcal{D}_n}(\alpha, \beta)$. We assume that* $\lim_{n\longrightarrow\infty} \text{size}(\mathcal{D}_n) = 0$.

Then there exist a subsequence of $(\mathcal{D}_n, \mu_n)_{n\in\mathbb{N}}$, again denoted $(\mathcal{D}_n, \mu_n)_{n\in\mathbb{N}}$, and a unique measurable function $M \in \mathcal{M}(\alpha, \beta, \Omega)$ (this set is defined in section 2*) such that*

- *for all $f \in L^2(\Omega)$, $F_{\mathcal{D}_n}(f, \mu_n)$ converges to $\mathcal{F}(f, M)$ in $L^2(\Omega)$ as $n \longrightarrow \infty$ and $\mathbf{G}_{\mathcal{D}_n}F_{\mathcal{D}_n}(f, \mu_n)$ weakly converges to $\nabla\mathcal{F}(f, M)$ in $(L^2(\Omega))^N$ as $n \longrightarrow \infty$ (the functions $F_{\mathcal{D}}(f, \mu)$, denoting the discrete solution of a finite volume scheme for an elliptic problem with the homogeneous Dirichlet boundary condition, the right-hand side f, and a discrete diffusion field μ, and $\mathbf{G}_{\mathcal{D}}F_{\mathcal{D}}(f, \mu)$, denoting a discrete gradient of this numerical solution, are defined in section* 3 *and the function $\mathcal{F}(f, M)$, denoting the solution of an elliptic problem with the homogeneous Dirichlet boundary condition, the right-hand side f, and a diffusion matrix field M, is defined in section* 2*);*
- *for all $f \in L^2(\Omega)$, $\mu_n\mathbf{G}_{\mathcal{D}_n}F_{\mathcal{D}_n}(f, \mu_n)$ weakly converges to $M\nabla\mathcal{F}(f, M)$ in $(L^2(\Omega))^N$ as $n \longrightarrow \infty$.*

We then say that the sequence $(\mathcal{D}_n, \mu_n)_{n\in\mathbb{N}}$ Hd-converges to M, called the Hd-limit of the sequence.

Some comments can be made on the relation between Hd-convergence and H-convergence. Let us first study the one-dimensional case. We take again the case and the notations of section 2.3. Let $\Omega = (0,1)$, α, and β be such that $0 < \alpha \leq \beta$. In order to define an admissible discretization of Ω, let $p \in \mathbb{N}_*$ and let $(y_k)_{k=0,\ldots,p}$ and $(x_k)_{k=1,\ldots,p}$ be real values such that

$$y_0 = 0 < x_1 < y_1 < x_2 \ldots < y_{k-1} < x_k < y_k \ldots < y_{p-1} < x_p < y_p = 1.$$

Then the discretization $\mathcal{D} = (\mathcal{T}, \mathcal{E}, \mathcal{P})$ defined by $\mathcal{T} = \{(y_{k-1}, y_k),\ k = 1,\ldots,p\}$, $\mathcal{E} = \{\{y_k\},\ k = 0,\ldots,p\}$, and $\mathcal{P} = \{x_k,\ k = 1,\ldots,p\}$ is an admissible discretization of Ω in the sense of Definition 1. Let $f \in L^2(\Omega)$ and $\mu \in \mathcal{M}_{\mathcal{D}}(\alpha,\beta)$ be given (recall that the function μ takes constant values in $(0,x_1)$, $\ldots$, (x_k, x_{k+1}),$\ldots$, $(x_p, 1)$). We again define the function $\hat{f}$ by $\hat{f}(x) = \int_{(0,x)} f(t)dt$ for all $x \in \Omega$, and we introduce the function $\hat{f}_{\mathcal{D}}$ defined by $\hat{f}_{\mathcal{D}}(x) = 0 = \hat{f}(y_0)$ for all $x \in (0,x_1)$, by $\hat{f}_{\mathcal{D}}(x) = \hat{f}(y_k)$ for all $x \in (x_k, x_{k+1})$, and by $\hat{f}_{\mathcal{D}}(x) = \hat{f}(1)$ for all $x \in (x_p, 1)$. Some calculations show that the solution of the finite volume scheme (11) is defined by

$$\begin{aligned}(15) \qquad & F_{\mathcal{D}}(f,\mu)(x) = \frac{\int_{(0,x_k)}(1/\mu(t))dt}{\int_{(0,1)}(1/\mu(t))dt}\int_{(0,1)}\frac{\hat{f}_{\mathcal{D}}(t)}{\mu(t)}dt - \int_{(0,x_k)}\frac{\hat{f}_{\mathcal{D}}(t)}{\mu(t)}dt \\ & \forall x \in (y_{k-1}, y_k), \quad \forall k = 1,\ldots,p.\end{aligned}$$

Let $(\mathcal{D}_n, \mu_n)_{n\in\mathbb{N}}$ be a sequence such that, for all $n \in \mathbb{N}$, $\mathcal{D}_n$ is an admissible discretization of Ω in the sense of Definition 1, and $\mu_n \in \mathcal{M}_{\mathcal{D}_n}(\alpha,\beta)$. We assume that $\lim_{n\longrightarrow\infty} \text{size}(\mathcal{D}_n) = 0$. Up to a subsequence, we can suppose that the sequence $(1/\mu_n)_{n\in\mathbb{N}}$ converges to a function $1/M$ for the weak $\star$ topology of $L^\infty(\Omega)$. Since the sequence $(\hat{f}_{\mathcal{D}_n})_{n\in\mathbb{N}}$ strongly converges to the continuous function $\hat{f}$ as $n \longrightarrow \infty$, we get, using (15) in which we let $\mathcal{D} = \mathcal{D}_n$ and $\mu = \mu_n$, that the limit of the sequence $(F_{\mathcal{D}_n}(f,\mu_n))_{n\in\mathbb{N}}$ is exactly the function $\mathcal{F}(f,M)$ defined, for all $x \in \Omega$, by

$$\mathcal{F}(f,M)(x) = \frac{\int_{(0,x)}(1/M(t))dt}{\int_{(0,1)}(1/M(t))dt}\int_{(0,1)}\frac{\hat{f}(t)}{M(t)}dt - \int_{(0,x)}\frac{\hat{f}(t)}{M(t)}dt.$$

This proves that the Hd-limit of $(\mathcal{D}_n, \mu_n)_{n\in\mathbb{N}}$ is the function M, and therefore coincides, when using the finite volume scheme (11), with the H-limit of $(\mu_n)_{n\in\mathbb{N}}$; the use of some convergence for the weak $\star$ topology of $L^\infty(\Omega)$ is again sufficient to pass to the limit.

REMARK 3. *Note that the coincidence of the discrete and the continuous H-limits is not true for all the one-dimensional numerical schemes which can be associated with the same function μ. Indeed, assume, in order to simplify, that $y_k - y_{k-1} = h$, for $k = 1,\ldots,p$ (with $h = 1/p$), $x_k = (y_k + y_{k-1})/2$, for $k = 1,\ldots,p$, and that the function μ takes constant values in $(0,x_1)$, $\ldots$, (x_k,x_{k+1}), $\ldots$, $(x_p,1)$ which are μ_r and μ_b in alternance. If we discretize the Dirichlet problem with this function μ as the diffusion coefficient and the piecewise linear finite element scheme with nodes located at the points $(y_k)_{k=0,\ldots,p}$, we obtain an approximate solution which is exactly the same as the one which is obtained by the same method (piecewise linear finite element) and a constant value of μ as the diffusion coefficient, namely the arithmetic average of μ_r and μ_b. Then, this approximate solution converges, as $h \longrightarrow 0$, towards*

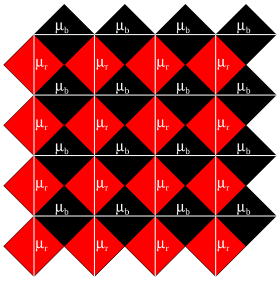


FIG. 3. *Case of the discrepancy between the H- and Hd-limits.*

the solution of the Dirichlet problem whose diffusion is this arithmetic average of μ_r and μ_b. However, the H-limit as $h \longrightarrow 0$ of the continuous operators is given by the harmonic average of μ_r and μ_b.

However, in the case $N > 1$ and even in the isotropic case, the obtention of the Hd-limit by passing to the limit for the weak $\star$ topology of $L^\infty(\Omega)$ is no longer possible. Indeed, let us consider the sequence of admissible discretizations $\mathcal{D}_n$ of $\Omega = (0,1)\times(0,1)$, where the control volumes are some $(k/n,(k{+}1)/n)\times(l/n,(l{+}1)/n)$, for integer values k and l between 0 and $n-1$ (see Figure 3). Assume that the function μ_n is defined by the value $\mu_r > 0$ on the vertical edges $\{k/n\}\times(l/n,(l+1)/n)$ and by the value $\mu_b > 0$ on the horizontal edges $(k/n,(k+1)/n)\times\{l/n\}$. Then the function $\mu_n \in \mathcal{M}_{\mathcal{D}_n}(\alpha,\beta)$ (with $\alpha = \min(\mu_r,\mu_b)$ and $\beta = \max(\mu_r,\mu_b)$) corresponds to the first two-dimensional example of section 2.4 (recall that the function μ_n is constant on subsets which, in this case, are the squares of lengthside equal to $1/(n\sqrt{2})$, tilted with an angle of measure $\pi/4$ with respect to the grid. As seen in section 2, the H-limit of $(\mu_n)_{n\in\mathbb{N}}$ is the field with constant value $\sqrt{\mu_r\mu_b}$. We then remark that for a given $f \in L^2(\Omega)$ the discrete values solutions of the finite volume scheme (11) are identical to those obtained from (11), written in the case where $\tilde{\Omega} = (0,1/\sqrt{\mu_r})\times(0,1/\sqrt{\mu_b})$, the grid is given by the subsets $(k/(n\sqrt{\mu_r}),(k+1)/(n\sqrt{\mu_r}))\times(l/(n\sqrt{\mu_b}),(l+1)/(n\sqrt{\mu_b}))$, $\mu = 1$, and the right-hand side $\tilde{f} = f(\cdot\sqrt{\mu_r},\cdot\sqrt{\mu_b})$. Thanks to Lemma 3 which states the convergence of the finite volume scheme for $\mu = 1$ we then get that $u_{\mathcal{D}_n}$ converges to $\overline{u} = \tilde{u}(\cdot/\sqrt{\mu_r},\cdot/\sqrt{\mu_b})$ with $\tilde{u} = \mathcal{F}_{\tilde{\Omega}}(\tilde{f},1)$, denoting here by $\mathcal{F}_{\tilde{\Omega}}$ the function $\mathcal{F}$ obtained when the Dirichlet problem is solved in the domain $\tilde{\Omega}$. An easy change of variable proves that $\overline{u} = \mathcal{F}(f,M)$, where M is the constant field, the value of which is the linear application defined by $(1,0)\to(\mu_r,0)$ and $(0,1)\to(0,\mu_b)$. This field M, which is homogeneous anisotropic and differs from the H-limit of $(\mu_n)_{n\in\mathbb{N}}$, is therefore the Hd-limit of $(\mathcal{D}_n,\mu_n)_{n\in\mathbb{N}}$.

The physical reason for this discrepancy is the fact that in this example the heterogeneous behavior and the grid are at the same scale: note that this occurs when solving the coupled two-phase flow in porous media problem using a coupled scheme on the same grid (see section 5). On the contrary, in the cases where it is possible to let the size of the mesh tend to zero faster than the size of the heterogeneities, the obtained H- and Hd-limits are equal.

REMARK 4. *Similar results to Theorem* 2 *can be obtained within the finite element framework, leading to the same distinction between the resulting Hd-limit and the H-limit (see Remark* 3 *for an example in the one-dimensional case).*

REMARK 5. *Exactly in the same manner as for the continuous case, it is possible to show the local character of Hd-convergence in the sense of Theorem* 2 *and the* *independence of the Hd-limit on the boundary conditions (see* [16] *for such results within the finite difference setting).*

4.2. Existence of limit operators. The first step leading to the proof of Theorem 2 is given by the results of the following lemma, which are similar to the continuous ones (see [14], [19]).

LEMMA 4. *Let Ω be an open bounded polygonal subset of $\mathbb{R}^N$, with $N \in \mathbb{N}_*$. Let two real numbers α and β be such that $0 < \alpha \leq \beta$. Let $(\mathcal{D}_n, \mu_n)_{n\in\mathbb{N}}$ be a sequence such that, for all $n \in \mathbb{N}$, $\mathcal{D}_n$ is an admissible discretization of Ω in the sense of Definition* 1*, and $\mu_n \in \mathcal{M}_{\mathcal{D}_n}(\alpha,\beta)$. We assume that $\lim_{n\longrightarrow\infty} \operatorname{size}(\mathcal{D}_n) = 0$.*

Then there exists a subsequence of $(\mathcal{D}_n, \mu_n)_{n\in\mathbb{N}}$, again denoted $(\mathcal{D}_n, \mu_n)_{n\in\mathbb{N}}$, that verifies that there exists an invertible continuous linear application $\overline{F} : H^{-1}(\Omega) \longrightarrow H_0^1(\Omega)$ and a continuous linear application $\overline{\mathbf{G}} : H^{-1}(\Omega) \longrightarrow (L^2(\Omega))^N$ such that

- *for all $f \in L^2(\Omega)$, the sequence $(F_{\mathcal{D}_n}(f,\mu_n))_{n\in\mathbb{N}}$ converges to $\overline{F}(f)$ in $L^2(\Omega)$ and the sequence $(\mathbf{G}_{\mathcal{D}_n} F_{\mathcal{D}_n}(f,\mu_n))_{n\in\mathbb{N}}$ weakly converges to $\nabla\overline{F}(f)$ in $(L^2(\Omega))^N$;*
- *for all $f \in L^2(\Omega)$, the sequence $(\mu_n \mathbf{G}_{\mathcal{D}_n} F_{\mathcal{D}_n}(f,\mu_n))_{n\in\mathbb{N}}$ weakly converges to $\overline{\mathbf{G}}(f)$ in $(L^2(\Omega))^N$;*
- *the following relation holds:*

$$(16) \qquad \int_\Omega \overline{\mathbf{G}}(b)(x) \cdot \nabla\bar{v}(x)dx = b(\bar{v}) \ \forall \bar{v} \in H_0^1(\Omega), \ \forall b \in H^{-1}(\Omega).$$

Proof. Let us assume the hypotheses of the lemma. Let $f \in L^2(\Omega)$. Thanks to (13), for all $n \in \mathbb{N}$, denoting $u_n = F_{\mathcal{D}_n}(f,\mu_n)$, we have

$$(17) \qquad \alpha|u_n|_{\mathcal{D}_n} \leq \operatorname{diam}(\Omega)\ \|f\|_{L^2(\Omega)}.$$

This shows that the hypotheses of Lemma 2 are satisfied. Therefore, there exists a subsequence of $(\mathcal{D}_n, \mu_n)_{n\in\mathbb{N}}$, again denoted $(\mathcal{D}_n, \mu_n)_{n\in\mathbb{N}}$, and $\overline{u} \in H_0^1(\Omega)$ such that the sequence $(F_{\mathcal{D}_n}(f,\mu_n))_{n\in\mathbb{N}}$ converges to $\overline{u}$ in $L^2(\Omega)$. We again denote $u_n = F_{\mathcal{D}_n}(f,\mu_n)$.

Let us introduce the functions $\bar{w} \in H_0^1(\Omega)$ defined by $\bar{w} = \mathcal{F}(f,1)$ and, for all $n \in \mathbb{N}$, $w_n = F_{\mathcal{D}_n}(f,1)$. For $n \in \mathbb{N}$, we deduce from (12) that

$$(18) \qquad \int_\Omega f(x)u_n(x)dx = [u_n,u_n]_{\mathcal{D}_n,\mu_n,1} \geq \alpha\left(|u_n|_{\mathcal{D}_n}\right)^2$$

and, thanks to the Cauchy–Schwarz inequality,

$$(19) \qquad \int_\Omega f(x)u_n(x)dx = [w_n,u_n]_{\mathcal{D}_n,1,1} \leq |w_n|_{\mathcal{D}_n}\ |u_n|_{\mathcal{D}_n}.$$

Therefore (18) and (19) yield

$$\alpha|u_n|_{\mathcal{D}_n} \leq |w_n|_{\mathcal{D}_n}.$$

Passing to the limit on $n \longrightarrow \infty$ in the above equation gives, using (14),

$$\alpha \limsup_{n\longrightarrow\infty} |u_n|_{\mathcal{D}_n} \leq \|\nabla\bar{w}\|_{(L^2(\Omega))^N},$$

which gives, since $\|f\|_{H^{-1}(\Omega)} = \|\nabla\bar{w}\|_{(L^2(\Omega))^N}$,

$$(20) \qquad \alpha \limsup_{n\longrightarrow\infty} |u_n|_{\mathcal{D}_n} \leq \|f\|_{H^{-1}(\Omega)}.$$

Thanks to (9), we get

$$\alpha\|\overline{u}\|_{H_0^1(\Omega)} \le \|f\|_{H^{-1}(\Omega)}. \tag{21}$$

Turning to the study of the sequence $\mathbf{g}_n = \mu_n \mathbf{G}_{\mathcal{D}_n} F_{\mathcal{D}_n}(f,\mu_n)$ for $n \in \mathbb{N}$, we have (in a similar way as the case $\mu_n = 1$ handled in Lemma 2),

$$\left(\|\mathbf{g}_n\|_{(L^2(\Omega))^N}\right)^2 \le N\beta^2 \left(|u_n|_{\mathcal{D}_n}\right)^2,$$

which yields, using (20),

$$\limsup_{n\longrightarrow\infty} \|\mathbf{g}_n\|_{(L^2(\Omega))^N} \le \frac{\sqrt{N}\beta}{\alpha}\|f\|_{H^{-1}(\Omega)}. \tag{22}$$

Thus there exists a subsequence of $(\mathcal{D}_n,\mu_n)_{n\in\mathbb{N}}$, again denoted $(\mathcal{D}_n,\mu_n)_{n\in\mathbb{N}}$ and $\bar{\mathbf{g}} \in (L^2(\Omega))^N$, such that $\mathbf{g}_n = \mu_n \mathbf{G}_{\mathcal{D}_n} F_{\mathcal{D}_n}(f,\mu_n)$ converges weakly to $\bar{\mathbf{g}}$ as $n \longrightarrow \infty$ in $(L^2(\Omega))^N$. Passing to the limit in (22), we then get

$$\|\bar{\mathbf{g}}\|_{(L^2(\Omega))^N} \le \frac{\sqrt{N}\beta}{\alpha}\|f\|_{H^{-1}(\Omega)}. \tag{23}$$

We then consider a sequence $(f_m)_{m\in\mathbb{N}}$ of functions of $L^2(\Omega)$ which is dense in $H^{-1}(\Omega)$. We can then extract a subsequence (using the classical diagonal process), again denoted $(\mathcal{D}_n,\mu_n)_{n\in\mathbb{N}}$, such that for all $m \in \mathbb{N}$ the sequence $(F(f_m,\mathcal{D}_n,\mu_n))_{n\in\mathbb{N}}$ converges to some function denoted $\overline{F}(f_m) \in H_0^1(\Omega)$ in $L^2(\Omega)$ and the sequence $(\mu_n \mathbf{G}_{\mathcal{D}_n} F(f_m,\mathcal{D}_n,\mu_n))_{n\in\mathbb{N}}$ converges to some function denoted $\overline{\mathbf{G}}(f_m) \in (L^2(\Omega)^N)$ weakly in $(L^2(\Omega)^N)$. The linear functions $\overline{F}$ (resp., $\overline{\mathbf{G}}$) can then be prolonged by continuity, thanks to (21) (resp., (23)) to a continuous linear function, again denoted $\overline{F}\ : H^{-1}(\Omega) \longrightarrow H_0^1(\Omega)$ (resp., $\overline{\mathbf{G}}\ : H^{-1}(\Omega) \longrightarrow (L^2(\Omega))^N$).

Let us now prove (16). Let $f \in L^2(\Omega)$. We set $\overline{u} = \overline{F}(f)$ and $\bar{\mathbf{g}} = \overline{\mathbf{G}}(f)$. Let $\varphi \in C_c^\infty(\Omega)$. For a given $n \in \mathbb{N}$, we denote $\mathcal{D}_n = (\mathcal{T}_n, \mathcal{E}_n, \mathcal{P}_n)$, $u_n = F_{\mathcal{D}_n}(f,\mu_n)$. Omitting the indexes n in the discrete expressions, we set, for all $K \in \mathcal{T}$ and $\sigma \in \mathcal{E}_K$,

$$R_{K,\sigma} = \frac{1}{d_{K,\sigma}}(\varphi(y_\sigma) - \varphi(x_K)) - \frac{N}{m_\sigma d_{K,\sigma}} \int_{D_{K,\sigma}} \nabla\varphi(x)\cdot \mathbf{n}_{K,\sigma} dx.$$

Then there exists $C_\varphi > 0$ which depends only on φ such that $|R_{K,\sigma}| \le \text{size}(\mathcal{D}_n) C_\varphi$. Setting

$$T_n = \sum_{K\in\mathcal{T}_n} \sum_{\sigma\in\mathcal{E}_K} \mu_\sigma \tau_{K,\sigma}(u_\sigma - u_K)(\varphi(y_\sigma) - \varphi(x_K)),$$

we then get $\lim_{n\longrightarrow\infty} \left|T_n - \int_\Omega \mathbf{g}_n(x)\cdot\nabla\varphi(x)dx\right| = 0$ which yields

$$\lim_{n\longrightarrow\infty} T_n = \int_\Omega \bar{\mathbf{g}}(x)\cdot\nabla\varphi(x)dx. \tag{24}$$

Since, using (11), we have

$$T_n = \sum_{K\in\mathcal{T}_n} \int_K f(x)dx\ \varphi(x_K), \tag{25}$$

we also get $\lim_{n\longrightarrow\infty} T_n = \int_\Omega f(x)\varphi(x)dx$. We thus get

$$\int_\Omega \bar{\mathbf{g}}(x)\cdot\nabla\varphi(x)dx = \int_\Omega f(x)\varphi(x)dx \;\forall\varphi\in C_c^\infty(\Omega). \tag{26}$$

Using (26) and the density of $C_c^\infty(\Omega)$ in $H_0^1(\Omega)$, we conclude (16).

Let us show that $\overline{F}$ is invertible. We consider the bilinear form $a\ :\ \left(H^{-1}(\Omega)\right)^2 \longrightarrow \mathbb{R}$ defined by

$$\forall (b,b')\in \left(H^{-1}(\Omega)\right)^2,\ a(b,b') = b(\overline{F}(b')).$$

Let again $f\in L^2(\Omega)$. We introduce the functions $\overline{u},\bar{w}\in H_0^1(\Omega)$ defined by $\overline{u}=\overline{F}(f)$, $\bar{w}=\mathcal{F}(f,1)$, and, for all $n\in\mathbb{N}$, $u_n = F_{\mathcal{D}_n}(f,\mu_n)$ and $w_n = F_{\mathcal{D}_n}(f,1)$. We have

$$a(f,f) = \int_\Omega f(x)\overline{F}(f)(x)dx = \lim_{n\longrightarrow\infty}\int_\Omega f(x)u_n(x)dx.$$

We can write, on the one hand,

$$\int_\Omega f(x)u_n(x)dx = [u_n,u_n]_{\mathcal{D}_n,\mu_n,1},$$

which yields

$$\int_\Omega f(x)u_n(x)dx \geq \alpha\left(|u_n|_{\mathcal{D}_n}\right)^2. \tag{27}$$

We have, on the other hand,

$$\int_\Omega f(x)w_n(x)dx = [u_n,w_n]_{\mathcal{D}_n,\mu_n,1}$$

and

$$\int_\Omega f(x)w_n(x)dx = \left(|w_n|_{\mathcal{D}_n}\right)^2.$$

This yields

$$\left(|w_n|_{\mathcal{D}_n}\right)^2 = [u_n,w_n]_{\mathcal{D}_n,\mu_n,1}.$$

Since

$$\begin{array}{rcl}([u_n,w_n]_{\mathcal{D}_n,\mu_n,1})^2 & \leq & [u_n,u_n]_{\mathcal{D}_n,\mu_n,1}\ [w_n,w_n]_{\mathcal{D}_n,\mu_n,1}\\ & \leq & \beta^2[u_n,u_n]_{\mathcal{D}_n,1,1}\ [w_n,w_n]_{\mathcal{D}_n,1,1},\end{array}$$

we therefore get

$$\left(|w_n|_{\mathcal{D}_n}\right)^2 \leq \beta|u_n|_{\mathcal{D}_n}|w_n|_{\mathcal{D}_n}. \tag{28}$$

From (27) and (28) we deduce

$$\int_\Omega f(x)u_n(x)dx \geq \frac{\alpha}{\beta^2}\left(|w_n|_{\mathcal{D}_n}\right)^2. \tag{29}$$

Letting $n \longrightarrow \infty$ in (29) gives

$$a(f,f) \geq \frac{\alpha}{\beta^2} \int_\Omega (\nabla \bar{w}(x))^2 dx, \tag{30}$$

which shows that

$$a(f,f) \geq \frac{\alpha}{\beta^2} \left(\|f\|_{H^{-1}(\Omega)} \right)^2. \tag{31}$$

By continuity of a, this property is available on $H^{-1}(\Omega)$, which shows the coercivity of a. Let $\bar{v} \in H_0^1(\Omega)$. The problem, find $b \in H^{-1}(\Omega)$ such that for all $b' \in H^{-1}(\Omega)$, $a(b', b) = b'(\bar{v})$, has a unique solution b, thanks to Lax–Milgram's theorem. It then satisfies $\overline{F}(b) = \bar{v}$.

REMARK 6. *The previous lemma could be stated using sequences $(F_{\mathcal{D}_n}(f, \mu_n))_{n\in\mathbb{N}}$ and $(\mathbf{G}_{\mathcal{D}_n} F_{\mathcal{D}_n}(f, \mu_n))_{n\in\mathbb{N}}$ with $f \in H^{-1}(\Omega)$ (see* [6] *for the definition of the finite volume scheme in this case).*

4.3. Proof of Theorem 2. We assume the hypotheses of Theorem 2, which are the same as those of Lemma 4. Therefore, let $(\mathcal{D}_n, \mu_n)_{n\in\mathbb{N}}$ denote a subsequence of $(\mathcal{D}_n, \mu_n)_{n\in\mathbb{N}}$, and let $\overline{F} : H^{-1}(\Omega) \longrightarrow H_0^1(\Omega)$ and $\overline{\mathbf{G}} : H^{-1}(\Omega) \longrightarrow (L^2(\Omega)^N)$ denote the linear continuous functions verifying the conclusions of Lemma 4. It suffices now to prove that there exists a function $M : \Omega \longrightarrow \mathcal{L}(\mathbb{R}^N, \mathbb{R}^N)$ such that for a.e. $x \in \Omega$, $\overline{\mathbf{G}}(b)(x) = M(x)\nabla \overline{F}(b)(x)$, for all $b \in H^{-1}(\Omega)$, and for all $(\xi, \chi) \in (\mathbb{R}^N)^2$, $\alpha|\xi|^2 \leq M(x)\xi \cdot \xi \leq \beta|\xi|^2$, and $M(x)\xi \cdot \chi = \xi \cdot M(x)\chi$. Let $f, g \in L^2(\Omega)$. We set $\overline{u} = \overline{F}(f)$ and $\bar{v} = \overline{F}(g)$. Let $\varphi \in C_c^\infty(\Omega)$. For a given $n \in \mathbb{N}$, we denote $\mathcal{D}_n = (\mathcal{T}_n, \mathcal{E}_n, \mathcal{P}_n)$, $u_n = F_{\mathcal{D}_n}(f, \mu_n)$, $v_n = F_{\mathcal{D}_n}(g, \mu_n)$, and we consider the expression

$$A_n = [u_n, v_n]_{\mathcal{D}_n, \mu_n, \varphi}. \tag{32}$$

We get $A_n = B_n - C_n$, where B_n and C_n are defined by (omitting indexes n in the right-hand sides)

$$B_n = \sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}_K} \mu_\sigma \tau_{K,\sigma}(u_\sigma - u_K)(\varphi(y_\sigma)v_\sigma - \varphi(x_K)v_K)$$

and

$$C_n = \sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}_K} v_\sigma \mu_\sigma \tau_{K,\sigma}(u_\sigma - u_K)(\varphi(y_\sigma) - \varphi(x_K)).$$

Since $B_n = \sum_{K\in\mathcal{T}} \int_K f(x)dx\ \varphi(x_K)v_K$, we then get

$$\lim_{n\longrightarrow\infty} B_n = \int_\Omega f(x)\varphi(x)\bar{v}(x)dx.$$

Let $\tilde{v} \in C_c^\infty(\Omega)$ be a function which is meant to tend to $\bar{v}$ in $H_0^1(\Omega)$. We set

$$\tilde{B}_n = \sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}_K} \mu_\sigma \tau_{K,\sigma}(u_\sigma - u_K)(\varphi(y_\sigma)\tilde{v}(y_\sigma) - \varphi(x_K)\tilde{v}(x_K))$$

and we again have $\tilde{B}_n = \sum_{K\in\mathcal{T}} \int_K f(x)dx\ \varphi(x_K)\tilde{v}(x_K)$, which yields

$$\lim_{n\longrightarrow\infty} \tilde{B}_n = \int_\Omega f(x)\varphi(x)\tilde{v}(x)dx.$$

Using (24), we have

$$\lim_{n\longrightarrow\infty} \tilde{B}_n = \int_\Omega \overline{\mathbf{G}}(f)(x)\cdot\nabla(\varphi(x)\tilde{v}(x))dx.$$

We thus get

$$\int_\Omega f(x)\varphi(x)\tilde{v}(x)dx = \int_\Omega \overline{\mathbf{G}}(f)(x)\cdot\nabla(\varphi(x)\tilde{v}(x))dx. \tag{33}$$

In (33), we let $\tilde{v}\longrightarrow\bar{v}$ in $H^1_0(\Omega)$. It gives

$$\int_\Omega f(x)\varphi(x)\bar{v}(x)dx = \int_\Omega \overline{\mathbf{G}}(f)(x)\cdot\nabla(\varphi(x)\bar{v}(x))dx, \tag{34}$$

and therefore

$$\lim_{n\longrightarrow\infty} B_n = \int_\Omega \overline{\mathbf{G}}(f)(x)\cdot\nabla(\varphi(x)\bar{v}(x))dx. \tag{35}$$

We now study C_n. Let $\hat{v}_n$ be the function defined by

$$\hat{v}_n(x) = v_\sigma\ \forall x\in D_{K,\sigma},\ \forall K\in\mathcal{T},\ \forall\sigma\in\mathcal{E}_K. \tag{36}$$

Since $|v_n|_{\mathcal{D}_n}$ remains bounded, it is easy to see that $\hat{v}_n - v_n$ converges to 0 in $L^2(\Omega)$. We set

$$\hat{C}_n = \int_\Omega \hat{v}_n(x)\ \mu_n\mathbf{G}_{\mathcal{D}_n}F_{\mathcal{D}_n}(f,\mu_n)(x)\cdot\nabla\varphi(x)dx.$$

We easily get, thanks to the Cauchy–Schwarz inequality,

$$|C_n - \hat{C}_n| \le C(\varphi,\beta)\text{size}(\mathcal{D}_n)\ |u_n|_{\mathcal{D}_n}\ \|\hat{v}_n\|_{L^2(\Omega)},$$

which shows that

$$\lim_{n\longrightarrow\infty} C_n = \lim_{n\longrightarrow\infty}\hat{C}_n = \int_\Omega \bar{v}(x)\ \overline{\mathbf{G}}(f)(x)\cdot\nabla\varphi(x)dx. \tag{37}$$

We thus get, gathering (35) and (37), recalling that $\bar{v} = \overline{F}(g)$,

$$\lim_{n\longrightarrow\infty} A_n = \int_\Omega \varphi(x)\ \overline{\mathbf{G}}(f)(x)\cdot\nabla\overline{F}(g)(x)dx. \tag{38}$$

Note that the preceding proof ((32)–(38)) also gives a discrete version of a compensated compactness lemma; see Remark 7 below. We can now exchange the roles of f and g in (38). We thus get

$$\lim_{n\longrightarrow\infty} A_n = \int_\Omega \varphi(x)\ \overline{\mathbf{G}}(g)(x)\cdot\nabla\overline{F}(f)(x)dx. \tag{39}$$

This yields

$$\int_\Omega \varphi(x)\overline{\mathbf{G}}(f)(x)\cdot\nabla\overline{F}(g)(x)dx = \int_\Omega \varphi(x)\overline{\mathbf{G}}(g)(x)\cdot\nabla\overline{F}(f)(x)dx. \tag{40}$$

In order to prove the existence of M as it is given in Theorem 2, we now proceed exactly as in the continuous setting. Since (40) is true for all $\varphi \in C_c^\infty(\Omega)$, we get

$$\overline{\mathbf{G}}(f)(x) \cdot \nabla \overline{F}(g)(x) = \overline{\mathbf{G}}(g)(x) \cdot \nabla \overline{F}(f)(x), \text{ for a.e. } x \in \Omega. \tag{41}$$

Since (41) is true for all f and g in $L^2(\Omega)$, by continuity of $\overline{F}$ and $\overline{\mathbf{G}}$, we get

$$\begin{aligned}&\overline{\mathbf{G}}(b)(x) \cdot \nabla \overline{F}(b')(x) = \overline{\mathbf{G}}(b')(x) \cdot \nabla \overline{F}(b)(x)\\ &\forall b,\ b' \in H^{-1}(\Omega), \quad \text{for a.e. } x \in \Omega.\end{aligned} \tag{42}$$

Since $\overline{F}$ is invertible, we can choose some $b_i \in H^{-1}(\Omega)$ such that, in an open set ω such that $\overline{\omega} \subset \Omega$, $\overline{F}(b_i)(x) = x \cdot \mathbf{e}_i$ and then $\nabla \overline{F}(b_i)(x) = \mathbf{e}_i$ (where $\mathbf{e}_i$ is the ith unit vector of $\mathbb{R}^N$), for $i = 1, \ldots, N$. Thus, for a.e. $x \in \omega$, we can define $M(x) \in \mathcal{L}(\mathbb{R}^N, \mathbb{R}^N)$ by $M^\star(x)\mathbf{e}_i = \overline{\mathbf{G}}(b_i)(x)$, for $i = 1, \ldots, N$, where $M^\star$ is the adjoint operator of M. We thus get

$$\overline{\mathbf{G}}(b)(x) = M(x)\nabla \overline{F}(b)(x) \ \forall b \in H^{-1}(\Omega) \text{ for a.e. } x \in \omega.$$

Taking $b = b_i$ and $b' = b_j$ in (42) proves that $M(x)$ is symmetric in ω. Since ω is arbitrary, we then obtain M a.e. in Ω such that

$$\overline{\mathbf{G}}(b)(x) = M(x)\nabla \overline{F}(b)(x) \ \forall b \in H^{-1}(\Omega) \text{ for a.e. } x \in \Omega. \tag{43}$$

The uniqueness of M is a direct consequence of the invertibility of $\overline{F}$. We now prove that $M \in \mathcal{M}(\alpha, \beta, \Omega)$. Letting $f = g$ in (32), and taking $\varphi \geq 0$, we get, using (38), that

$$\lim_{n \longrightarrow \infty} [u_n, u_n]_{\mathcal{D}_n, \mu_n, \varphi} = \int_\Omega \varphi(x) M(x) \nabla \overline{u}(x) \cdot \nabla \overline{u}(x) dx.$$

Since

$$\begin{aligned}\lim_{n \longrightarrow \infty} [u_n, u_n]_{\mathcal{D}_n, \mu_n, \varphi} &\geq \alpha \liminf_{n \longrightarrow \infty} [u_n, u_n]_{\mathcal{D}_n, 1, \varphi}\\ &\geq \alpha \int_\Omega \varphi(x) (\nabla \overline{u}(x))^2 dx,\end{aligned} \tag{44}$$

we get, for a.e. $x \in \Omega$, $M(x)\nabla \overline{F}(f)(x) \cdot \nabla \overline{F}(f)(x) \geq \alpha(\nabla \overline{F}(f)(x))^2$. By density and invertibility of $\overline{F}$, since f can be arbitrarily chosen, this proves that, for a.e. $x \in \Omega$ and for all $\xi \in \mathbb{R}^N$, $M(x)\xi \cdot \xi \geq \alpha(\xi)^2$. Let $\varphi \in C_c^\infty(\Omega, \mathbb{R}_+)$, and let $(f, g) \in (L^2(\Omega))^N$, $\overline{u} = \overline{F}(f)$, and $\bar{w} = \mathcal{F}(f, 1)$. For $n \in \mathbb{N}$, we define $u_n = F_{\mathcal{D}_n}(f, \mu_n)$ and $w_n = F_{\mathcal{D}_n}(g, 1)$. We define D_n by

$$D_n = [u_n, w_n]_{\mathcal{D}_n, \mu_n, \varphi}. \tag{45}$$

We study D_n in the same manner as A_n. Since w_n converges to $\bar{w}$ in $L^2(\Omega)$, we get

$$\lim_{n \longrightarrow \infty} D_n = \int_\Omega \varphi(x) M(x) \nabla \overline{u}(x) . \nabla \bar{w}(x) dx.$$

On the other hand, we have

$$(D_n)^2 \leq \quad [u_n, u_n]_{\mathcal{D}_n, \mu_n, \varphi} \quad [w_n, w_n]_{\mathcal{D}_n, \mu_n, \varphi}, \tag{46}$$

and therefore

$$(D_n)^2 \leq \beta \ [u_n, u_n]_{\mathcal{D}_n, \mu_n, \varphi} \quad [w_n, w_n]_{\mathcal{D}_n, 1, \varphi}. \tag{47}$$

We thus get

$$\lim_{n\longrightarrow\infty}(D_n)^2 \le \beta \int_\Omega \varphi(x)M(x)\nabla\overline{u}(x).\nabla\overline{u}(x)dx \quad \int_\Omega \varphi(x)(\nabla\bar{w}(x))^2dx,$$

which gives

$$\begin{aligned}(48)\qquad &\left(\int_\Omega \varphi(x)M(x)\nabla\overline{u}(x).\nabla\bar{w}(x)dx\right)^2\\ &\le \beta \quad \int_\Omega \varphi(x)M(x)\nabla\overline{u}(x).\nabla\overline{u}(x)dx \quad \int_\Omega \varphi(x)(\nabla\bar{w}(x))^2dx.\end{aligned}$$

Since g can be arbitrarily chosen, it is therefore possible to let $\bar{w} \longrightarrow \overline{u}$ (in $H_0^1(\Omega)$) in (48). We thus get

$$(49)\qquad \int_\Omega \varphi(x)M(x)\nabla\overline{u}(x).\nabla\overline{u}(x)dx \le \beta \quad \int_\Omega \varphi(x)(\nabla\overline{u}(x))^2dx,$$

which yields, for a.e. $x \in \Omega$, $M(x)\nabla\overline{u}(x)\cdot\nabla\overline{u}(x) \le \beta(\nabla\overline{u}(x))^2$. By density and invertibility of $\overline{F}$, since f can be arbitrarily chosen, we get that, for a.e. $x \in \Omega$ and for all $\xi \in \mathbb{R}^N$, $M(x)\xi\cdot\xi \le \beta\xi^2$.

This concludes the proof of Theorem 2.

REMARK 7. *Note that, as in the continuous setting (see* [14] *and* [19]*) and as in the finite difference framework (see* [12] *and* [16]*), an important step of the above proof (from* (32) *to* (38)*) consists of passing to the limit in some nonlinear terms. Indeed, the same proof as above also yields the following discrete version of a compensated compactness lemma (namely a discrete simplified "div-curl" lemma).*

LEMMA 5 (discrete compensated compactness lemma). *Let Ω be an open bounded polygonal subset of $\mathbb{R}^N$, with $N \in \mathbb{N}_*$. Let $(\mathcal{D}_n)_{n\in\mathbb{N}}$ be a sequence of admissible discretizations of Ω in the sense of Definition* 1 *such that $\lim_{n\longrightarrow\infty}\text{size}(\mathcal{D}_n) = 0$. Let us suppose that for all $n \in \mathbb{N}$ there exists $W_n, X_n \in V_{\mathcal{D}_n} \subset (L^2(\Omega))^N$ such that*

- *$W_n \longrightarrow W$ weakly in $(L^2(\Omega))^N$ as $n \longrightarrow \infty$;*
- *$X_n \longrightarrow X$ weakly in $(L^2(\Omega))^N$ as $n \longrightarrow \infty$;*
- *$\text{div}_{\mathcal{D}_n}W_n$ weakly converges in $L^2(\Omega)$ as $n \longrightarrow \infty$;*
- *there exists $u_n \in H_{\mathcal{D}_n}(\Omega)$ such that $X_n = \mathbf{G}_{\mathcal{D}_n}u_n$.*

Then $\lim_{n\longrightarrow\infty}\langle W_n, G_n\rangle_{\mathcal{D}_n} = \int_\Omega W(x)\cdot X(x)dx$.

In the preceding lemma, we denote, for any admissible discretization $\mathcal{D}$ of Ω in the sense of Definition 1*, by $V_\mathcal{D}$ the subset of $(L^2(\Omega))^N$ of functions W verifying that there exists a family of real values $(w_{K,\sigma})_{K\in\mathcal{T},\sigma\in\mathcal{E}_K}$ such that*

$$W(x) = Nw_{K,\sigma}\mathbf{n}_{K,\sigma}, \quad \textit{for a.e. } x \in D_{K,\sigma}\ \forall K \in \mathcal{T},\ \forall\sigma \in \mathcal{E}_K,$$

and $w_{K,\sigma} + w_{L,\sigma} = 0$ for all $\sigma = K|L$.

For all $W \in V_\mathcal{D}$, we denote by $\text{div}_\mathcal{D}W$ the piecewise constant function, whose value in $K \in \mathcal{T}$ is $\sum_{\sigma\in\mathcal{E}_K} m_\sigma w_{K,\sigma}$. We then get $\mathbf{G}_\mathcal{D}u \in V_\mathcal{D}$ for all $u \in H_\mathcal{D}(\Omega)$. For $(W,X) \in (V_\mathcal{D})^2$, we then define $\langle W,X\rangle_\mathcal{D} = \frac{1}{N}\int_\Omega W(x)\cdot X(x)dx$. It is interesting to notice that, under the hypotheses of the lemma, neither the sequence $(W_n)_{n\in\mathbb{N}}$ nor $(X_n)_{n\in\mathbb{N}}$ converges in $(L^2(\Omega))^N$, except if the limit is 0*. Note also that, contrary to the classical compensated compactness lemma, the sequence which converges in the distribution sense to $W\cdot X$ is not $(W_n\cdot X_n)_{n\in\mathbb{N}}$, but it is $\frac{1}{N}(W_n\cdot X_n)_{n\in\mathbb{N}}$.*

5. Application to a coupled problem.

5.1. A continuous system of equations. We now study the steady-state version of the evolution problem (1). We thus get the following system:

$$\left.\begin{aligned} -\mathrm{div}(\lambda(s)\nabla u) &= \bar{f} \\ -\mathrm{div}(\gamma(s)\lambda(s)\nabla u) &= (\bar{f})^{+}\gamma(\bar{s}) - (\bar{f})^{-}\gamma(s) \end{aligned}\right\} \text{ in } \Omega, \tag{50}$$

with the boundary conditions

$$\begin{aligned} &u = 0 \text{ on } \partial\Omega, \\ &s = \hat{s} \text{ on } \{x \in \partial\Omega, \nabla u(x)\cdot \mathbf{n}_{\partial\Omega}(x) \geq 0\}. \end{aligned} \tag{51}$$

We refer to the introduction for the physical meaning of the quantities appearing in (50) and (51). The following assumptions (denoted in the following Hypotheses (H)) are made on the data:

- the domain Ω is an open polygonal connected subset of $\mathbb{R}^N$, with $N = 2$ or $N = 3$;
- $\gamma \in C^0([0,1],[0,1])$ is a nondecreasing Lipschitz continuous function with $\gamma(0) = 0$ and $\gamma(1) = 1$, and Lipschitz constant $L_\gamma > 0$;
- there exists two real numbers α and β, with $0 < \alpha \leq \beta$, such that $\lambda \in C^0([0,1],[\alpha,\beta])$ (recall that λ is the "total mobility") verifies that $\gamma\lambda$ (the mobility of the phase 1, also denoted below k_1) is nondecreasing and $(1-\gamma)\lambda$ (the mobility of the phase 2) is nonincreasing;
- $\bar{f} \in L^2(\Omega)$ represents the rates at the wells;
- $\bar{s} \in L^\infty(\Omega)$ is such that $0 \leq \bar{s} \leq 1$ a.e. in Ω;
- $\hat{s} \in L^\infty(\partial\Omega)$ is such that $0 \leq \hat{s} \leq 1$ a.e. in $\partial\Omega$ (for the $N-1$-dimensional Lebesgue measure).

5.2. Finite volume coupled scheme. Let us assume Hypotheses (H). Let $\mathcal{D}$ be an admissible discretization of Ω in the sense of Definition 1. We set

$$\left.\begin{aligned} &\bar{f}_K = \int_K \bar{f}(x)dx,\ \bar{s}_K = \frac{1}{m_K}\int_K \bar{s}(x)dx \quad \forall K \in \mathcal{T}, \\ &\hat{s}_\sigma = \frac{1}{m_\sigma}\int_\sigma \hat{s}(x)dx \qquad\qquad\qquad\quad \forall \sigma \in \mathcal{E}_{\mathrm{ext}}. \end{aligned}\right\} \tag{52}$$

We introduce the set $L_{\mathcal{D}}(\Omega,[0,1])$ of the functions of $L^\infty(\Omega)$ whose value on each $K \in \mathcal{T}$ is a constant value belonging to $[0,1]$. For all $s \in L_{\mathcal{D}}(\Omega,[0,1])$ and $K \in \mathcal{T}$, we denote $s_K \in [0,1]$ the constant value of s in K. For all $u \in H_{\mathcal{D}}(\Omega)$ and $s \in L_{\mathcal{D}}(\Omega,[0,1])$, the upstream evaluation of the saturation at the edges $\sigma \in \mathcal{E}$ is defined by the functions $s_\sigma(u,s,\hat{s})$ such that

$$\begin{aligned} &\left.\begin{aligned} s_\sigma(u,s,\hat{s}) &= s_K && \text{if } u_K \geq u_L \\ s_\sigma(u,s,\hat{s}) &= s_L && \text{if } u_K < u_L \end{aligned}\right\} \quad \forall \sigma \in \mathcal{E}_{\mathrm{int}},\ \sigma = K|L, \\ &\left.\begin{aligned} s_\sigma(u,s,\hat{s}) &= s_K && \text{if } u_K \geq 0 \\ s_\sigma(u,s,\hat{s}) &= \hat{s}_\sigma && \text{if } u_K < 0 \end{aligned}\right\} \quad \forall \sigma \in \mathcal{E}_{\mathrm{ext}},\ \sigma \in \mathcal{E}_K, \end{aligned} \tag{53}$$

and the functions $\mu(u,s,\hat{s}) \in \mathcal{M}_{\mathcal{D}}(\alpha,\beta)$ by

$$\mu_\sigma(u,s,\hat{s}) = \lambda(s_\sigma(u,s,\hat{s}))\ \forall \sigma \in \mathcal{E}. \tag{54}$$

We consider the following scheme (classical in petroleum engineering), a solution of which is some $(u,s) \in H_{\mathcal{D}}(\Omega) \times L_{\mathcal{D}}(\Omega,[0,1])$:

$$u = F(\bar{f},\mathcal{D},\mu(u,s,\hat{s})), \tag{55}$$

$$-\sum_{\sigma\in\mathcal{E}_K}\gamma(s_\sigma(u,s,\hat{s}))\,\mu_\sigma(u,s,\hat{s})\,\tau_{K,\sigma}\,(u_\sigma-u_K)=\gamma(\bar{s}_K)(\bar{f}_K)^+-\gamma(s_K)(\bar{f}_K)^-\ \ \forall K\in\mathcal{T}.$$
(56)

Remark 8. *Note that the function λ is also evaluated in* (54) *using an upstream weighted scheme. This corresponds to the industrial scheme classically used in reservoir simulation, in which the mobility of each phase is upstream weighted. However, it would be natural to use a centered approximation in* (54) *and use an upstream weighted scheme for γ in the left-hand side of* (56)*, but in such a case the convergence results given in Theorem* 4 *should be weakened.*

5.2.1. Existence of a solution to the coupled scheme.

Lemma 6. *Let us assume Hypotheses (H). Let $\mathcal{D}$ be an admissible discretization of Ω in the sense of Definition* 1*. Then there exists at least one solution $(u,s)\in H_{\mathcal{D}}(\Omega)\times L_{\mathcal{D}}(\Omega,[0,1])$ to scheme* (52)*–*(56)*.*

Proof. We prove the existence of a solution of (52)–(56) using Brouwer's fixed point theorem. For all $K\in\mathcal{T}$, let us define $\bar{f}_K$, $\bar{s}_K$, $\hat{s}_K$ by (52). We denote by

$$E=\{u\in H_{\mathcal{D}}(\Omega),\alpha|u|_{\mathcal{D}}\le\operatorname{diam}(\Omega)\ \|\bar{f}\|_{L^2(\Omega)}\}.$$

We define the application $\mathcal{A}\ :E\times L_{\mathcal{D}}(\Omega,[0,1])\longrightarrow E\times L_{\mathcal{D}}(\Omega,[0,1])$ by $\mathcal{A}(u,s)=(u',s')$, with $u'=F(\bar{f},\mathcal{D},\mu(u,s,\hat{s}))$, and for a real value $k>0$ which will be chosen later we define $(s'_K)_{K\in\mathcal{T}}$ by

$$s'_K=s_K+\frac{k}{m_K}\left(\begin{array}{l}\displaystyle\sum_{\sigma\in\mathcal{E}_K}\gamma(s_\sigma(u',s,\hat{s}))\,\mu_\sigma(u,s,\hat{s})\,\tau_{K,\sigma}\,(u'_\sigma-u'_K)\\ +\ \gamma(\bar{s}_K)(\bar{f}_K)^+-\gamma(s_K)(\bar{f}_K)^-\end{array}\right)\tag{57}$$
$$\forall K\in\mathcal{T}.$$

Since $\lambda\ge\alpha$, one has, using (13), that $u'\in E$. Then, in order to prove that $\mathcal{A}(u,s)\in E\times L_{\mathcal{D}}(\Omega,[0,1])$, we have only to prove that we can choose $k>0$ such that, for all $K\in\mathcal{T}$, $0\le s'_K\le 1$ (the operator $\mathcal{A}$ is then defined with this value of k). Using (13), we get

$$|u'_\sigma-u'_K|\le\left(\frac{\operatorname{diam}(\Omega)\ \|f\|_{L^2(\Omega)}}{\alpha\displaystyle\inf_{K\in\mathcal{T},\sigma\in\mathcal{E}_K}\tau_{K,\sigma}}\right)^{1/2}\quad\forall K\in\mathcal{T},\forall\sigma\in\mathcal{E}_K.\tag{58}$$

Denoting by M_{du} the right-hand side of inequality (58), we then take $k>0$ such that

$$k\le\inf_{K\in\mathcal{T}}\frac{m_K}{L_\gamma\left(\beta M_{du}\displaystyle\sum_{\sigma\in\mathcal{E}_K}\tau_{K,\sigma}+|\bar{f}_K|\right)}\tag{59}$$

(recall that L_γ is a Lipschitz constant for γ). With such a choice for k, we can now prove that for all $K\in\mathcal{T}$, $s'_K\in[0,1]$. Indeed, for $K\in\mathcal{T}$, let us multiply (11) (in which we set $\mu=\mu(u,s,\hat{s})$ and $f=\bar{f}$) by $\gamma(s_K)$ and substract the result from (57). We then get

$$s'_K=s_K+\frac{k}{m_K}\left(\sum_{\sigma\in\mathcal{E}_K}T_{K,\sigma}(s_\sigma(u',s,\hat{s})-s_K)+T_K(\bar{s}_K-s_K)\right),$$

where we define the nonnegative values $T_{K,\sigma}$ and T_K by

$$\begin{cases} T_{K,\sigma} = \dfrac{\gamma(s_\sigma(u',s,\hat{s})) - \gamma(s_K)}{s_\sigma(u',s,\hat{s}) - s_K}\mu_\sigma(u,s,\hat{s})\tau_{K,\sigma}(u'_\sigma - u'_K)^+ & \text{if } s_\sigma(u',s,\hat{s}) \neq s_K, \\ T_{K,\sigma} = L_\gamma \mu_\sigma(u,s,\hat{s})\tau_{K,\sigma}(u'_\sigma - u'_K)^+ & \text{if } s_\sigma(u',s,\hat{s}) = s_K \end{cases}$$

and

$$\begin{cases} T_K = \dfrac{\gamma(\bar{s}_K) - \gamma(s_K)}{\bar{s}_K - s_K}(\bar{f}_K)^- & \text{if } \bar{s}_K \neq s_K, \\ T_K = L_\gamma \ (\bar{f}_K)^- & \text{if } \bar{s}_K = s_K. \end{cases}$$

Then, for $k > 0$ verifying (59), we have

$$0 \leq 1 - \frac{k}{m_K}\left(\sum_{\sigma\in\mathcal{E}_K} T_{K,\sigma} + T_K\right),$$

which ensures that s'_K is a convex combination of $(s_K)_{K\in\mathcal{T}}$, $(\hat{s}_\sigma)_{\sigma\in\mathcal{E}_{\text{ext}}}$, and $(\bar{s}_K)_{K\in\mathcal{T}}$. This proves that $0 \leq s'_K \leq 1$.

Since $\mathcal{A}$ is continuous, we can apply Brouwer's fixed point theorem. This gives the existence of $(u,s) \in E \times L_{\mathcal{D}}(\Omega,[0,1])$ such that $\mathcal{A}(u,s) = (u,s)$, which proves the existence of a solution to (52)–(56).

5.2.2. Convergence of the scheme. We have the following result, which appears to be very weak compared to the initial ambition of approximating problem (50).

Theorem 3. *Let us assume Hypotheses (H). Let $(\mathcal{D}_n)_{n\in\mathbb{N}}$ be a sequence such that, for all $n \in \mathbb{N}$, $\mathcal{D}_n$ is an admissible discretization of Ω in the sense of Definition* 1, *and* $\lim_{n\longrightarrow\infty} \text{size}(\mathcal{D}_n) = 0$.

Then there exists a subsequence of $(\mathcal{D}_n)_{n\in\mathbb{N}}$, again denoted $(\mathcal{D}_n)_{n\in\mathbb{N}}$, such that, denoting for all $n \in \mathbb{N}$, $(u_n, s_n, \mu_n) \in H_{\mathcal{D}_n}(\Omega)\times L_{\mathcal{D}_n}(\Omega,[0,1])\times\mathcal{M}_{\mathcal{D}_n}(\alpha,\beta)$ the solution given by the scheme (52)*–*(56) *with $\mathcal{D} = \mathcal{D}_n$, we have that*

- *the sequence $(\mathcal{D}_n, \mu_n)_{n\in\mathbb{N}}$ Hd-converges in the sense of Theorem 2 to a measurable function $M \in \mathcal{M}(\alpha,\beta,\Omega)$, which implies that u_n converges to $\bar{u} = \mathcal{F}(\bar{f}, M)$ in $L^2(\Omega)$ as $n \longrightarrow \infty$;*
- *there exists a function $s \in L^\infty(\Omega)$, with $0 \leq s \leq 1$ a.e. such that the sequence $(s_n)_{n\in\mathbb{N}}$ converges to s for the weak $\star$ topology of $L^\infty(\Omega)$ and there exists a function $\bar{\gamma} \in L^\infty(\Omega)$, with $0 \leq \bar{\gamma} \leq 1$ a.e. such that the sequence $(\gamma(s_n))_{n\in\mathbb{N}}$ converges to $\bar{\gamma}$ for the weak $\star$ topology of $L^\infty(\Omega)$;*

The first item of the conclusion of Theorem 3 is a direct consequence of Theorem 2. The second item is a consequence of the sequential weak $\star$ compactness of the closed balls of L^∞. Note that, since the way to handle the convergence of (56) does not seem to be clear, no relation is given in the previous theorem between the limit of $(\gamma(s_n)\lambda(s_n))_{n\in\mathbb{N}}$, which is a possibly degenerate diffusion if we consider the second equation of (50) as an elliptic equation on u, and the Hd-limit of $(\mathcal{D}_n, \mu_n)_{n\in\mathbb{N}}$. Such a relation can be found in the following particular case, where there exists a nondecreasing Lipschitz continuous function $k_1 \ : [0,1] \longrightarrow \mathbb{R}$, with $k_1(0) = 0$ and $k_1(1) > 0$, and a real $\Lambda \in (0,1)$ such that

$$\gamma(s) = \frac{k_1(s)}{k_1(s) + \Lambda(k_1(1) - k_1(s))}, \tag{60}$$
$$\lambda(s) = k_1(s) + \Lambda(k_1(1) - k_1(s)) \qquad \forall s \in [0,1].$$

Note that we can take in this case $\beta = k_1(1)$ and $\alpha = \Lambda k_1(1)$. This particular case corresponds to a mobility of the second phase defined by the function $\Lambda\ (k_1(1) - k_1(.))$ (this can be acceptable in some physical situations; recall that k_1 is the mobility of the first phase). We can then give the following result, which is more complete than Theorem 3 (as previously mentioned, the following theorem does not give the limit of the scheme as a solution of (50) since we could obtain such a result only within a strong convergence property for $(s_n)_{n\in\mathbb{N}}$).

THEOREM 4. *Let us assume Hypotheses (H) in the particular case* (60)*. Let $(\mathcal{D}_n)_{n\in\mathbb{N}}$ be a sequence such that, for all $n \in \mathbb{N}$, $\mathcal{D}_n$ is an admissible discretization of Ω in the sense of Definition* 1*, and* $\lim_{n\longrightarrow\infty} \mathrm{size}(\mathcal{D}_n) = 0$.

Then there exists a subsequence of $(\mathcal{D}_n)_{n\in\mathbb{N}}$, again denoted $(\mathcal{D}_n)_{n\in\mathbb{N}}$, such that, denoting for all $n \in \mathbb{N}$, $(u_n, s_n, \mu_n) \in H_{\mathcal{D}_n}(\Omega) \times L_{\mathcal{D}_n}(\Omega, [0,1]) \times \mathcal{M}_{\mathcal{D}_n}(\Lambda k_1(1), k_1(1))$ the solution given by the scheme (52)–(56) *with $\mathcal{D} = \mathcal{D}_n$, we have, in addition to the conclusions of Theorem* 3*, the existence of a function $\bar{\gamma} \in L^\infty(\Omega)$, with $0 \le \bar{\gamma} \le 1$ a.e. such that the sequence $(\gamma(s_n))_{n\in\mathbb{N}}$ converges to $\bar{\gamma}$ for the weak $\star$ topology of $L^\infty(\Omega)$ and*

$$\begin{aligned}&\int_\Omega \frac{1}{1-\Lambda}(M(x) - \Lambda k_1(1) I_N)\nabla\overline{u}(x)\cdot\nabla\bar{v}(x)dx \\ &= \int_\Omega (\gamma(\bar{s}(x))(\bar{f}(x))^+ - \bar{\gamma}(x)(\bar{f}(x))^-)\ \bar{v}(x)dx \qquad \forall \bar{v} \in H_0^1(\Omega).\end{aligned} \tag{61}$$

Note that $k_1(\cdot) = \lambda(\cdot)\gamma(\cdot) = \frac{1}{1-\Lambda}(\lambda(\cdot) - \Lambda k_1(1))$.

Proof. We have only to prove (61). Let us assume the hypotheses of Theorem 4. Let $\varphi \in C_c^\infty(\Omega)$. We define, for $n \in \mathbb{N}$ and omitting indexes n in the right-hand sides,

$$D_n = \sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{E}_K} \mu_\sigma \tau_{K,\sigma}(u_\sigma - u_K)(\varphi(y_\sigma) - \varphi(x_K))$$

and

$$E_n = \sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{E}_K} \tau_{K,\sigma}(u_\sigma - u_K)(\varphi(y_\sigma) - \varphi(x_K)).$$

We have, using the results of Theorem 2,

$$\lim_{n\longrightarrow\infty} D_n = \int_\Omega M(x)\nabla\overline{u}(x)\cdot\nabla\varphi(x)dx.$$

On the other hand, we have

$$\lim_{n\longrightarrow\infty} E_n = \int_\Omega \nabla\overline{u}(x)\cdot\nabla\varphi(x)dx.$$

Since we assume the particular case (60), we get that

$$\begin{aligned}\sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{E}_K} \gamma(s_\sigma(u,s,\hat{s}))\ \mu_\sigma(u,s,\hat{s})\ \tau_{K,\sigma}(u_\sigma - u_K)(\varphi(y_\sigma) - \varphi(x_K))& \\ = \frac{1}{1-\Lambda}(D_n - \Lambda k_1(1)E_n).&\end{aligned}$$

Using the fact that $\sum_{K\in\mathcal{T}}\gamma(s_K)(\bar{f}_K)^-\varphi(x_K) \longrightarrow \int_\Omega \bar{\gamma}(x)(\bar{f}(x))^-\varphi(x)dx$ as $n \longrightarrow \infty$ (thanks to the L^∞ weak $\star$ convergence of $\gamma(s_n)$ to $\bar{\gamma}$), we thus get (61) with $\varphi \in C_c^\infty(\Omega)$. Then we obtain (61) using a classical result of density.

6. Concluding remarks. The notion of Hd-convergence, developed in this paper, gives a useful tool for studying the convergence of a discrete finite volume scheme, used for the approximation of a two-phase flow in a porous medium. The proof of the Hd-convergence theorem mimics that of the H-convergence theorem; however, although the methods are similar, the limits can be different. This discrete tool is therefore adapted to the case of a coupled discretization: the discrete pressure field is the solution of a discrete scheme for an elliptic equation, the coefficients of which result from another discrete scheme in the same grid.

This tool thus helps to get the limit problem of which the limit of the approximate pressure is the solution. A weak limit also exists for the saturation since the discrete values are bounded, as well as the continuous ones. Unfortunately, we are not able to link the Hd-limit of the sequence of discrete total mobilities and a convenient limit of the sequence of saturations.

Finally, the time-dependent problem must now be studied. Following [18] in which the G-convergence notion is adapted to general parabolic time-dependent operators, it is then possible to develop a discrete H-convergence in the case of a two-phase flow in compressible porous media (see [8]).

REFERENCES

[1] H.W. Alt, S. Luckhaus, and A. Visintin, *On nonstationary flow through porous media*, Ann. Mat. Pura. Appl. (4), 136 (1984), pp. 303–316.

[2] H.W. Alt and E. Di Benedetto, *Nonsteady flow of water and oil through inhomogeneous porous media*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 12 (1985), pp. 335–392.

[3] K. Aziz and A. Settari, *Petroleum Reservoir Simulation*, Applied Science, London, 1979.

[4] J. Carrillo, *Entropy solutions for nonlinear degenerate problems*, Arch. Ration. Mech. Anal., 147 (1999), pp. 269–361.

[5] Z. Chen, *Degenerate two-phase incompressible flow*, J. Differential Equations, 171 (2001), pp. 203–232.

[6] J. Droniou and T. Gallouët, *Finite volume methods for convection-diffusion equations with right-hand side in H^{-1}*, M2AN Math. Model. Numer. Anal, 36 (2002), pp. 705–724.

[7] R. Eymard and T. Gallouët, *Convergence d'un schéma de type éléments finis-volumes finis pour un système formé d'une équation elliptique et d'une équation hyperbolique*, RAIRO Modél. Math. Anal. Numér., 27 (1993), pp. 843–861.

[8] R. Eymard and T. Gallouët, *Finite volume schemes for two-phase flows in porous media*, Comput. Vis. Sci., submitted.

[9] R. Eymard, T. Gallouët, and R. Herbin, *Finite volume methods*, in Handbook of Numerical Analysis, Vol. VII, P.G. Ciarlet and J.L. Lions, eds., North-Holland, Amsterdam, 2000, pp. 723–1020.

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

[11] G. Gagneux and M. Madaune-Tort, *Analyse mathématique de modèles non linéaires de l'ingénierie pétrolière*, Math. Appl. 22, Springer-Verlag, Berlin, 1996.

[12] S.M. Kozlov, *Averaging of difference schemes*, Math. USSR-Sb., 57 (1987), pp. 351–369.

[13] S.N. Krushkov, *First order quasilinear equations with several space variables*, Math. USSR-Sb., 10 (1970), pp. 217–243.

[14] F. Murat, *H-Convergence*, Séminaire d'analyse fonctionnelle et numérique de l'Université d'Alger, Alger, Algeria, 1977–1978.

[15] M. Ohlberger, *Convergence of a mixed finite element-finite volume method for the two phase flow in porous media*, East-West J. Numer. Math., 5 (1997), pp. 183–210.

[16] A. Piatnitski and E. Rémy, *Homogenization of elliptic difference operators*, SIAM J. Math. Anal., 33 (2001), pp. 53–83.

[17] S. Spagnolo, *Sulla convergenza di soluzioni di equazioni paraboliche ed ellittiche*, Ann. Scuola Norm. Sup. Pisa (3), 22 (1968), pp. 571–597.

[18] N. Svanstedt, *G-convergence of parabolic operators*, Nonlinear Anal., 7 (1999), pp. 807–842.

[19] L. Tartar, *Cours Peccot,* Collège de France, Paris, 1977.

[20] L. Tartar, *Homogénéisation et H-Mesures,* ESAIM Proc. 6, Paris, 1999, pp. 111–131.

[21] M.H. Vignal, *Convergence of a finite volume scheme for an elliptic-hyperbolic system*, RAIRO Modél. Math. Anal. Numér., 30 (1996), pp. 841–872.

SIAM J. NUMER. ANAL.
Vol. 41, No. 2, pp. 563–588

NUMERICAL APPROXIMATION OF THE LIFSHITZ–SLYOZOV–WAGNER EQUATION*

FRANCIS FILBET[†] AND PHILIPPE LAURENÇOT[‡]

Abstract. The Lifshitz–Slyozov–Wagner theory of coarsening (Ostwald ripening) describes the late stages of the growth by diffusional mass transfer of the grains of a new phase from a supersaturated solution. It results in a nonlinear transport equation with a nonlocal nonlinearity for the volume distribution function of the grains. A time explicit finite volume numerical scheme is proposed to solve this equation in self-similar variables and is shown to converge under a CFL condition. Numerical simulations are also presented.

Key words. Lifshitz–Slyozov–Wagner model, Ostwald ripening, finite volume method, convergence

AMS subject classifications. 35L60, 65M60, 82C21

PII. S0036142902407599

1. Introduction.

The theory of coarsening (Ostwald ripening) in alloys describes the late stages of the formation and growth of grains of a new phase from a supersaturated solution. During these stages, no new grains can form, and the determining process is the growth of the grains by diffusional mass exchange [13, 25]. More precisely, the grains of the new phase that are larger than some critical size grow at the expense of smaller ones, the critical size varying in time as a function of the degree of supersaturation. A mean-field approach for this process has been formulated by Lifshitz and Slyozov [13] and Wagner [25]. For very dilute solutions at large times, the variation in the degree of supersaturation may be neglected, and the time evolution of the volume distribution function f of the grains is given by

$$\partial_t f + \partial_x (\mathcal{V}\ f) = 0, \quad (t,x) \in \mathbb{R}_+^2, \tag{1.1}$$

with the constraint (total volume conservation)

$$\int_0^\infty x\ f(t,x)\ dx = \text{ const.}, \quad t \in \mathbb{R}_+. \tag{1.2}$$

Here $x \in \mathbb{R}_+ := (0, +\infty)$ is the volume of the grains, $t \in \mathbb{R}_+$ is the time variable, and $\mathcal{V} = \mathcal{V}(t,x)$ denotes the rate of growth of the grains, which is determined by the mechanism of mass transfer between the grains, e.g., volume diffusion [13, 25] or grain-boundary diffusion [22]. In general, one has $\mathcal{V}(t,x) = k(x)u(t) - q(x)$, where k and q are computed from the modeling of the mechanism of mass transfer between the grains [13, 22, 25]. For instance, in the model considered in [13], where the mass transfer is driven by diffusion, $\mathcal{V}$ is explicitly computable and $k(x) = 3\ x^{1/3}$, $q(x) = 3$,

*Received by the editors May 14, 2002; accepted for publication (in revised form) November 13, 2002; published electronically May 6, 2003. This research was partially supported by the ACI "Jeunes Chercheurs" *Analyse mathématique et simulation numérique de particules chargées*, grant 2000–2526.

http://www.siam.org/journals/sinum/41-2/40759.html

[†] Institut de Recherche en Mathématique Avancée, CNRS UMR 7501, Université Louis Pasteur, 7 rue René Descartes, F–67084 Strasbourg, France (filbet@math.u-strasbg.fr).

[‡] Mathématiques pour l'Industrie et la Physique, CNRS UMR 5640, Université Paul Sabatier – Toulouse 3, 118 route de Narbonne, F–31062 Toulouse cedex 4, France (laurenco@mip.ups-tlse.fr).

$x \in \mathbb{R}_+$. The function u is then determined by requiring that the solution f to (1.1) comply with (1.2), that is,

$$u(t) \int_0^\infty k(x)\ f(t,x)\ dx = \int_0^\infty q(x)\ f(t,x)\ dx, \quad t \in \mathbb{R}_+. \tag{1.3}$$

The main purpose of this work is to present a numerical scheme for solving (1.1)–(1.2) and to study the properties and the convergence of this scheme when the functions k and q determining the rate of growth of the grains $\mathcal{V}$ are given by

$$k(x) = x^\alpha \quad \text{and} \quad q(x) = 1, \quad x \in \mathbb{R}_+, \tag{1.4}$$

for some $\alpha \in (0,1)$. (Recall that $\alpha = 1/3$ is the case considered in [13].) We will not, however, study (1.1)–(1.2) directly but will first perform a couple of transformations to obtain an equivalent formulation more suitable for our purposes. As in [17], we first introduce the number $F(t,x)$ of grains of size larger than x at time t, that is,

$$F(t,x) = \int_x^\infty f(t,x')\ dx', \quad (t,x) \in \mathbb{R}_+^2. \tag{1.5}$$

The constraint (1.2) then straightforwardly translates to the conservation of the L^1-norm of $F(t)$ throughout time evolution, and F solves

$$\partial_t F + \mathcal{V}\ \partial_x F = 0, \qquad \int_0^\infty F(t,x)\ dx = \text{ const.}, \quad (t,x) \in \mathbb{R}_+^2. \tag{1.6}$$

The second transformation we shall perform is related to the large time behavior of solutions to (1.1)–(1.2) and is motivated by the following fact: formal asymptotic expansions performed in [13] for $\alpha = 1/3$ indicate that the pair (f,u) approaches a self-similar form as time increases to infinity with the following scaling:

$$f(t,x) \sim t^{-2}\ f_\infty\left(\frac{x}{t}\right) \quad \text{and} \quad u(t) \sim u_\infty\ t^{-\alpha}.$$

Observing that convergence to a self-similar profile translates to convergence to a steady state in self-similar variables, we introduce

$$\left\{ \begin{aligned} F(t,x) &= (1+t)^{-1}\ G\left(\ln(1+t), x/(1+t)\right), \\ u(t) &= (1+t)^{-\alpha}\ v\left(\ln(1+t)\right), \end{aligned} \right. \qquad (t,x) \in \mathbb{R}_+^2. \tag{1.7}$$

It then follows from (1.6) that (G,v) satisfies

$$\partial_t G + \mathcal{W}\ \partial_x G = G, \qquad \int_0^\infty G(t,x) = \text{ const.}, \tag{1.8}$$

where $\mathcal{W}(t,x) = x^\alpha\ v(t) - 1 - x$, $(t,x) \in \mathbb{R}_+^2$. In this paper, we will focus on this alternative formulation of the Lifshitz–Slyozov–Wagner (LSW) equation (1.1)–(1.2). We investigate the properties and the convergence of a numerical scheme for (1.8) built upon an explicit Euler discretization with respect to the time variable t and a finite volume discretization with respect to the volume variable x. Finally, numerical simulations will be presented, allowing us to check the numerical convergence of the scheme and to compare the large time behavior of our approximation with that expected for the solution to (1.8).

Let us provide some comments about the behavior of our numerical scheme, referring to section 5 for a more complete discussion. Unlike what was conjectured by Lifshitz and Slyozov [13], the large time behavior of solutions to (1.8) is complex and very sensitive to perturbations. In particular, according to the analysis in [14, 20], the behavior for large times of solutions to (1.8) with compactly supported initial data changes drastically in the presence of a small diffusion (say, an additional term $\eta\, \partial_x^2 G$ on the right-hand side of (1.8) with $\eta > 0$). Since our numerical scheme is a classical upwind method, a small numerical diffusion comes into play during the simulations. It is then unlikely that our scheme reproduces the correct large time behavior for compactly supported initial data, and this is exactly what we observe in the numerical simulations. On the other hand, our scheme gives the correct limit for noncompactly supported initial data. In order to capture the expected behavior for large times for arbitrary initial data, it seems that a less diffusive numerical scheme is needed. One possibility is to use a higher-order scheme, and this is the approach developed by Carrillo and Goudon in [2], where a WENO (weight essentially nonoscillatory)-type scheme is used to numerically compute the solutions to (1.8). (The main focus of [2] is actually the variant of (1.8) described in Remark 1.1 below.) The numerical simulations reported in [2] show that such a scheme gives the expected behavior for intermediate times, providing better results than our scheme. Still, for larger times, some numerical diffusion effects also come into play and drive the numerical solution away from the theoretical predictions. Another approach relies on a nonlinear and antidissipative scheme [5]. It has been recently considered by Lagoutière and seems to successfully compute the correct behavior, even for large times [8]. Let us point out, however, that no convergence proof seems to be available for these schemes.

Before describing our results more precisely, let us recall that the LSW equation (1.1)–(1.2) has been the object of several studies recently; existence and uniqueness of weak solutions have been proved in [10, 17, 19] for the initial value problem (1.1)–(1.2) under various assumptions on the functions k and q determining the growth rate of the grains $\mathcal{V}$ and the initial data. Also, the large time asymptotics have been investigated in [1, 16] by analytical means and in [2, 6] by numerical simulations.

Remark 1.1. A different version of the LSW equation (originally introduced in [13]), in which the constraint (1.2) is replaced by

$$u(t) + A \int_0^\infty x\, f(t,x)\, dx = Q, \quad t \in \mathbb{R}_+, \tag{1.9}$$

is actually the main concern of [2, 18]. In (1.9), Q is the total initial supersaturation and A is a physical constant [13]. Still, the large time behavior of solutions to (1.1), (1.9) is expected to be the same as that of (1.1), (1.2), provided that $u(t)$ defined by (1.9) converges to zero, which is true for initial data with a sufficiently wide support [2, 18]. Let us also mention that the well-posedness of the initial value problem (1.1), (1.9) has been studied in [3, 9, 17].

We now briefly outline the contents of the paper. In the next section, we introduce the numerical approximation of (1.8) and state the convergence result, which we prove in sections 3 and 4. Two points are worth mentioning here. First, it readily follows from (1.5) and the nonnegativity of f that $x \mapsto F(t,x)$ is nonincreasing and so is $x \mapsto G(t,x)$ by (1.8). At the discrete level, our approximation of G also enjoys this property. Secondly, since the definition of v involves the inverse of a moment of G (recall the definition (1.3) of u), an important step of the convergence proof is the derivation of a uniform L^∞-estimate on the approximations of v. For compactly supported initial data, such a bound has been obtained by estimating the time evolution of the support of the solution [10, 17], but this method does not seem to work here because of the (small) viscosity induced by the numerical approximation. We therefore use a different approach and obtain a new L^∞-bound in terms of the first moment of G. Since the proof of this estimate is quite technical at the discrete level, we also provide a (formal) proof for (1.1)–(1.2) in the appendix, hoping to clarify the underlying idea. The final section (section 5) is devoted to some numerical simulations performed with the numerical scheme presented in section 2.

2. Main results. Before describing our numerical scheme and stating a convergence result, we first introduce some notation and assumptions and recall previous results on (1.8). As already mentioned, we focus on the approximation of the initial value problem

$$\partial_t G + \partial_x (\mathcal{W}\ G) = S, \quad (t,x) \in \mathbb{R}_+^2, \tag{2.1}$$

$$\alpha\ v(t) \int_0^\infty x^{\alpha-1}\ G(t,x)\ dx = G(t,0), \quad t \in \mathbb{R}_+, \tag{2.2}$$

$$G(0,x) = G_0(x), \quad x \in \mathbb{R}_+, \tag{2.3}$$

where $\alpha \in (0,1)$ is fixed,

$$\begin{aligned} \mathcal{W}(t,x) &= x^\alpha\ v(t) - 1 - x, \quad (t,x) \in \mathbb{R}_+^2, \\ S(t,x) &= \alpha\ x^{\alpha-1}\ v(t)\ G(t,x), \quad (t,x) \in \mathbb{R}_+^2, \end{aligned} \tag{2.4}$$

and we assume that the initial datum G_0 satisfies

(2.5) $G_0 \in W^{1,1}(\mathbb{R}_+) \cap L^1(\mathbb{R}_+, xdx)$ is a nonnegative and nonincreasing function and $G_0 \not\equiv 0$.

Here and below, the notation $L^1(\mathbb{R}_+, xdx)$ stands for the space of the Lebesgue measurable real-valued functions on $\mathbb{R}_+$ which are integrable with respect to the measure xdx.

Observe that (2.1) is nothing but $\partial_t G + \mathcal{W}\ \partial_x G = G$ written in conservative form. Next, as a consequence of [10, Theorem 2] and Proposition A.1, there are at least a pair of nonnegative functions (G, v) satisfying

$$G \in \mathcal{C}([0,T]; L^1(\mathbb{R}_+)) \cap \mathcal{C}^1(0,T; L^1(\mathbb{R}_+, \min\{x,1\}dx)), \qquad v \in L^\infty(0,T)$$

for each $T \in \mathbb{R}_+$ and (2.1), (2.2), (2.3) with $\mathcal{W}$ given by (2.4). In addition, $x \mapsto G(t,x)$ is a nonincreasing function for each $t \geq 0$ and

$$\int_0^\infty G(t,x)\ dx = \int_0^\infty G_0(x)\ dx, \quad t \geq 0, \tag{2.6}$$

the identity (2.6) being actually equivalent to (2.2). Furthermore, the uniqueness of the pair (G, v) follows from [18] if G_0 is compactly supported, and from [19] in the general case. (Only the case $\alpha = 1/3$ is actually considered in [18, 19], but their proofs extend to $\alpha \in (0,1)$.) Let us finally point out that the integrability assumption $G_0 \in L^1(\mathbb{R}_+; xdx)$ is not needed for the existence of a solution to (2.1), (2.2), (2.3). It can probably be dispensed with herein also but allows us to avoid many technicalities in the proof of the L^∞-bound for v and its approximations.

Next, let $h \in (0,1)$ denote the mesh size and set

$$x_{-1/2} = 0, \quad x_i = x_{i-1/2} + \frac{h}{2}, \quad x_{i+1/2} = x_{i-1/2} + h, \tag{2.7}$$

and $\Lambda_i^h = [x_{i-1/2}, x_{i+1/2})$ for $i \geq 0$. Since x ranges in the unbounded domain $\mathbb{R}_+$, the numerical solution will actually be computed on the bounded domain $[0, x_{I^h+1/2})$, where I^h is a large integer depending on h. We shall of course require that $h\ I^h \to +\infty$ as $h \to 0$. We then define the approximation $G^{0,h}$ of the initial datum G_0 as usual by

$$G^{0,h} = \sum_{i=0}^{I^h} G_i^{0,h}\ \mathbf{1}_{\Lambda_i^h} \quad \text{with} \quad G_i^{0,h} = \frac{1}{h} \int_{\Lambda_i^h} G_0(x)\ dx, \tag{2.8}$$

and recall that

$$\|G^{0,h}\|_{L^1} \leq \|G_0\|_{L^1} \quad \text{and} \quad \lim_{h \to 0} \|G^{0,h} - G_0\|_{L^1} = 0. \tag{2.9}$$

Here and below, $\mathbf{1}_E$ denotes the characteristic function of the subset E of $\mathbb{R}_+$.

Finally, let $T \in \mathbb{R}_+$ be some final time and N the number of time iterations, and set

$$\Delta t = \frac{T}{N}, \qquad t^n = n\ \Delta t, \quad 0 \leq n \leq N.$$

The data h, Δt, and I^h have to fulfill the following conditions: we first require that the domain of computation approach $[0, +\infty)$ and the discrete initial data be close enough to G_0, that is,

$$\lim_{h \to 0} h\ I^h = +\infty, \qquad \|G^{0,h}\|_{L^1} \geq \frac{1}{2}\ \|G_0\|_{L^1}. \tag{2.10}$$

We also impose the following CFL condition:

$$10\ \frac{\Delta t}{h}\ \left(h\ I^h\right) \leq 1. \tag{2.11}$$

Observe that the above constraints are satisfied when $I^h \sim h^{-1-\vartheta}$ for some $\vartheta > 0$ and when $\Delta t\ h^{-1-\vartheta}$ is sufficiently small.

Denoting by $G_i^{n,h}$ an approximation of the mean value of $G(t^n)$ on Λ_i^h for $i \in \{0, \ldots, I^h\}$, and by v^n an approximation of $v(t^n)$, the numerical scheme to be studied in this paper reads

$$G_i^{n+1,h} = G_i^{n,h} - \frac{\Delta t}{h}\ \left(F_{i+1/2}^{n,h} - F_{i-1/2}^{n,h}\right) + \Delta t\ S_i^{n,h}, \quad 0 \leq i \leq I^h, \tag{2.12}$$

$$G_{-1}^{n,h} = G_{I^h+1}^{n,h} = 0, \tag{2.13}$$

$$h\ v^{n+1}\ \left(\sum_{i=0}^{I^h} a_i^h\ G_i^{n+1,h}\right) = G_0^{n+1,h}, \tag{2.14}$$

for $n \in \{0, \ldots, N-1\}$, with initial data $(G_i^{0,h})_{0 \leq i \leq I^h}$ defined in (2.8) and v^0 given by (2.14) with $n = -1$. In (2.12) the approximate flux $F_{i+1/2}^{n,h}$ is given by

$$F_{i+1/2}^{n,h} = \nu_+^n(x_{i+1/2})\ G_i^{n,h} - \nu_-^n(x_{i+1/2})\ G_{i+1}^{n,h}, \quad -1 \leq i \leq I^h, \tag{2.15}$$

with

$$\nu^n(x) = x^\alpha\ v^n - 1 - x, \quad x \in \mathbb{R}_+, \tag{2.16}$$

$\nu_+^n(x) = \max\{0, \nu^n(x)\}$, $\nu_-^n(x) = \max\{0, -\nu^n(x)\}$, and the source term $S_i^{n,h}$ is given by

$$S_i^{n,h} = a_i^h \; v^n \; G_i^{n,h}, \quad \text{with} \quad a_i^h = \frac{\alpha}{h} \int_{\Lambda_i^h} x^{\alpha-1} \, dx, \quad 0 \le i \le I^h. \tag{2.17}$$

Remark 2.1. Note that the boundary condition $G_{I^h+1}^{n,h} = 0$ in (2.13) is needed only when $\nu^n(x_{I^h+1/2}) < 0$.

Before stating some properties on the scheme (2.12)–(2.17), let us briefly comment on its derivation, which relies obviously on an explicit Euler scheme for the time variable and on a finite volume approach for the volume variable (see, e.g., [7, 12]). Concerning the latter, the formula (2.15) comes from the approximation by a classical upwind scheme of the fluxes $\mathcal{W}(t^n, x_{i+1/2})$ and $\mathcal{W}(t^n, x_{i-1/2})$ arising from the integration of (2.1) over the cell Λ_i^h. As for (2.14), it is a discrete version of (2.2), which guarantees the conservation of the L^1-norm of G^h; see (2.20) below.

Under the conditions (2.10), (2.11) and if hI^h is large enough, the solution $(G_i^{n,h})$ to the scheme (2.12)–(2.17) enjoys properties similar to those of G, which we gather in Proposition 2.2 below.

PROPOSITION 2.2. *There is a positive constant $x_\star$ depending only on α, G_0, and T such that, if*

$$h \; I^h \ge x_\star \tag{2.18}$$

and the conditions (2.10), (2.11) *are fulfilled, the solution* $(G_i^{n,h})$ *to the scheme* (2.12)–(2.17) *satisfies the following:*

- *nonnegativity and monotonicity:*

$$0 \le G_{i+1}^{n,h} \le G_i^{n,h} \le G_0^{n,h}, \quad 0 \le i \le I^h - 1, \tag{2.19}$$

- *conservation of the total volume:*

$$\sum_{i=0}^{I^h} h \; G_i^{n,h} = \sum_{i=0}^{I^h} h \; G_i^{0,h} \tag{2.20}$$

for $n \in \{0, \ldots, N\}$.

We next define the numerical approximation (G^h, v^h) of (G, v) by

$$G^h(t,x) = \sum_{i=0}^{I^h} G_i^{n,h} \; \mathbf{1}_{\Lambda_i^h}(x), \quad v^h(t) = v^n, \quad x \in \mathbb{R}_+, \tag{2.21}$$

for $t \in [t^n, t^{n+1})$ and $n \in \{0, \ldots, N-1\}$, and

$$G^h(T,x) = \sum_{i=0}^{I^h} G_i^{N,h} \; \mathbf{1}_{\Lambda_i^h}(x), \quad v^h(T) = v^N, \quad x \in \mathbb{R}_+. \tag{2.22}$$

We may now state our main result.

THEOREM 2.3. *Assume that the conditions* (2.10), (2.11), *and* (2.18) *are fulfilled and that G_0 satisfies* (2.5). *Then*

$$G^h \longrightarrow G \quad in \quad L^\infty(0,T; L^1(\mathbb{R}_+)), \tag{2.23}$$

$$v^h \overset{*}{\rightharpoonup} v \quad in \quad L^\infty(0,T), \tag{2.24}$$

where (G,v) is the weak solution to (2.1), (2.2), (2.3) *on* $[0,T]$ *with initial datum* G_0. *More precisely,* (G,v) *is a pair of nonnegative functions satisfying*

$$(2.25)\qquad \begin{cases} G \in \mathcal{C}([0,T];L^1(\mathbb{R}_+)) \cap L^\infty\big(0,T;W^{1,1}(\mathbb{R}_+;(1+x)dx)\big), \\ v \in L^\infty(0,T) \end{cases}$$

and

$$(2.26)\qquad \int_0^\infty (G(t)-G_0)\ \varphi\ dx = \int_0^t \int_0^\infty (G(s) - \mathcal{W}(s)\ \partial_x G(s))\ \varphi\ dxds$$

for each $t \in [0,T]$ *and* $\varphi \in L^\infty(\mathbb{R}_+)$, *where* v *and* $\mathcal{W}$ *are given by* (2.2) *and* (2.4), *respectively. Equivalently,* G *satisfies* (2.6). *In addition,* $x \mapsto G(t,x)$ *is nonincreasing for each* $t \in [0,T]$.

Observe that each term in (2.26) makes sense since, by (2.25), $\mathcal{W}$ and $\partial_x G$ belong to $L^\infty((0,T)\times\mathbb{R}_+,(1+x)^{-1}dtdx)$ and $L^\infty(0,T;L^1(\mathbb{R}_+,(1+x)dx))$, respectively.

Let us finally explain the main steps of the proof of Theorem 2.3. Similarly to the existence proof in [10], the proof of Theorem 2.3 relies on estimates of G^h and its discrete gradient in $L^\infty(0,T;L^1(\mathbb{R}_+,(1+x)dx))$ and on an $L^\infty(0,T)$-estimate on v^h. On the continuous equation (2.1), (2.2), (2.3), these bounds are obtained as follows: uniform estimates for G in $L^1(\mathbb{R}_+)$ and $L^\infty(\mathbb{R}_+)$ are straightforward consequences of (2.6) and (2.1), respectively. The main new observation then is that an upper bound on v can be obtained from (2.2) and the previous estimates in terms of the $L^1(\mathbb{R}_+;xdx)$-norm of G (Lemma 3.1). Inserting this estimate into (2.1) yields a uniform estimate for G in $L^1(\mathbb{R}_+;xdx)$, and thus an upper bound for v in $L^\infty(0,T)$ (Lemma 3.2). The equation satisfied by $\partial_x G$ then reveals an $L^1(\mathbb{R}_+)$-weak compactness estimate on $\partial_x G$, which, in turn, implies some time equicontinuity on G (see Lemmas 3.5 and 3.7 and (4.12) below). From these estimates, one deduces that G lies in compact subsets of $\mathcal{C}([0,T];L^1(\mathbb{R}_+))$ and $L^1(0,T;W^{1,1}(\mathbb{R}_+))$. At the discrete level, we perform the same steps for (G^h,v^h), which is possible thanks to the conditions (2.10), (2.11), and (2.18).

3. Properties of $(G_i^{n,h})$.

This section is devoted to the proof of Proposition 2.2 and the uniform bounds satisfied by $(G_i^{n,h})$. The parameters h, Δt, and I^h being fixed such that (2.10), (2.11), and (2.18) are fulfilled, we omit the superscript h throughout this section. Also, owing to (2.10), we may assume without loss of generality that $h \in (0,1)$ and $x_{I+1/2} \geq 2$.

Lemma 3.1. *Let* $n \in \{0,\dots,N\}$ *be such that*

$$(3.1)\qquad \sum_{i=0}^{I} h\ G_i^n = \sum_{i=0}^{I} h\ G_i^0 \quad \text{and} \quad G_i^n \geq 0 \quad \text{for} \quad i \in \{0,\dots,I\}.$$

Then there is a positive constant C_1 *depending only on* α *and* $\|G_0\|_{L^1}$ *such that*

$$(3.2)\qquad 0 \leq v^n \leq C_1\ G_0^n\ \left(1 + \left(\sum_{i=\ell+1}^{I} h\ x_{i-1/2}\ G_i^n\right)^{1-\alpha}\right),$$

where $\ell \in \{0,\dots,I/2\}$ *denotes the integer such that* $1 \in \Lambda_\ell$.

Proof. We infer from (2.10), (3.1), and the Hölder inequality that

$$\begin{aligned}\frac{\|G_0\|_{L^1}}{2} &\le \sum_{i=0}^{I} h\ G_i^0 = \sum_{i=0}^{I} h\ G_i^n \\ &\le \left(\sum_{i=0}^{I} h\ x_{i+1/2}^{\alpha-1}\ G_i^n\right)^{1/(2-\alpha)} \left(\sum_{i=0}^{I} h\ x_{i+1/2}\ G_i^n\right)^{(1-\alpha)/(2-\alpha)}, \\ \left(\frac{\|G_0\|_{L^1}}{2}\right)^{2-\alpha} &\le \left(\frac{1}{\alpha}\ \sum_{i=0}^{I} h\ a_i\ G_i^n\right) \left(\sum_{i=0}^{I} h\ x_{i+1/2}\ G_i^n\right)^{1-\alpha}.\end{aligned}$$

Multiplying both sides of the above inequality by v^n and using (2.14) yields

$$\left(\frac{\|G_0\|_{L^1}}{2}\right)^{2-\alpha}\ v^n \le \frac{G_0^n}{\alpha}\ \left(\sum_{i=0}^{I} h\ x_{i+1/2}\ G_i^n\right)^{1-\alpha}.$$

Since $x_{i+1/2} \le x_{\ell+1/2} \le 2$ for $0 \le i \le \ell$, and $x_{i+1/2} \le 2\ x_{i-1/2}$ for $i \ge \ell+1$, we deduce from (2.9) and (3.1) that

$$\begin{aligned}\sum_{i=0}^{I} h\ x_{i+1/2}\ G_i^n &\le 2\ \sum_{i=0}^{\ell} h\ G_i^n + 2\ \sum_{i=\ell+1}^{I} h\ x_{i-1/2}\ G_i^n \\ &\le 2\ \|G_0\|_{L^1} + 2\ \sum_{i=\ell+1}^{I} h\ x_{i-1/2}\ G_i^n.\end{aligned}$$

Combining the previous two inequalities, we end up with

$$\begin{aligned}v^n &\le \left(\frac{2}{\|G_0\|_{L^1}}\right)^{2-\alpha}\ \frac{2^{1-\alpha}\ G_0^n}{\alpha}\ \left(\|G_0\|_{L^1}^{1-\alpha} + \left(\sum_{i=\ell+1}^{I} h\ x_{i-1/2}\ G_i^n\right)^{1-\alpha}\right) \\ &\le C_1\ G_0^n\ \left(1 + \left(\sum_{i=\ell+1}^{I} h\ x_{i-1/2}\ G_i^n\right)^{1-\alpha}\right),\end{aligned}$$

whence (3.2). □

LEMMA 3.2. *Let $n \in \{0,\dots,N-1\}$ be such that* (3.1) *holds true and*

$$\nu_+^n(x_{I+1/2}) = 0. \tag{3.3}$$

Then

$$\sum_{i=0}^{I} h\ G_i^{n+1} = \sum_{i=0}^{I} h\ G_i^0, \tag{3.4}$$

$$0 \le G_i^{n+1} \le (1+\Delta t)\ \sup_j \{G_j^n\}, \quad 0 \le i \le I, \tag{3.5}$$

$$\sum_{i=\ell+1}^{I} h\ x_{i-1/2}\ G_i^{n+1} \le \left(1 + C_2\ \left(1 + \sup_i \{G_i^n\}\right)\ \Delta t\right)\ \sum_{i=\ell+1}^{I} h\ x_{i-1/2}\ G_i^n$$
$$+ C_2\ \left(1 + \left(\sup_i \{G_i^n\}\right)^{(1+\alpha)/\alpha}\right)\ \Delta t, \tag{3.6}$$

where C_2 is a positive constant depending only on α and $\|G_0\|_{L^1}$.

Proof. We first check (3.4). For $n \in \{0, \dots, N-1\}$, it follows from (2.12), (2.13), (2.14), and (2.15) that

$$\begin{aligned}\sum_{i=0}^{I} h\, G_i^{n+1} &= \sum_{i=0}^{I} h\, G_i^n - \Delta t \sum_{i=1}^{I+1} F_{i-1/2}^n + \Delta t \sum_{i=0}^{I} F_{i-1/2}^n + h\, \Delta t\, v^n \sum_{i=0}^{I} a_i\ G_i^n \\ &= \sum_{i=0}^{I} h\, G_i^n - \Delta t\, G_0^n - \Delta t\, \nu_+^n(x_{I+1/2})\ G_I^n + h\, \Delta t\, v^n \sum_{i=0}^{I} a_i\ G_i^n \\ &= \sum_{i=0}^{I} h\, G_i^n - \Delta t\, \nu_+^n(x_{I+1/2})\ G_I^n,\end{aligned}$$

whence (3.4) by (3.3). Before completing the proof of Lemma 3.2, we provide an alternative formulation of (2.12). By (2.15) we have

$$\begin{aligned}F_{i+1/2}^n - F_{i-1/2}^n &= \nu_+^n(x_{i+1/2})\ G_i^n - \nu_-^n(x_{i+1/2})\ G_{i+1}^n \\ &\quad - \nu_+^n(x_{i-1/2})\ G_{i-1}^n + \nu_-^n(x_{i-1/2})\ G_i^n \\ &= (\nu^n(x_{i+1/2}) - \nu^n(x_{i-1/2}))\ G_i^n \\ &\quad + \nu_-^n(x_{i+1/2})\ (G_i^n - G_{i+1}^n) + \nu_+^n(x_{i-1/2})\ (G_i^n - G_{i-1}^n).\end{aligned}$$

Since

$$\nu^n(x_{i-1/2}) - \nu^n(x_{i+1/2}) + h\ a_i\ v^n = h, \tag{3.7}$$

we insert the above formula for $F_{i+1/2}^n - F_{i-1/2}^n$ into (2.12) and obtain

$$G_i^{n+1} = (1 + \Delta t)\ G_i^n + \frac{\Delta t}{h}\ \nu_-^n(x_{i+1/2})\ (G_{i+1}^n - G_i^n)$$
$$+ \frac{\Delta t}{h}\ \nu_+^n(x_{i-1/2})(G_{i-1}^n - G_i^n). \tag{3.8}$$

Now, since $\nu_+^n(x_{I+1/2}) = 0$ by (3.3) and $x_{I+1/2} \ge 1$, we have $v^n \le 2\ x_{I+1/2}^{1-\alpha}$, and (2.11) ensures that

$$\left|\nu^n(x_{i+1/2})\right| \le x_{i+1/2}^\alpha\ v^n + 1 + x_{i+1/2} \le 2\ x_{I+1/2} + 2\ x_{I+1/2} \le \frac{h}{2\ \Delta t} \tag{3.9}$$

for $0 \le i \le I$. Owing to (3.9) and the nonnegativity (3.1) of $(\Delta t\ G_i^n)$, it follows from (3.8) that G_i^{n+1} lies above a convex combination of G_{i-1}^n, G_i^n, and G_{i+1}^n, which are nonnegative by (3.1), whence the nonnegativity of G_i^{n+1} for $0 \le i \le I$. Similarly, we infer from (3.8) and (3.9) that $G_i^{n+1}/(1 + \Delta t)$ is a convex combination of G_{i-1}^n, G_i^n, and G_{i+1}^n, from which we deduce (3.5).

We next turn to (3.6). We infer from (2.13), (3.1), and (3.8) that

$$\begin{aligned}
&\sum_{i=\ell+1}^{I} h\ x_{i-1/2}\ \left(G_i^{n+1} - (1+\Delta t)\ G_i^n\right) \\
&= \Delta t \sum_{i=\ell+2}^{I+1} x_{i-3/2}\ \nu_-^n(x_{i-1/2})\ G_i^n + \Delta t \sum_{i=\ell}^{I-1} x_{i+1/2}\ \nu_+^n(x_{i+1/2})\ G_i^n \\
&\quad - \Delta t \sum_{i=\ell+1}^{I} x_{i-1/2}\ \left(\nu_-^n(x_{i+1/2}) + \nu_+^n(x_{i-1/2})\right)\ G_i^n \\
&\le \Delta t \sum_{i=\ell+1}^{I} (x_{i-3/2} - x_{i-1/2})\ \nu_-^n(x_{i-1/2})\ G_i^n \\
&\quad + \Delta t \sum_{i=\ell+1}^{I} x_{i-1/2}\ \left(\nu_-^n(x_{i-1/2}) - \nu_-^n(x_{i+1/2})\right)\ G_i^n \\
&\quad + \Delta t \sum_{i=\ell+1}^{I} (x_{i+1/2} - x_{i-1/2})\ \nu_+^n(x_{i-1/2})\ G_i^n \\
&\quad + \Delta t \sum_{i=\ell+1}^{I} x_{i+1/2}\ \left(\nu_+^n(x_{i+1/2}) - \nu_+^n(x_{i-1/2})\right)\ G_i^n \\
&\quad + \Delta t\ x_{\ell+1/2}\ \nu_+^n(x_{\ell+1/2})\ G_\ell^n \\
&\le \Delta t \sum_{i=\ell+1}^{I} h\ \nu^n(x_{i-1/2})\ G_i^n \\
&\quad + \Delta t \sum_{i=\ell+1}^{I} x_{i-1/2}\ \left(\nu^n(x_{i-1/2}) - \nu^n(x_{i+1/2})\right)_-\ G_i^n \\
&\quad + \Delta t \sum_{i=\ell+1}^{I} x_{i+1/2}\ \left(\nu^n(x_{i+1/2}) - \nu^n(x_{i-1/2})\right)_+\ G_i^n \\
&\quad + \Delta t\ x_{\ell+1/2}\ \nu_+^n(x_{\ell+1/2})\ G_\ell^n,
\end{aligned}$$

the last inequality being a consequence of the subadditivity of $r \mapsto r_+$ and $r \mapsto r_-$. Since $h \in (0,1)$, the choice of ℓ guarantees that $x_{\ell+1/2} \le 2$ and $x_{i+1/2} \le 2\ x_{i-1/2}$ for $i \ge \ell+1$. Consequently, for $i \ge \ell+1$,

$$\begin{aligned}
x_{i-1/2}\ \left(\nu^n(x_{i-1/2}) - \nu^n(x_{i+1/2})\right)_- &\le v^n\ x_{i-1/2}\ \left(x_{i-1/2}^\alpha - x_{i+1/2}^\alpha\right)_- \\
&\le v^n\ x_{i-1/2}\ \alpha\ h\ x_{i-1/2}^{\alpha-1} \\
&\le h\ v^n\ x_{i-1/2}^\alpha,
\end{aligned}$$

$$x_{i+1/2}\ \left(\nu^n(x_{i+1/2}) - \nu^n(x_{i-1/2})\right)_+ \le 2\ h\ v^n\ x_{i-1/2}^\alpha,$$

and $\nu^n(x_{i-1/2}) \le v^n\ x_{i-1/2}^\alpha$, while

$$x_{\ell+1/2}\ \nu_+^n(x_{\ell+1/2})\ \le v^n\ x_{\ell+1/2}^{1+\alpha} \le 4\ v^n.$$

Therefore, since $G_i^n \geq 0$ by (3.1),

$$\begin{aligned} &\sum_{i=\ell+1}^{I} h\ x_{i-1/2}\ \left(G_i^{n+1} - (1+\Delta t)\ G_i^n\right) \\ &\leq \Delta t\ v^n\ \sum_{i=\ell+1}^{I} h\ x_{i-1/2}^{\alpha}\ G_i^n + 3\ \Delta t\ v^n\ \sum_{i=\ell+1}^{I} h\ x_{i-1/2}^{\alpha}\ G_i^n + 4\ \Delta t\ v^n\ G_\ell^n \\ &\leq 4\ \Delta t\ v^n\ \left(\sum_{i=\ell+1}^{I} h\ G_i^n\right)^{1-\alpha}\ \left(\sum_{i=\ell+1}^{I} h\ x_{i-1/2}\ G_i^n\right)^{\alpha} + 4\ \Delta t\ v^n\, G_\ell^n \end{aligned}$$

by the Hölder inequality. Using (3.1) once more yields

$$\begin{aligned} &\sum_{i=\ell+1}^{I} h\ x_{i-1/2}\ \left(G_i^{n+1} - (1+\Delta t)\ G_i^n\right) \\ &\leq 4\ \Delta t\ \|G_0\|_{L^1}^{1-\alpha}\ v^n\ \left(\sum_{i=\ell+1}^{I} h\ x_{i-1/2}\ G_i^n\right)^{\alpha} + 4\ \Delta t\ v^n\ G_\ell^n. \end{aligned} \tag{3.10}$$

Owing to (3.1), we may use Lemma 3.1 and insert (3.2) into (3.10) to obtain, with the help of the Young inequality,

$$\begin{aligned} &\sum_{i=\ell+1}^{I} h\ x_{i-1/2}\ \left(G_i^{n+1} - (1+\Delta t)\ G_i^n\right) \\ &\leq C_1'\ \Delta t\ G_0^n\ \left(\sum_{i=\ell+1}^{I} h\ x_{i-1/2}\ G_i^n\right)^{\alpha} + C_1'\ \Delta t\ G_0^n\ \sum_{i=\ell+1}^{I} h\ x_{i-1/2}\ G_i^n \\ &\quad + C_1'\ \Delta t\ G_0^n\ \sup_i\ \{G_i^n\} + C_1'\ \Delta t\ G_0^n\ \sup_i\ \{G_i^n\}\ \left(\sum_{i=\ell+1}^{I} h\ x_{i-1/2}\ G_i^n\right)^{1-\alpha} \\ &\leq 2\ C_1'\ \Delta t\ \sup_i\ \{G_i^n\}\ \left(\sum_{i=\ell+1}^{I} h\ x_{i-1/2}\ G_i^n + 1 + \sup_i\ \{G_i^n\}^{1/\alpha}\right), \end{aligned}$$

with $C_1' = 4C_1 \max\{1, \|G_0\|_{L^1}^{1-\alpha}\}$, whence (3.6), and the proof of Lemma 3.2 is complete. □

We now introduce

$$\begin{aligned} &K_1 := 2\ C_1\ \|G_0\|_{L^\infty}\ e^T, \qquad K_2 := C_2\ \left(1 + \|G_0\|_{L^\infty}\ e^T\right), \\ &K_3 := C_2\ \left(1 + \|G_0\|_{L^\infty}^{(1+\alpha)/\alpha}\ e^{(1+\alpha)T/\alpha}\right), \\ &x_\star := \left\{K_1\ e^{K_2 T}\ \left(1 + \int_0^\infty x\ G_0(x)\ dx + K_3\ T\right)\right\}^{1/(1-\alpha)}. \end{aligned}$$

Proposition 3.3. *Assume that* (2.18) *holds true. For* $n \in \{0, \ldots, N\}$, *we have*

$$\sum_{i=0}^{I} h\ G_i^n = \sum_{i=0}^{I} h\ G_i^0, \tag{3.11}$$

$$0 \le G_i^n \le (1+\Delta t)^n \ \|G_0\|_{L^\infty}, \quad 0 \le i \le I, \tag{3.12}$$

$$\begin{aligned}\sum_{i=\ell+1}^{I} h\ x_{i-1/2}\ G_i^n \le (1+K_2\ \Delta t)^n \ \int_0^\infty x\ G_0(x)\ dx \\ + K_3\ \Delta t\ \sum_{j=0}^{n-1} (1+K_2\ \Delta t)^j\ ,\end{aligned} \tag{3.13}$$

$$0 \le v^n \le K_1\ \left(1+\sum_{i=\ell+1}^{I} h\ x_{i-1/2}\ G_i^n\right) \le x_\star^{1-\alpha}. \tag{3.14}$$

Proof. We proceed by induction on $n \in \{0,\dots,N\}$ and first consider the case $n=0$. The assertion (3.11) is obvious in that case, while (3.12) and (3.13) readily follow from (2.5) and (2.8). We then infer from Lemma 3.1, (3.12), (3.13), and the Young inequality that

$$\begin{aligned} v^0 &\le 2\ C_1\ G_0^0\ \left(1+\sum_{i=\ell+1}^{I} h\ x_{i-1/2}\ G_i^0\right) \\ &\le K_1\ \left(1+\sum_{i=\ell+1}^{I} h\ x_{i-1/2}\ G_i^0\right) \\ &\le K_1\ e^{K_2 T}\ \left(1+\int_0^\infty x\ G_0(x)\ dx + K_3\ T\right) = x_\star^{1-\alpha}, \end{aligned}$$

and we have checked that Proposition 3.3 is valid for $n=0$. Consider now $n \in \{0,\dots,N-1\}$ such that the assertions (3.11)–(3.14) hold true. By (2.18) and (3.14), we have

$$\nu^n(x_{I+1/2}) \le (h\ I)^\alpha\ v^n - (h\ I) \le (h\ I)^\alpha\ \left(x_\star^{1-\alpha} - (h\ I)^{1-\alpha}\right) \le 0,$$

and thus $\nu_+^n(x_{I+1/2}) = 0$. This fact, together with (3.11) and (3.12), allows us to use Lemma 3.2 to conclude that (3.4)–(3.6) hold true. Then, (3.11) for $n+1$ follows at once from (3.4), while (3.12) for $n+1$ is a consequence of (3.5) and (3.12) for n. In addition, inserting (3.12) into (3.6) yields

$$\sum_{i=\ell+1}^{I} h\ x_{i-1/2}\ G_i^{n+1} \le (1+K_2\ \Delta t)\ \sum_{i=\ell+1}^{I} h\ x_{i-1/2}\ G_i^n + K_3\ \Delta t.$$

Taking into account (3.13) for n, we deduce (3.13) for $n+1$. A straightforward consequence of (3.13) for $n+1$ is that

$$\sum_{i=\ell+1}^{I} h\ x_{i-1/2}\ G_i^{n+1} \le e^{K_2\ T}\ \left(\int_0^\infty x\ G_0(x)\ dx + K_3\ T\right). \tag{3.15}$$

Since (3.11) and (3.12) hold true for $n+1$ by the previous analysis, we are in a position to apply Lemma 3.1 and conclude that

$$v^{n+1} \le C_1\ G_0^{n+1}\ \left(1+\left(\sum_{i=\ell+1}^{I} h\ x_{i-1/2}\ G_i^{n+1}\right)^{1-\alpha}\right).$$

Using the Young inequality and (3.12) for $n+1$, we are led to

$$\begin{aligned} v^{n+1} &\le C_1\ \|G_0\|_{L^\infty}\ e^T\ \left(2+\sum_{i=\ell+1}^{I} h\ x_{i-1/2}\ G_i^{n+1}\right) \\ &\le K_1\ \left(1+\sum_{i=\ell+1}^{I} h\ x_{i-1/2}\ G_i^{n+1}\right), \end{aligned}$$

whence the first inequality in (3.14) for $n+1$. Combining this last inequality with (3.15) finally entails that $v^{n+1} \le x_\star^{1-\alpha}$, and the proof of Proposition 3.3 is complete. □

Summarizing the outcome of Proposition 3.3, we have proved that (G_i^n) satisfies the following estimate.

Corollary 3.4. *There is a positive constant C_3 depending only on α, G_0, and T such that*

$$\sup_i\ \{G_i^n\} + v^n + \sum_{i=0}^{I} h\ (1+x_{i-1/2})\ G_i^n \le C_3$$

for $n \in \{0,\dots,N\}$.

Recalling that the solution $G(t)$ to (2.1), (2.2), (2.3) is nonincreasing with respect to the variable x for each $t \ge 0$, we now show that this property is also enjoyed by (G_i^n).

Lemma 3.5. *For $n \in \{0,\dots,N\}$,*

$$G_{i+1}^n \le G_i^n, \quad 0 \le i \le I, \tag{3.16}$$

$$\sum_{i=0}^{I} \left|G_{i+1}^n - G_i^n\right| \le \|\partial_x G_0\|_{L^1}\ e^T. \tag{3.17}$$

Proof. For $n \in \{0,\dots,N\}$ and $i \in \{0,\dots,I\}$, we set $g_{i+1/2}^n = (G_{i+1}^n - G_i^n)/h$ and use (3.8) to compute $g_{i+1/2}^n$:

$$\begin{aligned} h\ g_{i+1/2}^{n+1} &= (1+\Delta t)\ G_{i+1}^n + \Delta t\ \nu_-^n(x_{i+3/2})\ g_{i+3/2}^n - \Delta t\ \nu_+^n(x_{i+1/2})\ g_{i+1/2}^n \\ &\quad - (1+\Delta t)\ G_i^n - \Delta t\ \nu_-^n(x_{i+1/2})\ g_{i+1/2}^n + \Delta t\ \nu_+^n(x_{i-1/2})\ g_{i-1/2}^n, \end{aligned}$$

$$\begin{aligned} g_{i+1/2}^{n+1} &= \left(1+\Delta t - \frac{\Delta t}{h}\ \left|\nu^n(x_{i+1/2})\right|\right)\ g_{i+1/2}^n \\ &\quad + \frac{\Delta t}{h}\ \nu_-^n(x_{i+3/2})\ g_{i+3/2}^n + \frac{\Delta t}{h}\ \nu_+^n(x_{i-1/2})\ g_{i-1/2}^n. \end{aligned} \tag{3.18}$$

Since $r \mapsto r_+$ and $r \mapsto r_-$ are subadditive, we realize that

$$\begin{aligned} &\nu_-^n(x_{i+3/2}) + \nu_+^n(x_{i-1/2}) - |\nu^n(x_i+1/2)| \\ &\quad = \nu_-^n(x_{i+3/2}) - \nu_-^n(x_{i+1/2}) + \nu_+^n(x_{i-1/2}) - \nu_+^n(x_{i+1/2}) \\ &\quad \le \big(\nu^n(x_{i+3/2}) - \nu^n(x_{i+1/2})\big)_- + \big(\nu^n(x_{i-1/2}) - \nu^n(x_{i+1/2})\big)_+ \\ &\quad \le \big((x_{i+3/2}^\alpha - x_{i+1/2}^\alpha)\ v^n - h\big)_- + \big((x_{i-1/2}^\alpha - x_{i+1/2}^\alpha)\ v^n + h\big)_+. \end{aligned}$$

Then,

$$\nu_-^n(x_{i+3/2}) + \nu_+^n(x_{i-1/2}) - |\nu^n(x_{i+1/2})| \le 2h. \tag{3.19}$$

Introducing

$$\lambda_{2,i}^n = \frac{\Delta t}{(1+3\ \Delta t)h}\ \nu_-^n(x_{i+3/2}), \qquad \lambda_{3,i}^n = \frac{\Delta t}{(1+3\ \Delta t)h}\ \nu_+^n(x_{i-1/2}),$$
$$\lambda_{1,i}^n = \frac{1}{(1+3\ \Delta t)}\ \left(1+\Delta t - \frac{\Delta t}{h}\ |\nu^n(x_{i+1/2})|\right), \qquad \lambda_{4,i}^n = 1-\sum_{j=1}^{3}\lambda_{j,i}^n,$$

we clearly have $\lambda_{2,i}^n \geq 0$, $\lambda_{3,i}^n \geq 0$, while (3.9) ensures that $\lambda_{1,i}^n \geq 0$. In addition, it follows from (3.19) that

$$1-\lambda_{4,i}^n \leq \frac{1}{(1+3\ \Delta t)}\ \left(1+\Delta t+\frac{\Delta t}{h}\ 2\ h\right) \leq 1,$$

whence $\lambda_{4,i}^n \geq 0$. Consequently, $\lambda_{j,i}^n \in [0,1]$ for $1 \leq j \leq 4$, and $g_{i+1/2}^{n+1}/(1+3\Delta t)$ is a convex combination of $g_{i+1/2}^n$, $g_{i+3/2}^n$, $g_{i-1/2}^n$, and 0.

Now, let $\Psi : \mathbb{R} \to [0,+\infty)$ be a nonnegative and convex function with $\Psi(0)=0$ and such that

$$\Psi(\lambda\ r) \leq \lambda^\gamma\ \Psi(r), \quad (r,\lambda) \in [0,+\infty) \times [1,+\infty), \tag{3.20}$$

for some $\gamma \geq 1$. The convexity of Ψ then entails that

$$\sum_{i=0}^{I} \Psi\left(\frac{g_{i+1/2}^{n+1}}{(1+3\ \Delta t)}\right) \leq \sum_{i=0}^{I}\left(\lambda_{1,i}^n\ \Psi(g_{i+1/2}^n) + \lambda_{2,i}^n\ \Psi(g_{i+3/2}^n) + \lambda_{3,i}^n\ \Psi(g_{i-1/2}^n)\right).$$

Since $\nu_+^n(x_{-1/2}) = 0$, we have $\lambda_{3,0}^n = 0$ and

$$\begin{aligned}(1+3\ \Delta t)\ \sum_{i=0}^{I}\Psi\left(\frac{g_{i+1/2}^{n+1}}{(1+3\ \Delta t)}\right) &\leq \sum_{i=0}^{I}\left(1+\Delta t-\frac{\Delta t}{h}\ |\nu^n(x_{i+1/2})|\right)\ \Psi(g_{i+1/2}^n)\\ &\quad + \sum_{i=1}^{I}\frac{\Delta t}{h}\ \nu_-^n(x_{i+1/2})\ \Psi(g_{i+1/2}^n)\\ &\quad + \sum_{i=0}^{I-1}\frac{\Delta t}{h}\ \nu_+^n(x_{i+1/2})\ \Psi(g_{i+1/2}^n)\\ &\leq (1+\Delta t)\ \sum_{i=0}^{I}\Psi(g_{i+1/2}^n).\end{aligned}$$

Owing to (3.20), we end up with

$$\begin{aligned}\sum_{i=0}^{I}\Psi(g_{i+1/2}^{n+1}) &= \sum_{i=0}^{I}\Psi\left(\frac{(1+3\ \Delta t)\ g_{i+1/2}^{n+1}}{(1+3\ \Delta t)}\right)\\ &\leq (1+3\ \Delta t)^{\gamma-1}\ (1+\Delta t)\ \sum_{i=0}^{I}\Psi(g_{i+1/2}^n).\end{aligned}$$

The discrete Gronwall lemma yields

$$\sum_{i=0}^{I}\Psi(g_{i+1/2}^n) \leq e^{(3(\gamma-1)+1)T}\ \sum_{i=0}^{I}\Psi(g_{i+1/2}^0). \tag{3.21}$$

We first take $\Psi(r) = r_+$, which obviously satisfies (3.20) with $\gamma = 1$. The assertion (3.16) then readily follows from (3.21) since $g^0_{i+1/2} \le 0$ for $i \in \{0, \dots, I\}$ by (2.5) and (2.8). Similarly, $\Psi(r) = |r|$ satisfies (3.20) with $\gamma = 1$, and (3.17) is a straightforward consequence of (3.21), taking into account that

$$\sum_{i=0}^{I} h \, \left| g^0_{i+1/2} \right| \le \|\partial_x G_0\|_{L^1},$$

and the proof of Lemma 3.5 is complete. □

Remark 3.6. Note that $\Psi(r) = r^p$ satisfies (3.20) for $p \in [1, \infty)$ with $\gamma = p$. In that case, (3.21) simply means that the discrete L^p-norm of $(g^n_{i+1/2})$ remains finite if it is initially finite.

At this point, note that Proposition 2.2 is a consequence of Proposition 3.3 and Lemma 3.5.

We end this section with the time equicontinuity of (G^n_i).

LEMMA 3.7. *For each $R \ge 1$ there is a constant $C_4(R)$ depending only on α, G_0, T, and R such that, for $n \in \{0, \dots, N-1\}$,*

$$\sum_{i=0}^{\ell_R} h \, \left| G^{n+1}_i - G^n_i \right| \le C_4(R) \, \Delta t, \tag{3.22}$$

where ℓ_R denotes the integer such that $R \in \Lambda_{\ell_R}$.

Proof. By Corollary 3.4, Lemma 3.5, and (3.8), we have for $n \in \{0, \dots, N-1\}$ and $i \in \{0, \dots, \ell_R\}$

$$\begin{aligned} \left| G^{n+1}_i - G^n_i \right| &\le \Delta t \, \left\{ C_3 + x^\alpha_{i+1/2} \, v^n \, \frac{G^n_i - G^n_{i+1}}{h} + x^\alpha_{i-1/2} \, v^n \, \frac{G^n_{i-1} - G^n_i}{h} \right\} \\ &\le \Delta t \, \left\{ C_3 + (1+R)^\alpha \, C_1 \, \left(\frac{G^n_i - G^n_{i+1}}{h} + \frac{G^n_{i-1} - G^n_i}{h} \right) \right\}. \end{aligned}$$

We sum up the above inequalities for $i \in \{0, \dots, \ell_R\}$ and use (3.17) to conclude that (3.22) holds true. □

4. Convergence. As a consequence of the analysis of the previous section, the sets $\{G^h\}_h$ and $\{v^h\}_h$ enjoy the following compactness properties.

LEMMA 4.1. *There are a subsequence of (G^h, v^h) (not relabeled) and a pair of nonnegative functions $G \in \mathcal{C}([0,T]; L^1(\mathbb{R}_+))$ and $v \in L^\infty(0,T)$ such that*

$$G^h \longrightarrow G \quad in \quad L^\infty(0,T; L^1(\mathbb{R}_+)), \tag{4.1}$$

$$v^h \overset{*}{\rightharpoonup} v \quad in \quad L^\infty(0,T), \tag{4.2}$$

and $G \in L^\infty(0,T; L^1(\mathbb{R}_+, xdx))$ satisfies (2.6).

Proof. We introduce the auxiliary function

$$\mathcal{G}^h(t,x) = \sum_{i=0}^{I^h} \left\{ G^{n,h}_i + \frac{(t-t^n)}{\Delta t} \, \left(G^{n+1,h}_i - G^{n,h}_i \right) \right\} \mathbf{1}_{\Lambda^h_i}(x), \quad (t,x) \in [t^n, t^{n+1}] \times \mathbb{R}_+,$$

for $n \in \{0, \dots, N-1\}$. Clearly, $\mathcal{G}^h \in \mathcal{C}([0,T]; L^1(\mathbb{R}_+))$, and it readily follows from Lemma 3.7 that

$$\sup_{[0,T]} \left\| G^h(t) - \mathcal{G}^h(t) \right\|_{L^1(0,R)} \le C_4(R) \, \Delta t \tag{4.3}$$

for any $R \ge 1$.

We next fix $R \geq 1$. On the one hand, it follows from Corollary 3.4 and (3.17) that $(\mathcal{G}^h)$ is bounded in $L^\infty(0,T;BV(0,R))$. On the other hand, an easy computation shows that (3.22) implies

$$\left\| \mathcal{G}^h(t) - \mathcal{G}^h(s) \right\|_{L^1(0,R)} \leq C_4(R)\ |t-s|$$

for $(s,t) \in [0,T] \times [0,T]$. Since $BV(0,R)$ is compactly embedded in $L^1(0,R)$, a classical compactness result [21, Theorem 5] entails that

$$(\mathcal{G}^h) \quad \text{is relatively compact in} \quad \mathcal{C}([0,T];L^1(0,R)), \tag{4.4}$$

and (4.4) is valid for every $R \geq 1$. Also, by Corollary 3.4, there is a constant C depending only on α, G_0, and T such that

$$\int_0^\infty \left(G^h(t,x) + \mathcal{G}^h(t,x) \right)\ x\ dx \leq C, \quad t \in [0,T]. \tag{4.5}$$

Thanks to (4.5), we may improve the compactness (4.4) of $(\mathcal{G}^h)$ and conclude that $(\mathcal{G}^h)$ is relatively compact in $\mathcal{C}([0,T];L^1(\mathbb{R}_+))$. Consequently, there are a subsequence of $(\mathcal{G}^h)$ (not relabeled) and a function G in $\mathcal{C}([0,T];L^1(\mathbb{R}_+))$ such that

$$\mathcal{G}^h \longrightarrow G \quad \text{in} \quad \mathcal{C}([0,T];L^1(\mathbb{R}_+)).$$

Recalling (4.3) and (4.5), we readily conclude that (4.1) holds true and that G is a nonnegative function in $L^\infty(0,T;L^1(\mathbb{R}_+,xdx))$. Moreover, the convergence (4.1) of (G^h), (2.9), and (3.11) imply that G satisfies (2.6). The convergence (4.2) and the nonnegativity of v are then straightforward consequences of Corollary 3.4 and the nonnegativity of v^h. □

We next show that (3.11), (3.21), and the integrability of $\partial_x G_0$ guarantee that G enjoys the regularity properties claimed in Theorem 2.3.

LEMMA 4.2. *We have* $\partial_x G \in L^\infty(0,T;L^1(\mathbb{R}_+,(1+x)dx))$*, and* $t \mapsto G(t,x)$ *is nonincreasing for each* $t \geq 0$*.*

Proof. For $\varphi \in L^\infty(\mathbb{R}_+)$, we define the discrete gradient $D_h\varphi$ by

$$D_h\varphi(x) := \frac{\varphi(x+h)-\varphi(x)}{h}, \quad x \in \mathbb{R}_+.$$

We first observe that

$$D_h G^h(t,x) = \sum_{i=0}^{I^h} \left(\frac{G_{i+1}^{n,h} - G_i^{n,h}}{h} \right) \mathbf{1}_{\Lambda_i^h}(x), \quad x \in \mathbb{R}_+, \tag{4.6}$$

for $t \in [t^n, t^{n+1})$ and $n \in \{1,\ldots,N\}$ and recall that Lemma 3.5 and in particular (3.21) provide some information on $((G_{i+1}^{n,h} - G_i^{n,h})/h)$. It turns out that the available information allows us to show the weak compactness of $(D_h G^h)$ in $L^1((0,T)\times\mathbb{R}_+)$. Indeed, we first notice that

$$\sup_{t\in[0,T]} \int_0^\infty (1+x)\ \left| D_h G^h(t,x) \right|\ dx \leq C \tag{4.7}$$

for some constant C depending only on α, G_0, and T. Indeed, (4.7) readily follows from (3.11), (3.16), and (3.17), thanks to the identity

$$\sum_{i=0}^{I^h} x_{i+1/2}\ \left| G_{i+1}^{n,h} - G_i^{n,h} \right| = \sum_{i=0}^{I^h} h\ G_i^{n,h}.$$

We next recall that, since $\partial_x G_0 \in L^1(\mathbb{R}_+)$, a refined version of the de la Vallée–Poussin theorem [11, Proposition I.1.1] ensures that there is a nonnegative and convex function $\Psi_0 \in \mathcal{C}^1([0,+\infty)) \cap W^{2,\infty}_{\text{loc}}(\mathbb{R}_+)$ satisfying

$$\lim_{r\to+\infty} \frac{\Psi_0(r)}{r} = +\infty, \tag{4.8}$$

with $\Psi_0(0) = 0$, $\Psi_0'(0) \geq 0$, Ψ_0' a concave function on $[0,+\infty)$, and such that $\Psi_0(|\partial_x G_0|) \in L^1(\mathbb{R}_+)$. (See also, e.g., [4, p. 38] for the construction of such a function Ψ_0 without the requirement that Ψ_0' be concave.) Thanks to the concavity of Ψ_0' and the nonnegativity of $\Psi_0'(0)$, we have $\Psi_0'(\lambda r) \leq \lambda\ \Psi_0'(r)$ for $r \geq 0$ and $\lambda \geq 1$. Integrating this inequality yields

$$\Psi_0(\lambda r) \leq \lambda^2\ \Psi_0(r), \quad (r,\lambda) \in [0,+\infty) \times [1,+\infty). \tag{4.9}$$

Since Ψ_0 is nondecreasing, the function Ψ defined by $\Psi(r) = \Psi_0(|r|)$ is a nonnegative and convex function with $\Psi(0) = 0$, which satisfies (3.20) with $\gamma = 2$ by (4.9). Consequently, we infer from (3.21) and (4.6) that

$$\sup_{t\in[0,T]} \int_0^\infty \Psi_0\left(\left|D_h G^h(t,x)\right|\right)\ dx \leq e^{4T} \int_0^\infty \Psi_0\left(\left|D_h G^h(0,x)\right|\right)\ dx. \tag{4.10}$$

However,

$$\int_0^\infty \Psi_0\left(\left|D_h G^h(0,x)\right|\right)\ dx = h\ \sum_{i=0}^{I^h} \Psi_0\left(\left|\frac{G^{0,h}_{i+1} - G^{0,h}_i}{h}\right|\right), \tag{4.11}$$

and it follows from (2.5), (2.8), (4.9), the convexity of Ψ_0, and the Jensen inequality that

$$\begin{aligned}
&\Psi_0\left(\left|\frac{G^{0,h}_{i+1} - G^{0,h}_i}{h}\right|\right) \\
&\quad \leq \Psi_0\left(\frac{1}{h}\ \int_{\Lambda_i^h} |\partial_x G_0|\ dx + \frac{1}{h}\ \int_{\Lambda_{i+1}^h} |\partial_x G_0|\ dx\right) \\
&\quad \leq 2\ \left\{\Psi_0\left(\frac{1}{h}\ \int_{\Lambda_i^h} |\partial_x G_0|\ dx\right) + \Psi_0\left(\frac{1}{h}\ \int_{\Lambda_{i+1}^h} |\partial_x G_0|\ dx\right)\right\} \\
&\quad \leq \frac{2}{h}\ \left\{\int_{\Lambda_i^h} \Psi_0\left(|\partial_x G_0|\right)\ dx + \int_{\Lambda_{i+1}^h} \Psi_0\left(|\partial_x G_0|\right)\ dx\right\},
\end{aligned}$$

whence

$$h\ \sum_{i=0}^{I^h} \Psi_0\left(\left|\frac{G^{0,h}_{i+1} - G^{0,h}_i}{h}\right|\right) \leq 4\ \|\partial_x G_0\|_{L^1}.$$

Inserting this estimate into (4.11), we deduce from (4.10) that

$$\sup_{t\in[0,T]} \int_0^\infty \Psi_0\left(\left|D_h G^h(t,x)\right|\right)\ dx \leq C \tag{4.12}$$

for some constant C depending on α, G_0, and T. Owing to (4.7), (4.8), and (4.12), we are in a position to apply the Dunford–Pettis theorem and conclude that (D_hG^h) is relatively weakly sequentially compact in $L^1((0,T)\times\mathbb{R}_+)$. We may thus extract a subsequence of (D_hG^h) (not relabeled) such that

$$D_hG_h \rightharpoonup g \quad \text{in} \quad L^1((0,T)\times\mathbb{R}_+) \tag{4.13}$$

for some function $g \in L^1((0,T)\times\mathbb{R}_+)$. Now, it follows from (4.7) and (4.13) that g belongs to $L^\infty(0,T;L^1(\mathbb{R}_+,(1+x)dx))$, while a classical computation entails that $g = \partial_x G$ in the sense of distributions. In addition, $D_hG^h \le 0$ by (3.16), whence $\partial_x G \le 0$, and the proof of Lemma 4.2 is complete. □

We are now in a position to complete the proof of Theorem 2.3.

Proof of Theorem 2.3. We first observe that Lemmas 4.1 and 4.2 ensure that (G,v) enjoy the regularity properties (2.25).

We next consider $\varphi \in \mathcal{C}_0^\infty([0,T)\times[0,+\infty))$ with supp $\varphi \subset [0,\tau)\times[0,R)$ for some $\tau \in [0,T)$ and $R \in \mathbb{R}_+$ and set

$$\varphi_i^{n,h} = \frac{1}{h\ \Delta t}\int_{t^n}^{t^{n+1}}\int_{\Lambda_i^h}\varphi(t,x)\ dxdt$$

for $i \ge 0$ and $n \in \{0,\dots,N-1\}$. We also assume that h and Δt are sufficiently small so that $\tau \le t^{N-2}$ and $R \le x_{I^h-1/2}$, and we denote by ℓ_R^h the integer such that $R \in \Lambda^h_{\ell_R^h}$. In the following we denote by C_φ any nonnegative constant depending only on α, G_0, T, and φ.

We multiply (3.8) by $\varphi_i^{n,h}$ and sum up the resulting identities to obtain

$$Y_1^h = Y_2^h,$$

where

$$Y_1^h := \sum_{n=0}^{N-1}\sum_{i=0}^{I^h} h\ (G_i^{n+1,h} - G_i^{n,h})\ \varphi_i^{n,h}$$

and

$$\begin{aligned}Y_2^h := h\ \Delta t\ &\sum_{n=0}^{N-1}\sum_{i=0}^{I^h} G_i^{n,h}\ \varphi_i^{n,h} + \Delta t\ \sum_{n=0}^{N-1}\sum_{i=0}^{I^h}\nu_-^n(x_{i+1/2})\ \left(G_{i+1}^{n,h} - G_i^{n,h}\right)\ \varphi_i^{n,h}\\ &+ \Delta t\ \sum_{n=0}^{N-1}\sum_{i=0}^{I^h}\nu_+^n(x_{i-1/2})\ \left(G_{i-1}^{n,h} - G_i^{n,h}\right)\ \varphi_i^{n,h}.\end{aligned}$$

We next introduce

$$Z_1^h := -\int_0^T\int_0^\infty G^h(t,x)\ \partial_t\varphi(t,x)\ dtdx - \int_0^\infty G_0(x)\ \varphi(0,x)\ dx,$$

$$\begin{aligned}Z_2^h := &\int_0^T\int_0^\infty G^h(t,x)\ \varphi(t,x)\ dtdx - \int_0^T G^h(t,0)\ \varphi(t,0)\ dt\\ &+ \int_0^T\int_0^\infty G^h(t,x)\ \partial_x\left(\mathcal{W}^h\ \varphi\right)(t,x)\ dxdt,\end{aligned}$$

where

$$\mathcal{W}^h(t,x) = x^\alpha \ v^h(t) - 1 - x, \quad (t,x) \in [0,T] \times \mathbb{R}_+.$$

On the one hand, since φ is compactly supported, it follows at once from (4.1) and (4.2) that

$$\lim_{h,\Delta t \to 0} Z_1^h = -\int_0^T \int_0^\infty G(t,x) \ \partial_t \varphi(t,x) \ dtdx - \int_0^\infty G_0(x) \ \varphi(0,x) \ dx, \tag{4.14}$$

and

$$\begin{aligned} &\lim_{h,\Delta t \to 0} \left(Z_2^h + \int_0^T G^h(t,0) \ \varphi(t,0) \ dt \right) \\ &\qquad = \int_0^T \int_0^\infty G(t,x) \ \left(\varphi + \partial_x \left(\mathcal{W} \ \varphi\right)\right)(t,x) \ dxdt. \end{aligned} \tag{4.15}$$

On the other hand, we have

$$\begin{aligned} &\int_0^T \left(G^h(t,0) - G(t,0)\right) \ \varphi(t,0) \ dt \\ &\quad = \int_0^T \sum_{i=0}^{I^h} \left(G^h(t,x_{i-1/2}) - G^h(t,x_{i+1/2})\right) \ \varphi(t,0) \ dt + \int_0^\infty \partial_x G(t,x) \ \varphi(t,0) \ dxdt \\ &\quad = \int_0^T \int_0^\infty \left(\partial_x G(t,x) - D_h G^h(t,x)\right) \ \varphi(t,0) \ dxdt. \end{aligned}$$

We then infer from (4.13) that the right-hand side of the above identity converges to zero as $h \to 0$. Inserting this result into (4.15), we end up with

$$\begin{aligned} \lim_{h,\Delta t \to 0} Z_2^h = &-\int_0^T G(t,0) \ \varphi(t,0) \ dt \\ &+ \int_0^T \int_0^\infty G(t,x) \ \left(\varphi + \partial_x \left(\mathcal{W} \ \varphi\right)\right)(t,x) \ dxdt. \end{aligned} \tag{4.16}$$

Having identified the limits of (Z_1^h) and (Z_2^h) as $h \to 0$, we next aim at comparing the terms Y_k^h and Z_k^h, $k = 1, 2$, in order to show that $Z_1^h - Z_2^h$ converges to zero as $(h, \Delta t) \to 0$.

We first compute $(Z_1^h - Y_1^h)$. Since G^h is constant on $[t^n, t^{n+1}) \times \Lambda_i^h$ for $i \geq 0$ and $n \in \{0, \ldots, N-1\}$, we have

$$\begin{aligned} Z_1^h &= -\sum_{n=0}^{N-1} \int_0^\infty G^h(t^n, x) \ \left(\varphi(t^{n+1}, x) - \varphi(t^n, x)\right) \ dx - \int_0^\infty G_0(x) \ \varphi(0,x) \ dx \\ &= \sum_{n=0}^{N-1} \int_0^\infty \left(G^h(t^{n+1}, x) - G^h(t^n, x)\right) \ \varphi(t^{n+1}, x) \ dx \\ &\quad + \int_0^\infty \left(G^h(0,x) - G_0(x)\right) \ \varphi(0,x) \ dx, \end{aligned}$$

from which we deduce that

$$\begin{aligned}\left|Z_1^h - Y_1^h\right| &\le \sum_{n=0}^{N-1} \sum_{i=0}^{\ell_R^h} \left|G_i^{n+1,h} - G_i^{n,h}\right| \int_{t^n}^{t^{n+1}} \int_{\Lambda_i^h} |\partial_t \varphi| \ dxdt \\ &\quad + \|G^h(0,.) - G_0(.)\|_{L^1} \ \|\varphi\|_{L^\infty} \\ &\le T \ \|\partial_t\varphi\|_{L^\infty} \ \sup_{0\le n\le N-1} \sum_{i=0}^{\ell_R^h} h \ \left|G_i^{n+1,h} - G_i^{n,h}\right| \\ &\quad + \|G^h(0,.) - G_0(.)\|_{L^1} \ \|\varphi\|_{L^\infty}.\end{aligned}$$

We now use (2.9) and Lemma 3.7 to conclude that

$$\left|Z_1^h - Y_1^h\right| \le C_\varphi \ \Delta t. \tag{4.17}$$

We next turn to $(Z_2^h - Y_2^h)$. Since $\mathcal{W}^h(t^n, x_{i+1/2}) = \nu^n(x_{i+1/2})$, G^h is constant on $[t^n, t^{n+1}) \times \Lambda_i^h$ for $i \ge 0$, and $n \in \{0, \ldots, N-1\}$, we have

$$\begin{aligned}Z_2^h = &\sum_{n=0}^{N-1} \sum_{i=0}^{I^h} G_i^{n,h} \int_{t^n}^{t^{n+1}} \int_{\Lambda_i^h} \varphi(t,x) \ dxdt - \sum_{n=0}^{N-1} G_0^n \int_{t^n}^{t^{n+1}} \varphi(t,0) \ dt \\ &+ \sum_{n=0}^{N-1} \sum_{i=0}^{I^h} G_i^{n,h} \int_{t^n}^{t^{n+1}} \left(\nu^n(x_{i+1/2}) \ \varphi(t, x_{i+1/2}) - \nu^n(x_{i-1/2}) \ \varphi(t, x_{i-1/2})\right) \ dt.\end{aligned}$$

Since $\nu^n(x) = \nu_+^n(x) - \nu_-^n(x)$ and $\nu^n(x_{-1/2}) = -1$, a discrete integration by parts yields

$$\begin{aligned}Z_2^h = &\sum_{n=0}^{N-1} \sum_{i=0}^{I^h} G_i^{n,h} \int_{t^n}^{t^{n+1}} \int_{\Lambda_i^h} \varphi(t,x) \ dxdt \\ &+ \sum_{n=0}^{N-1} \sum_{i=0}^{I^h} \nu_+^n(x_{i-1/2}) \ \left(G_{i-1}^{n,h} - G_i^{n,h}\right) \int_{t^n}^{t^{n+1}} \varphi(t, x_{i-1/2}) \ dt \\ &+ \sum_{n=0}^{N-1} \sum_{i=0}^{I^h} \nu_-^n(x_{i+1/2}) \ \left(G_{i+1}^{n,h} - G_i^{n,h}\right) \int_{t^n}^{t^{n+1}} \varphi(t, x_{i+1/2}) \ dt.\end{aligned}$$

It is then easy to compute $(Z_2^h - Y_2^h)$ and deduce from Corollary 3.4 that

$$\begin{aligned}\left|Z_2^h - Y_2^h\right| &\le h \ \Delta t \ \sum_{n=0}^{N-1} \sum_{i=0}^{\ell_R^h} \nu_+^n(x_{i-1/2}) \ \left|G_{i-1}^{n,h} - G_i^{n,h}\right| \ \|\partial_x\varphi\|_{L^\infty} \\ &\quad + h \ \Delta t \ \sum_{n=0}^{N-1} \sum_{i=0}^{\ell_R^h} \nu_-^n(x_{i+1/2}) \ \left|G_{i+1}^{n,h} - G_i^{n,h}\right| \ \|\partial_x\varphi\|_{L^\infty} \\ &\le C_\varphi \ h \ \sup_{0\le n\le N-1} \sum_{i=0}^{I^h} \left|G_{i+1}^{n,h} - G_i^{n,h}\right|,\end{aligned}$$

whence

$$\left|Z_2^h - Y_2^h\right| \le C_\varphi \ h \tag{4.18}$$

by (3.17). Since $Y_1^h = Y_2^h$, we infer from (4.17) and (4.18) that $(Z_1^h - Z_2^h)$ converges to zero as $(h, \Delta t) \to 0$. This fact, together with (4.14) and (4.16), immediately ensures that G satisfies

$$\int_0^T \int_0^\infty G(t,x)\ \partial_t\varphi(t,x)\ dtdx + \int_0^\infty G_0(x)\ \varphi(0,x)\ dx$$
$$= -\int_0^T G(t,0)\ \varphi(t,0)\ dt + \int_0^T \int_0^\infty G(t,x)\ \left(\varphi + \partial_x\left(\mathcal{W}\ \varphi\right)\right)(t,x)\ dxdt.$$

Owing to the regularity of G, standard approximation arguments allow us to conclude from the previous identity that G actually satisfies (2.26).

To conclude the proof, it remains to show that v is given by (2.2). The easiest way to see it is to take $\varphi \equiv 1$ in (2.26), from which (2.2) readily follows, since we already know that G satisfies (2.6). We may, however, prove it directly by passing to the limit in (2.14). Indeed, consider $\varphi \in \mathcal{C}(0,T)$. Arguing as in the proof of (4.16), we realize that

$$\lim_{(h,\Delta t)\to 0} \int_0^T \left(G^h(t,0) - G(t,0)\right)\ \varphi(t)\ dt = 0. \tag{4.19}$$

We next claim that

$$\lim_{(h,\Delta t)\to 0} \sup_{t\in[0,T]} \int_0^\infty x^{\alpha-1}\ \left|G^h(t,x) - G(t,x)\right|\ dx = 0. \tag{4.20}$$

Indeed, it follows from Corollary 3.4 and (2.25) that, for $\delta \in (0,1)$ and $t \in [0,T]$,

$$\begin{aligned}\int_0^\infty x^{\alpha-1}\ \left|G^h(t,x) - G(t,x)\right|\ dx &\le \frac{\delta^\alpha}{\alpha}\ \left(\|G^h(t)\|_{L^\infty} + \|G(t)\|_{L^\infty}\right)\\ &\quad + \delta^{\alpha-1}\ \int_\delta^\infty \left|G^h(t,x) - G(t,x)\right|\ dx\\ &\le C\ \delta^\alpha + \delta^{\alpha-1}\ \sup_{s\in[0,T]} \left\|G^h(s) - G(s)\right\|_{L^1}\end{aligned}$$

for some constant C depending only on α, G_0, T, and G. Thanks to (4.1), we may pass to the limit as $(h,\Delta t) \to 0$ and obtain

$$\limsup_{(h,\Delta t)\to 0} \sup_{t\in[0,T]} \int_0^\infty x^{\alpha-1}\ \left|G^h(t,x) - G(t,x)\right|\ dx \le C\ \delta^\alpha.$$

As $\delta \in (0,1)$ is arbitrary, we let $\delta \to 0$ to obtain the claim (4.20).

Now, owing to (4.2) and (4.20), it is straightforward to check that

$$\lim_{(h,\Delta t)\to 0} \int_0^T \int_0^\infty x^{\alpha-1}\ \left(v^h(t)\ G^h(t,x) - v(t)\ G(t,x)\right)\ \varphi(t)\ dxdt = 0. \tag{4.21}$$

Thanks to (4.19) and (4.21), we may pass to the limit in (2.14) and conclude that v is given by (2.2). □

5. Numerical simulations. In this section, we perform numerical experiments with $\alpha = 1/3$, which corresponds to the original model of Lifshitz and Slyozov [13]. Our aim is twofold: first, to study the numerical accuracy of the scheme analyzed in the previous sections and second, to see its behavior for large times.

We first check the order of the scheme with the following explicit stationary solution to (2.1)–(2.2):

$$G_{LS}(x) := \frac{6}{\left(1-(2x)^{1/3}\right)^{5/3}\ \left(1+(x/4)^{1/3}\right)^{4/3}}\ \exp\left(-\frac{(2x)^{1/3}}{1-(2x)^{1/3}}\right) \tag{5.1}$$

Table 1
Relative errors in the L^1- and L^∞-norms with respect to the number of grid points I^h.

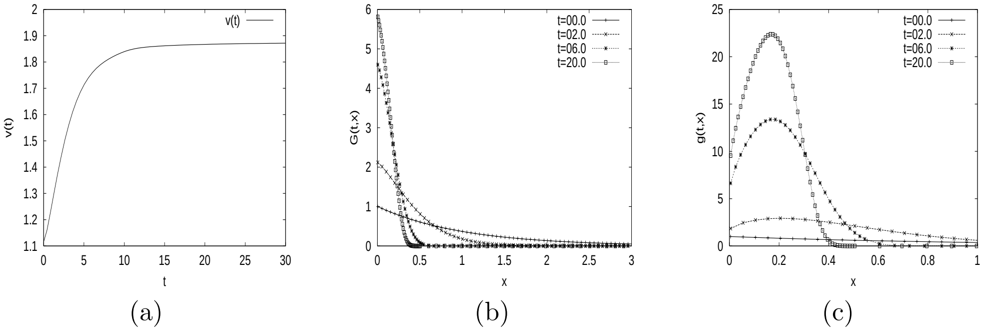
Number of points	Number of iterations	L^1 Error	L^∞ Error
200	221	$1.4\ 10^{-3}$	$1.5\ 10^{-3}$
400	443	$7.5\ 10^{-4}$	$7.5\ 10^{-4}$
800	887	$4.0\ 10^{-4}$	$3.8\ 10^{-4}$
1600	1776	$2.0\ 10^{-4}$	$2.0\ 10^{-4}$

for $x \in [0,1/2]$ and $G_{LS}(x) = 0$ if $x \geq 1/2$. Since G_{LS} is compactly supported in $[0,1/2]$, we take $hI^h = 1$ and $T = 1$. We compute the relative errors at $T = 1$ in the L^1- and L^∞-norms for different values of I^h, which are reported in Table 1. As expected, the scheme is first-order; that is, the error is proportional to h.

We next turn to the large time behavior and first recall that, for (2.1)–(2.3), it is much more complex that originally conjectured by Lifshitz and Slyozov in [13]. As already mentioned, formal asymptotic expansions performed in [13] indicate that the pair (G,v) converges towards a stationary solution (G_∞, v_∞) to (2.1)–(2.3), and it was further conjectured in [13] (and also in [25], but for a different choice of functions k and q) that the asymptotic profile (G_∞, v_∞) does not depend on the shape of the initial data G_0 but only on $\|G_0\|_{L^1}$. More precisely, the conjecture in [13] states that $G_\infty = a\, G_{LS}$ (defined by (5.1) above), with $a = \|G_0\|_{L^1}\ \|G_{LS}\|_{L^1}^{-1}$, while $v_\infty = V_{LS} := 3/2^{2/3}$. It was, however, noticed in [13] that (2.1)–(2.2) actually has a continuum $(G_V)_{V\geq V_{LS}}$ of stationary solutions (with $G_{V_{LS}} = G_{LS}$) satisfying $\|G_V\|_{L^1} = \|G_{LS}\|_{L^1}$, but it was argued that G_V is "unstable" for $V > V_{LS}$. This conjecture turns out to be false, as noticed on the ground of physical arguments in [6, 15] and confirmed by numerical simulations performed in [6]. More precisely, if the initial datum is compactly supported, the asymptotic profile (G_∞, v_∞) is determined by the way in which the initial datum vanishes at the edge of its support. Mathematical proofs of these facts have subsequently been supplied in [1] for $\alpha = 1$ by means of the Laplace transform. Though a convergence proof is still lacking in the general case $\alpha \in (0,1)$, necessary conditions for convergence are provided in [16] when $\alpha = 1/3$. In addition, it is established in [16] that, if one can prove that $v(t)$ converges to some $V > V_{LS}$ as $t \to +\infty$, then $G(t)$ converges towards G_V as $t \to +\infty$. Analogous results for the variant (1.1), (1.9) have subsequently been obtained in [18], still in the case $\alpha = 1/3$. It is also shown in [1, 16, 18] that there are initial data for which convergence towards a stationary solution does not hold at all. In addition, several numerical simulations have been performed in [2] with an accurate numerical method. The results in [2] provide further numerical evidence that the solutions to (2.1)–(2.3) with compactly supported initial data do not converge to the asymptotic profile (G_{LS}, V_{LS}) conjectured in [13] but to the one determined by the way the initial datum vanishes at the edge of its support; that is, (G_V, V) for some $V > V_{LS}$. Still, it is expected that the Lifshitz–Slyozov conjecture is valid for noncompactly supported initial data (with a "smooth" behavior for large x), and the aim of our first computations is to provide some numerical evidence of this fact. We thus choose

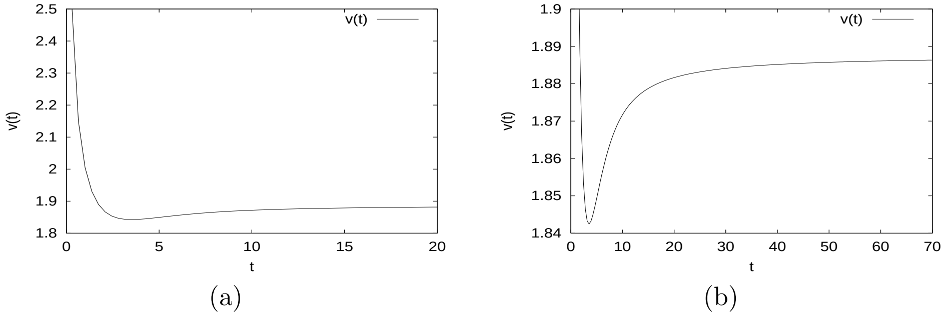
$$G_0(x) = \|G_{LS}\|_{L^1}\ \exp(-x), \quad x \in \mathbb{R}_+, \tag{5.2}$$

and report in Figure 1 the time evolution of v, G, and $g = -\partial_x G$ obtained by the scheme (2.12)–(2.17). For this simulation, we take the number of grid points $I^h = 1000$ with $h\, I^h = 10$, and the final time is $T = 30$. Noticing that $v(0) < V_{LS}$, we see that the function v first increases rapidly towards V_{LS} and then stabilizes to this value



(a) (b) (c)

FIG. 1. *Evolution of* (a) $v(t)$, (b) $G(t,x)$, *and* (c) $g(t,x)$ *corresponding to the initial datum* (5.2).



(a) (b)

FIG. 2. (a) *Evolution of* $v(t)$, (b) *zoom on the small variations of* $v(t)$ *corresponding to the initial datum* (5.3).

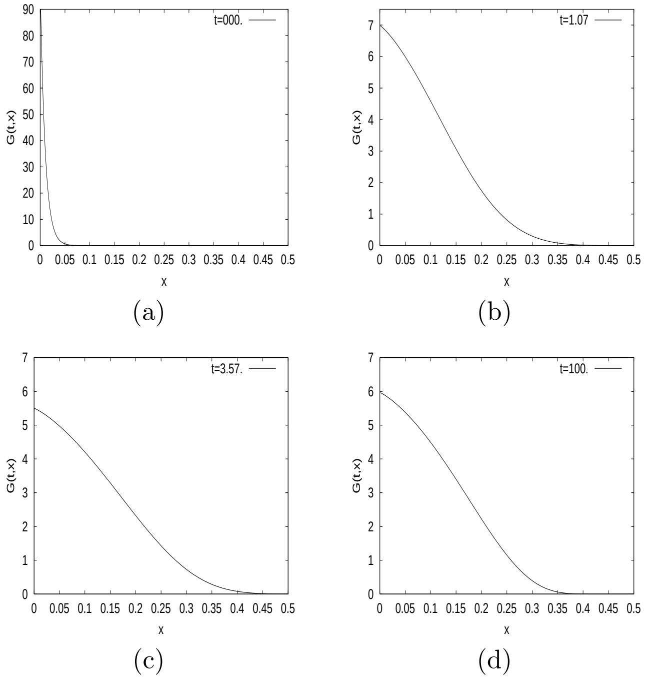
as conjectured. As for G, its support decreases with time towards $[0,1/2]$, and $G(t)$ converges to the stationary solution G_{LS} with a good accuracy.

We next investigate what happens to a similar initial datum but with $v(0) > V_{LS}$. More precisely, we take

$$G_0(x) = 100\,\|G_{LS}\|_{L^1}\,\exp(-100\,x), \quad x \in \mathbb{R}_+, \tag{5.3}$$

and observe that G_0 has the same L^1-norm as (5.2) but decreases faster for large x. In this case, we take $I^h = 1000$ with $h\,I^h = 1$, and the final time is $T = 100$. We observe that the behavior of v differs from that in the previous simulation. Indeed, $v(0)$ being greater than V_{LS}, v first decreases with time but to a smaller value than V_{LS}, as shown in Figure 2. It then increases again towards V_{LS} and finally stabilizes to V_{LS}. The evolution of G in that case is presented in Figure 3, which shows the convergence of G to G_{LS}.

Finally, following the previous discussion on the large time behavior for compactly supported initial data, one may wonder what the behavior of our scheme in that case might be. We have performed numerical simulations with $G_0(x) = (1-x)_+$ and observe that, in this case also, the numerical solution converges to (G_{LS}, V_{LS}), which is definitely not the behavior predicted by the theory [16]. It is, however, not surprising, as the numerical scheme induces some small diffusive effects, and diffusion is known to significantly modify large time behavior. More precisely, it is conjectured that solutions to a diffusive perturbation of (2.1)–(2.3) with a time-dependent diffusion coefficient vanishing for large times should converge to (G_{LS}, V_{LS}) (see [14, 20, 23]



(a) (b) (c) (d)

Fig. 3. *Evolution of $G(t,x)$ at time $t = 0, 1.07, 3.57, 100$ corresponding to the initial datum* (5.3).

and the references therein). The initial datum $G_0(x) = (1-x)_+$ being quite far from its expected limit, the diffusive effects of the scheme become not negligible after some time and thus induce a difference between the behavior of the numerical and the exact solutions. In an attempt to quantify the time of appearance of diffusive effects, it is interesting to look at the behavior of the scheme if the initial datum is one of the stationary solutions (G_V, V) for some $V > V_{LS}$. Given $V > V_{LS}$, this solution G_V is given by

$$G_V(x) := \frac{6\,\left(1-(x/x_0)^{1/3}\right)^{-\lambda_0}}{\left(1-(x/x_-)^{1/3}\right)^{\lambda_-}\,\left(1-(x/x_+)^{1/3}\right)^{\lambda_+}}, \tag{5.4}$$

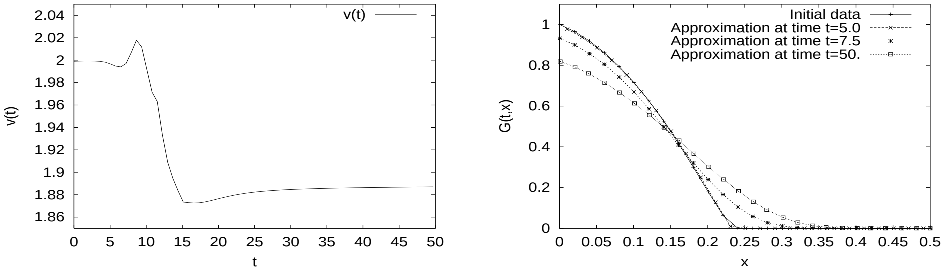
where λ_-, λ_0, λ_+ satisfy

$$\lambda_* = \frac{3\,x_*^{2/3}}{3\,x_*^{2/3} - V} \qquad \text{for} \qquad * \in \{-, 0, +\}$$

and $x_-^{1/3}$, $x_0^{1/3}$, $x_+^{1/3}$ are solutions of the following equation:

$$X^3 - V\,X + 1 = 0 \quad \text{with} \quad x_- \leq 0 \leq x_0 \leq x_+.$$

We choose $V = 2$ and $G_0 = G_2$ with $I^h = 1000$, $h\,I^h = 1$, and $T = 50$. The numerical simulations are reported in Figure 4, and we observe that, for $t \in [0,5]$, the computed solution remains close to G_2. After that time, diffusive effects come into play and the numerical solution evolves towards (G_{LS}, V_{LS}). In order to capture the expected behavior for compactly supported initial data for larger times, one thus needs a less



(a) (b)

FIG. 4. *Evolution of* (a) $v(t)$ *and* (b) $G(t,x)$ *corresponding to the initial datum* G_2*.*

diffusive numerical scheme such as the one used in [2], to which we refer for a more complete discussion on that issue.

Appendix. An L^∞-estimate for u. We consider a nonnegative function $f_0 \in L^1(\mathbb{R}_+, (1+x^2)dx)$, $f_0 \not\equiv 0$, and denote by f a weak solution to (1.1)–(1.2) (in the sense of [10, Theorem 2]) with initial datum f_0, the functions k and q being still given by (1.4) with $\alpha \in (0,1)$, so that $\mathcal{V}(t,x) = x^\alpha\ u(t) - 1$. We have the following result.

PROPOSITION A.1. *There is a constant C depending only on α and f_0 such that, for each $t \geq 0$,*

$$u(t) + \int_0^\infty x^2\ f(t,x)\ dx \leq C\ \exp(Ct). \tag{A.1}$$

Proof. For $\lambda \in [0,2]$ and $t \geq 0$, we set

$$M_\lambda(t) := \int_0^\infty x^\lambda\ f(t,x)\ dx.$$

Consider $t \in \mathbb{R}_+$. Since $\mathcal{V}(t,0) = -1$, we infer from (1.1) that $M_0(t) \leq M_0(0)$ and

$$\frac{dM_2}{dt}(t) \leq 2\ \int_0^\infty x\ \mathcal{V}(t,x)\ f(t,x)\ dx \leq 2\ u(t)\ M_{1+\alpha}(t). \tag{A.2}$$

We next infer from the Hölder inequality and (1.2) that

$$M_{1+\alpha}(t) \leq M_1(t)^{1-\alpha}\ M_2(t)^\alpha \leq C\ M_2(t)^\alpha,$$

$$0 < M_1(0) = M_1(t) \leq M_\alpha(t)^{1/(2-\alpha)}\ M_2(t)^{(1-\alpha)/(2-\alpha)}.$$

As a consequence of the first inequality and (A.2), we deduce that

$$\frac{dM_2}{dt}(t) \leq C\ u(t)\ M_2(t)^\alpha, \tag{A.3}$$

while the second inequality and (1.3) yield

$$u(t) = \frac{M_0(t)}{M_\alpha(t)} \leq \frac{M_0(0)}{M_1(0)^{2-\alpha}}\ M_2(t)^{1-\alpha}. \tag{A.4}$$

Combining (A.3) and (A.4), we end up with

$$\frac{dM_2}{dt}(t) \leq C\ M_2(t)$$

and use the Gronwall lemma to conclude that $M_2(t) \leq C \exp{(Ct)}$ for $t \geq 0$. A similar bound for u then follows by (A.4). □

Remark A.2. Observe that (A.3) and (A.4) are the continuous analogues of (3.10) and (3.2), respectively.

Acknowledgments. We thank Professor J. A. Carrillo for enlightening discussions on the computational issues of the LSW model. We are also grateful to the referees for their careful reading of the manuscript and their helpful suggestions.

REFERENCES

[1] J. Carr and O. Penrose, *Asymptotic behavior of solutions to a simplified Lifshitz–Slyozov equation*, Phys. D, 124 (1998), pp. 166–176.

[2] J. A. Carrillo and T. Goudon, *Numerical investigation of the behavior of solutions of the Lifshitz–Slyozov equations*, J. Sci. Comput., to appear.

[3] J. F. Collet and T. Goudon, *On solutions of the Lifshitz–Slyozov model*, Nonlinearity, 13 (2000), pp. 1239–1262.

[4] C. Dellacherie and P. A. Meyer, *Probabilités et Potentiel,* Vols. I–IV, Hermann, Paris, 1975.

[5] B. Després and F. Lagoutière, *Contact discontinuity capturing schemes for linear advection and compressible gas dynamics*, J. Sci. Comput., 16 (2001), pp. 479–524.

[6] B. Giron, B. Meerson, and P. V. Sasorov, *Weak selection and stability of localized distributions in Ostwald ripening*, Phys. Rev. E, 58 (1998), pp. 4213–4216.

[7] R. Eymard, T. Gallouët, and R. Herbin, *Finite volume methods*, in Handb. Numer. Anal. 7, North-Holland, Amsterdam, 2000, pp. 713–1020.

[8] F. Lagoutière, *personal communication.*

[9] Ph. Laurençot, *Weak solutions to the Lifshitz–Slyozov–Wagner equation*, Indiana Univ. Math. J., 50 (2001), pp. 1319–1346.

[10] Ph. Laurençot, *The Lifshitz–Slyozov–Wagner equation with conserved total volume*, SIAM J. Math. Anal., 34 (2002), pp. 257–272.

[11] Lê Châu-Hoàn, *Etude de la classe des opérateurs m-accrétifs de $L^1(\Omega)$ et accrétifs dans $L^\infty(\Omega)$*, Thèse de 3ème cycle, Université de Paris VI, Paris, 1977.

[12] R. J. LeVeque, *Numerical Methods for Conservation Laws*, 2nd ed., Lectures Math. ETH Zürich, Birkhäuser-Verlag, Basel, 1992.

[13] I. M. Lifshitz and V. V. Slyozov, *The kinetics of precipitation from supersaturated solid solutions*, J. Phys. Chem. Solids, 19 (1961), pp. 35–50.

[14] B. Meerson, *Fluctuations provide strong selection in Ostwald ripening*, Phys. Rev. E, 60 (1999), pp. 3072–3075.

[15] B. Meerson and P. V. Sasorov, *Domain stability, competition, growth, and selection in globally constrained bistable systems*, Phys. Rev. E, 53 (1996), pp. 3491–3494.

[16] B. Niethammer and R. L. Pego, *Non-self-similar behavior in the LSW theory of Ostwald ripening*, J. Statist. Phys., 95 (1999), pp. 867–902.

[17] B. Niethammer and R. L. Pego, *On the initial-value problem in the Lifshitz–Slyozov–Wagner theory of Ostwald ripening*, SIAM J. Math. Anal., 31 (2000), pp. 467–485.

[18] B. Niethammer and R. L. Pego, *The LSW model for domain coarsening: Asymptotic behavior for conserved total mass*, J. Statist. Phys., 104 (2001), pp. 1113–1144.

[19] B. Niethammer and R. L. Pego, in preparation.

[20] I. Rubinstein and B. Zaltzman, *Diffusional mechanism of strong selection in Ostwald ripening*, Phys. Rev. E, 61 (2000), pp. 709–717.

[21] J. Simon, *Compact sets in the space $L^p(0,T;B)$*, Ann. Mat. Pura Appl., 146 (1987), pp. 65–96.

[22] V. V. Slezov and V. V. Sagalovich, *Diffusive decomposition of solid solutions*, Soviet Phys. Uspekhi, 30 (1987), pp. 23–45.

[23] J. J. L. Velázquez, *The Becker–Döring equations and the Lifshitz–Slyozov theory of coarsening*, J. Statist. Phys., 92 (1998), pp. 195–236.

[24] I. I. Vrabie, *Compactness Methods for Nonlinear Evolutions*, 2nd ed., Pitman Monogr. Surveys Pure Appl. Math. 75, Longman, Harlow, UK, 1995.

[25] C. Wagner, *Theorie der Alterung von Niederschlägen durch Umlösen (Ostwald-Reifung)*, Z. Elektrochem., 65 (1961), pp. 581–591.

SIAM J. Numer. Anal.
Vol. 41, No. 2, pp. 589–604

A PRECONDITIONED CONJUGATE GRADIENT METHOD FOR NONSELFADJOINT OR INDEFINITE ORTHOGONAL SPLINE COLLOCATION PROBLEMS[*]

RAKHIM AITBAYEV[†] AND BERNARD BIALECKI[‡]

Abstract. We study the computation of the orthogonal spline collocation solution of a linear Dirichlet boundary value problem with a nonselfadjoint or an indefinite operator of the form $Lu = \sum a_{ij}(x)u_{x_ix_j} + \sum b_i(x)u_{x_i} + c(x)u$. We apply a preconditioned conjugate gradient method to the normal system of collocation equations with a preconditioner associated with a separable operator, and prove that the resulting algorithm has a convergence rate independent of the partition step size. We solve a problem with the preconditioner using an efficient direct matrix decomposition algorithm. On a uniform $N \times N$ partition, the cost of the algorithm for computing the collocation solution within a tolerance ϵ is $\mathrm{O}(N^2 \ln N |\ln \epsilon|)$.

Key words. nonselfadjoint or indefinite elliptic boundary value problem, orthogonal spline collocation, conjugate gradient method, preconditioner, matrix decomposition algorithm

AMS subject classifications. 65N35, 65N22, 65F10

PII. S0036142901391396

1. Introduction.

On $\Omega = (0,1) \times (0,1)$ with boundary $\partial\Omega$, we consider the Dirichlet boundary value problem (BVP)

$$Lu = f(x),\ x \in \Omega, \quad u(x) = 0,\ x \in \partial\Omega, \tag{1.1}$$

where $x = (x_1, x_2)$ and

$$Lu \ = \ \sum_{i,j=1}^{2} a_{ij}(x)\, u_{x_ix_j} + \sum_{i=1}^{2} b_i(x)\, u_{x_i} + c(x)\, u. \tag{1.2}$$

We assume that a_{ij}, c_i, b, and f are sufficiently smooth, $a_{12}(x) = a_{21}(x)$, $x \in \Omega$, and the a_{ij} satisfy the uniform ellipticity condition

$$\nu \sum_{i=1}^{2} \eta_i^2 \ \leq \ \sum_{i,j=1}^{2} a_{ij}(x)\, \eta_i\, \eta_j, \qquad x \in \Omega, \qquad \eta_1,\ \eta_2 \in R, \qquad \nu > 0. \tag{1.3}$$

In general, the operator L of (1.2) is nonselfadjoint and could be indefinite with respect to the L^2 inner product. The principal part of L is given in nondivergence form rather than the divergence form $\sum_{i,j=1}^{2} \left(a_{ij}(x)\, u_{x_i}\right)_{x_j}$. While the divergence form is natural for the standard finite element Galerkin method, the nondivergence form is more appropriate for the orthogonal spline collocation (OSC) method since, in this case, the implementation of the OSC method requires neither partial derivatives

[*] Received by the editors June 25, 2001; accepted for publication (in revised form) November 12, 2002; published electronically May 6, 2003. This work was supported in part by National Science Foundation grant DMS-9805827.

http://www.siam.org/journals/sinum/41-2/39139.html

[†] Department of Mathematics, New Mexico Institute of Mining and Technology, Socorro, NM 87801 (aitbayev@nmt.edu).

[‡] Department of Mathematical and Computer Sciences, Colorado School of Mines, Golden, CO 80401 (bbialeck@mines.edu).

of the equation coefficients nor their approximations. Also, in comparison with finite element methods, the OSC method requires no integrals or their approximations to set up the corresponding linear system.

An analysis of OSC for the BVP (1.1)–(1.3) was given in [7], where optimal order L^2 and H^1 error estimates and an optimal H^2 error estimate were obtained. The solution of the resulting linear system by banded Gaussian elimination requires $\mathrm{O}(N^4)$ operations on $N \times N$ partition [24, 25, 34]. The application of iterative methods reduces this cost. Classical iterative methods, such as Jacobi, Gauss–Seidel, or SOR, for the OSC solution of Poisson's equation on a uniform partition were studied in [23, 29, 37]. ADI methods for solving OSC problems with separable operators were investigated in [5, 13, 19].

Since the operator in the BVP is nonseparable, nonselfadjoint, or indefinite, one can attempt to solve the corresponding linear system by preconditioned BICGSTAB, QMR, CGS, and GMRES methods described in [35]. On the other hand, the preconditioned conjugate gradient (PCG) method is an effective method for solving a linear system with a symmetric and positive-definite matrix. We solve the OSC problem by a PCG method applied to a linear system of normal equations. This method is called PCGNR (see section 5.2 in [4] and section 9.5 in [35]).

Preconditioning of a selfadjoint positive-definite operator by a spectrally equivalent operator was suggested by Kantorovich [26]. This idea, used first by D'yakonov [17, 18] for the finite difference solution of a BVP by Richardson's method, was later extended to the PCG method for the finite element and finite difference solutions of nonselfadjoint or indefinite BVPs [10, 11, 20, 31]. Preconditioning for some nonseparable OSC problems was studied in [6, 27, 28, 38].

Before describing our approach to solving the OSC problem, let us discuss some common techniques used in other discretization methods. Let L_h be a finite difference or a finite element operator associated with a nonselfadjoint or indefinite BVP, and let L_h^* be the adjoint of L_h with respect to an appropriate inner product $(\cdot,\cdot)_h$. Two well-known approaches for solving the equation $L_h u_h = f_h$ are based on preconditioned normal equations:

$$L_h^* M_h^{-2} L_h u_h = L_h^* M_h^{-2} f_h, \tag{1.4}$$

$$M_h^{-1} L_h^* M_h^{-1} L_h u_h = M_h^{-1} L_h^* M_h^{-1} f_h, \tag{1.5}$$

where a selfadjoint and positive-definite operator M_h is a preconditioner for L_h [10, 11, 20, 30, 31]. The operators $L_h^* M_h^{-2} L_h$ and $M_h^{-1} L_h^* M_h^{-1} L_h$ are selfadjoint with respect to $(\cdot,\cdot)_h$-inner product and M_h-inner product, respectively. Therefore, the equations (1.4) and (1.5) can be solved by the CG method with the corresponding inner products. Analyses of the CG solution of (1.4) and (1.5) are, respectively, related to L^2-norm and H^1-norm analyses of a finite difference or a finite element discretization. The finite element equation $L_h u_h = f_h$ can also be solved by modern preconditioned domain decomposition and multilevel methods (see, for example, [33, 36]). However, since these methods are not well developed for OSC, in this article we consider the solution of the OSC problem $L_h u_h = f_h$ approximating BVP (1.1)–(1.2) based on the normal equation

$$(M_h^* M_h)^{-1} L_h^* L_h u_h \;=\; (M_h^* M_h)^{-1} L_h^* f_h, \tag{1.6}$$

where a nonselfadjoint or indefinite OSC operator M_h is associated with a separable operator $\tilde{L}$ which is "close" to L. Following an H^2-norm analysis of [7], we show that $M_h^* M_h$ and $L_h^* L_h$ are spectrally equivalent with respect to $(\cdot,\cdot)_h$-inner product. Since the operator $(M_h^* M_h)^{-1} L_h^* L_h$ is selfadjoint and positive-definite with respect to $M_h^* M_h$-inner product, we solve (1.6) by the corresponding CG method. This method is equivalent to solving the equation $L_h^* L_h u_h = L_h^* f_h$ by PCG with $M_h^* M_h$ as a preconditioner. At each iteration of PCG, a new matrix decomposition algorithm allows us to solve the equation $M_h^* M_h w = r$ in one step rather than in two separate steps $M_h^* z = r$ and $M_h w = z$. On a uniform $N \times N$ partition, the total cost of our PCG algorithm with a tolerance ϵ is $\mathrm{O}(N^2 \ln N |\ln \epsilon|)$. The approach presented in this paper was used in [1] for the solution of a nonlinear OSC Dirichlet BVP by Newton's method.

An outline of this paper is as follows. Notation and auxiliary results are introduced in section 2. We prove spectral equivalence of the OSC operators in section 3 and discuss the matrix-vector form of the OSC problem in section 4. In section 5, we prove convergence of the PCG algorithm, and in section 6, we formulate matrix decomposition algorithms for the solution of an equation with the preconditioner. The implementation and the cost are discussed in section 7. In section 8, we present results of our numerical tests, and finally, section 9 is devoted to our conclusions.

2. Preliminaries. For $k = 1, 2$, let $\pi_k = \{x_{k,i}\}_{i=0}^{N_k}$ be a partition of the interval $[0, 1]$ such that

$$0 = x_{k,0} < x_{k,1} < \cdots < x_{k,N_k} = 1,$$

and let $h_{k,i} = x_{k,i} - x_{k,i-1}$ for $i = 1, \ldots, N_k$. Let

$$\underline{h}_k = \min_i h_{k,i}, \quad \overline{h}_k = \max_i h_{k,i}, \quad h = \max(\overline{h}_1, \overline{h}_2).$$

Throughout we assume that the partitions $\pi_h = \pi_1 \times \pi_2$ are regular; that is, there exist positive constants σ_1, σ_2, and σ_3, all independent of h, such that $\sigma_1 \overline{h}_1 \le \underline{h}_1$, $\sigma_1 \overline{h}_2 \le \underline{h}_2$, and $\sigma_2 \le \overline{h}_1/\overline{h}_2 \le \sigma_3$.

For an integer $r \ge 3$, let P_r be the set of all polynomials of degree $\le r$. For $k = 1, 2$, let

$$V_k = \{v \in C^1[0,1] : \, v|_{[x_{k,i-1},x_{k,i}]} \in P_r, \; i = 1, \ldots, N_k\}$$

be the space of Hermite splines of degree r associated with the partition π_k, and let $V_k^0 = \{v \in V_k : \, v(0) = v(1) = 0\}$. It is easy to verify that the dimension of V_k^0 is $K_k = (r-1)N_k$. Let $V^0 = V_1^0 \otimes V_2^0$, where $\otimes$ denotes the tensor product of vector spaces. Note that V^0 is the set of all functions that are finite linear combinations of products $v_1(x_1)\, v_2(x_2)$, where $v_1 \in V_1^0$ and $v_2 \in V_2^0$. The dimension of V^0 is $K = K_1 K_2$.

Let $\{\eta_l\}_{l=1}^{r-1}$ and $\{\omega_l\}_{l=1}^{r-1}$ be, respectively, the nodes and the weights of the $(r-1)$-point Gauss quadrature rule on $(0, 1)$. For $k = 1, 2$, let $\mathcal{G}_k$ consist of the points

$$\xi_{k,i,l} \; = \; x_{k,i-1} + h_{k,i}\eta_l, \quad i = 1, \ldots, N_k, \quad l = 1, \ldots, r-1. \tag{2.1}$$

Then $\mathcal{G} = \mathcal{G}_1 \times \mathcal{G}_2$ is the set of Gauss points in Ω associated with the partition π_h. Corollary 5.3 of [32] implies that any $v \in V^0$ is uniquely defined by its values on $\mathcal{G}$.

For v and z defined on $\mathcal{G}$, let

$$(v, z)_h = \sum_{i=1}^{N_1} h_{1,i} \sum_{k=1}^{r-1} \omega_k \sum_{j=1}^{N_2} h_{2,j} \sum_{l=1}^{r-1} \omega_l \, (vz)(\xi_{1,i,k}, \xi_{2,j,l}) \tag{2.2}$$

and let $\|v\|_h = \sqrt{(v,v)_h}$. Let $\rho(x)$ be a continuous positive function on $\overline{\Omega}$, and let $\rho_{\min} = \min_{x\in\overline{\Omega}} \rho(x)$ and $\rho_{\max} = \max_{x\in\overline{\Omega}} \rho(x)$. We shall also use

$$(v,z)_{h,\rho} \;=\; (\rho v, z)_h \tag{2.3}$$

and $\|v\|_{h,\rho} = \sqrt{(v,v)_{h,\rho}}$. We note that $(\cdot,\cdot)_{h,\rho}$ and $\|\cdot\|_{h,\rho}$ are, respectively, an inner product and a norm in V^0. It is easy to see that

$$\rho_{\min}\|v\|_h \;\le\; \|v\|_{h,\rho} \;\le\; \rho_{\max}\|v\|_h \tag{2.4}$$

for any v defined on $\mathcal{G}$.

Throughout, $H^l(\Omega)$ denotes the Sobolev space with the standard norm $\|\cdot\|_{H^l(\Omega)}$ [12]. We write $\partial_k^l = \partial^l/\partial x_k^l$ and $\partial^{(i,j)} = \partial^{i+j}/(\partial x_1^i \partial x_2^j)$. In the following, C denotes a generic positive constant independent of h.

The OSC problem for (1.1)–(1.2) consists of finding $u_h \in V^0$ such that

$$Lu_h(\xi) \;=\; f(\xi), \qquad \xi \in \mathcal{G}. \tag{2.5}$$

The following result was proved in [7, Theorem 3.3].

Theorem 2.1. *Let operator L of* (1.2) *be one-to-one from $\{v \in H^2(\Omega) : v = 0$ on $\partial\Omega\}$ to $L^2(\Omega)$, and let h be sufficiently small. Then the OSC problem* (2.5) *has a unique solution $u_h \in V^0$. Moreover, if $u \in H^{r+1}(\Omega)$ is the solution of* (1.1), *then*

$$\|u-u_h\|_{H^2(\Omega)} \;\le\; C\, h^{r-1}\, \|u\|_{H^{r+1}(\Omega)}.$$

3. Spectral equivalence of the OSC operators.

The following is the key result of this paper.

Theorem 3.1. *Let the assumptions of Theorem* 2.1 *be satisfied. Then there are positive constants γ_1 and γ_2 independent of h such that*

$$\gamma_1\|v\|_{H^2(\Omega)} \;\le\; \|Lv\|_{h,\rho} \;\le\; \gamma_2\|v\|_{H^2(\Omega)}, \qquad v\in V^0. \tag{3.1}$$

Proof. We note that the inequality

$$C\,\|v\|_{H^2(\Omega)} \;\le\; \|Lv\|_h + \|v\|_{L^2(\Omega)}, \quad v \in V^0,$$

was proved in [7, (3.20)]. Also, it follows from Lemma 3.2 and (3.21) of [7] that

$$C\,\|v\|_{L^2(\Omega)} \;\le\; h\,\|v\|_{H^2(\Omega)} + \|Lv\|_h, \quad v\in V^0.$$

Thus, for h sufficiently small, we have $C\|v\|_{H^2(\Omega)} \le \|Lv\|_h$, $v \in V^0$, which, along with the first inequality in (2.4), gives the first inequality in (3.1).

Using (1.2) and the boundedness of the coefficients of L, we obtain

$$\|Lv\|_h \;\le\; C \sum_{0\le i+j\le 2} \left\|\partial^{(i,j)}v\right\|_h, \quad v\in V^0. \tag{3.2}$$

Applying the inverse inequality of Theorem 3.2.6 in [12], we have

$$\left\|\partial^{(i,j)}v\right\|_h \;\le\; C\,\left\|\partial^{(i,j)}v\right\|_{L^2(\Omega)}, \quad v\in V^0, \quad 0\le i+j\le 2. \tag{3.3}$$

Using the second inequality in (2.4), (3.2), (3.3), and the Cauchy–Schwarz inequality, we obtain the second inequality in (3.1). ∎

We also consider the separable differential operator

$$\tilde{L} \;=\; \tilde{L}_1 + \tilde{L}_2, \tag{3.4}$$

where, for $k = 1, 2$,

$$\tilde{L}_k v \;=\; \tilde{a}_k(x_k)\, v_{x_k x_k} + \tilde{b}_k(x_k)\, v_{x_k} + \tilde{c}_k(x_k)\, v, \tag{3.5}$$

$\tilde{a}_k$, $\tilde{b}_k$, and $\tilde{c}_k$ are sufficiently smooth, and

$$\tilde{a}_k(x) \geq \nu > 0, \quad x \in \Omega, \quad k = 1, 2.$$

LEMMA 3.1. *Let the assumptions of Theorem* 2.1 *be satisfied and let the operator* $\tilde{L}$ *be one-to-one from* $\{v \in H^2(\Omega) : v = 0$ *on* $\partial\Omega\}$ *to* $L^2(\Omega)$. *Then there are positive constants* α *and* β, *independent of* h, *such that*

$$\sqrt{\alpha}\, \|\tilde{L}v\|_{h,\rho} \;\leq\; \|Lv\|_{h,\rho} \;\leq\; \sqrt{\beta}\, \|\tilde{L}v\|_{h,\rho}, \quad v \in V^0. \tag{3.6}$$

Proof. Since $\tilde{L}$ is a special case of L, Theorem 3.1 implies that

$$\tilde{\gamma}_1\, \|v\|_{H^2(\Omega)} \;\leq\; \|\tilde{L}v\|_{h,\rho} \;\leq\; \tilde{\gamma}_2\, \|v\|_{H^2(\Omega)}, \qquad v \in V^0, \tag{3.7}$$

where the positive constants $\tilde{\gamma}_1$ and $\tilde{\gamma}_2$ are independent of h. Using (3.1) and (3.7), we obtain (3.6) with $\sqrt{\alpha} = \gamma_1/\tilde{\gamma}_2$ and $\sqrt{\beta} = \gamma_2/\tilde{\gamma}_1$. □

If L_h and M_h are OSC operators from V^0 to V^0 associated with L of (1.2) and $\tilde{L}$ of (3.4)–(3.5), respectively, then (3.6) shows that $L_h^* L_h$ and $M_h^* M_h$ are spectrally equivalent with respect to the inner product $(\cdot,\cdot)_{h,\rho}$. This is equivalent to L_h and M_h being uniformly L^2-norm equivalent (see (1.15) in [30]). Consequently, our Lemma 3.1 is the OSC counterpart of Lemma 3.1 in [30] for continuous operators.

4. Matrix-vector form of the OSC problem. For $k = 1, 2$, let $\{\phi_{k,j}\}_{j=1}^{K_k}$ be a basis for V_k^0. Then $\{\phi_j(x)\}_{j=1}^{K}$, where

$$\phi_{K_2(j_1-1)+j_2}(x) \;=\; \phi_{1,j_1}(x_1)\phi_{2,j_2}(x_2), \quad j_k = 1, \ldots, K_k, \quad k = 1, 2, \tag{4.1}$$

is a basis for V^0. Thus, for any $v \in V^0$, there exists a unique vector $[v]_{\mathcal{H}} = [v_1, \ldots, v_K]^T \in R^K$ such that

$$v(x) \;=\; \sum_{j=1}^{K} v_j\, \phi_j(x), \quad x \in \overline{\Omega}. \tag{4.2}$$

Let $\mathcal{G} = \{\xi_i\}_{i=1}^{K}$, where

$$\xi_{(i_1-1)K_2+i_2} = (\xi_{1,i_1}, \xi_{2,i_2}), \quad i_k = 1, \ldots, K_k, \quad k = 1, 2, \tag{4.3}$$

$$\xi_{k,(i-1)(r-1)+l} = \xi_{k,i,l}, \qquad i = 1, \ldots, N_k, \qquad l = 1, \ldots, r-1, \tag{4.4}$$

and $\xi_{k,i,l}$ are given by (2.1). For any v defined on $\mathcal{G}$, we introduce the vector $[v]_{\mathcal{G}} = [v(\xi_1), \ldots, v(\xi_K)]^T \in R^K$.

For a matrix A, its (i,j) entry is denoted by $(A)_{ij}$. Let M_L be the matrix defined by

$$(M_L)_{ij} \;=\; (L\phi_j)(\xi_i), \quad i, j = 1, \ldots, K. \tag{4.5}$$

Then using (4.2) and (4.5), we have

$$[Lv]_{\mathcal{G}} \;=\; M_L[v]_{\mathcal{H}}, \qquad v \in V^0. \tag{4.6}$$

Let $\rho(x)$ be a continuous positive function on $\overline{\Omega}$. We introduce

$$D \;=\; \operatorname{diag}(\rho(\xi_1), \ldots, \rho(\xi_K)), \tag{4.7}$$

$$W \;=\; W_1 \otimes W_2, \tag{4.8}$$

where $\otimes$ denotes the matrix tensor product and, for $k = 1, 2$,

$$W_k \;=\; \operatorname{diag}\,(h_{k,1}, \ldots, h_{k,N_k}) \otimes \operatorname{diag}\,(\omega_1, \ldots, \omega_{r-1}). \tag{4.9}$$

From (2.3), (2.2), (4.3), (4.4), and (4.7)–(4.9), we have

$$(v, z)_{h,\rho} \;=\; [v]_{\mathcal{G}}^T W D [z]_{\mathcal{G}}. \tag{4.10}$$

Using (4.6), the OSC problem (2.5) can be rewritten in the matrix-vector form

$$M_L[u_h]_{\mathcal{H}} \;=\; [f]_{\mathcal{G}}. \tag{4.11}$$

Multiplying this equation by $M_L^T W D$ on the left, we obtain

$$A\,\vec{u} \;=\; \vec{f}, \tag{4.12}$$

where

$$A = M_L^T W D M_L, \quad \vec{u} = [u_h]_{\mathcal{H}}, \quad \text{and} \quad \vec{f} = M_L^T W D [f]_{\mathcal{G}}. \tag{4.13}$$

5. PCG algorithm. Let the operator $\tilde{L}$ be as in (3.4)–(3.5), and let

$$\tilde{A} \;=\; M_{\tilde{L}}^T\, W\, D\, M_{\tilde{L}}, \tag{5.1}$$

where $M_{\tilde{L}}$ is defined by

$$(M_{\tilde{L}})_{ij} \;=\; (\tilde{L}\phi_j)(\xi_i), \quad i, j = 1, \ldots, K. \tag{5.2}$$

It follows easily from (4.13) and (5.1) that A and $\tilde{A}$ are symmetric.

LEMMA 5.1. *Let the assumptions of Lemma* 3.1 *be satisfied. Then the matrices* A *of* (4.13) *and* $\tilde{A}$ *of* (5.1) *are positive-definite. Moreover,*

$$\alpha\, \vec{v}^T \tilde{A} \vec{v} \;\le\; \vec{v}^T A \vec{v} \;\le\; \beta\, \vec{v}^T \tilde{A} \vec{v}, \qquad \vec{v} \in R^K, \tag{5.3}$$

where the positive constants α *and* β *are the same as in* (3.6).

Proof. Using (4.10), (4.6), and (4.13), we obtain, for $v \in V^0$,

$$\|Lv\|_{h,\rho}^2 = (Lv, Lv)_{h,\rho} = [Lv]_{\mathcal{G}}^T W D [Lv]_{\mathcal{G}} = [v]_{\mathcal{H}}^T M_L^T W D M_L [v]_{\mathcal{H}} = [v]_{\mathcal{H}}^T A [v]_{\mathcal{H}}. \tag{5.4}$$

Hence the first inequality in (3.1) and $\gamma_1 > 0$ imply that A is positive-definite. Similarly, we have $\|\tilde{L}v\|_{h,\rho}^2 = [v]_{\mathcal{H}}^T \tilde{A}\,[v]_{\mathcal{H}}$. Therefore, (5.3) follows from (3.6) and (5.4). The second inequality in (5.3) and $\beta > 0$ imply that $\tilde{A}$ is also positive-definite. □

We solve (4.12) by the PCG method (see Algorithm 9.4.14 in [22]) with $\tilde{A}$ as a preconditioner.

THEOREM 5.1. *Let the assumptions of Lemma* 3.1 *be satisfied. For an iterate $\vec{u}_k$ generated by the PCG method, let $u_{h,k} \in V^0$ be such that $[u_{h,k}]_{\mathcal{H}} = \vec{u}_k$. Then*

$$\|f - Lu_{h,k}\|_{h,\rho} \;\le\; 2\delta^k \|f - Lu_{h,0}\|_{h,\rho}, \quad k = 0, 1, 2, \ldots, \tag{5.5}$$

where $\delta = (\sqrt{\beta/\alpha} - 1)/(\sqrt{\beta/\alpha} + 1)$ and α and β are the same as in (3.6).

Proof. Since A and $\tilde{A}$ are symmetric positive-definite, (5.5) follows from (5.3), Theorem 9.4.14 in [22], (5.4), and (2.5). □

COROLLARY 5.1. *With δ of Theorem* 5.1*, we have*

$$\|u_h - u_{h,k}\|_{H^2(\Omega)} \;\le\; C\delta^k \|u_h - u_{h,0}\|_{H^2(\Omega)}, \quad k = 0, 1, 2, \ldots . \tag{5.6}$$

Proof. Inequality (5.6) follows from (5.5), (2.5), and (3.1). □

Let $\tilde{r}_k = [f]_{\mathcal{G}} - M_L \vec{u}_k$, $k = 0, 1, \ldots$. Then (2.5), (5.4), and (4.11) give $\|\tilde{r}_k\|_{WD} = \|f - Lu_{h,k}\|_{h,\rho}$. Hence, if $\tilde{r}_k$ is required at each iteration and the iterations are terminated when

$$\|f - Lu_{h,k}\|_{h,\rho} \;\le\; \epsilon \|f - Lu_{h,0}\|_{h,\rho}, \tag{5.7}$$

then the PCG method can be rewritten in the following form (cf. Algorithm 9.7 in [35]).

ALGORITHM 5.1.

```
select $\vec{u}_0$, $\tilde{r}_0 = [f]_{\mathcal{G}} - M_L\vec{u}_0$, $\vec{r}_0 = M_L^T WD\tilde{r}_0$, solve $\tilde{A}\vec{p}_0 = \vec{r}_0$, $\rho_0 = \vec{r}_0^T\vec{p}_0$,
for $k = 0, 1, 2, \ldots$ (as long as $\|\tilde{r}_k\|_{WD} > \epsilon\|\tilde{r}_0\|_{WD}$):
    $\vec{w}_k = M_L\vec{p}_k$, $\alpha_k = \rho_k/(\vec{w}_k^T WD\vec{w}_k)$, $\vec{u}_{k+1} = \vec{u}_k + \alpha_k\vec{p}_k$,
    $\tilde{r}_{k+1} = \tilde{r}_k - \alpha_k\vec{w}_k$, $\vec{r}_{k+1} = M_L^T WD\tilde{r}_k$,
    solve $\tilde{A}\vec{z}_{k+1} = \vec{r}_{k+1}$, $\rho_{k+1} = \vec{r}_{k+1}^T\vec{z}_{k+1}$, $\vec{p}_{k+1} = \vec{z}_{k+1} + (\rho_{k+1}/\rho_k)\vec{p}_k$.
```

6. Preconditioning.

At each iteration of Algorithm 5.1, a linear system

$$\tilde{A}\vec{w} \;=\; \vec{r} \tag{6.1}$$

must be solved, where $\tilde{A}$ is defined by (5.1)–(5.2). If $\tilde{b}_1 = 0$ or $\tilde{b}_2 = 0$, then (6.1) can be solved by matrix decomposition algorithms which we describe assuming $\tilde{b}_1 = 0$.

For $k = 1, 2$, let I_k be the identity matrix of order K_k, and let the matrices A_k and B_k be defined by

$$(A_k)_{ij} \;=\; (\tilde{L}_k\phi_{k,j})\,(\xi_{k,i}), \qquad (B_k)_{ij} \;=\; \phi_{k,j}\,(\xi_{k,i}), \qquad i, j = 1, 2, \ldots, K_k, \tag{6.2}$$

where $\tilde{L}_k$ is given by (3.5). It follows from (5.2), (3.4), (4.1), (4.3), and (6.2) that

$$M_{\tilde{L}} \;=\; A_1 \otimes B_2 + B_1 \otimes A_2. \tag{6.3}$$

With $\tilde{a}_1$ of (3.5) for $k = 1$, let

$$D_1 = \operatorname{diag}(1/\tilde{a}_1(\xi_{1,1}), \ldots, 1/\tilde{a}_1(\xi_{1,K_1})). \tag{6.4}$$

We introduce the $K_1 \times K_1$ matrices

$$G = B_1^T W_1 D_1 A_1, \quad F = B_1^T W_1 D_1 B_1, \tag{6.5}$$

where W_1 is given by (4.9). It was proved in [8, Lemma 3.1] that F is symmetric positive-definite and G is symmetric. Therefore, it follows from [21, Corollary 8.7.2] that there exists a real diagonal matrix

$$\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_{K_1}) \tag{6.6}$$

and a real nonsingular matrix Z such that

$$Z^T G Z = \Lambda, \quad Z^T F Z = I_1. \tag{6.7}$$

Now we discuss two approaches to solving (6.1). In the first, we take $\rho = 1$ in $(\cdot,\cdot)_{h,\rho}$ of (2.3) and obtain $D = I$ by (4.7). Hence, by (5.1), the linear system (6.1) becomes

$$M_{\tilde{L}}^T W M_{\tilde{L}} \vec{w} \;=\; \vec{r},$$

where the diagonal matrix W is defined by (4.8)–(4.9). Thus, the system in (6.1) can be solved as follows.

Algorithm 6.1.

Step 1. Determine Λ and Z satisfying (6.7).

Step 2. Solve $M_{\tilde{L}}^T \vec{z} = \vec{r}$ by a modification of Algorithm I in [8] (see below).

Step 3. Solve the diagonal system $W\vec{v} = \vec{z}$.

Step 4. Solve $M_{\tilde{L}} \vec{w} = \vec{v}$ by Algorithm I in [8].

We note that the matrix decomposition Algorithm I of [8] is based on the decomposition

$$(Z^T B_1^T W_1 D_1 \otimes I_2) M_{\tilde{L}} (Z \otimes I_2) = \Lambda \otimes B_2 + I_1 \otimes A_2,$$

which follows easily from (6.3), (6.5), and (6.7). By taking the transpose of both sides, we also have

$$(Z^T \otimes I_2) M_{\tilde{L}}^T (W_1 D_1 B_1 Z \otimes I_2) = \Lambda \otimes B_2^T + I_1 \otimes A_2^T.$$

Therefore, Step 2 is implemented in a way similar to Step 4 (see [8] for details).

In the second approach to solving (6.1), we take $\rho(x_1, x_2) = 1/\tilde{a}_1(x_1)$, $(x_1, x_2) \in \overline{\Omega}$, which, by (4.7), gives

$$D \;=\; D_1 \otimes I_2, \tag{6.8}$$

where D_1 is given by (6.4). Then the system in (6.1) can be solved in one step by a matrix decomposition algorithm which we describe in the following. Since $Z \otimes I_2$ is nonsingular, the system in (6.1) is equivalent to

$$S\vec{y} = \vec{d}, \tag{6.9}$$

where $\vec{y} = (Z \otimes I_2)^{-1}\vec{w}$, $\vec{d} = (Z^T \otimes I_2)\vec{r}$, and

$$S \;=\; (Z^T \otimes I_2)\tilde{A}(Z \otimes I_2). \tag{6.10}$$

Lemma 6.1. *Assume that $\tilde{L}$ satisfies the assumptions in Lemma* 3.1 *and that h is sufficiently small. Then S of* (6.10) *is a real block diagonal matrix with $K_2 \times K_2$ symmetric positive-definite diagonal blocks*

$$S_i \;=\; (A_2 + \lambda_i B_2)^T\, W_2\, (A_2 + \lambda_i B_2), \quad i = 1, \dots, K_1. \tag{6.11}$$

Proof. Using (5.1), (6.3), (4.8), (6.8), (6.5), and $G = G^T$, we obtain

$$\begin{aligned} \tilde{A} &= (A_1^T \otimes B_2^T + B_1^T \otimes A_2^T)(W_1 D_1 \otimes W_2)(A_1 \otimes B_2 + B_1 \otimes A_2) \\ &= A_1^T W_1 D_1 A_1 \otimes B_2^T W_2 B_2 + G \otimes B_2^T W_2 A_2 \\ &\quad + G \otimes A_2^T W_2 B_2 + F \otimes A_2^T W_2 A_2. \end{aligned} \tag{6.12}$$

The second equations in (6.5) and (6.7) give

$$B_1 Z Z^T B_1^T W_1 D_1 \;=\; I_1. \tag{6.13}$$

Using (6.13), the first equations in (6.7) and (6.5), and $G^T = G$, we obtain

$$Z^T A_1^T W_1 D_1 A_1 Z = Z^T A_1^T W_1 D_1 B_1 Z Z^T B_1^T W_1 D_1 A_1 Z = \Lambda^2. \tag{6.14}$$

Thus, (6.10), (6.12), (6.14), and (6.7) give

$$S \;=\; \Lambda^2 \otimes B_2^T W_2 B_2 + \Lambda \otimes B_2^T W_2 A_2 + \Lambda \otimes A_2^T W_2 B_2 + I_1 \otimes A_2^T W_2 A_2. \tag{6.15}$$

It follows from (6.15) and (6.6) that S is real block diagonal with the diagonal blocks given by (6.11).

We see from (6.11) that each matrix S_i is symmetric and $\vec{v}^T S_i \vec{v} \geq 0$ for any $\vec{v} \in R^{K_2}$. Since Z is nonsingular and $\tilde{A}$ is positive-definite (see Lemma 5.1), it follows from (6.10) that S is nonsingular. This implies that S_i is nonsingular and hence positive-definite. □

For $\vec{v} \in R^K$, let $[\vec{v}]_i = [v_{(i-1)K_2+1}, \ldots, v_{iK_2}]^T \in R^{K_2}$, $i = 1, \ldots, K_1$. Based on (6.9) and Lemma 6.1, we can formulate the following matrix decomposition algorithm for the solution of (6.1).

Algorithm 6.2.

```
Step 1. Determine $\Lambda$ and $Z$ satisfying (6.7).
Step 2. Compute $\vec{d} = (Z^T \otimes I_2)\vec{r}$.
Step 3. Solve $S_i[\vec{y}]_i = [\vec{d}]_i$ for $i = 1, \ldots, K_1$.
Step 4. Compute $\vec{w} = (Z \otimes I_2)\vec{y}$.
```

7. Implementation and cost. To discuss the implementation and cost, we assume that the basis functions $\{\phi_{k,j}\}_{j=1}^{K_k}$ for V_k^0, $k = 1, 2$, are B-splines or Hermite-type functions ordered in the standard way. Then matrices A_k and B_k, $k = 1, 2$, in (6.2) are almost block diagonal and have the structure described in [3], depending on the type of basis functions.

Step 1 of Algorithms 6.1 and 6.2 involves solving the symmetric generalized eigenproblem (6.7). This can be done by one of the following three algorithms.

Algorithm 7.1.

```
Step 1. Compute $G$ and $F$ of (6.5).
Step 2. Compute band Cholesky factorization $F = LL^T$.
Step 3. Compute full symmetric $C = L^{-1} G L^{-T}$.
Step 4. Use QR algorithm to compute the diagonal $\Lambda$ and an orthogonal $Q$
            such that $Q^T C Q = \Lambda$.
Step 5. Compute $Z = L^{-T} Q$.
```

Algorithm 7.2.

```
Step 1. Compute $G$ and $F$ of (6.5).
Step 2. Compute band Cholesky factorization $F = LL^T$.
Step 3. Use Crawford's algorithm to compute $C$ and $X$.
Step 4. Use band QR algorithm to compute the diagonal $\Lambda$ and
            an orthogonal $Q$ such that $Q^T C Q = \Lambda$.
Step 4. Compute $Z = XQ$.
```

Algorithm 7.3.

```
Step 1. Compute full symmetric $C = (W_1 D_1)^{1/2} A_1 B_1^{-1} (W_1 D_1)^{-1/2}$.
Step 2. Use QR algorithm to compute the diagonal $\Lambda$ and an orthogonal $Q$
            such that $Q^T C Q = \Lambda$.
Step 3. Compute $Z = B_1^{-1} (W_1 D_1)^{-1/2} Q$.
```

TABLE 7.1
Costs of Algorithms 7.1–7.3.

Compute	Algorithm 7.1	Algorithm 7.2	Algorithm 7.3
F, G	$\mathrm{O}(K_1)$	$\mathrm{O}(K_1)$	–
L	$\mathrm{O}(K_1)$	$\mathrm{O}(K_1)$	–
C (X for Alg. 7.2)	$\mathrm{O}(K_1^2)$	$\mathrm{O}(K_1^2)$	$\mathrm{O}(K_1)$
Λ, Q	$9K_1^3$	$6K_1^3$	$9K_1^3$
Z	$\mathrm{O}(K_1^2)$	$2K_1^3$	$\mathrm{O}(K_1^2)$
total cost	$9K_1^3+\mathrm{O}(K_1^2)$	$8K_1^3+\mathrm{O}(K_1^2)$	$9K_1^3+\mathrm{O}(K_1^2)$

Algorithm 7.1 is the standard Wilkinson's algorithm (see Algorithm 8.7.1 in [21]). Algorithm 7.2 is based on Crawford's algorithm (see [2] and [14] for details). If $F = LL^T$ is the band Cholesky factorization of F, Crawford's algorithm computes band symmetric C and $X = L^{-T}P$, with P orthogonal, such that $C = X^TGX$ and C is orthogonally similar to Λ. Algorithm 7.3 is based on Step 1 of Algorithm II of [8]. The algorithm uses the factorization $F = LL^T$ with $L = B_1^T(W_1D_1)^{1/2}$. The costs of Algorithms 7.1–7.3 are given in Table 7.1.

The implementation of Step 4 of Algorithm 6.1 and its cost of $4K_1^2K_2$ are discussed in [8]. The implementation of Step 2 of Algorithm 6.1 is similar and its cost is also $4K_1^2K_2$.

Step 2 and Step 4 of Algorithm 6.2 involve K_2 multiplications by $K_1\times K_1$ matrices Z^T and Z, respectively, and hence each step requires $2K_1^2K_2$ operations. Each $K_2\times K_2$ matrix S_i in (6.11) is symmetric, positive-definite, and block tridiagonal with $r-1$ by $r-1$ blocks. A linear system with S_i can be solved by a direct block tridiagonal solver (for example, by the routine BLKTRI from the package FISHPACK described in [39]) at the cost $\mathrm{O}(K_2)$. Therefore, the cost of Step 3 of Algorithm 6.2 is $\mathrm{O}(K_1K_2)$. Since A_2 and B_2 are almost block diagonal, so is S_i. Hence a linear system with S_i can also be solved by two calls to the routine COLROW [15, 16].

Of course, when Algorithms 6.1 and 6.2 are used in Algorithm 5.1 to solve a linear system with the coefficient matrix $\tilde{A}$, the matrices Λ and Z are precomputed first. For Algorithm 6.1, the remaining cost is $8K_1^2K_2$ since this is the cost of all multiplications by Z^T and Z. For Algorithm 6.2, the remaining cost is half of that of Algorithm 6.1.

In a special case of $r = 3$, a uniform partition π_1, constant coefficients $\tilde{a}_1$, $\tilde{c}_1$ of $\tilde{L}$ in (3.5), and $\tilde{b}_1 = 0$, the matrices Λ and Z in (6.7) are known in a closed form. Moreover, it follows from Theorem 2.3 in [9] that matrix Z is given in terms of sines and cosines. Therefore, all multiplications by Z^T and Z in Algorithm 6.1 can be performed using FFTs with the cost $\mathrm{O}(K_1K_2\log K_1)$. Thus, the total cost of Algorithm 6.1 is $\mathrm{O}(K_1K_2\log K_1)$. In this case, it follows from (5.6) that the cost of our PCG Algorithm 5.1 with a tolerance ϵ on an $N\times N$ partition is $\mathrm{O}(N^2\ln N|\ln\epsilon|)$.

8. Numerical tests.

Before considering our numerical tests, we present an additional result which was used in the tests. Let $\hat{a}_1(x)$, $\hat{a}_2(x)$, and $\hat{c}(x)$ be sufficiently smooth functions on Ω, and for $i = 1, 2$, let $\hat{a}_i(x) \geq \nu > 0$, $x \in \Omega$. Let

$$\hat{L}v \;=\; (\hat{a}_1(x)v_{x_1})_{x_1} + (\hat{a}_2(x)v_{x_2})_{x_2} + \hat{c}(x)v. \tag{8.1}$$

The operator $\hat{L}$ is selfadjoint with respect to the standard L^2-inner product, and it is negative-definite if $\hat{c}(x) \leq 0$, $x \in \Omega$. We prove that $\hat{L}$ is indefinite if

$$\min_{x\in\Omega}\{\hat{c}(x)\} > 2\pi^2 \max_{x\in\Omega}\{\hat{a}_1(x),\, \hat{a}_2(x)\}. \tag{8.2}$$

Using Green's formula and (8.2), we have, for $v \neq 0$,

$$\begin{aligned}\int_\Omega \hat{L}v(x)\,v(x)\,dx &= -\int_\Omega (\hat{a}_1 v_{x_1}^2 + \hat{a}_2 v_{x_2}^2)\,dx + \int_\Omega \hat{c}v^2\,dx \\ &\geq -\max_{x\in\Omega}\{\hat{a}_1(x),\, \hat{a}_2(x)\}\|\nabla v\|^2_{L^2(\Omega)} + \min_{x\in\Omega}\{\hat{c}(x)\}\|v\|^2_{L^2(\Omega)} \\ &> \max_{x\in\Omega}\{\hat{a}_1(x),\, \hat{a}_2(x)\}\left(2\pi^2\|v\|^2_{L^2(\Omega)} - \|\nabla v\|^2_{L^2(\Omega)}\right), \end{aligned} \tag{8.3}$$

where $\|\nabla v\|^2_{L^2(\Omega)} = \int_\Omega (v_{x_1}^2 + v_{x_2}^2)dx$. It is easy to see that, for

$$v_{k,l}(x) \;=\; 2\sin(k\pi x_1)\sin(l\pi x_2), \quad x \in \Omega,$$

with integers k and l, we have

$$\|\nabla v_{k,l}\|^2_{L^2(\Omega)} \;=\; \pi^2(k^2 + l^2)\|v_{k,l}\|^2_{L^2(\Omega)}. \tag{8.4}$$

Thus, from (8.3) and (8.4), we obtain $\int_\Omega \hat{L}v_{1,1}(x)\,v_{1,1}(x)\,dx > 0$.

On the other hand, by (8.4), we have

$$\begin{aligned}\int_\Omega \hat{L}v_{k,l}(x)\,v_{k,l}(x)\,dx &\leq -\nu\|\nabla v_{k,l}\|^2_{L^2(\Omega)} + \max_{x\in\Omega}\{\hat{c}(x)\}\|v_{k,l}\|^2_{L^2(\Omega)} \\ &= \left(\max_{x\in\Omega}\{\hat{c}(x)\} - \nu\pi^2(k^2+l^2)\right)\|v_{k,l}\|^2_{L^2(\Omega)}.\end{aligned}$$

Hence, $\int_\Omega \hat{L}v_{k,l}(x)\,v_{k,l}(x)\,dx < 0$ for sufficiently large k^2 or l^2. Thus, under the condition (8.2), $\hat{L}$ is indefinite.

Now we describe our numerical tests. The operator L in (1.2) was taken with the coefficients

$$\begin{aligned}&a_{11}(x) = e^{x_1x_2}, \quad a_{12}(x) = \alpha/(1+x_1+x_2), \quad a_{22}(x) = e^{-x_1x_2}, \\ &b_1(x) = x_2e^{x_1x_2} + \beta_1\cos[\pi(x_1+x_2)], \quad b_2(x) = -x_1e^{-x_1x_2} + \beta_2\sin(2\pi x_1x_2), \\ &c(x) = \gamma[1 + 1/(1+x_1+x_2)],\end{aligned} \tag{8.5}$$

where α, β_1, β_2, and γ are parameters. In BVP (1.1), we set $f(x) = Lu(x)$ for $u(x) = e^{x_1+x_2}x_1x_2(1-x_1)(1-x_2)$.

We note that, for the coefficients given by (8.5) with $\alpha = \beta_1 = \beta_2 = 0$, we have $b_1 = (a_{11})_{x_1}$ and $b_2 = (a_{22})_{x_2}$. Therefore, in this case, the operator L in (1.2) can be written in the form of (8.1) with $\hat{a}_i(x) = a_{ii}(x)$, $i = 1, 2$, and $\hat{c}(x) = c(x)$. It follows from (8.5) with $\gamma \geq 0$ that

$$\min_{x\in\Omega}\{c(x)\} = (4/3)\gamma, \quad \max_{x\in\Omega}\{a_{11}(x),\, a_{22}(x)\} = e.$$

Hence, if $\gamma > (3/2)\pi^2 e \approx 40.243$, then the operator L is indefinite by (8.2).

In our numerical tests, we considered the case of $r = 3$, that is, V^0 is the space of Hermite bicubic splines on a uniform $N \times N$ partition of $\overline{\Omega}$ with the step size $h = 1/N$ (hence, $K_1 = K_2 = 2N$ and $K = 4N^2$). In this case, the standard basis for V^0 is defined as follows. For $k = 1, 2$, let $v_{k,j}, s_{k,j} \in V_k$, $j = 0, \dots, N$, be the "value function" and the "scaled slope function" associated with the node $x_{k,j}$ and defined respectively by

$$v_{k,j}(x_{k,i}) \;=\; \delta_{ij}, \quad [v_{k,j}]'(x_{k,i}) \;=\; 0, \quad i = 0, \dots, N,$$

Table 8.1

Iteration numbers for Algorithm 5.1. $\epsilon = 10^{-10}$ *in the stopping condition* (5.7). LP = *Laplace preconditioner,* VCP = *variable coefficient preconditioner.* 1 = *selfadjoint negative-definite* L, 2 = *selfadjoint indefinite* L, 3 = *nonselfadjoint* L, 4 = *general* L.

	1		2		3		4	
N	LP	VCP	LP	VCP	LP	VCP	LP	VCP
8	37	22	136	43	133	31	103	59
16	50	26	165	46	163	34	116	68
32	61	30	185	51	173	38	128	75
64	68	33	196	54	178	40	137	81
128	72	34	203	55	184	42	143	84

and

$$s_{k,j}(x_{k,i}) \;=\; 0, \qquad [s_{k,j}]'(x_{k,i}) \;=\; \delta_{ij}/h, \qquad i = 0, \ldots, N,$$

where δ_{ij} is the Kronecker delta. Then

$$\{\phi_{k,1}, \ldots, \phi_{k,K_k}\} \;=\; \{s_{k,0}, v_{k,1}, s_{k,1}, \ldots, v_{k,N_k-1}, s_{k,N_k-1}, s_{k,N_k}\}$$

is a basis for V_k^0, $k = 1, 2$, and basis functions for V^0 are given by (4.1).

The OSC problem (2.5) was solved by Algorithm 5.1 with the initial approximation $\vec{u}_0 = \vec{0}$ and the stopping condition (5.7) with $\epsilon = 10^{-10}$. A linear system with a preconditioner was solved by Algorithm 6.2. Since the partition is uniform, Step 1 of Algorithm 6.2 need not be performed, and we used FFTs to implement Steps 2 and 4.

We tested convergence properties of Algorithm 5.1 with two choices of the preconditioner $\tilde{A}$ of (5.1)–(5.2), the first corresponding to $\tilde{L} = \partial^2/\partial x_1^2 + \partial^2/\partial x_2^2$ and the second to the variable coefficient operator $\tilde{L}$ of (3.4)–(3.5) with

$$\begin{array}{llll} (8.6) & \tilde{a}_1(x_1) = a_{11}(0.5, 0.5), & \tilde{b}_1(x_1) = 0, & \tilde{c}_1(x_1) = 0, \\ & \tilde{a}_2(x_2) = a_{22}(0.5, x_2), & \tilde{b}_2(x_2) = b_2(0.5, x_2), & \tilde{c}_2(x_2) = c(0.5, x_2). \end{array}$$

We refer to these two preconditioners as the Laplacian preconditioner and the variable coefficient preconditioner, respectively. The following cases were tested:

1. selfadjoint negative-definite L ($\alpha = \beta_1 = \beta_2 = \gamma = 0$);
2. selfadjoint indefinite L ($\alpha = \beta_1 = \beta_2 = 0$ and $\gamma = 100$);
3. nonselfadjoint L ($\beta_2 = 100$ and $\alpha = \beta_1 = \gamma = 0$);
4. general L ($\alpha = 0.5$, $\beta_1 = 10$, $\beta_2 = \gamma = 50$).

The numerical results are shown in Table 8.1. We see that PCG with the variable coefficient preconditioner requires fewer iterations than PCG with the Laplacian preconditioner. Moreover, as N increases, the number of PCG iterations grows much slower with the variable coefficient preconditioner than with the Laplacian preconditioner.

In Figure 8.1, we present logarithmic plots of the relative residual curves. The vertical axis represents the values of

$$\log_{10}\left(\|f - Lu_{h,k}\|_{h,\rho}/\|f\|_{h,\rho}\right).$$

For both the Laplacian and the variable coefficient preconditioners, we observe monotone convergence. Curve LP1 shows that the Laplacian preconditioner works quite well for the selfadjoint negative-definite problem, but the Laplacian preconditioner

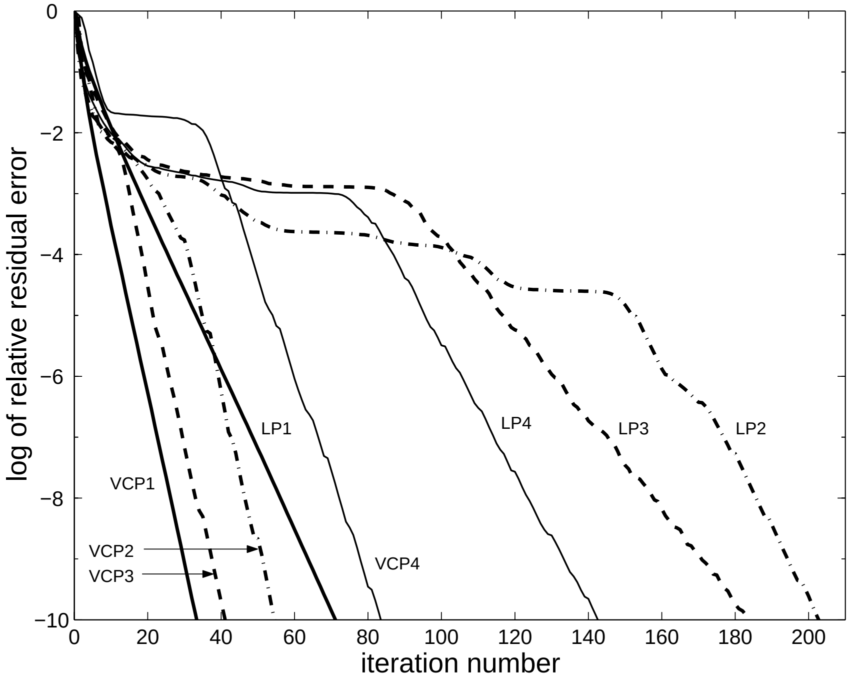


Fig. 8.1. *Logarithmic plots of relative residual curves for Algorithm* 5.1 ($N = 128$)*.* LP# = *Laplacian preconditioner,* VCP# = *variable coefficient preconditioner.* LP1, VCP1 = *selfadjoint negative-definite L;* LP2, VCP2 = *selfadjoint indefinite L;* LP3, VCP3 = *nonselfadjoint L;* LP4, VCP4 = *general L.*

is not so good for the indefinite, the nonselfadjoint, and the general problems (see the plateaus of curves LP2–LP4). The variable coefficient preconditioner does well in all cases (see curves VCP1–VCP4), although, for the general operator L, the curve VCP4 has a plateau as well. We see that curves VCP1–VCP3 are nearly parallel for large iteration numbers, which indicates that the convergence rates of the PCG with the variable coefficient preconditioner are about the same for different types of L. We observe the same behavior for the Laplacian preconditioner for the selfadjoint and the general problems (see curves LP1, LP2, and LP4). We see from curves LP1 and VCP1 that the convergence of PCG for the selfadjoint negative-definite problem is almost linear. We note that, for the first few iterations, the Laplacian preconditioner works well in all cases; for the general problem, even better than the variable coefficient preconditioner. However, for the larger iteration numbers, curve LP4 has a longer plateau and a smaller slope than curve VCP4.

Next we tested convergence properties of OSC for the general L. Let u_h be the computed OSC solution, and let $e_h = u - u_h$. For any v defined on the partition π_h, let $\|v\|_{\pi_h} = \max_{x\in\pi_h} |v(x)|$. We computed the maximal nodal errors $\|\partial^{(i,j)}e_h\|_{\pi_h}$ for $i, j = 0, 1$ and the Sobolev norms $\|e_h\|_{H^i(\Omega)}$ for $i = 0, 1, 2$, and determined approximate convergence orders using

$$\log_2\left(\|\partial^{(i,j)}e_h\|/\|\partial^{(i,j)}e_{h/2}\|\right),$$

Table 8.2
Maximal nodal errors and convergence orders.

N	e_h		$(e_h)_{x_1}$		$(e_h)_{x_2}$		$(e_h)_{x_1x_2}$	
4	4.64E–04		5.79E–03		1.34E–02		5.14E–02	
8	9.51E–05	2.29	9.13E–04	2.67	2.26E–03	2.57	9.10E–03	2.50
16	7.63E–07	6.96	6.38E–06	7.16	4.69E–05	5.59	3.79E–04	4.58
32	3.99E–08	4.26	4.47E–07	3.83	3.09E–06	3.92	3.50E–05	3.44
64	2.48E–09	4.01	2.73E–08	4.04	1.93E–07	4.00	3.17E–06	3.46
128	1.55E–10	4.00	1.68E–09	4.02	1.21E–08	4.00	3.21E–07	3.30

Table 8.3
Sobolev norm errors and convergence orders.

N	L^2		H^1		H^2	
4	2.11E–04		2.19E–03		4.63E–02	
8	3.56E–04	2.56	3.34E–04	2.72	9.31E–03	2.31
16	2.98E–07	6.90	1.63E–05	4.36	1.72E–03	2.44
32	1.83E–08	4.02	1.88E–06	3.11	3.94E–04	2.12
64	1.13E–09	4.02	2.31E–07	3.03	9.60E–05	2.04
128	7.01E–11	4.01	2.87E–08	3.01	2.38E–05	2.01

where $\|\cdot\|$ is $\|\cdot\|_{\pi_h}$ or $\|\cdot\|_{H^i(\Omega)}$. From the results presented in Table 8.2, we observe the expected order 4 for $\|e_h\|_{\pi_h}$ and the orders 4, 4, and 3 for $\|(e_h)_{x_1}\|_{\pi_h}$, $\|(e_h)_{x_2}\|_{\pi_h}$, and $\|(e_h)_{x_1x_2}\|_{\pi_h}$, respectively. The last three orders indicate a superconvergence property of OSC. The results in Table 8.3 demonstrate the expected optimal convergence orders for the Sobolev norms.

9. Conclusions. We have shown that PCG is an efficient algorithm for solving the OSC problem (2.5). The convergence analysis of this algorithm is carried out using an H^2 norm analysis of OSC. The convergence rate of PCG is independent of the partition step size h. The approach allows us to use a preconditioner associated with a nonselfadjoint or an indefinite separable operator $\tilde{L}$. A linear system with the preconditioner can be solved very efficiently using a matrix decomposition algorithm. On a uniform $N \times N$ partition, PCG with a tolerance ϵ requires $\mathrm{O}(N^2 \ln N |\ln \epsilon|)$ operations to obtain the Hermite bicubic spline solution of the OSC problem.

Our future work will involve the construction of nonselfadjoint or indefinite OSC domain decomposition and multilevel preconditioners for the OSC problem (2.5).

Acknowledgment. The authors wish to thank Graeme Fairweather for his assistance during the preparation of this paper.

REFERENCES

[1] A. Aitbayev, *Orthogonal Spline Collocation for Nonlinear Elliptic Dirichlet Problems*, Ph.D. thesis, University of Kentucky, Lexington, KY, 1998.

[2] E. Anderson, Z. Bai, C. Bischof, S. Blackford, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, and D. Sorensen, *LAPACK Users' Guide*, 3rd ed., SIAM, Philadelphia, 1999.

[3] U. Ascher, S. Pruess, and R. D. Russell, *On spline basis selection for solving differential equations*, SIAM J. Numer. Anal., 20 (1983), pp. 121–142.

[4] S. F. Ashby, T. A. Manteuffel, and P. E. Saylor, *A taxonomy for conjugate gradient methods*, SIAM J. Numer. Anal., 27 (1990), pp. 1542–1568.

[5] B. Bialecki, *An alternating direction implicit method for orthogonal spline collocation linear systems*, Numer. Math., 59 (1991), pp. 413–429.

[6] B. Bialecki, *Preconditioned Richardson and minimal residual iterative methods for piecewise Hermite bicubic orthogonal spline collocation equations*, SIAM J. Sci. Comput., 15 (1994), pp. 668–680.

[7] B. Bialecki, *Convergence analysis of orthogonal spline collocation for elliptic boundary value problems*, SIAM J. Numer. Anal., 35 (1998), pp. 617–631.

[8] B. Bialecki and G. Fairweather, *Matrix decomposition algorithms in orthogonal spline collocation for separable elliptic boundary value problems*, SIAM J. Sci. Comput., 16 (1995), pp. 330–347.

[9] B. Bialecki, G. Fairweather, and K. R. Bennett, *Fast direct solvers for piecewise Hermite bicubic orthogonal spline collocation equations*, SIAM J. Numer. Anal., 29 (1992), pp. 156–173.

[10] J. H. Bramble, Z. Leyk, and J. E. Pasciak, *Iterative schemes for nonsymmetric and indefinite elliptic boundary value problems*, Math. Comp., 60 (1993), pp. 1–22.

[11] J. H. Bramble and J. E. Pasciak, *Preconditioned iterative methods for nonselfadjoint or indefinite elliptic boundary value problems*, in Unification of Finite Element Methods, H. Kardestuncer, ed., North-Holland, Amsterdam, 1984, pp. 167–183.

[12] P. Ciarlet, *The Finite Element Method for Elliptic Problems*, North-Holland, Amsterdam, 1978.

[13] K. D. Cooper and P. M. Prenter, *Alternating direction collocation for separable elliptic partial differential equations*, SIAM J. Numer. Anal., 28 (1991), pp. 711–727.

[14] C. R. Crawford, *Reduction of a band-symmetric generalized eigenvalue problem*, Comm. ACM, 16 (1973), pp. 41–44.

[15] J. C. Diaz, G. Fairweather, and P. Keast, *Algorithm* 603 *COLROW and ARCECO: FORTRAN packages for solving certain almost block diagonal linear systems by modified alternate row and column elimination*, ACM Trans. Math. Software, 9 (1983), pp. 376–380.

[16] J. C. Diaz, G. Fairweather, and P. Keast, *FORTRAN packages for solving certain almost block diagonal linear systems by modified alternate row and column elimination*, ACM Trans. Math. Software, 9 (1983), pp. 358–375.

[17] Y. G. D'yakonov, *An iteration method for solving systems of finite difference equations*, Soviet Math. Dokl., 2 (1961), pp. 647–650.

[18] Y. G. D'yakonov, *The construction of iterative methods based on the use of spectrally equivalent operators*, USSR Comp. Math. and Math. Phys., 6 (1966), pp. 14–46.

[19] W. R. Dyksen, *Tensor product generalized ADI methods for separable elliptic problems*, SIAM J. Numer. Anal., 24 (1987), pp. 59–76.

[20] H. C. Elman and M. H. Schultz, *Preconditioning by fast direct methods for nonself-adjoint nonseparable elliptic equations*, SIAM J. Numer. Anal., 23 (1986), pp. 44–57.

[21] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed., The John Hopkins University Press, Baltimore, 1996.

[22] W. Hackbusch, *Iterative Solution of Large Sparse Systems of Equations*, Springer-Verlag, New York, 1994.

[23] A. Hadjidimos, E. N. Houstis, J. R. Rice, and E. Vavalis, *Modified successive overrelaxation (MSOR) and equivalent* 2*-step iterative methods for collocation matrices*, J. Comput. Appl. Math., 42 (1992), pp. 375–393.

[24] E. N. Houstis, W. F. Mitchell, and J. R. Rice, *Algorithms INTCOL and HERMCOL: Collocation on rectangular domains with bicubic Hermite polynomials*, ACM Trans. Math. Software, 11 (1985), pp. 416–418.

[25] E. N. Houstis, W. F. Mitchell, and J. R. Rice, *Collocation software for second-order partial differential equations*, ACM Trans. Math. Software, 11 (1985), pp. 379–412.

[26] L. Kantorovich, *Functional analysis and applied mathematics*, Usp. Mat. Nauk, 3 (1948), pp. 89–185 (in Russian).

[27] H. O. Kim, S. D. Kim, and Y. H. Lee, *Finite difference preconditioning cubic spline collocation method of elliptic equations*, Numer. Math., 77 (1997), pp. 83–103.

[28] S. D. Kim and S. V. Parter, *Preconditioning cubic spline collocation discretizations of elliptic equations*, Numer. Math., 72 (1995), pp. 39–72.

[29] Y.-L. Lai, A. Hadjidimos, E. N. Houstis, and J. R. Rice, *On the iterative solution of Hermite collocation equations*, SIAM J. Matrix Anal. Appl., 16 (1995), pp. 254–277.

[30] T. A. Manteuffel and S. V. Parter, *Preconditioning and boundary conditions*, SIAM J. Numer. Anal., 27 (1990), pp. 656–694.

[31] S. V. Parter and S.-P. Wong, *Preconditioning second-order elliptic operators: condition numbers and the distribution of the singular values*, J. Sci. Comput., 6 (1991), pp. 129–157.

[32] P. Percell and M. F. Wheeler, *A C^1 finite element collocation method for elliptic equations*,

SIAM J. Numer. Anal., 17 (1980), pp. 605–622.

[33] A. Quarteroni and A. Valli, *Domain Decomposition Methods for Partial Differential Equations*, Clarendon Press, Oxford, 1999.

[34] J. R. Rice and R. F. Boisvert, *Solving Elliptic Problems Using ELLPACK*, Springer-Verlag, New York, 1985.

[35] Y. Saad, *Iterative Methods for Sparse Linear Systems*, PWS Publishing, New York, 1996.

[36] B. F. Smith, P. E. Bjørstad, and W. D. Gropp, *Domain Decomposition: Parallel Multilevel Methods for Elliptic Partial Differential Equations*, Cambridge University Press, Cambridge, UK, 1996.

[37] W. Sun, *Block iterative algorithms for solving Hermite bicubic collocation equations*, SIAM J. Numer. Anal., 33 (1996), pp. 589–601.

[38] W. Sun, W. Huang, and R. D. Russell, *Finite difference preconditioning for solving orthogonal collocation equations for boundary value problems*, SIAM J. Numer. Anal., 33 (1996), pp. 2268–2285.

[39] P. N. Swarztrauber and R. Sweet, *Efficient Fortran Subprograms for the Solution of Elliptic Equations*, NCAR TN/IA-109, July 1975.

SIAM J. Numer. Anal.
Vol. 41, No. 2, pp. 605–623

MONOTONICITY-PRESERVING LINEAR MULTISTEP METHODS*

WILLEM HUNDSDORFER†, STEVEN J. RUUTH‡, AND RAYMOND J. SPITERI§

Abstract. In this paper we provide an analysis of monotonicity properties for linear multistep methods. These monotonicity properties include positivity and the diminishing of total variation. We also pay particular attention to related boundedness properties such as the total variation bounded (TVB) property. In the analysis the multistep methods are considered in combination with suitable starting procedures. This allows for monotonicity statements for classes of methods which are important and often used in practice but which were thus far not covered by theoretical results.

Key words. multistep schemes, monotonicity, positivity, TVD, TVB, strong stability

AMS subject classifications. 65L06, 65M06, 65M20

PII. S0036142902406326

1. Introduction.

In this paper we shall be concerned with preservation of certain monotonicity properties for systems of ordinary differential equations (ODEs) in $\mathbb{R}^m$, $m \geq 1$,

$$w'(t) = F(w(t)), \quad w(0) = w_0. \tag{1.1}$$

Specifically we are interested in the discrete preservation of these properties by numerical approximations $w_n \approx w(t_n)$, $t_n = n\Delta t$, generated by linear multistep methods. The multistep methods will be considered in combination with suitable starting procedures. Hence for a given problem (1.1) and step size Δt, we can regard the sequence $\{w_n\}_{n\geq 1}$ as being determined by the initial value w_0 only, just as for the true solution of (1.1).

There are a number of closely related monotonicity concepts. In this paper we shall mainly consider the property

$$\|w_n\| \ \leq \ \|w_0\| \qquad \text{for all } n \geq 1,\ w_0 \in \mathbb{R}^m, \tag{1.2}$$

where $\|\cdot\|$ is a given seminorm, such as the total variation over the components; see, for instance, (2.1). Related concepts, such as positivity and contractivity, are considered in the next section. Note that for one-step methods, such as Runge–Kutta methods, property (1.2) is equivalent to

$$\|w_n\| \ \leq \ \|w_{n-1}\| \qquad \text{for all } n \geq 1 \text{ with arbitrary } w_0 \in \mathbb{R}^m. \tag{1.3}$$

The relevant monotonicity property should hold for the ODE system (1.1) itself, of course. In the following we assume that there is a maximal step size $\Delta t_{FE} > 0$, under which (1.3) holds for the forward Euler method,

$$\|v + \Delta t F(v)\| \ \leq \ \|v\| \qquad \text{for all } 0 < \Delta t \leq \Delta t_{FE},\ v \in \mathbb{R}^m, \tag{1.4}$$

*Received by the editors April 23, 2002; accepted for publication (in revised form) October 16, 2002; published electronically May 6, 2003.

http://www.siam.org/journals/sinum/41-2/40632.html

†CWI, P.O. Box 94079, 1090 GB Amsterdam, The Netherlands (Willem.Hundsdorfer@cwi.nl).

‡Department of Mathematics, Simon Fraser University, Burnaby, BC, V5A 1S6 Canada (sruuth@sfu.ca). The work of this author was partially supported by a grant from NSERC Canada.

§Department of Computer Science, Dalhousie University, Halifax, NS, B3H 1W5 Canada (spiteri@cs.dal.ca). The work of this author was partially supported by a grant from NSERC Canada.

and we shall determine constants C_{LM} such that the property is valid for a multistep method with suitable starting procedure under the step size restriction

$$\Delta t \;\le\; C_{LM}\,\Delta t_{FE}. \tag{1.5}$$

In our analysis it is crucial to consider the linear multistep method in combination with suitable starting procedures. If a linear k-step method is considered with *arbitrary* starting vectors $w_1, \ldots, w_{k-1}$, in addition to w_0, then a natural generalization of (1.3) is

$$\|w_n\| \;\le\; \max_{0\le j\le k-1} \|w_j\| \qquad \text{for all } n \ge k,\;\; w_0, w_1, \ldots, w_{k-1} \in \mathbb{R}^m. \tag{1.6}$$

This is a common generalization for the purpose of analyzing multistep methods. However, there is no direct analogy with the analysis of (1.1), where the solution is determined by w_0 only. More importantly, it turns out that the insistence on arbitrary starting vectors severely limits the class of methods for which monotonicity can be demonstrated; for example, the familiar backward differentiation formulae (BDF) and Adams methods are then excluded. Consequently, relevant properties of many popular methods used in practice have not been covered yet by theoretical results.

In section 2 some related monotonicity properties are briefly discussed, together with existing results on multistep methods of the type (1.6) that were obtained in [2, 4, 12, 13, 15, 17]. Then in section 3 we analyze the time-step restrictions for (1.2) of both explicit and implicit two-step methods with various starting procedures. Apart from the monotonicity property (1.2), we also consider related boundedness properties $\|w_n\| \le M\|w_0\|$ with constant $M \ge 1$. In section 4 we extend our analysis to linear multistep methods of higher order. Finally, in section 5 we provide some numerical examples to illustrate our results.

2. Background.

2.1. Related monotonicity concepts. If the ODE system (1.1) is derived from a spatial discretization of a one-dimensional partial differential equation (PDE), then the components $w_j(t)$ of $w(t)$ will approximate the PDE solution $u(x,t)$ at grid points $x = x_j$ or surrounding cells. In that case $w_n = (w_j^n)$ contains the fully discrete method-of-lines approximation to $u(x_j, t_n)$. Consider for vectors $v = (v_j)$ the seminorm $\|v\| = \mathrm{TV}(v)$ given by

$$\mathrm{TV}(v) = \sum_j |v_j - v_{j-1}|. \tag{2.1}$$

We note that this is a seminorm, and not a norm, because $\mathrm{TV}(v)$ may vanish for $v \ne 0$; viz. v_j constant. For a pure initial-value PDE on an unbounded domain, the index j will run over all integers; but in general, boundary or periodicity conditions will result in a finite-dimensional ODE system. If (1.3) is valid, that is, $\mathrm{TV}(w_n) \le \mathrm{TV}(w_{n-1})$, $n \ge 1$, the scheme is called *total variation diminishing* (TVD). With property (1.2) we have

$$\mathrm{TV}(w_n) \;\le\; M\,\mathrm{TV}(w_0), \quad n \ge 1, \tag{2.2}$$

with constant $M = 1$. A scheme satisfying (2.2) with some $M \ge 1$ is called *total variation bounded* (TVB). Although this is formally weaker than TVD, conservative schemes with this boundedness property are known to converge to the correct entropy solutions for hyperbolic conservation laws; see, for instance, [5] for details. If (2.2) holds with $M = 1$, no spatial oscillations can be introduced during the time stepping; such spatial oscillations can be regarded as local overshoots and undershoots. Moreover, the scheme will then also be *monotonicity-preserving* in the sense that if the initial data w_j^0 is monotonically increasing or decreasing in j, then this will be preserved over time [14].

Another property related to avoiding undershoots is *positivity* [2], where it is required that

$$w_n \;\geq\; 0 \quad \text{whenever } w_0 \;\geq\; 0. \tag{2.3}$$

Here inequalities for vectors are to be interpreted componentwise. The corresponding requirement on the forward Euler method then reads

$$v + \Delta t F(v) \;\geq\; 0 \qquad \text{for all } 0 < \Delta t \leq \Delta t_{FE},\;\; v \geq 0. \tag{2.4}$$

Although we shall mainly focus on the relations (1.2), (1.4), results for positivity with (2.3), (2.4) can be derived in the same way. Positivity is often a natural requirement for general ODE systems, not necessarily semidiscrete PDEs, especially if the components $w_j(t)$ represent physical quantities such as mass or chemical concentrations that must be nonnegative by definition.

Further, related to (1.3) one can consider the *contractivity* property where the difference $\|\tilde{w}_n - w_n\|$ between any two sequences is required to be nonincreasing with increasing n [10, 17]. If we are dealing with a genuine norm, this is a strong stability requirement. Recently [4], methods satisfying (1.3) have also been called *strong stability-preserving*, and there is a tradition in the computational gas dynamics literature of referring to TVB, TVD, monotonicity preservation, and other nonlinear conditions as *nonlinear stability* [11]. However, properties like TVD or positivity are not directly related to the classical numerical concept of stability, which deals with growth between two sequences, one of which is viewed as a perturbation of the other. For linear problems we could well associate (1.3) with numerical stability, whereas for general nonlinear problems it may be viewed as a (strong) boundedness property.

Finally, we note that the term *monotonicity* appears in the numerical analysis literature for a variety of related concepts. For example, it is sometimes also used for properties like maximum principles ($\min_j w_j^0 \leq w_j^n \leq \max_j w_j^0$) or comparison principles ($\tilde{w}_0 \leq w_0$ implies $\tilde{w}_n \leq w_n$). In this paper we restrict ourselves to (1.2) and (2.3), but related properties could be studied in a similar way.

2.2. Monotonicity with arbitrary starting values. In this paper we mainly consider explicit linear multistep methods

$$w_n \;=\; \sum_{j=1}^{k} \big(a_j w_{n-j} + b_j \Delta t F(w_{n-j})\big), \quad n \geq k, \tag{2.5}$$

where starting vectors $w_0, w_1, \ldots, w_{k-1}$ are either given or computed by an appropriate starting procedure. Consistency of the method implies that

$$\sum_{j=1}^{k} a_j \;=\; 1. \tag{2.6}$$

Assume for the moment that all $a_j, b_j \geq 0$. By regarding the step (2.5) as a linear combination of scaled forward Euler steps,

$$w_n \;=\; \sum_{j=1}^{k} a_j \,\big(w_{n-j} + c_j \Delta t F(w_{n-j})\big), \qquad c_j = \frac{b_j}{a_j}, \tag{2.7}$$

it easily follows that

$$\|w_n\| \;\leq\; \max_{1\leq j\leq k} \|w_{n-j}\| \tag{2.8}$$

will hold under (1.4) with the step size restriction

$$\Delta t \;\leq\; K_{\scriptscriptstyle LM}\,\Delta t_{\scriptscriptstyle FE}, \qquad K_{\scriptscriptstyle LM} = \min_{1\leq j\leq k}\left(\frac{a_j}{b_j}\right) \quad \text{if } a_j, b_j \geq 0 \text{ for all } j. \tag{2.9}$$

By convention, terms of the form 0/0 should be omitted in the minimization, and if some coefficient a_j, b_j is negative, we leave $K_{\scriptscriptstyle LM}$ undefined. Note that (2.8) can also be formulated equivalently as (1.6).

This result, obtained with scaled forward Euler steps, is due to Shu [15], where it was formulated with total variations. Related results for multistep methods were derived by Bolley and Crouzeix [2] in terms of positivity for linear systems. Contractivity for linear systems was studied by Spijker [17] and Lenferink [12, 13]. The results in [2, 4, 13, 17] also cover implicit methods; we discuss implicit methods in some detail in section 3.

However, these results exclude many schemes that are useful in practice, and also may give unnecessary step size restrictions. This is due to the fact that (2.8) should hold with *arbitrary* initial vectors $w_0, w_1, \ldots, w_{k-1}$. As a simple example, consider the familiar BDF2 method applied to the trivial problem $w'(t) = 0$. Then

$$w_2 = \frac{4}{3}w_1 - \frac{1}{3}w_0.$$

It is obvious that one cannot have $w_2 \geq 0$ for arbitrary $w_0, w_1 \geq 0$. Likewise, it is not always possible to have $\|w_2\| \leq \|w_0\|$ whenever $\|w_1\| \leq \|w_0\|$. On the other hand, it is also clear that only the choice $w_1 = w_0$ makes sense for this trivial problem, in which case the inequality $\|w_2\| \leq \|w_0\|$ trivially holds. For this reason we shall analyze the monotonicity properties of multistep schemes with suitable starting procedures. As a result, schemes like BDF2 can be included in the theory.

Remark 2.1. To arrive at (2.9), the assumption $a_j \geq 0$ is necessary to have a convex combination of scaled forward Euler steps. The assumption $b_j \geq 0$ is then needed to ensure that the scaled step sizes $c_j\Delta t$ are nonnegative. As noted in [15, 16], the latter assumption can be avoided for discretizations of the conservation law

$$u_t + f(u)_x = 0,$$

by first applying the discretization in time followed by the spatial discretization (i.e., a transverse method-of-lines discretization), instead of starting with the semidiscrete system $w' = F(w)$. The only modification to our previous treatment is that if some $b_j < 0$, then $F(w_{n-j})$ in (2.5) should be replaced by $\tilde{F}(w_{n-j})$, where $w' = -\tilde{F}(w)$ is the semidiscretization of

$$u_t - f(u)_x = 0,$$

that is, of the equation with reversed time. Its realization in practice is simply a reversal of the upwind direction in the spatial discretization. Along with (1.4), we then also assume

$$\|v - \Delta t\tilde{F}(v)\| \;\le\; \|v\| \qquad \text{for all } 0 < \Delta t \le \Delta t_{\scriptscriptstyle FE},\;\; v \in \mathbb{R}^m, \tag{2.10}$$

and this counteracts the negativity of a_j/b_j in (2.7). Instead of (2.9), this modification will give the step size restriction

$$\Delta t \;\le\; \tilde{K}_{\scriptscriptstyle LM}\,\Delta t_{\scriptscriptstyle FE}, \qquad \tilde{K}_{\scriptscriptstyle LM} = \min_{1\le j\le k}\left(\frac{a_j}{|b_j|}\right) \quad \text{if } a_j \ge 0 \text{ for all } j \tag{2.11}$$

to achieve (2.8). □

3. Two-step methods.

3.1. Reformulations. In this section we derive monotonicity results for two-step methods, including some familiar implicit methods [1, 7]. The standard form is written as

$$w_n - b_0\Delta tF_n \;=\; a_1w_{n-1} + a_2w_{n-2} + b_1\Delta tF_{n-1} + b_2\Delta tF_{n-2}, \qquad n \ge 2, \tag{3.1}$$

where $F_{n-j} = F(w_{n-j})$. To obtain precise results, this recursion will be fully written out to include the starting values. Let $\theta \ge 0$ be a parameter to be specified later. Then the two-step recursion can be written in three-step form as

$$\begin{aligned} w_n - b_0\Delta tF_n \;=\; &(a_1-\theta)w_{n-1} + (b_1+\theta b_0)\Delta tF_{n-1} + (a_2+\theta a_1)w_{n-2} \\ &+ (b_2+\theta b_1)\Delta tF_{n-2} + \theta a_2w_{n-3} + \theta b_2\Delta tF_{n-3}, \qquad n \ge 3. \end{aligned}$$

Continuing this way, by subtracting and adding $\theta^j w_{n-j}$ and using (3.1), we arrive at

$$\begin{aligned} w_n - b_0\Delta tF_n \;=\; &(a_1-\theta)w_{n-1} + (b_1+\theta b_0)\Delta tF_{n-1} \\ &+ \sum_{j=2}^{n-2}\theta^{j-2}\big((a_2+\theta a_1-\theta^2)w_{n-j} + (b_2+\theta b_1+\theta^2 b_0)\Delta tF_{n-j}\big) \\ &+ \;\theta^{n-3}\big((a_2+\theta a_1)w_1 + (b_2+\theta b_1)\Delta tF_1 + \theta a_2w_0 + \theta b_2\Delta tF_0\big). \end{aligned} \tag{3.2}$$

This formula is valid for all $n \ge 3$, with empty sums naturally defined as zero. The reformulation (3.2) will be the basis for our derivations. To bound the last term in (3.2), together with w_1, w_2, appropriate starting procedures will be considered. Further, we shall determine θ so as to obtain nonnegative coefficients

$$a_1-\theta \ge 0, \quad a_2+\theta(a_1-\theta) \ge 0, \quad b_1+\theta b_0 \ge 0, \quad b_2+\theta(b_1+\theta b_0) \ge 0, \tag{3.3}$$

with optimal ratio $r(\theta)$ given by

$$r_1(\theta) = \frac{a_1-\theta}{b_1+\theta b_0}, \quad r_2(\theta) = \frac{a_2+\theta(a_1-\theta)}{b_2+\theta(b_1+\theta b_0)}, \quad r(\theta) = \min(r_1(\theta), r_2(\theta)). \tag{3.4}$$

As before, values 0/0 are ignored when taking the minimum.

3.2. Explicit second-order two-step methods. The maximal size of the threshold factor K_{LM} in (2.9) for explicit k-step methods of order p has been analyzed by Lenferink [12]. For explicit methods of order $p = 1$, we have $K_{LM} \leq 1$, a bound which is already attained by Euler's method. For explicit methods with $k \geq 2$, Lenferink showed that

$$K_{LM} \leq \frac{k-p}{k-1}. \tag{3.5}$$

Hence $K_{LM} > 0$ is not possible for second-order explicit two-step methods. By allowing $b_2 < 0$, Shu [15] obtained an explicit two-step method with $p = 2$, $\tilde{K}_{LM} = \frac{1}{2}$. However, this result is applicable only to semidiscretizations of conservation laws. Moreover, with $b_1 > 0$, $b_2 < 0$, both F_j and $\tilde{F}_j$ have to be calculated in the process, making the scheme twice as expensive computationally as the standard form (3.1).

Here we consider the monotonicity property (1.2) for schemes (3.1), and optimal threshold factors C_{LM} will be derived for second-order explicit two-step methods combined with suitable starting procedures. The main assumptions on the starting procedures will be

$$\begin{aligned} &\|w_1\| \leq M\|w_0\|, \qquad \|w_2\| = \|a_1 w_1 + b_1 \Delta t F_1 + a_2 w_0 + b_2 \Delta t F_0\| \leq M\|w_0\|, \\ &\|(a_2 + \theta a_1) w_1 + (b_2 + \theta b_1)\Delta t F_1 + \theta a_2 w_0 + \theta b_2 \Delta t F_0\| \leq (a_2 + \theta) M \|w_0\| \end{aligned} \tag{3.6}$$

for a given step size $\Delta t > 0$ and with $M = 1$. To derive weaker properties, such as (2.2), constants $M > 1$ will also be allowed.

Lemma 3.1. *Assume that* (1.4) *holds. Let $\theta \geq 0$ satisfy* (3.3), *and let $r(\theta)$ be given by* (3.4) *with $b_0 = 0$. Suppose that $\Delta t \leq r(\theta)\Delta t_{FE}$ and* (3.6) *holds with $M \geq 1$. Then*

$$\|w_n\| \leq M\|w_0\| \quad \textit{for all } n \geq 1. \tag{3.7}$$

Proof. From (3.2) and (3.3) we obtain

$$\|w_n\| \leq (a_1 - \theta)\|w_{n-1}\| + \sum_{j=2}^{n-2} \theta^{j-2}\big(a_2 + \theta a_1 - \theta^2\big)\|w_{n-j}\| + \theta^{n-3}(a_2 + \theta) M \|w_0\|.$$

By assumption, the lemma is valid for $n = 1, 2$. Since we have, in view of (2.6), the relation

$$(a_1 - \theta) + \sum_{j=2}^{n-2} \theta^{j-2}\big(a_2 + \theta a_1 - \theta^2\big) + \theta^{n-3}(a_2 + \theta) = 1, \quad n \geq 3,$$

the proof now follows easily by induction. □

To apply this lemma to specific methods, we shall determine θ to obtain an optimal constant

$$C^*_{LM} = \max\{\, r(\theta) : \theta \text{ satisfies } (3.3)\,\}. \tag{3.8}$$

This will give a step size restriction $\Delta t \leq C^*_{LM}\Delta t_{FE}$, which is intrinsic for the specific two-step method. The requirement (3.6) with $M = 1$ may give an additional restriction, say $\Delta t \leq C^0_{LM}\Delta t_{FE}$, depending on the starting procedure and the coefficients of the multistep method. For the combined scheme we then obtain the monotonicity property (1.2) under the step size restriction (1.5) with

$$C_{LM} \;=\; \min\{C^0_{LM}, C^*_{LM}\}. \tag{3.9}$$

The above derivation will be applied to explicit second-order two-step methods, which constitute a one-parameter family given by (3.1) with $b_0 = 0$ and

$$a_1 = 2-\xi, \quad a_2 = \xi - 1, \quad b_1 = 1 + \frac{1}{2}\xi, \quad b_2 = \frac{1}{2}\xi - 1. \tag{3.10}$$

The methods in this class are zero-stable if and only if $0 < \xi \le 2$, and we shall restrict ourselves to these parameter values. Examples of practical interest are the two-step Adams–Bashforth method ($\xi = 1$) and the extrapolated BDF2 method ($\xi = \frac{2}{3}$). With this class of second-order methods it follows, by a straightforward but somewhat tedious calculation, that optimality in (3.8) is attained by setting $b_2 + \theta b_1 = 0$, which gives

$$\theta \;=\; \frac{2-\xi}{2+\xi}, \qquad C^*_{LM} \;=\; \frac{2(1+\xi)(2-\xi)}{(2+\xi)^2}. \tag{3.11}$$

To obtain a complete bound (3.9), various starting procedures will be considered next.

Remark 3.2. In the remainder of this section we shall focus primarily on condition (3.6) with $M = 1$. For these results all coefficients in the occurring expressions will be required to be nonnegative. Consequently, results on positivity (2.3) with (2.4) can be derived under the same assumptions.

We shall also derive results with $M > 1$. These will only be qualitative. Precise bounds for M can be obtained by using

$$\max_{\Delta t \le C \Delta t_{FE}} \|v + \gamma \Delta t F(v)\| \;\le\; \max\{1, |2\gamma C - 1|\}\, \|v\| \tag{3.12}$$

for arbitrary $C > 0$, $\gamma \in \mathbb{R}$, and $v \in \mathbb{R}^m$. This relation is an obvious consequence of (1.4) if $0 \le \gamma C \le 1$. For values γC outside the interval $[0, 1]$, it follows by using in addition the implication $\Delta t_{FE}\, \|F(v)\| \le 2\, \|v\|$ from (1.4). □

3.2.1. Starting with the forward Euler method. The natural candidate to compute the starting vector w_1 for an explicit two-step method of order $p = 2$ is the forward Euler method

$$w_1 = w_0 + \Delta t F_0.$$

Of course, the forward Euler method itself is only first-order accurate; but because it is applied only once, the accuracy of the combined scheme will still be of order two.

With this starting procedure the first condition in (3.6) holds with $M = 1$ for $\Delta t \le \Delta t_{FE}$, of course. The second condition, $\|w_2\| \le M\|w_0\|$, can be written as

$$\|(a_1 - \tilde\theta) w_1 + b_1 \Delta t F_1 + (a_2 + \tilde\theta) w_0 + (b_2 + \tilde\theta)\Delta t F_0\| \le M \|w_0\|,$$

where an optimal $\tilde\theta$ should be selected. With $M = 1$ it is easily seen that the optimal value is $\tilde\theta = \frac{1}{2}(2-\xi)$, under which the inequality holds for all step sizes

$$\Delta t \;\le\; \frac{2-\xi}{2+\xi}\, \Delta t_{FE}. \tag{3.13}$$

With larger step sizes we will have a bound with $M > 1$. We note that this step size restriction for $M = 1$ is more stringent than $\Delta t \le C^*_{LM} \Delta t_{FE}$ for any $\xi \in (0, 2)$. Finally, with θ given by (3.11), the third condition in (3.6) reads

$$\|(a_2 + \theta)w_0 + (a_2 + \theta a_1 + \theta b_2)\Delta t F_0\| \;\le\; (a_2 + \theta)M\,\|w_0\|,$$

which can be written here as

$$\left\| w_0 + \frac{1}{2\xi}(3\xi - 2)\Delta t F_0 \right\| \;\le\; M\,\|w_0\|.$$

Hence $M = 1$ now requires

$$\Delta t \;\le\; \frac{2\xi}{3\xi - 2}\,\Delta t_{FE}, \qquad \xi \ge \frac{2}{3}. \tag{3.14}$$

If either the step size is allowed to be larger or $0 < \xi < \frac{2}{3}$, then we obtain a bound with $M > 1$ (see Remark 3.2), where it should be mentioned that we will have $M \sim \xi^{-1}$ for $\xi \downarrow 0$. We note that for $\xi \ge \frac{2}{3}$ the restriction (3.14) is less stringent than (3.13). The above results can be summarized as follows.

THEOREM 3.3. *Consider the explicit second-order two-step method* (3.1), (3.10), *and let w_1 be computed by the forward Euler method. Then the monotonicity property* (1.2) *will hold under* (1.4) *with the restriction*

$$\Delta t \;\le\; \frac{2 - \xi}{2 + \xi}\,\Delta t_{FE}\,, \qquad \frac{2}{3} \le \xi \le 2\,.$$

Under (1.4) *with the weaker restriction*

$$\Delta t \;\le\; \frac{2(1 + \xi)(2 - \xi)}{(2 + \xi)^2}\,\Delta t_{FE}, \qquad 0 < \xi \le 2,$$

the boundedness property (3.7) *will hold with $M \ge 1$.*

3.2.2. A modified two-step starting procedure. The use of the forward Euler method as starting procedure for the second-order two-step methods (3.1), (3.10) leads to a step size restriction for monotonicity that is more stringent than $\Delta t \le C^*_{LM} \Delta t_{FE}$. Similar restrictions were obtained with standard two-stage Runge–Kutta methods.

As an alternative starting procedure that can be used for semidiscrete conservation laws following Remark 2.1, we compute w_1 with the forward Euler method but use for the second step a modified scheme,

$$w_1 = w_0 + \Delta t F_0\,, \;\; w_2 = a_1 w_1 + a_2 w_0 + b_1 \Delta t F_1 + \alpha b_2 \Delta t F_0 + (1 - \alpha) b_2 \Delta t \tilde{F}_0, \tag{3.15}$$

where $\tilde{F}_0 = \tilde{F}(w_0)$ and $\alpha \in [0, 1]$ is to be determined later. We assume $\tilde{F}$ satisfies (2.10). We note that because $\tilde{F}$ is evaluated only once (for w_0), the computational costs will not increase significantly.

Consider the optimal θ value (3.11), for which $b_2 + \theta b_1 = 0$. With the modified second step, it follows that (3.2) with $b_0 = 0$ will change accordingly to

$$\begin{aligned} w_n = (a_1 - \theta)w_{n-1} + b_1 \Delta t F_{n-1} + \sum_{j=2}^{n-1} \theta^{j-2}(a_2 + \theta a_1 - \theta^2) w_{n-j} \\ + \theta^{n-2}\Big(\theta w_1 + a_2 w_0 + \alpha b_2 \Delta t F_0 + (1 - \alpha) b_2 \Delta t \tilde{F}_0\Big). \end{aligned}$$

If the last term can be bounded by $\theta^{n-2}(a_2+\theta)\|w_0\|$ and if $\|w_1\| \le \|w_0\|$, the result of Lemma 3.1 will remain valid with $M=1$. With the forward Euler approximation w_1, we thus get the requirement

$$\|(a_2+\theta)w_0 + (\alpha b_2+\theta)\Delta t F_0 + (1-\alpha)b_2 \Delta t \tilde{F}_0\| \;\le\; (a_2+\theta)\|w_0\| \tag{3.16}$$

for $\Delta t \le C^0_{\scriptscriptstyle LM}\Delta t_{\scriptscriptstyle FE}$, where an optimal $C^0_{\scriptscriptstyle LM} \in (0,1]$ will be selected by a favorable choice of the parameter α.

The contribution of F_0 in this inequality is minimized by taking $\alpha = -\theta/b_2 = 1/b_1$. By using (2.10), it then follows that $\|w_1\| \le \|w_0\|$ and (3.16) will be satisfied for $\Delta t \le C^0_{\scriptscriptstyle LM}\Delta t_{\scriptscriptstyle FE}$ with

$$C^0_{\scriptscriptstyle LM} = \min\left\{1, \frac{2\xi}{2-\xi}\right\}.$$

Taking $C_{\scriptscriptstyle LM} = \min\{C^0_{\scriptscriptstyle LM}, C^*_{\scriptscriptstyle LM}\}$, we can summarize this result as follows.

THEOREM 3.4. *Consider the explicit second-order two-step method* (3.1), (3.10) *for* $n \ge 3$*, and let* w_1, w_2 *be computed by* (3.15) *with* $\alpha = 1/b_1$*. Then the monotonicity property* (1.2) *will hold under* (1.4) *and* (2.10) *with the step size restriction*

$$\Delta t \;\le\; C_{\scriptscriptstyle LM}\,\Delta t_{\scriptscriptstyle FE}, \qquad C_{\scriptscriptstyle LM} \;=\; \begin{cases} \dfrac{2\xi}{2-\xi} & \text{if} \quad 0<\xi<\dfrac{1}{\frac12+\sqrt{2}}, \\[2ex] \dfrac{2(1+\xi)(2-\xi)}{(2+\xi)^2} & \text{if} \quad \dfrac{1}{\frac12+\sqrt{2}} \le \xi \le 2. \end{cases}$$

If the step size restriction in this theorem for $\xi < 1/(\frac12+\sqrt{2})$ is not satisfied, but still $\Delta t \le C^*_{\scriptscriptstyle LM}\Delta t_{\scriptscriptstyle FE}$, then we will have, as with other starting procedures, the boundedness property (3.7) with $M>1$.

A somewhat related result was obtained in [9] for the extrapolated BDF2 method ($\xi = \frac23$) in the so-called one-leg form. For that particular method it was demonstrated that (1.2) holds for all $\Delta t \le C^*_{\scriptscriptstyle LM}\Delta t_{\scriptscriptstyle FE}$, provided that an appropriate two-stage Runge–Kutta starting method is used. Also it was observed in [9] that this implies boundedness of $\|w_n\|$ for the standard multistep form (3.1) of the extrapolated BDF2 method if a special starting procedure is used. From the above we see that the boundedness property holds for any starting procedure and all $0<\xi\le 2$.

3.3. Implicit second-order two-step methods. In this section we consider the implicit two-step methods of order 2. These methods form a two-parameter family with coefficients

$$a_1 = 2-\xi, \;\; a_2 = \xi-1, \;\; b_0=\eta, \;\; b_1 = 1+\frac12\xi-2\eta, \;\; b_2 = \eta+\frac12\xi-1. \tag{3.17}$$

As for the explicit methods (3.10), we need $0<\xi\le 2$ for zero-stability. The methods are A-stable if and only if in addition $\eta \ge \frac12$. If $\eta=\frac12$, these methods are reducible to the trapezoidal rule, in the sense that if w_1 is calculated by the trapezoidal rule, then the whole sequence $\{w_n\}$ will satisfy the trapezoidal rule recurrence; see [3, 7]. Two interesting subclasses in (3.17) are $\xi=\frac23$, giving BDF2-type methods, and $\xi=1$, giving implicit two-step Adams methods.

In order to deal with implicit terms in (3.1), we shall use, in addition to (1.4),

$$\|v\| \;\le\; \|v-\Delta t F(v)\| \qquad \text{for all } \Delta t>0,\;\; v\in\mathbb{R}^m. \tag{3.18}$$

This can be interpreted as a condition on the implicit Euler method: $\|v_1\| \le \|v_0\|$ if $v_1 = v_0 + \Delta t F(v_1)$. It might appear that (3.18) should be imposed as an additional assumption to (1.4), but it is in fact a simple consequence. From $v_1 = v_0 + \Delta t F(v_1)$ it follows that

$$\left(1 + \frac{\Delta t}{\Delta t_{FE}}\right) v_1 \;=\; v_0 + \frac{\Delta t}{\Delta t_{FE}} \left(v_1 + \Delta t_{FE} F(v_1)\right) \,,$$

$$\left(1 + \frac{\Delta t}{\Delta t_{FE}}\right) \|v_1\| \;\le\; \|v_0\| + \frac{\Delta t}{\Delta t_{FE}} \|v_1\| \,,$$

and hence $\|v_1\| \le \|v_0\|$ for any $\Delta t > 0$. Thus under (1.4), the implicit Euler method gives the monotonicity property (1.2) without any step size restriction. However, this is only a first-order method, and for practical applications higher accuracy is often required. In the following we therefore concentrate on the class of second-order methods (3.17).

It was shown by Lenferink [13], in terms of contractivity for linear systems, that the threshold value K_{LM} in (2.9) is bounded by 2 for all two-step methods of order $p > 1$. The optimal $K_{LM} = 2$ is attained by the trapezoidal rule. In view of the results for explicit methods, one might hope that such severe restrictions could be circumvented in our formulation (1.2) with suitable starting procedures. Using (3.18), we can follow the derivation of Lemma 3.1, just as for explicit methods. Depending on the starting procedure, according to (3.6), this will give monotonicity (1.2) or boundedness (3.7) for $\Delta t \le r(\theta)\Delta t_{FE}$. We now consider the factors C^*_{LM} that are obtained by optimal values for θ in (3.8).

Determination of the optimal factors C^*_{LM} in analytical form is cumbersome, even if we restrict ourselves to subclasses such as $\xi = \frac{2}{3}$ and $\xi = 1$. On the other hand, numerically it is easy to compute the optimal θ values in (3.8). The corresponding threshold values C^*_{LM} are given in Figure 3.1 for $\xi = \frac{2}{3}, 1$ as function of η. We note that $C^*_{LM} = \frac{1}{2}$ for the familiar implicit BDF2 method ($\xi = \eta = \frac{2}{3}$).

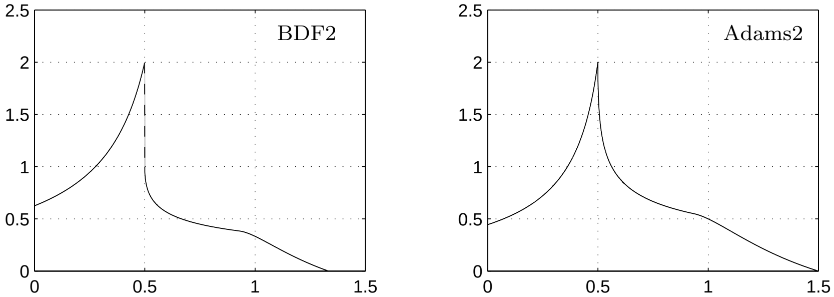


FIG. 3.1. *Threshold values C^*_{LM} versus $\eta \in [0, 1.5]$, with $\xi = \frac{2}{3}$ (left) and $\xi = 1$ (right).*

The results are rather disappointing. The largest numbers $C^*_{LM} = 2$ are found for the values $\eta = \frac{1}{2}$, and numerical verification shows that the same also holds with other choices of $\xi \in (0, 2]$. With $\eta = \frac{1}{2}$ the function r_2 in (3.4) has a removable singularity, which is related to the reducibility of the method, and this is the reason why the curves for $\eta < \frac{1}{2}$ and $\eta > \frac{1}{2}$ are very different, even having a discontinuity for the case $\xi = \frac{2}{3}$.

Of course, for a complete bound, suitable starting procedures such as the implicit Euler method should also be taken into account. However, the main result is that we obtain restrictions that are hardly better than those for explicit methods, and such restrictions have been confirmed in numerical experiments [8]. For practical purposes this means that the implicit schemes are not competitive with the explicit ones if monotonicity properties like (1.2) or (2.3) are crucial in an application. For this reason we shall restrict ourselves in the following section to explicit methods.

Remark 3.5. For the class of BDF2-type methods, threshold values for monotonicity were calculated analytically in [8] for linear, constant-coefficient problems $w'(t) = Aw(t)$. The curve in Figure 3.1 with $\xi = \frac{2}{3}$ almost coincides with the linear result for $\eta \lesssim 0.9$, but for larger η an extra condition sets in due to nonlinearity. For linear systems the restrictions in (3.3) can be relaxed by allowing negative values for $b_0 + \theta b_1$ and $b_2 + \theta(b_0 + \theta b_1)$.

The essential difference between linear and nonlinear results is most easily illustrated by the simple one-step method

$$w_n - \eta \Delta t F_n = w_{n-1} + (1-\eta)\Delta t F_{n-1}, \tag{3.19}$$

with parameter $\eta \geq 0$, whose stability function is given by

$$R(z) = (1-\eta z)^{-1}(1 + (1-\eta)z).$$

Let γ be the largest number such that R and all its derivatives are nonnegative on $[-\gamma, 0]$. It has been shown in [2, 17] (in terms of positivity and contractivity) that the monotonicity property (1.2) will hold under (1.4) for linear systems $w'(t) = Aw(t)$, provided that $\Delta t \leq \gamma \Delta t_{FE}$. Thus for linear problems we get the restriction

$$\Delta t \ \leq \ \gamma \, \Delta t_{FE}, \qquad \gamma = \begin{cases} (1-\eta)^{-1} & \text{if } \eta < 1, \\ \infty & \text{if } \eta \geq 1. \end{cases}$$

On the other hand, for nonlinear problems the optimal condition is seen to be

$$\Delta t \ \leq \ C \, \Delta t_{FE}, \qquad C = \begin{cases} (1-\eta)^{-1} & \text{if } \eta < 1, \\ \infty & \text{if } \eta = 1, \\ 0 & \text{if } \eta > 1. \end{cases}$$

Note that for $\eta > 1$ the coefficient in front of F_{n-1} becomes negative. For linear problems this can be counteracted by the implicit term, but for general nonlinear problems we need this coefficient to be nonnegative. □

4. Higher-order methods. This section contains derivations of boundedness results for various important higher-order explicit linear multistep methods: the extrapolated BDF and explicit Adams methods of order three or greater. To study the boundedness property (3.7), with $M \geq 1$, it is not necessary to specify the starting schemes: although the value of M may vary according to the choice of starting procedure, the boundedness property itself is independent of this choice.

4.1. Reformulations. We begin with a reformulation for the explicit multistep schemes. This is similar to formula (3.2) for two-step schemes, but to obtain proper step size restrictions different θ_j will be used in the various stages. To keep the presentation concise we give the reformulation here in detail only for three-step schemes. Consider (2.5) with $k = 3$. Then by subtracting and adding $\theta_1 \ldots \theta_j w_{n-j}$,

$j = 1, 2, \ldots, n-3$, substituting w_{n-j} in terms of $w_{n-j-1}, \ldots, w_{n-j-3}$, and collecting terms, it follows that w_n can be expressed as

$$w_n = \sum_{j=1}^{n-3} \big(\alpha_j w_{n-j} + \beta_j \Delta t F_{n-j}\big) + \sum_{i=0}^{2} \big(\rho_{i,n} w_i + \sigma_{i,n} \Delta t F_i\big), \tag{4.1}$$

where the coefficients α_j, β_j are given by

$$\alpha_1 = a_1 - \theta_1, \quad \alpha_2 = a_2 + a_1\theta_1 - \theta_1\theta_2, \quad \alpha_3 = a_3 + a_2\theta_1 + a_1\theta_1\theta_2 - \theta_1\theta_2\theta_3,$$

$$\alpha_j = \left(\prod_{k=1}^{j-3} \theta_k\right) (a_3 + a_2\theta_{j-2} + a_1\theta_{j-2}\theta_{j-1} - \theta_{j-2}\theta_{j-1}\theta_j), \quad j \geq 4,$$

$$\beta_1 = b_1, \quad \beta_2 = b_2 + b_1\theta_1, \quad \beta_3 = b_3 + b_2\theta_1 + b_1\theta_1\theta_2,$$

$$\beta_j = \left(\prod_{k=1}^{j-3} \theta_k\right) (b_3 + b_2\theta_{j-2} + b_1\theta_{j-2}\theta_{j-1}), \quad j \geq 4.$$

We shall take $\theta_i \geq 0$ such that

$$\alpha_j \geq 0, \quad \beta_j \geq 0 \qquad \text{for all } j \geq 1, \tag{4.2}$$

and we define

$$C^*_{LM} = \max_{\{\theta_i\}_{i\geq 1}} \min_{j\geq 1} \frac{\alpha_j}{\beta_j}. \tag{4.3}$$

Then it follows, similar to Lemma 3.1, that the boundedness property (3.7) will hold with $M \geq 1$ under the step size restriction $\Delta t \leq C^*_{LM} \Delta t_{FE}$. To obtain results on monotonicity (1.2), that is, $M = 1$, it is also necessary to study the coefficients $\rho_{i,n}$, $\sigma_{i,n}$ of the remainder term in (4.1) and to include specific starting procedures.

For k-step methods with $k \geq 4$ we can proceed similarly. In the above reformulation (4.1) we get the same expressions for α_1, α_2, α_3 and β_1, β_2, β_3; the other α_j, β_j will then involve more terms.

4.2. Boundedness and TVB. First we give the step size restrictions for boundedness and the related TVB property for the third-order extrapolated BDF3 scheme

$$w_n = \frac{18}{11} w_{n-1} - \frac{9}{11} w_{n-2} + \frac{2}{11} w_{n-3} + \frac{18}{11} \Delta t F_{n-1} - \frac{18}{11} \Delta t F_{n-2} + \frac{6}{11} \Delta t F_{n-3} \tag{4.4}$$

and the fourth-order extrapolated BDF4 scheme

$$\begin{aligned} w_n \ = \ & \frac{48}{25} w_{n-1} - \frac{36}{25} w_{n-2} + \frac{16}{25} w_{n-3} - \frac{3}{25} w_{n-4} \\ & + \frac{48}{25} \Delta t F_{n-1} - \frac{72}{25} \Delta t F_{n-2} + \frac{48}{25} \Delta t F_{n-3} - \frac{12}{25} \Delta t F_{n-4}. \end{aligned} \tag{4.5}$$

THEOREM 4.1. *Assume that* (1.4) *holds. The extrapolated BDF*3 (4.4) *scheme satisfies the boundedness property* (3.7) *with* $M \geq 1$, *provided* $\Delta t \leq \frac{7}{18} \Delta t_{FE}$. *For the extrapolated BDF*4 (4.5) *scheme the boundedness property will hold if* $\Delta t \leq \frac{7}{32} \Delta t_{FE}$. *These values* $\frac{7}{18}, \frac{7}{32}$ *are optimal within* (4.2), (4.3).

Proof. Consider (4.4). We first maximize α_1/β_1 over the constraint $\beta_2 \geq 0$ to get $\theta_1 = 1$. This also maximizes α_2/β_2; so next we maximize α_3/β_3 over the constraints

$\alpha_2 \geq 0, \beta_3 \geq 0$ to get $\theta_2 = \frac{2}{3}$. Maximizing α_4/β_4 over the constraints $\alpha_3 \geq 0, \beta_4 \geq 0$ gives $\theta_3 = \frac{1}{2}$. We can now set the remaining $\theta_j = \frac{1}{2}$, $j \geq 4$, because this choice is admissible in the sense of (4.2) and does not contribute to the step size restriction; indeed, $\frac{1}{2}$ is the value that minimizes the factor $(b_3 + b_2\theta_{j-2} + b_1\theta_{j-2}\theta_{j-1})$ in β_j, $j \geq 4$. Since

$$\min_{j\geq 1} \frac{\alpha_j}{\beta_j} \;=\; \min\left\{\frac{\alpha_1}{\beta_1}, \frac{\alpha_2}{\beta_2}, \frac{\alpha_3}{\beta_3}, \frac{\alpha_4}{\beta_4}, \frac{\alpha_5}{\beta_5}\right\} \;=\; \frac{\alpha_1}{\beta_1} \;=\; \frac{7}{18},$$

and we first optimized over α_1/β_1, we see that $C^*_{LM} = \frac{7}{18}$.

The result for the extrapolated BDF4 scheme follows in a similar manner, except that an admissible value for θ_3 is more difficult to find; for this we used a numerical search. We remark that in both cases the constant M will depend on the choice of starting procedure used. □

Another popular class of methods is formed by the explicit k-step Adams methods with order $p = k$. The third-order method is

$$w_n = w_{n-1} + \frac{23}{12}\Delta t F(w_{n-1}) - \frac{16}{12}\Delta t F(w_{n-2}) + \frac{5}{12}\Delta t F(w_{n-3}), \tag{4.6}$$

and the coefficients a_j, b_j for the higher-order methods can be found in [6], for example. For these methods the results are less favorable than for the extrapolated BDF schemes.

Theorem 4.2. *Assume that* (1.4) *holds. The explicit three-step Adams method* (4.6) *satisfies the boundedness property* (3.7) *with* $M \geq 1$*, provided* $\Delta t \leq \frac{84}{529}\Delta t_{FE}$. *For the explicit Adams methods with* $k \geq 4$*, no positive* C^*_{LM} *value in* (4.3) *exists.*

Proof. To have $\beta_2 \geq 0$ we need $\theta_1 \geq -b_2/b_1$, and consequently

$$\frac{\alpha_1}{\beta_1} \;\leq\; \frac{1 + b_2/b_1}{b_1} \;=\; \frac{1}{b_1^2}\big(b_1 + b_2\big).$$

If $k = 3$ we have $b_1 = \frac{23}{12}$ and $b_2 = -\frac{16}{12}$, leading to $C^*_{LM} \leq \frac{84}{529}$. Moreover, it follows by some simple calculations that this upper bound is attained by taking all $\theta_i = \frac{16}{23}$.

To show that we cannot have $C^*_{LM} > 0$ if $k \geq 4$, note that the k-step explicit Adams method may be written as

$$w_n \;=\; w_{n-1} + \Delta t \sum_{j=0}^{k-1} \gamma_j \nabla^j F_{n-1},$$

where ∇^j represent the usual backward differences and the γ_j are positive constants given in [6, section III.1]. A straightforward calculation for $k \geq 4$ shows that

$$b_1 \;=\; \sum_{j=0}^{k-1} \gamma_j \;=\; \frac{55}{24} + \sum_{j=4}^{k-1} \gamma_j, \qquad b_2 \;=\; -\sum_{j=0}^{k-1} j\gamma_j \;=\; -\frac{59}{24} - \sum_{j=4}^{k-1} j\gamma_j.$$

Therefore $b_1 + b_2 \leq \frac{-4}{24} < 0$, implying that $\alpha_1/\beta_1 < 0$. Hence the scheme does not possess a positive threshold value C^*_{LM}. □

Remark 4.3. Following the same lines, it is also straightforward to show that none of the explicit Nyström methods [6] has a positive threshold value C^*_{LM}. □

The generation of monotonicity results for high-order multistep schemes such as extrapolated BDF3 by means of optimized strong-stability-preserving Runge–Kutta starting procedures [4, 18] is part of our current research.

5. Numerical illustrations.

5.1. Linear positivity test. As a first numerical test we consider the positivity property (2.3) for the linear advection problem $u_t + u_x = 0$, $0 \le x \le 1$, with inflow boundary condition $u(0,t) = 0$ and initial mass $u(x,0)$ concentrated at the inflow boundary. The semidiscrete system is obtained with first-order upwind discretization in space and constant mesh width $\Delta x = 1/m$. The resulting linear ODE system in $\mathbb{R}^m$ is

$$w'(t) = Aw(t)\,, \qquad A = \frac{1}{\Delta x}\begin{pmatrix} -1 & & & \\ 1 & -1 & & \\ & \ddots & \ddots & \\ & & 1 & -1 \end{pmatrix}, \qquad w_0 = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}. \tag{5.1}$$

The dimension of the system is taken to be $m = 100$. For this system we determined experimentally the largest Courant number $\nu = \Delta t/\Delta x$ for which $w_n \ge 0$ is maintained up to $n = 1000$. We note that with the forward Euler method this will hold up to $\nu = 1$. Further we note that, by changing $w_j(t)$ in (5.1) into $1-w_j(t)$, identical results can be obtained with the condition $\|w_n\|_\infty \le \|w_0\|_\infty$.

First we consider the class of explicit two-step methods (3.10) with parameter values $\xi = j/20$, $j = 0, 1, \ldots, 40$. Along with the forward Euler method and the modified two-step procedure (3.15), we also consider the exact starting value $w_1 = \exp(\Delta t A)w_0$. The results are plotted in Figure 5.1, in combination with the theoretical values C^*_{LM} from (3.11).

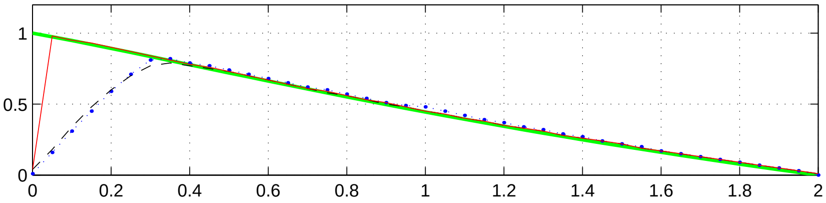


FIG. 5.1. *Positivity test for the explicit two-step methods* (3.10). *Courant numbers versus* $\xi \in [0,2]$*, starting with exact solution values [dots], forward Euler [solid line], and the modified two-step procedure* (3.15) *[dashed line]. The thick gray line is the* C^*_{LM}*-curve from* (3.11).

The influence of the starting values as given in the Theorems 3.3, 3.4 does not show up accurately in Figure 5.1. We note, however, that the test problem here is linear, whereas the theoretical results were obtained for nonlinear problems.

In a similar manner the behavior of the implicit two-step Adams and BDF-type methods has been tested. The results are shown in Figure 5.2. The starting value w_1 was computed with the implicit Euler method and with method (3.19), where the parameter η is the same as in (3.17); taking an exact starting value for w_1 did give results close to the latter starting procedure. For $\xi > \frac{1}{2}$ the results with implicit Euler and (3.19) also almost coincide.

The Courant numbers in Figure 5.2 are close to the theoretical bound C^*_{LM} in Figure 3.1 for $\eta \lesssim 0.9$. In particular, the different behavior for $\eta < \frac{1}{2}$ and $\eta > \frac{1}{2}$ shows up very clearly. Quantitatively, only the results with the BDF-type methods with $\eta \le \frac{1}{2}$ and the implicit Euler method as starting procedure are somewhat more favorable than the bound C^*_{LM}. The difference between the curves in Figure 3.1 for the larger η

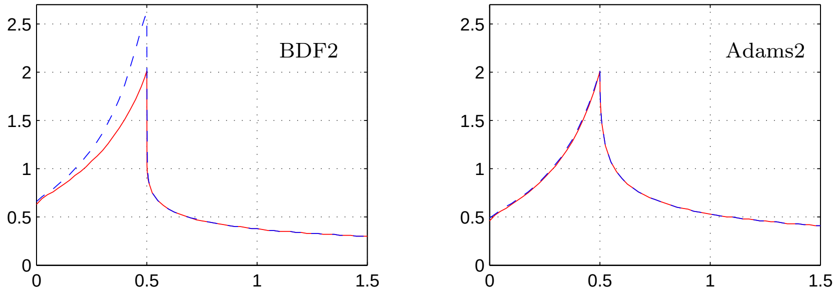


FIG. 5.2. *Positivity test for the implicit two-step methods* (3.17) *with* $\xi = \frac{2}{3}$ *[left] and* $\xi = 1$ *[right]. Courant numbers versus* $\eta \in [0, \frac{3}{2}]$*, starting with implicit Euler [dashed lines] and method* (3.19) *[solid lines].*

values is due to the fact that the values C^*_{LM} were obtained for nonlinear problems; see Remark 3.5. As noted previously, the rather small Courant numbers allowed with the implicit methods in practice mean that these implicit second-order two-step methods are not competitive with the explicit ones for problems where monotonicity is crucial.

In Table 5.1 the experimental positivity results are presented for the k-step extrapolated BDF schemes (eBDFk) and the k-step explicit Adams methods, which are also known as the Adams–Bashforth methods (ABk), $k = 3, 4$. Here we also list the theoretical bounds on the Courant numbers for these methods that were obtained in section 4. The experimental bounds were found with exact starting values and with high-order Runge–Kutta starting procedures, giving approximately the same values.

TABLE 5.1
Positivity test for higher-order methods. Experimental Courant numbers and theoretical bounds.

	eBDF3	AB3	eBDF4	AB4
Theoretical	$\frac{7}{18} \approx 0.39$	$\frac{84}{529} \approx 0.16$	$\frac{7}{32} \approx 0.22$	0
Experimental	0.43	0.23	0.30	0.11

5.2. Nonlinear accuracy test. To compare the explicit linear multistep methods for a nonlinear example, we consider the Burgers equation

$$u_t + (u^2)_x = 0, \quad 0 \le x \le 1,\ 0 \le t \le \frac{1}{4},$$

with periodic boundary conditions and initial profile $u(x, 0)$ given by the block function which equals 0 for $x \in (0, \frac{1}{2}]$ and 1 for $x \in (\frac{1}{2}, 1]$. For increasing time the solution $u(x, t)$ consists of a shock at $x = t$ and a rarefaction wave between $\frac{1}{2} \le x \le \frac{1}{2} + 2t$; see Figure 5.3.

Spatial discretization is performed with the flux-limited scheme of van Leer [19], which combines a second-order upwind-biased discretization (in smooth solution regions) with first-order upwind fluxes; see also [14, p. 180] and [8]. For this test, with $u \in [0, 1]$, it can be shown that the forward Euler method is TVD and positive for Courant numbers $\nu = 2\Delta t/\Delta x \le \frac{1}{2}$. However, to achieve a reasonable accuracy the Courant number should be taken significantly smaller than $\frac{1}{2}$, because otherwise the

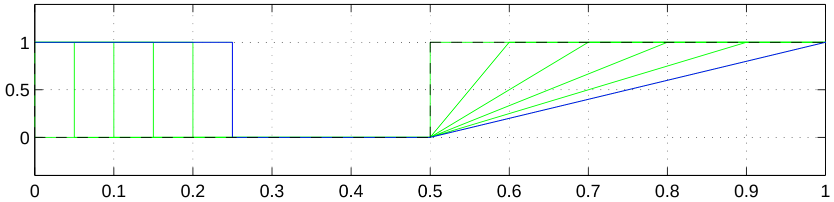


FIG. 5.3. *Solution of Burgers equation for $0 \leq x \leq 1$ at $t = 0$ [dashed] and $t = \frac{1}{4}$ [solid line]. The light gray lines indicate the time evolution.*

rarefaction wave suffers from compression due to linear instability of the forward Euler method with the second-order discretization. For Courant numbers in the range $[\frac{1}{2}, 1]$ the forward Euler method is no longer strictly TVD, but the oscillations are quite small. This can be understood heuristically by the observation that with the first-order upwind discretization the forward Euler method is TVD up to $\nu = 1$, and in nonsmooth regions, where monotonicity matters most, the flux-limited scheme becomes close to first-order upwinding.

The same observations apply to the multistep methods used in this test; the theoretical limits for monotonicity can be nearly doubled without introducing large temporal errors. Still, the theoretical predictions based on the threshold values C^*_{LM} show up when compared with forward Euler. The choice of starting procedures did have only minor significance; for the results presented here the first step was taken with the forward Euler method. In this test the discrete L_1-errors

$$\Delta x \sum_{j=1}^{m} |u(x_j, t_n) - w_j^n|, \qquad m\,\Delta x = 1,$$

were measured for different Courant numbers in the range $[0, 1]$ at time $t_n = \frac{1}{4}$. The test was performed on a fixed grid with mesh width $\Delta x = 10^{-2}$. The results for various second-order two-step methods (3.10) are shown in Figure 5.4.

The methods in Figure 5.4(a) are the extrapolated BDF2 scheme (eBDF2, $\xi = \frac{2}{3}$), the two-step Adams–Bashforth method (AB2, $\xi = 1$), and the second-order modified two-step method (Sh2) of Shu [15],

$$w_n = \frac{4}{5}w_{n-1} + \frac{1}{5}w_{n-2} + \frac{8}{5}\Delta t F(w_{n-1}) - \frac{2}{5}\Delta t \tilde{F}(w_{n-2}), \tag{5.2}$$

which is the modified form of (3.10), $\xi = \frac{6}{5}$, with threshold factor $\tilde{K}_{LM} = \frac{1}{2}$; see Remark 2.1. This scheme is more expensive in CPU time, and in this test it does not perform as well as the other two, of which the extrapolated BDF2 scheme has a slight advantage over the explicit two-step Adams method.

In Figure 5.4(b) the results are given for the methods (3.10) with $\xi = \frac{1}{5}, \frac{6}{5}, \frac{9}{5}$. For comparison, results for the forward Euler method are also included. As predicted by the bound C^*_{LM} of (3.11), the method with $\xi = \frac{9}{5}$ can only be used with small Courant numbers. The method with $\xi = \frac{1}{5}$ does provide results for larger Courant numbers, but its accuracy deteriorates for large ν. The results with $\xi = \frac{6}{5}$ are intermediate, where it should be noted that this method is competitive with the more expensive modified method (5.2), which is based on the same parameter choice.

In Figure 5.4 we have also indicated the spatial errors of the flux-limited van Leer discretization with $\Delta x = 10^{-2}$; that is, the L_1 difference between a semidiscrete

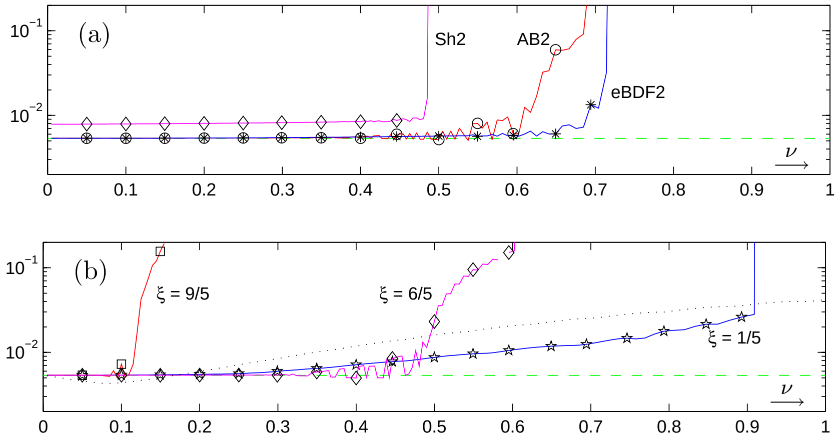


FIG. 5.4. *Burgers equation, L_1-errors versus Courant numbers ν for explicit two-step methods* (3.10). (a) *eBDF*2$[\xi = \frac{2}{3}]$, *AB*2$[\xi = 1]$, *and Sh*2$[\tilde{\xi} = \frac{6}{5}]$; (b) $\xi = \frac{1}{5}$, $\xi = \frac{6}{5}$, *and* $\xi = \frac{9}{5}$, *together with forward Euler results [dotted line]. The light dashed horizontal line indicates the spatial error.*

solution and PDE solution at $t = \frac{1}{4}$ on this spatial grid. The modified scheme (5.2) gives larger errors for $\nu \to 0$. This is due to the use of $\tilde{F}$, which introduces some extra numerical dissipation, in particular at the bottom and top of the rarefaction wave. With most of the methods the L_1-errors show oscillations as a function of ν before becoming unbounded. This is an onset of instability, due to spatial oscillations at the top of the shock or rarefaction wave.

Method (3.10) with $\xi = \frac{1}{5}$ could be used with relatively large Courant numbers ν without becoming unstable, but for the larger values ν the results are no longer accurate, due to compression of the rarefaction wave. With the forward Euler method this compression is much more pronounced. Time-stepping methods with high order will mostly be beneficial for smooth solutions. The present test is primarily intended to show the relevance of monotonicity. This should also be kept in mind with the results for the third-order methods below.

The fact that the starting procedures did not matter significantly in this test is somewhat more surprising than with the previous linear example. In the derivation of our theoretical results for nonlinear problems no relation at all was assumed between terms like $F(w_n)$ and $F(w_{n-1})$. For grid points x_j adjacent to the shock the spatial discretization becomes close to the first-order upwind scheme and elsewhere we will have $F_j(w_n) = F_j(w_{n-1}) + \mathcal{O}(\Delta t)$. It is not clear, however, how such arguments could be used in a rigorous mathematical fashion.

In the same way some three-step methods were tested. The results are shown in Figure 5.5. Here we selected the extrapolated BDF3 scheme (eBDF3) and the three-step Adams–Bashforth method (AB3). Also included are the results for the second-order three-step method

$$w_n = \frac{3}{4} w_{n-1} + \frac{1}{4} w_{n-3} + \frac{3}{2} \Delta t F(w_{n-1}) \tag{5.3}$$

of Shu [15] with threshold value $K_{LM} = \frac{1}{2}$, which is optimal among the three-step methods of order 2; see also Lenferink [12]. In the figure this method is indicated as Sh2,3. Since this is a second-order method, comparison with AB2 or eBDF2 is actually more appropriate. As expected from the theoretical bounds, the eBDF3 scheme does

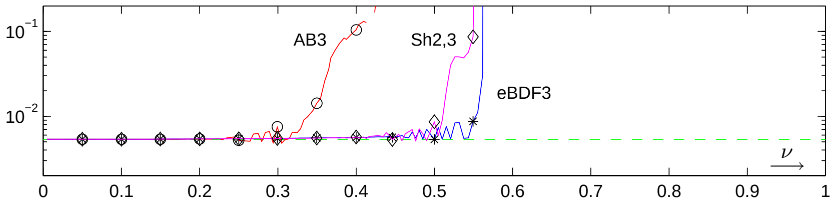


FIG. 5.5. *Burgers equation, L_1-errors versus Courant numbers ν for the explicit three-step methods eBDF*3*, AB*3*, and Sh*2, 3*. The light dashed horizontal line indicates the spatial error.*

perform better than the AB3 method. Also with these three-step methods, starting procedures turned out not to be very influential. Here a standard two-stage second-order Runge–Kutta method was used.

In summary we can say that the theoretical step-size restrictions of section 3 for the monotonicity property (1.2) are probably somewhat pessimistic, but the step-size restrictions under which the more general boundedness property (3.7) could be proved give a good indication of the applicability of the various methods.

6. Summary and conclusions.

We have shown that inclusion of starting procedures in multistep schemes allows for statements on monotonicity (1.2) and boundedness (3.7) with classes of methods that are important in practice (such as the Adams and BDF-type methods). We find that the standard second-order two-step methods AB2 (second-order Adams–Bashforth),

$$w_n = w_{n-1} + \frac{3}{2}\Delta t F(w_{n-1}) - \frac{1}{2}\Delta t F(w_{n-2}),$$

and eBDF2 (second-order extrapolated BDF),

$$w_n = \frac{4}{3}w_{n-1} - \frac{1}{3}w_{n-2} + \frac{4}{3}\Delta t F(w_{n-1}) - \frac{2}{3}\Delta t F(w_{n-2}),$$

have more relaxed time-stepping restrictions than schemes with positive coefficients when monotonicity-preservation is required. Similarly, the well-known higher-order schemes AB3 (4.6), eBDF3 (4.4), and eBDF4 (4.5) have more relaxed time-stepping restrictions than schemes with positive coefficients when related boundedness properties are required. Numerical tests confirm the results of these theoretical studies: these standard methods performed better than specially constructed methods with positive coefficients. We particularly recommend the well-known extrapolated BDF schemes.

Finally we have found that implicit second-order two-step schemes are not competitive with the explicit ones when monotonicity properties are crucial, generalizing one of the results in [8] for the two-step BDF-type schemes. The restrictions on the step sizes for monotonicity and boundedness with the implicit schemes were shown to be only marginally better than for the explicit schemes, and in practical computations this will not be enough to justify the increase in computational work with the implicit schemes.

REFERENCES

[1] U.M. Ascher, S.J. Ruuth, and B.T.R. Wetton, *Implicit-explicit methods for time-dependent partial differential equations*, SIAM J. Numer. Anal., 32 (1995), pp. 797–823.

[2] C. Bolley and M. Crouzeix, *Conservation de la positivité lors de la discrétisation des problèmes d'évolution paraboliques*, RAIRO Anal. Numer., 12 (1978), pp. 237–245.

[3] G. Dahlquist, *Error analysis for a class of methods for stiff nonlinear initial value problems*, in Proceedings of the Dundee Conference 1975, Lectures Notes in Mathematics 506, G.A. Watson, ed., Springer-Verlag, Berlin, 1976, pp. 60–74.

[4] S. Gottlieb, C.-W. Shu, and E. Tadmor, *Strong stability-preserving high-order time discretization methods*, SIAM Rev., 43 (2001), pp. 89–112.

[5] A. Harten, *On a class of high resolution total-variation-stable finite-difference schemes*, SIAM J. Numer. Anal., 21 (1984), pp. 1–23.

[6] E. Hairer, S.P. Nørsett, and G. Wanner, *Solving Ordinary Differential Equations* I—*Nonstiff Problems*, 2nd ed., Springer Ser. Comput. Math. 8, Springer-Verlag, New York, 1993.

[7] E. Hairer and G. Wanner, *Solving Ordinary Differential Equations* II—*Stiff and Differential-Algebraic Problems*, 2nd ed., Springer Ser. Comput. Math. 14, Springer, New York, 1996.

[8] W. Hundsdorfer, *Partially implicit BDF*2 *blends for convection dominated flows*, SIAM J. Numer. Anal., 38 (2001), pp. 1763–1783.

[9] W. Hundsdorfer and J. Jaffré, *Implicit-explicit time stepping with spatial discontinuous finite elements*, Appl. Num. Math., 45 (2003), pp. 231–254.

[10] J.F.B.M. Kraaijevanger, *Contractivity of Runge–Kutta methods*, BIT, 31 (1991), pp. 482–528.

[11] C.B. Laney, *Computational Gasdynamics*, Cambridge University Press, London, Cambridge, 1998.

[12] H.W.J. Lenferink, *Contractivity preserving explicit linear multistep methods*, Numer. Math., 55 (1989), pp. 213–223.

[13] H.W.J. Lenferink, *Contractivity preserving implicit linear multistep methods*, Math. Comp., 56 (1991), pp. 177–199.

[14] R.J. LeVeque, *Numerical Methods for Conservation Laws*, Lectures Math., ETH Zürich, Birkhäuser Verlag, Basel, 1992.

[15] C.-W. Shu, *Total-variation-diminishing time discretizations*, SIAM J. Sci. Stat. Comp., 9 (1988), pp. 1073–1084.

[16] C.-W. Shu and S. Osher, *Efficient implementation of essentially non-oscillatory shock-capturing schemes*, J. Comput. Phys., 77 (1988), pp. 439–471.

[17] M.N. Spijker, *Contractivity in the numerical solution of initial value problems*, Numer. Math., 42 (1983), pp. 271–290.

[18] R.J. Spiteri and S.J. Ruuth, *A new class of optimal high-order strong-stability-preserving time discretization methods*, SIAM J. Numer. Anal., 40 (2002), pp. 469–491.

[19] B. van Leer, *Towards the ultimate conservative difference scheme* II. *Monotonicity and conservation combined in a second order scheme*, J. Comput. Phys., 14 (1974), pp. 361–370.

SIAM J. Numer. Anal.
Vol. 41, No. 2, pp. 624–640

P_1-NONCONFORMING QUADRILATERAL FINITE ELEMENT METHODS FOR SECOND-ORDER ELLIPTIC PROBLEMS*

CHUNJAE PARK† AND DONGWOO SHEEN‡

Abstract. A P_1-nonconforming quadrilateral finite element is introduced for second-order elliptic problems in two dimensions. Unlike the usual quadrilateral nonconforming finite elements, which contain quadratic polynomials or polynomials of degree greater than 2, our element consists of only piecewise linear polynomials that are continuous at the midpoints of edges. One of the benefits of using our element is convenience in using rectangular or quadrilateral meshes with the least degrees of freedom among the nonconforming quadrilateral elements. An optimal rate of convergence is obtained. Also a nonparametric reference scheme is introduced in order to systematically compute stiffness and mass matrices on each quadrilateral. An extension of the P_1-nonconforming element to three dimensions is also given. Finally, several numerical results are reported to confirm the effective nature of the proposed new element.

Key words. nonconforming finite elements, quadrilateral, elliptic problems

AMS subject classifications. 65N30, 65N12, 65N15

PII. S0036142902404923

1. Introduction.

We are concerned with nonconforming finite element methods for second-order elliptic problems. Nonconforming elements have been used effectively especially in fluid and solid mechanics due to their stability. Recently, these elements have attracted increasing attention from scientists and engineers in more wide areas, as this type of element is potentially useful in parallel computing.

The use of finite elements for Stokes problems, which is fundamental in fluid mechanics, usually requires the discrete Babuška–Brezzi condition (inf-sup condition) to be satisfied by the velocity and pressure variables, generally set in the mixed finite element formulation; for instance, the standard P_1-P_0 pair for triangular decompositions or the Q_1-P_0 pair for quadrilateral decompositions of the computational domain lead to checkerboard solutions for pressure. However, if the nonconforming elements introduced in [3, 8, 15, 5] are used to approximate the velocity part instead of the usual P_1 or Q_1 elements, the Babuška–Brezzi condition is easily satisfied, and thus stable solutions are obtained. Nonconforming finite element methods have been proved to be effective for several parameter dependent elasticity problems in a stable fashion such that the methods converge independently of the Lamé parameters, while standard conforming methods fail to converge as the parameters tend to a locking limit; see [2, 12, 13].

Moreover, in view of domain decomposition methods, the use of nonconforming elements facilitates the exchange of information across each subdomain and provides spectral radius estimates for the iterative domain decomposition operator [9].

The nonconforming simplicial finite element space of lowest degree introduced by Crouzeix and Raviart [8] is identical to the corresponding conforming one (that

*Received by the editors March 29, 2002; accepted for publication (in revised form) October 21, 2002; published electronically May 6, 2003. This research was supported in part by the Korea Research Foundation (KRF-2000-DS0004).

http://www.siam.org/journals/sinum/41-2/40492.html

†Institute for Pure and Applied Mathematics, University of California, Los Angeles, CA 90095. Current address: Impedance Imaging Research Center, Kyunghee University, Seoul, Korea (cpark@nasc.snu.ac.kr).

‡Department of Mathematics, Seoul National University, Seoul 151–747, Korea (sheen@snu.ac.kr).

is, P_1 in both cases), and thus it is rather simple to understand. Although the triangular meshes are popular to use, in many cases one wishes to use quadrilateral meshes with appropriate elements instead, when the problem geometry is of quadrilateral nature, especially in three dimensions. Concerning rectangular nonconforming elements, Han [11] introduced a rectangular element with local degrees of freedom being five, and Rannacher and Turek [15] introduced the rotated Q_1 nonconforming elements of two types: the first set of local degrees of freedom consists of the four values at the midpoints of the edges, while the second one is composed of the four average values over the edges. Recently, new nonconforming elements, which use only the four values at the midpoints of the edges as degrees of freedom, have been announced by Douglas et al. [9], who in a sense combined and improved the two types of local degrees of freedom for rotated Q_1 elements, using high-order polynomials with the degrees of freedom still being four. These elements were successfully applied to solve Navier–Stokes problems by Cai, Douglas, and Ye [5]. A recent observation by Arnold, Boffi, and Falk [1] implies that where the rectangular elements are applied to truly quadrilateral meshes, the optimality in convergence will be lost. Thus for the truly quadrilateral case, an extra element should be added [4], with the local degrees of freedom being five; the extra degrees of freedom can be eliminated easily at an element level since they are essentially bubble functions.

The purpose of this paper is to introduce P_1-nonconforming finite element spaces on quadrilateral meshes which have the lowest degrees of freedom. The motivation for our new element comes from the observation that any P_1 function on a quadrilateral can be uniquely determined at any three of the four midpoints of edges.

The degrees of freedom for our P_1-nonconforming quadrilateral element are about half of those for the other rectangular nonconforming elements, and about a third of those for the P_1 triangular nonconforming space on the mesh with each quadrilateral being divided into two triangles. Indeed, our P_1-nonconforming quadrilateral element space turns out to be a subspace of P_1-nonconforming triangular element spaces by dividing each quadrilateral into two triangles.

In the Q_1-conforming quadrilateral element case, it is convenient to use a fixed reference rectangle and basis from which, corresponding to each quadrilateral, a bilinear transformation can be used to calculate stiffness and mass matrices by pulling back to the reference rectangle without losing the order of convergence. However, as mentioned above, such a reference system does not guarantee optimal convergence any more with existing nonconforming quadrilateral elements with only four degrees of freedom [1]. We present a nonparametric reference scheme in section 4, which provides an efficient way of calculating the stiffness and mass matrices from a reference rectangle without losing the order of convergence.

As discussed earlier, one of the motivations for seeking the P_1-nonconforming quadrilateral element space is to try to use it for the approximation of the velocities and the P_0 space for the pressure as in [8, 11, 15, 4]. However, we remark that with this combination the discrete inf-sup condition is not fulfilled, as there are only three degrees of freedom for the normal components at the midpoints of a quadrilateral. But the current element works well as a locking-free element for elasticity problems [14].

The organization of the paper is as follows. In the next section we present two P_1-nonconforming element spaces on quadrilateral meshes. Then section 3 describes an interpolation operator and also deals with a brief analysis of convergence in the cases of Dirichlet and Robin problems. Then a nonparametric reference scheme is introduced in section 4. The analysis carried out in the current paper has a counterpart in three dimensions: P_1-nonconforming hexahedral finite elements will be briefly discussed in section 5, detailed analyses being treated in a forthcoming paper. Finally, numerical examples are illustrated in section 6.

2. The P_1-nonconforming element on quadrilateral meshes.

2.1. The P_1-nonconforming quadrilateral element.

Let Ω be a simply connected polygonal domain in $\mathbb{R}^2$ with boundary Γ. Let $(\mathcal{T}_h)_{h>0}$ be a *regular* family of decompositions (or triangulations) of Ω into convex quadrilaterals, where $h = \max_{Q\in\mathcal{T}_h} h_Q$ with $h_Q = \operatorname{diam}(Q)$. For the standard definition of regular decomposition, we refer to [10]. Henceforth, in this paper, a quadrilateral will be implicitly assumed to be convex.

For a general quadrilateral Q, denote by $v_j, 1 \le j \le 4$, its four vertices and by $m_j, 1 \le j \le 4$, the midpoints of edges of Q such that $m_j = \frac{v_{j-1}+v_j}{2}, 1 \le j \le 4$, with the identification $v_0 = v_4$. Let $P_1(Q) = \operatorname{Span}\{1, x, y\}$. The following lemmas are trivial but useful in what follows.

Lemma 2.1. *If $u \in P_1(Q)$, then $u(m_1) + u(m_3) = u(m_2) + u(m_4)$. Conversely, if u_j is a given value at m_j, for $1 \le j \le 4$, satisfying $u_1 + u_3 = u_2 + u_4$, then there is a unique $u \in P_1(Q)$ such that $u(m_j) = u_j, 1 \le j \le 4$.*

Proof. The first half is trivial:

$$\begin{aligned} u(m_1) + u(m_3) &= \frac{u(v_4) + u(v_1)}{2} + \frac{u(v_2) + u(v_3)}{2} \\ &= \frac{u(v_1) + u(v_2)}{2} + \frac{u(v_3) + u(v_4)}{2} = u(m_2) + u(m_4). \end{aligned}$$

For the latter half, suppose that $u_1 + u_3 = u_2 + u_4$ and then choose a $u \in P_1(Q)$ such that $u(m_j) = u_j, j = 1, 2, 3$. Then by the first half of the lemma, $u_1 + u_3 = u_2 + u(m_4)$, which implies that $u(m_4) = u_4$, so that $u(m_j) = u_j, 1 \le j \le 4$. Uniqueness is obvious. □

Lemma 2.2. *For $1 \le j \le 4$, let $\widehat{\varphi}_j \in P_1(Q)$ be defined such that*

$$\widehat{\varphi}_j(m_k) = \begin{cases} 1, & k = j, j+1 \mod 4, \\ 0, & \textit{otherwise}. \end{cases}$$

Then $\operatorname{Span}\{\widehat{\varphi}_1, \widehat{\varphi}_2, \widehat{\varphi}_3, \widehat{\varphi}_4\} = P_1(Q)$. Indeed, any three of $\widehat{\varphi}_1, \widehat{\varphi}_2, \widehat{\varphi}_3, \widehat{\varphi}_4$ span $P_1(Q)$.

Proof. Clearly $\operatorname{Span}\{\widehat{\varphi}_1, \widehat{\varphi}_2, \widehat{\varphi}_3, \widehat{\varphi}_4\} \subset P_1(Q)$. To show the other direction of inclusion, it suffices to show that $P_1(Q) \subset \operatorname{Span}\{\widehat{\varphi}_1, \widehat{\varphi}_2, \widehat{\varphi}_3\}$; then rotational symmetry would imply that any three of $\widehat{\varphi}_1, \widehat{\varphi}_2, \widehat{\varphi}_3, \widehat{\varphi}_4$ span $P_1(Q)$. Let $u \in P_1(Q)$ be arbitrary. Set

$$\psi = u(m_1)\widehat{\varphi}_1 + [u(m_2) - u(m_1)]\widehat{\varphi}_2 + [u(m_3) - u(m_2) + u(m_1)]\widehat{\varphi}_3.$$

Then it is immediate to see that $\psi(m_j) = u(m_j), j = 1, 2, 3$. Lemma 2.1 implies that $\psi(m_4) = u(m_4)$ and therefore ψ is identical to u. This proves that $P_1(Q) \subset \operatorname{Span}\{\widehat{\varphi}_1, \widehat{\varphi}_2, \widehat{\varphi}_3\}$. □

Given a decomposition $\mathcal{T}_h$ of Ω into quadrilaterals, let N_Q, N_V, and N_E denote

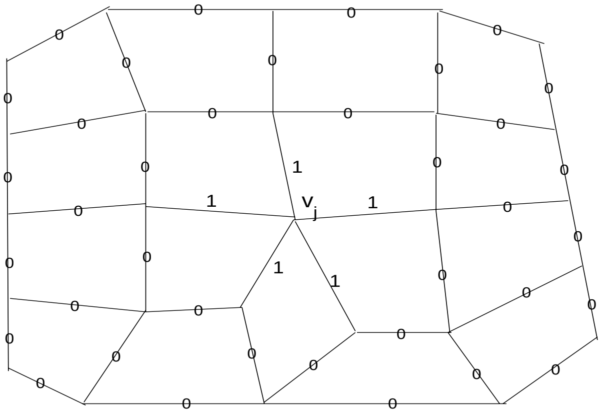


FIG. 1. *Values at the midpoints of the basis function φ_j associated with the vertex v_j.*

the numbers of quadrilaterals, vertices, and edges, respectively. Then set

$$\begin{aligned}
\mathcal{T}_h &= \{Q_1, Q_2, \ldots, Q_{N_Q}\}; \quad \bigcup_{j=1}^{N_Q} Q_j = \overline{\Omega},\\
\mathcal{V} &= \{v_1, v_2, \ldots, v_{N_V}\}: \text{ the set of all vertices of } Q \in \mathcal{T}_h,\\
\mathcal{E} &= \{e_1, e_2, \ldots, e_{N_E}\}: \text{ the set of all edges of } Q \in \mathcal{T}_h,\\
\mathcal{M} &= \{m_1, m_2, \ldots, m_{N_E}\}: \text{ the set of all midpoints of } e \in \mathcal{E}.
\end{aligned}$$

In particular, let N_V^i, N_E^i, and N_M^i denote the numbers of interior vertices, edges, and midpoints of $Q \in \mathcal{T}_h$, respectively. Our objective is to introduce a P_1-nonconforming finite element space associated with the quadrilateral decomposition $\mathcal{T}_h$.

Set

$$\begin{aligned}
\mathcal{NC}^h &= \{v_h : \Omega \to \mathbb{R} |\ v_h|_Q \in P_1(Q) \text{ for all } Q \in \mathcal{T}_h,\\
&\qquad v_h \text{ is continuous at every } m \in \mathcal{M} \setminus \Gamma\},\\
\mathcal{NC}_0^h &= \{v_h \in \mathcal{NC}^h |\ v_h(m) = 0 \text{ for all } m \in \Gamma \cap \mathcal{M}\}.
\end{aligned}$$

For each vertex $v_j \in \mathcal{V}$, denote by $\mathcal{E}(j)$ the set of all edges $e \in \mathcal{E}$ such that one of the endpoints is v_j, and by $\mathcal{M}(j)$ the set of all midpoints m of edges in $\mathcal{E}(j)$. Let $\varphi_j \in \mathcal{NC}^h$ be such that

$$\varphi_j(m) = \begin{cases} 1 & \text{if } m \in \mathcal{M}(j), \\ 0 & \text{if } m \in \mathcal{M} \setminus \mathcal{M}(j). \end{cases} \tag{2.1}$$

An example of such a function φ_j is shown in Figure 1. Notice that $\widehat{\varphi}_k, 1 \le k \le 4$, given in Lemma 2.2 belong to the restriction of $\varphi_j, j = 1, \ldots, N_V$, to Q.

Remark 2.3. Obviously $\varphi_j|_Q(v_j) < 2$ for all $Q \in \mathcal{T}_h$, with v_j being one of its vertices; moreover, $\varphi_j|_Q(v_j) = 3/2$ if Q is a parallelogram. Therefore, if $\mathcal{T}_h$ is decomposed into parallelograms, φ_j is continuous at v_j for all j. However, φ_j may not be continuous in general. Examples of basis functions in conforming and nonconforming cases are depicted in Figure 3(b),(c) for a simple mesh given in Figure 3(a).

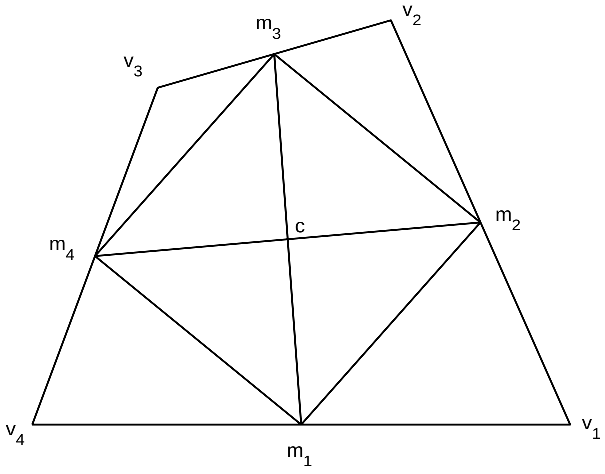


FIG. 2. *The midpoints $m_j, 1 \leq j \leq 4$, form a parallelogram in the quadrilateral with vertices $v_j, 1 \leq j \leq 4$.*

2.2. The dimension and basis for $\mathcal{NC}_0^h$. We proceed to investigate in the dimension of $\mathcal{NC}_0^h$; that of $\mathcal{NC}^h$ will be discussed in the next subsection. Implicitly the following assumption will be imposed on the decomposition in the rest of this article, especially for Dirichlet problems, in order to exclude pathological cases.

ASSUMPTION I. *Each interior edge has at least one interior vertex as its endpoint.*

There are cases in which Assumption I is violated. For instance, some decompositions $\mathcal{T}_h$ of Ω may contain elements whose four vertices lie on the boundary of Ω; in these cases, the reduced decomposition $\mathcal{T}_h'$, obtained by eliminating such elements from $\mathcal{T}_h$, fulfills Assumption I.

An upper bound of $\dim(\mathcal{NC}_0^h)$ is given in the following lemma.

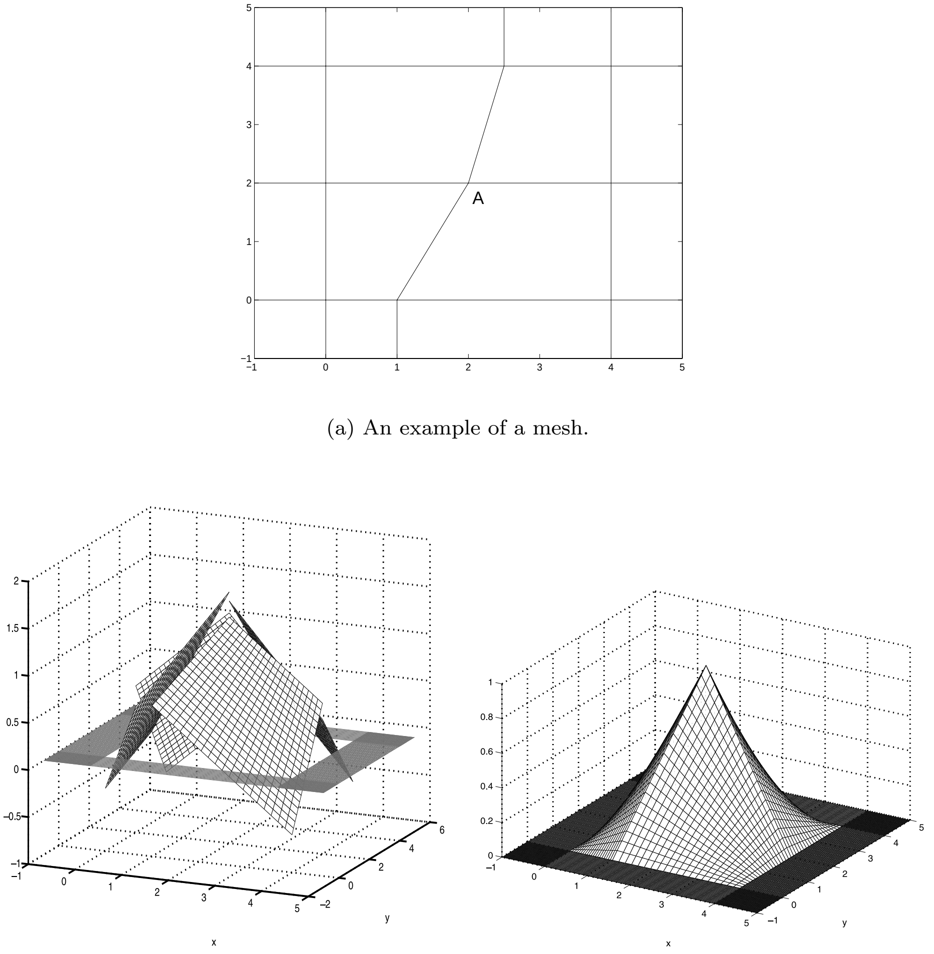
LEMMA 2.4. $\dim(\mathcal{NC}_0^h) \leq N_V^i$.

Proof. For the degrees of freedom for $\mathcal{NC}_0^h$ define $\mathbf{d} : \mathcal{NC}_0^h \to \mathbb{R}^{N_E^i}$ by

$$\mathbf{d}(\varphi) := (d_1(\varphi), \ldots, d_{N_E^i}(\varphi))^t, \quad \varphi \in \mathcal{NC}_0^h,$$

with $d_j(\varphi) = \varphi(m_j)$ for each interior midpoint $m_j, j = 1, \ldots, N_E^i$. If $\varphi \in \mathcal{NC}_0^h$ satisfies $d_j(\varphi) = 0$ for all $j = 1, \ldots, N_E^i$, clearly $\varphi = 0$. This implies that $\{d_j\}_{j=1}^{N_E^i}$ spans $(\mathcal{NC}_0^h)'$, the dual of $\mathcal{NC}_0^h$. Note that, for any $j = 1, \ldots, N_E^i$, the component d_j of $\mathbf{d}$ is nontrivial, since for each $m_j \in \mathcal{M} \setminus \Gamma$ there exists a function φ such that $\varphi(m_j) \neq 0$, as defined in (2.1) by Assumption I.

Due to Lemma 2.1, for each $Q_j \in \mathcal{T}_h$ having $m_{j_1}, m_{j_2}, m_{j_3}$, and m_{j_4} as its midpoints of edges as in Figure 2, one of the following linear restrictions should be



(a) An example of a mesh.

(b) The P_1-nonconforming basis whose values are equal to 1 at the midpoints of edges which meet with the vertex A, and 0 at all the other midpoints.

(c) The Q_1-conforming basis whose values are 1 at the vertex A, and 0 at all the other vertices.

FIG. 3. *Shapes of* P_1*-nonconforming* (b) *and* Q_1*-conforming* (c) *basis functions in the quadrilateral mesh shown in* (a).

imposed:

$$\begin{aligned} d_{j_1} + d_{j_3} - d_{j_2} - d_{j_4} &= 0 \qquad \text{if } m_{j_k} \notin \Gamma, 1 \le k \le 4, \\ d_{j_1} + d_{j_3} - d_{j_2} &= 0 \qquad \text{if } m_{j_1}, m_{j_2}, m_{j_3} \notin \Gamma \text{ and } m_{j_4} \in \Gamma, \\ d_{j_1} - d_{j_2} &= 0 \qquad \text{if } m_{j_1}, m_{j_2} \notin \Gamma \text{ and } m_{j_3}, m_{j_4} \in \Gamma, \end{aligned}$$

which can be written formally as

$$A_j \mathbf{d} = 0. \tag{2.2}$$

Here, $A_j = (A_{j,1}, \dots, A_{j,N_E^i})$ is a row vector in $\mathbb{R}^{N_E^i}$ with at most four nontrivial entries such that

$$\begin{aligned} A_{j,j_1} = A_{j,j_3} = -A_{j,j_2} = -A_{j,j_4} = 1 \qquad & \text{if } m_{j_k} \notin \Gamma, 1 \le k \le 4,\\ A_{j,j_1} = A_{j,j_3} = -A_{j,j_2} = 1 \qquad & \text{if } m_{j_1}, m_{j_2}, m_{j_3} \notin \Gamma \text{ and } m_{j_4} \in \Gamma,\\ A_{j,j_1} = -A_{j,j_2} = 1 \qquad & \text{if } m_{j_1}, m_{j_2} \notin \Gamma \text{ and } m_{j_3}, m_{j_4} \in \Gamma. \end{aligned}$$

We therefore see that

$$\begin{aligned} \dim(\mathcal{N}C_0^h) &= \dim((\mathcal{N}C_0^h)') \\ &\le \dim\{\mathbf{d} = (d_1, \dots, d_{N_E^i})^t;\ A_j\mathbf{d} = 0, j = 1, \dots, N_Q\}. \end{aligned} \tag{2.3}$$

We proceed to see whether $A_j, j = 1, \dots, N_Q$, are linearly independent vectors or not. For this, assume that for some proper subset $J \subsetneqq \{1, 2, \dots, N_Q\}$,

$$\sum_{j\in J} c_j A_j = 0, \tag{2.4}$$

with $c_j \neq 0$ for all $j \in J$. Set $\overline{\Omega}_J = \bigcup_{j\in J} \overline{Q}_j$. Then there exist an interior midpoint $m_l \in \partial\Omega_J \cap \mathcal{M} \setminus \Gamma$ and $Q_k \subset \Omega_J$ for which m_l is a midpoint of an edge of Q_k, since $\Omega_J \subsetneqq \Omega$. From the linear restriction concerning Q_k, $A_k\mathbf{d} = 0$, we see that A_k has a nonzero entry in the lth column; moreover, A_k is the unique vector that has a nonzero value in the lth entry among all $A_j, j \in J$, since m_l is at the boundary of Ω_J. Thus (2.4) implies that $c_k = 0$, which is a contradiction. Therefore, we have

$$\text{any } N_Q - 1 \text{ elements from } \{A_1, A_2, \dots, A_{N_Q}\} \text{ are linearly independent in } \mathbb{R}^{N_E^i}.$$

Let $A = (A_1^t, \dots, A_{N_Q-1}^t)^t$ be the $(N_Q - 1) \times N_E^i$ matrix whose jth row is A_j. Then the collection of (2.2) for $j = 1, \dots, N_Q - 1$ can be written formally in the matrix form

$$A\mathbf{d} = 0,$$

with the rank of A being $N_Q - 1$. Notice from (2.3) that

$$\dim(\mathcal{N}C_0^h) \le \dim\{\mathbf{d} = (d_1, \dots, d_{N_E^i})^t;\ A\mathbf{d} = 0\}. \tag{2.5}$$

Let B be an $(N_E^i - (N_Q - 1)) \times N_E^i$ matrix such that $\bar{A} = \left(\begin{smallmatrix} A \\ B \end{smallmatrix}\right)$ is invertible: such a matrix exists as $\operatorname{rank}(A) = N_Q - 1$. Setting $(\psi_1, \dots, \psi_{N_E^i-(N_Q-1)})^t = B\mathbf{d}$, we have

$$\bar{A}\mathbf{d} = (0, \dots, 0, \psi_1, \dots, \psi_{N_E^i-(N_Q-1)})^t \in \mathbb{R}^{N_E^i}.$$

This implies that $\{\psi_j\}_{j=1}^{N_E^i-(N_Q-1)}$ spans $\{\mathbf{d} = (d_1, \dots, d_{N_E^i})^t;\ A\mathbf{d} = 0\}$, since $\bar{A}$ is invertible. Therefore, from (2.5), we see that

$$\dim(\mathcal{N}C_0^h) \le N_E^i - (N_Q - 1).$$

Recall Euler's formula for a simply connected domain, $N_V - N_E + N_Q = 1$, which is equivalent to $N_V^i - N_E^i + N_Q = 1$. The following lemma is thus obtained: $\dim(\mathcal{N}C_0^h) \le N_E^i - (N_Q - 1) = N_V^i$. ◻

The dimension and basis functions for $\mathcal{N}C_0^h$ are given in the following theorem.

THEOREM 2.5. *Let φ_j be the function defined in* (2.1) *with interior vertex $v_j \in \mathcal{V} \setminus \Gamma$, $j = 1, \dots, N_V^i$. Then $\{\varphi_1, \varphi_2, \dots, \varphi_{N_V^i}\}$ forms a basis for $\mathcal{N}C_0^h$. Therefore,*

$\dim(\mathcal{N}C_0^h) = N_V^i$. *That is, the degrees of freedom for* $\mathcal{N}C_0^h$ *is equal to the number of interior vertices in* $\mathcal{T}_h$.

Proof. Suppose $\sum_{j=1}^{N_V^i} c_j\varphi_j = 0$. Choose a vertex v_l located at the boundary Γ. Since Ω is connected, there exists an interior vertex v_k adjacent to $v_l \in \Gamma$. Let m be the midpoint of $\overline{v_l v_k}$. Then we have

$$0 = \sum_{j=1}^{N_V^i} c_j\varphi_j(m) = c_k.$$

The coefficients c_j of the φ_j's corresponding to all the vertices adjacent to Γ will vanish in this manner. Then, stripping out all the boundary elements, we continue the above argument to the next layer to show again that all the coefficients c_j of the φ_j's corresponding to all the vertices adjacent to that boundary layer vanish. We can continue the argument to show that all the coefficients vanish until the domain is exhausted. Thus $\{\varphi_1, \varphi_2, \dots, \varphi_{N_V^i}\}$ is linearly independent. Moreover, $\{\varphi_1, \varphi_2, \dots, \varphi_{N_V^i}\}$ forms a basis for $\mathcal{N}C_0^h$ since $\dim(\mathcal{N}C_0^h) \le N_V^i$ by Lemma 2.4, and therefore $\dim(\mathcal{N}C_0^h) = N_V^i$. This completes the proof. □

Remark 2.6. Let $\widetilde{\mathcal{T}_h}$ be the triangulation of Ω into triangles by dividing each quadrilateral into two triangles. Consider the P_1-nonconforming simplicial element space $\widetilde{\mathcal{N}C_0^h}$ on $\widetilde{\mathcal{T}_h}$. We then observe that $\mathcal{N}C_0^h \subset \widetilde{\mathcal{N}C_0^h}$. Moreover,

$$\dim(\widetilde{\mathcal{N}C_0^h}) = N_E^i + N_Q = N_V^i + 2N_Q - 1 = \dim(\mathcal{N}C_0^h) + 2N_Q - 1.$$

2.3. The dimension and basis for $\mathcal{N}C^h$. The dimension and basis for $\mathcal{N}C^h$ is then obtained by the arguments in the previous subsection with slight modifications. Indeed, we have the following result.

LEMMA 2.7. $\dim(\mathcal{N}C^h) \le N_E - N_Q = N_V - 1$.

Proof. The arguments of the proof are essentially identical to those for Lemma 2.4 with minor modifications, but for the sake of the reader's convenience we repeat most of the arguments with proper modifications.

First, define $\mathbf{d} : \mathcal{N}C^h \to \mathbb{R}^{N_E}$ by

$$\mathbf{d}(\varphi) := (d_1(\varphi), \dots, d_{N_E}(\varphi))^t, \quad \varphi \in \mathcal{N}C^h,$$

with $d_j(\varphi) = \varphi(m_j)$ for each midpoint m_j, $j = 1, \dots, N_E$. Then one sees that $\{d_j\}_{j=1}^{N_E}$ spans $(\mathcal{N}C^h)'$, the dual of $\mathcal{N}C^h$.

For each $Q_j \in \mathcal{T}_h$ having $m_{j_1}, m_{j_2}, m_{j_3}$, and m_{j_4} as its midpoints of edges, the following linear restriction should be imposed:

$$d_{j_1} + d_{j_3} - d_{j_2} - d_{j_4} = 0,$$

which can be written formally as

$$A_j\mathbf{d} = 0,$$

where $A_j = (A_{j,1}, \dots, A_{j,N_E})$ is a row vector in $\mathbb{R}^{N_E}$ with at most four nontrivial entries such that

$$A_{j,j_1} = A_{j,j_3} = -A_{j,j_2} = -A_{j,j_4} = 1.$$

Consequently,

$$\dim(\mathcal{N}C^h) = \dim((\mathcal{N}C^h)') \leq \dim\{\mathbf{d} = (d_1, \ldots, d_{N_E})^t;\ A_j\mathbf{d} = 0, j = 1, \ldots, N_Q\}.$$

Next, assume that for a subset $J \subsetneqq \{1, 2, \ldots, N_Q\}$,

$$\sum_{j\in J} c_j A_j = 0, \tag{2.6}$$

with $c_j \neq 0$ for all $j \in J$. Set $\overline{\Omega}_J = \bigcup_{j\in J} \overline{Q}_j$. Then there exist a midpoint $m_l \in \partial\Omega_J \cap \mathcal{M}$ and $Q_k \subset \Omega_J$ for which m_l is a midpoint of an edge of Q_k. From the linear restriction concerning Q_k, $A_k\mathbf{d} = 0$, we see that A_k has a nonzero entry in the lth column; moreover, A_k is the unique vector that has a nonzero value in the lth entry among all $A_j, j \in J$, since m_l is at the boundary of Ω_J. Thus (2.6) implies that $c_k = 0$, which is a contradiction. Therefore, we have

$$\{A_1, A_2, \ldots, A_{N_Q}\} \text{ are linearly independent in } \mathbb{R}^{N_E}.$$

Then by an argument quite identical to the proof of Lemma 2.4, we see that

$$\dim(\mathcal{N}C^h) \leq N_E - N_Q.$$

Recall Euler's formula for a simply connected domain, $N_V - N_E + N_Q = 1$. The lemma thus is obtained: $\dim(\mathcal{N}C^h) \leq N_E - N_Q = N_V - 1$. □

The dimension and a basis functions for $\mathcal{N}C^h$ are given in the following theorem.

THEOREM 2.8. *Let φ_j be the function defined in* (2.1) *with each vertex $v_j \in \mathcal{V}$, $j = 1, \ldots, N_V$. Choose any vertex $v_{j_0} \in \mathcal{V}$. Then $\{\varphi_1, \varphi_2, \ldots, \varphi_{N_V}\} \setminus \{\varphi_{j_0}\}$ forms a basis for $\mathcal{N}C^h$. Moreover,* $\dim(\mathcal{N}C^h) = N_V - 1$. *That is, the degrees of freedom for $\mathcal{N}C^h$ is equal to the number of vertices in $\mathcal{T}_h$ minus* 1.

Proof. Without loss of generality, assume $v_{j_0} = v_{N_V}$. Suppose $\sum_{j=1}^{N_V-1} c_j\varphi_j = 0$. Let v_k be any vertex adjacent to v_{N_V}, and let m be the midpoint of $\overline{v_k v_{N_V}}$. Then

$$0 = \sum_{j=1}^{N_V-1} c_j\varphi_j(m) = c_k,$$

since $m \in \mathcal{E}(j)$ only if $j = N_V$ or k. Therefore we see that $c_{k_1} = 0$ for all k_1 such that v_{k_1} is a vertex of an edge $e \in \mathcal{E}(N_V)$. From all such v_{k_1}'s, we then proceed to show that that $c_{k_2} = 0$ for all k_2 such that v_{k_2} is a vertex of an edge $e \in \mathcal{E}(k_1)$. Since Ω is connected, by a finite repetition of the argument, we can conclude that all $c_j, j = 1, \ldots, N_V - 1$, are zeroes. Thus $\{\varphi_1, \varphi_2, \ldots, \varphi_{N_V-1}\}$ is linear independent and forms a basis for $\mathcal{N}C^h$ since $\dim(\mathcal{N}C^h) \leq N_V - 1$ by Lemma 2.7. □

3. The interpolation operator and convergence analysis. In this section we define an interpolation operator and analyze convergence. The case of Dirichlet problems is considered and convergence results are obtained by using standard arguments. The case of Neumann problems, which is analogous to that of Dirichlet problems, is then discussed in brief.

We first consider the following Dirichlet problem:

$$-\nabla \cdot \alpha\nabla u + \beta u = f, \quad \Omega, \tag{3.1a}$$

$$u = 0, \quad \Gamma, \tag{3.1b}$$

with $\alpha = (\alpha_{jk}), \alpha_{jk}, \beta \in L^\infty(\Omega), j,k = 1,2,\ 0 < \alpha_*|\xi|^2 \leq \xi^t\alpha(x)\xi \leq \alpha^*|\xi|^2 < \infty$, $\xi \in \mathbb{R}^2, \beta(x) \geq 0, x \in \Omega$, and $f \in H^{-1}(\Omega)$. The weak problem is given as usual: find $u \in H_0^1(\Omega)$ such that

$$a(u,v) = \langle f, v\rangle\,, \quad v \in H_0^1(\Omega), \tag{3.2}$$

where $a(u,v) = (\alpha\nabla u, \nabla v) + (\beta u, v)$, with $(\cdot,\cdot)$ being the $L^2(\Omega)$ inner product and $\langle\cdot,\cdot\rangle$ the duality pairing between $H^{-1}(\Omega)$ and $H_0^1(\Omega)$.

Our P_1-nonconforming method for problem (3.1a) is stated as follows: find $u_h \in \mathcal{NC}_0^h$ such that

$$a_h(u_h, v_h) = \langle f, v_h\rangle\,, \quad v_h \in \mathcal{NC}_0^h, \tag{3.3}$$

where

$$a_h(u,v) = \sum_{Q\in\mathcal{T}_h} a_Q(u,v),$$

with $a_Q : H^1(Q) \times H^1(Q) \to \mathbb{R}$ being the restriction of a to Q.

For our convergence analysis, define the projection $\Pi_h : H^2(\Omega) \cap H_0^1(\Omega) \to \mathcal{NC}_0^h$ such that, for $\varphi \in H^2(\Omega) \cap H_0^1(\Omega)$,

$$\Pi_h\varphi(m) = \frac{1}{2}(\varphi(v_1) + \varphi(v_2)) \qquad \text{for all } m \in \mathcal{M},$$

where v_1 and v_2 are the two vertices of the edge in $\mathcal{T}_h$ whose midpoint is m. Notice that Π_h is well defined. Indeed, with $Q \in \mathcal{T}_h$, $v_j, m_j, 1 \leq j \leq 4$, given as in Figure 2, one has

$$\Pi_h\varphi(m_1) + \Pi_h\varphi(m_3) = \frac{1}{2}(\varphi(v_1) + \varphi(v_2) + \varphi(v_3) + \varphi(v_4)) = \Pi_h\varphi(m_2) + \Pi_h\varphi(m_4).$$

Thus by Lemma 2.1, $\Pi_h\varphi \in P_1(Q)$. Clearly $\Pi_h\varphi$ is continuous at all midpoints of edges of $\mathcal{T}_h$. Therefore $\Pi_h\varphi \in \mathcal{NC}_0^h$.

Since Π_h preserves $P_1(Q)$ for all $Q \in \mathcal{T}_h$, standard interpolation approximation results, not by using a reference element but by applying the Bramble–Hilbert lemma to each actual element, lead to the finding that

$$\sum_{Q\in\mathcal{T}_h} ||\varphi - \Pi_h\varphi||_{L^2(Q)} + h\sum_{Q\in\mathcal{T}_h} ||\varphi - \Pi_h\varphi||_{H^1(Q)} \leq Ch^2||\varphi||_{H^2(\Omega)}, \quad \varphi \in H^2(\Omega) \cap H_0^1(\Omega). \tag{3.4}$$

(For instance, a slight modification to Exercise 3.1.2 in [6] using the result of [7] would give the estimate.)

Also, letting $\gamma_j = \partial\Omega \cap \partial Q_j, \gamma_{jk} = \partial Q_j \cap \partial Q_k$, and denoting the midpoint of γ_j and γ_{jk} by m_j and m_{jk}, respectively, define

$$\Lambda^h = \{\lambda \in \Pi_{j,k}\mathcal{P}_0(\gamma_{jk}) \times \Pi_j\mathcal{P}_0(\gamma_j) \mid \lambda_{jk} + \lambda_{kj} = 0, \text{ where } \lambda_{jk} = \lambda|_{\gamma_{jk}},\, \lambda_j = \lambda|_{\gamma_j}\},$$

where $\mathcal{P}_0(S)$ denotes the set of constant functions on a set S. Then define the projection $P_0 : H^2(\Omega) \to \Lambda^h$ so that if $v \in H^2(\Omega)$,

$$\left\langle \alpha\frac{\partial v_j}{\partial \nu_j} - P_0 v_j, z\right\rangle_\gamma = 0 \quad \text{for all } z \in \mathcal{P}_0(\gamma),\ \gamma = \gamma_{jk} \text{ or } \gamma_j, \tag{3.5}$$

where $v_j = v|_{Q_j}$ and ν_j is the unit outward normal to Q_j. One then has

$$\left\{ \sum_j \left\| \alpha \frac{\partial v_j}{\partial \nu_j} - P_0 v \right\|_{L^2(\partial Q_j)} \right\}^2 \leq C h^{\frac{1}{2}} ||v||_2. \tag{3.6}$$

With the broken energy norm

$$||\varphi||_h = a_h(\varphi, \varphi)^{\frac{1}{2}},$$

we are now in a position to state the usual second Strang lemma [16, 17, 6].

LEMMA 3.1. *Let $u \in H^1(\Omega)$ and $u_h \in \mathcal{N}C_0^h$ be the solutions of* (3.2) *and* (3.3), *respectively. Then,*

$$\|u - u_h\|_h \leq C \left\{ \inf_{v \in \mathcal{N}C_0^h} \|u - v\|_h + \sup_{w \in \mathcal{N}C_0^h} \frac{|a_h(u, w) - \langle f, w \rangle \,|}{\|w\|_h} \right\}.$$

Notice that (3.4) implies that

$$\inf_{v \in \mathcal{N}C_0^h} \|u - v\|_h \leq C \|u\|_2 h. \tag{3.7}$$

Next, for the consistency error term, by a simple calculation one has

$$a_h(u, w) - \langle f, w \rangle = \sum_j \left\langle \alpha \frac{\partial u_j}{\partial \nu_j}, w \right\rangle_{\partial Q_j \setminus \gamma_j}.$$

Since a function w in $\mathcal{N}C_0^h$ is linear on each γ_{jk} and continuous at the midpoints, the following useful orthogonality holds:

$$\langle P_0 u_j, w_j \rangle_{\gamma_{jk}} + \langle P_0 u_k, w_k \rangle_{\gamma_{kj}} = \langle P_0 u_j, w_j - w_k \rangle_{\gamma_{jk}} = 0 \qquad \text{for all } w \in \mathcal{N}C_0^h. \tag{3.8}$$

From the two orthogonalities (3.5) and (3.8),

$$a_h(u, w) - \langle f, w \rangle = \sum_j \left\langle \alpha \frac{\partial u_j}{\partial \nu_j} - P_0 u_j, w - m_j \right\rangle_{\partial Q_j \setminus \gamma_j}, \tag{3.9}$$

where m_j is chosen to be the average of w on Q_j. Due to (3.4), (3.6), and a trace theorem,

$$\begin{aligned} &\left| \sum_j \left\langle \alpha \frac{\partial u_j}{\partial \nu_j} - P_0 u_j, w - m_j \right\rangle_{\partial Q_j} \right| \\ &\qquad \leq C ||u||_2 h^{\frac{1}{2}} \left(\sum_j ||w - m_j||_{L^2(Q_j)} ||\nabla (w - m_j)||_{L^2(Q_j)} \right)^{\frac{1}{2}} \\ &\qquad \leq C ||u||_2 h \left(\sum_j ||\nabla w||^2_{L^2(Q_j)} \right)^{\frac{1}{2}} \leq C ||u||_2 ||w||_h h. \end{aligned} \tag{3.10}$$

Consequently, applying the estimates (3.7) and (3.10), combined with (3.9), in Lemma 3.1 gives the usual energy-norm error estimate. The use of a duality argument is

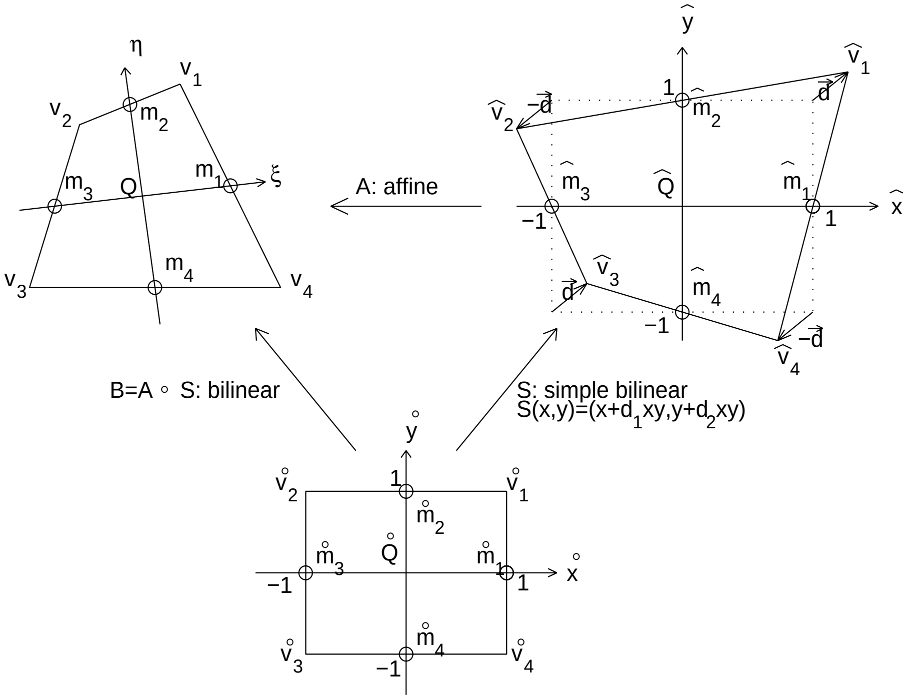


FIG. 4. *The nonparametric reference scheme: a general bilinear mapping B can be regarded as the composition of a simple bilinear map S and an affine map A.*

analogous to that in [9], and therefore we omit the details. To sum up, we have the following theorem.

THEOREM 3.2. *Let $u \in H^1(\Omega)$ and $u_h \in \mathcal{NC}_0^h$ be the solutions of* (3.2) *and* (3.3)*, respectively. Then we have*

$$||u - u_h||_h \leq Ch||u||_{H^2(\Omega)}.$$

Moreover, if Ω is convex and $f \in L^2(\Omega)$, then we have

$$||u - u_h||_{L^2(\Omega)} \leq Ch^2||u||_{H^2(\Omega)}.$$

Remark 3.3. The case of Robin problems is similar to that of Dirichlet problems, replacing the space $H_0^1(\Omega)$ and $\mathcal{NC}_0^h$ $H^1(\Omega)$ and $\mathcal{NC}^h$, as usual.

Remark 3.4. For the case of mixed boundary value problems, the dimension and basis functions can be computed and constructed analogously. Indeed, the dimension and basis functions are between those for the Dirichlet and Robin boundary problems.

4. A nonparametric reference scheme. In this section we introduce a nonparametric reference scheme with which finite elements in general quadrilaterals can be easily built from a fixed reference basis function space defined on a reference domain.

For given $Q \in \mathcal{T}_h$ with vertices $v_j, 1 \leq j \leq 4$, and midpoints of edges $m_j, 1 \leq j \leq 4$, as in Figure 4, there is a unique affine transformation $A : \mathbb{R}^2 \rightarrow Q$ such that

$$A(1,0) = m_1,\ A(0,1) = m_2,\ A(-1,0) = m_3,\ A(0,-1) = m_4,$$

since the four midpoints of any quadrilateral form a parallelogram. In fact, A is given by

$$A(\hat{x},\hat{y}) = \frac{v_1+v_2+v_3+v_4}{4} + \frac{v_1-v_2-v_3+v_4}{4}\hat{x} + \frac{v_1+v_2-v_3-v_4}{4}\hat{y}.$$

Denote $\widehat{Q} = A^{-1}(Q)$ and let $\widehat{m}_j, 1 \le j \le 4$, indicate the points $(1,0)$, $(0,1)$, $(-1,0)$, $(0,-1)$, respectively. Define $\widehat{\varphi}_j \in \mathrm{Span}\{1,\hat{x},\hat{y}\}, 1 \le j \le 4$, such that

$$\widehat{\varphi}_j(\widehat{m}_k) = \begin{cases} 1, & k = j, j+1 \mod 4, \\ 0, & \text{otherwise.} \end{cases}$$

Then, by Lemma 2.2, $P_1(Q) = \mathrm{Span}\{\widehat{\varphi}_j \circ A^{-1}; 1 \le j \le 4,\}$. This enables us to construct a basis function space by using this fixed reference basis function space $\{\widehat{\varphi_j}\}_{j=1}^4$, although $\widehat{Q}$ may vary.

A possible drawback due to the variance of $\widehat{Q}$ may come from difficulty in calculating the integrals of products of basis functions and their gradients on $\widehat{Q}$. However, this will be overcome easily as follows. Let $\mathring{Q}= [-1,1]^2$ and denote its vertices by $\mathring{v}_j, 1 \le j \le 4$, as in Figure 4. Then there is a unique bilinear transformation $B : \mathring{Q} \rightarrow Q$,

$$B(\mathring{x},\mathring{y}) = v_1 + (v_2 - v_1)\frac{1-\mathring{x}}{2} + (v_4 - v_1)\frac{1-\mathring{y}}{2} + (v_3 + v_1 - v_2 - v_4)\frac{1-\mathring{x}}{2}\frac{1-\mathring{y}}{2},$$

so that $B(\mathring{v}_j) = v_j, 1 \le j \le 4$. Indeed, $S = A^{-1} \circ B$ is given by

$$S(\mathring{x},\mathring{y}) = (\mathring{x} + d_1 \mathring{x}\mathring{y}, \mathring{y} + d_2 \mathring{x}\mathring{y}),$$

where

$$(d_1,d_2) = (v_1 + v_3 - v_2 - v_4)\begin{pmatrix} v_1 - v_3 - v_2 + v_4 \\ v_1 - v_3 + v_2 - v_4 \end{pmatrix}^{-1}.$$

Now, we can pull back the integrals on $\widehat{Q}$ to those on $\mathring{Q}$ by a change of variables, using the transformation S. For example, suppose $\varphi_j = \widehat{\varphi}_j \circ A^{-1}, j = 1,2$, to be two basis functions on Q. Then the integral on Q can be calculated as follows:

$$\int_Q \beta(x,y)\varphi_1(x,y)\varphi_2(x,y)\; dxdy$$
$$= \int_{\mathring{Q}} \beta(B(\mathring{x},\mathring{y}))\widehat{\varphi}_1(S(\mathring{x},\mathring{y}))\widehat{\varphi}_2(S(\mathring{x},\mathring{y}))\; |\det DB|\; d\,\mathring{x}\; d\,\mathring{y}.$$

5. The extension to three dimensions.

We give only a brief remark to extend the results in sections 2, 3, and 4, to three dimensions. For the sake of simplicity, let R be a three-dimensional hexahedron, with $m_j, j = 1,\dots,6$, being the barycenters of the six faces such that m_j and m_k are barycenters of opposite faces if $j+k=7$. Analogously to Lemma 2.1, if $u \in P_1(R)$, then

$$u(m_1) + u(m_6) = u(m_2) + u(m_5) = u(m_3) + u(m_4).$$

Conversely, if u_j is a given value at m_j, for $1 \le j \le 6$, satisfying $u_1 + u_6 = u_2 + u_5 = u_3 + u_4$, then there is a unique $u \in P_1(R)$ such that $u(m_j) = u_j, 1 \le j \le$

TABLE 1
Degrees of freedom for Q_1-conforming, P_1-nonconforming, and other nonconforming elements.

Elements	4^2	8^2	16^2	32^2	64^2	128^2	256^2
Q_1-conforming element	9	49	225	961	3969	16129	65025
P_1-NC element	9	49	225	961	3969	16129	65025
Other NC elements	24	112	480	1984	8064	32512	130560

6. This fact therefore leads to the conclusion that the local degrees of freedom for the three-dimensional nonconforming hexahedral element is four. Indeed, the space Span$\{1, x, y, z\}$ serves as the basis for the local nonconforming hexahedral element space for each hexahedron.

Concerning the global basis, consider a standard decomposition $\mathcal{T}_h$ of a three-dimensional domain Ω into the union of hexahedrons R_j with vertices p_k and barycenters m_l. At each vertex p_k, the global basis function φ_k is then defined analogously to the two-dimensional case: $\varphi_k|_{R_j} \in P_1(R_j)$, $\varphi_k(m_l) = 1$ if m_l is the barycenter of a face whose vertex contains p_k; $\varphi_k(m_l) = 0$ otherwise.

Then extensions of the rest of sections 2, 3, and 4 to three dimensions will be valid with suitable modifications.

6. Numerical results. In this section we present several numerical results to compare lowest-order quadrilateral elements which are either conforming or nonconforming. More precisely, six different elements are examined here including the P_1-nonconforming quadrilateral element and the standard Q_1-conforming element. We also test the two rotated Q_1-nonconforming elements introduced by Rannacher and Turek [15] with the degrees of freedom being the four midpoint values at the midpoints of edges and the four average values over edges. In addition, comparisons are made with the elements given by Douglas et al. [9], the local basis of which is of the form Span$\{1, x, y, \theta_l(x) - \theta_l(y)\}$, $l = 1, 2$, where the θ_l is given by

$$\theta_l(t) = \begin{cases} t^2 - \frac{5}{3}t^4, & l = 1, \\ t^2 - \frac{25}{6}t^4 + \frac{7}{2}t^6, & l = 2. \end{cases}$$

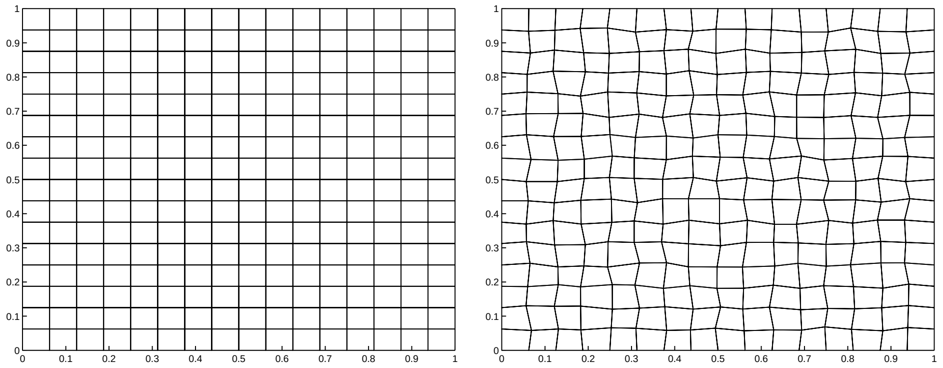
The following Dirichlet boundary problem is employed:

$$\begin{cases} -\triangle u = f, & \Omega, \\ u = 0, & \partial\Omega, \end{cases}$$

with the domain $\Omega = [0, 1]^2$ and the exact solution $u(x, y) = \sin(2\pi x)\sin(2\pi y)(x^3 - y^4 + x^2y^3)$, the function f being generated.

In every figure the logarithmic errors with base 2 are plotted against the logarithmic values of degrees of freedom again with base 2. With the uniform mesh as in Figure 5(a), the numerical errors are given in Figure 6. Convergence behaves more or less in optimal fashion for every element. Notice that the degrees of freedom for P_1-nonconforming and Q_1-conforming are nearly half of those of other nonconforming elements, as shown in Table 1.

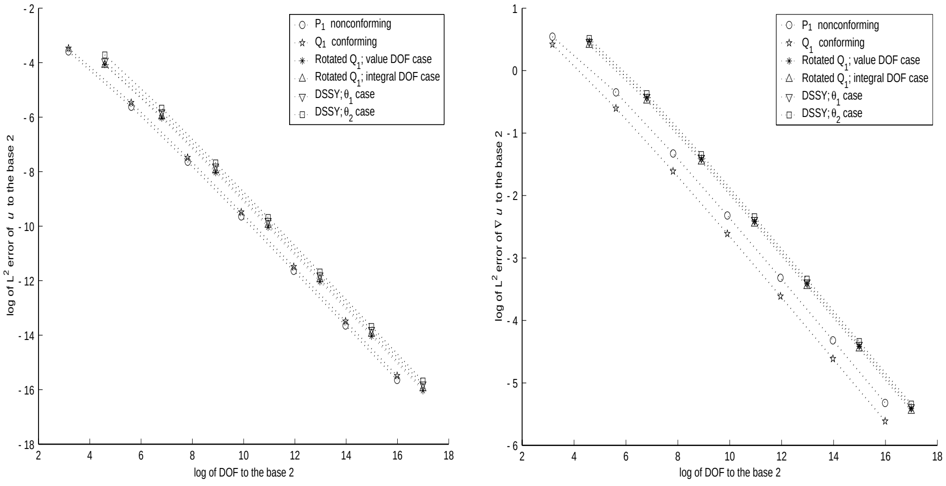
We observed that the optimal convergence patterns break for nonconforming elements if the nonuniform mesh depicted in Figure 5(b) is used with the standard bilinear reference scheme, since the nonconforming spaces do not contain the linear space as explained in [1]. In Figure 7 we show the error behaviors for the P_1-nonconforming element method, using the nonparametric reference scheme introduced in section 4,



(a) Uniform mesh.

(b) Nonuniform mesh.

FIG. 5. *The uniform and nonuniform meshes on* Ω*.*



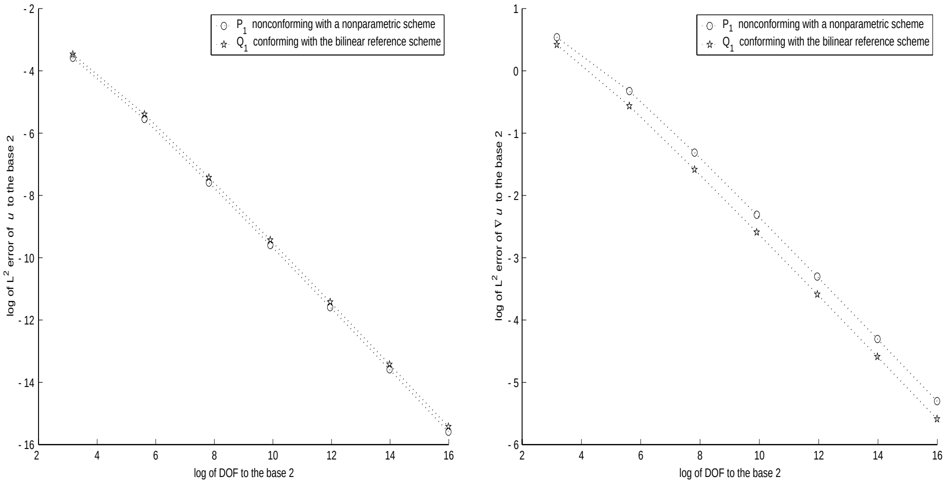
(a) $||u - u_h||_{L^2(\Omega)}$.

(b) $||\nabla u - \nabla u_h||_{L^2(\Omega)}$.

FIG. 6. $L^2(\Omega)$ *errors of* u_h *and* ∇u_h *(in logarithmic scale) on the uniform mesh.*

and compare them with those for the Q_1-conforming element method with the standard bilinear reference scheme applied. These two cases perform as well as we can expect, and the convergence rates are drawn in Figure 7. Our nonparametric reference scheme, which seems to be specific to the P_1-nonconforming quadrilateral element, does not work for the other known nonconforming quadrilateral elements mentioned in the paper; hence it does not seem fair to report such results here, some of which can be found in [14].

Several experiments were performed with the Robin problem. The errors, omitted here, behave quite similarly to those for the case of Dirichlet problems, as discussed above. Some reports can be found in [14].



(a) $||u - u_h||_{L^2(\Omega)}$.

(b) $||\nabla u - \nabla u_h||_{L^2(\Omega)}$.

FIG. 7. *$L^2(\Omega)$ errors of u_h and ∇u_h (in logarithmic scale) on the nonuniform mesh.*

REFERENCES

[1] D. N. ARNOLD, D. BOFFI, AND R. S. FALK, *Approximation by quadrilateral finite elements*, Math. Comp., 71 (2002), pp. 909–922.

[2] S. BRENNER AND L. SUNG, *Linear finite element methods for planar elasticity*, Math. Comp., 59 (1992), pp. 321–338.

[3] F. BREZZI AND M. FORTIN, *Mixed and Hybrid Finite Element Methods*, Springer-Verlag, New York, 1991.

[4] Z. CAI, J. DOUGLAS, JR., J. E. SANTOS, D. SHEEN, AND X. YE, *Nonconforming quadrilateral finite elements: A correction*, Calcolo, 37 (2000), pp. 253–254.

[5] Z. CAI, J. DOUGLAS, JR., AND X. YE, *A stable nonconforming quadrilateral finite element method for the stationary Stokes and Navier–Stokes equations*, Calcolo, 36 (1999), pp. 215–232.

[6] P. G. CIARLET, *The Finite Element Method for Elliptic Problems*, North–Holland, Amsterdam, 1978.

[7] P. G. CIARLET AND P.-A. RAVIART, *General Lagrange and Hermite interpolation in R^n with applications to finite element methods*, Arch. Ration. Mech. Anal., 46 (1972), pp. 177–199.

[8] M. CROUZEIX AND P.-A. RAVIART, *Conforming and nonconforming finite element methods for solving the stationary Stokes equations*, RAIRO Math. Model. Numer. Anal., 3 (1973), pp. 33–75.

[9] J. DOUGLAS, JR., J. E. SANTOS, D. SHEEN, AND X. YE, *Nonconforming Galerkin methods based on quadrilateral elements for second order elliptic problems*, RAIRO Math. Model. Numer. Anal., 33 (1999), pp. 747–770.

[10] V. GIRAULT AND P.-A. RAVIART, *Finite Element Methods for Navier–Stokes Equations*, Springer-Verlag, New York, 1986.

[11] H. HAN, *Nonconforming elements in the mixed finite element method*, J. Comput. Math., 2 (1984), pp. 223–233.

[12] G.-W. JANG, J. JEONG, Y. Y. KIM, M.-N. KIM, C. PARK, AND D. SHEEN, *Nonconforming finite element methods for checkerboard-free topology optimization*, Internat. J. Numer. Methods Engrg., to appear; available at http://www.nasc.snu.ac.kr/sheen/shortpub.html.

[13] C.-O. LEE, J. LEE, AND D. SHEEN, *A locking-free nonconforming finite element method for planar linear elasticity*, Adv. Comput. Math. (2003), to appear; available at http://www.nasc.snu.ac.kr/sheen/shortpub.html.

[14] C. Park, *A Study on Locking Phenomena in Finite Element Methods*, Ph.D. thesis, Department of Mathematics, Seoul National University, Seoul, Korea, 2002; available at http://www.nasc.snu.ac.kr/cpark/papers/phdthesis.ps.gz.

[15] R. Rannacher and S. Turek, *Simple nonconforming quadrilateral Stokes element*, Numer. Methods Partial Differential Equations, 8 (1992), pp. 97–111.

[16] G. Strang, *Variational crimes in the finite element method*, in The Mathematical Foundations of the Finite Element Method with Applications to Partial Differential Equations, A. K. Aziz, ed., Academic Press, New York, 1972, pp. 689–710.

[17] G. Strang and G. J. Fix, *An Analysis of the Finite Element Method*, Prentice–Hall, Englewood Cliffs, NJ, 1973.

SIAM J. NUMER. ANAL.
Vol. 41, No. 2, pp. 641–658

SPACE LOCALIZATION AND WELL-BALANCED SCHEMES FOR DISCRETE KINETIC MODELS IN DIFFUSIVE REGIMES[*]

LAURENT GOSSE[†] AND GIUSEPPE TOSCANI[‡]

Abstract. We derive and study well-balanced schemes for quasi-monotone discrete kinetic models. By means of a rigorous localization procedure, we reformulate the collision terms as nonconservative products and solve the resulting Riemann problem, whose solution is self-similar. The construction of an asymptotic preserving (AP) Godunov scheme is straightforward, and various compactness properties are established within different scalings. Finally, some computational results are supplied to show that this approach is realizable and efficient on concrete 2×2 models.

Key words. kinetic equations, diffusive relaxation schemes, nonconservative products

AMS subject classifications. 65M06, 65M12, 35F25

PII. S0036142901399392

1. Introduction.

In this paper, we are interested in the numerical analysis of the forthcoming one-dimensional system of semilinear equations,

$$\partial_t f^\pm \pm \partial_x f^\pm = \mp G(f^+, f^-), \qquad x \in \mathbb{R}, \qquad t > 0, \tag{1.1}$$

in both rarefied and diffusive regimes, the latter being obtained through the transformation $t \to t/\varepsilon^2$, $x \to x/\varepsilon$; see (3.4). The unknowns $0 \leq f^\pm$ are supposed to be at least *bounded variation (BV) functions* [39] in the space variable.

One motivation comes from the study of classical Boltzmann models,

$$\partial_t f + \vec{\xi} \cdot \nabla f = Q(f, f), \tag{1.2}$$

where $0 \leq f(t, x, \xi)$ stands for a density of particles moving with velocity $\vec{\xi}$ in the ambient space and $Q(f, f)$ is a collision operator satisfying some structural assumptions; see, e.g., [7, 38]. Such a model relaxes under certain variables scale

$$\varepsilon \partial_t f + \vec{\xi} \cdot \nabla f = \frac{Q(f, f)}{\varepsilon}, \qquad \varepsilon \to 0+,$$

towards the incompressible Navier–Stokes system [28], whereas for $\varepsilon \simeq 1$, it describes the flow of some cloudy bulk of rarefied molecules.

Coming back to our simplified model (3.4), we consider only particles moving with velocity ±1. Therefore, the density f in (1.2) boils down to a two-component vector satisfying the system (1.1), and the collision term takes a simple form [36]. In order to ensure some stability properties, namely $L^1(\mathbb{R})$-contraction [24, 33, 38], we ask for the so-called *quasimonotonicity* of the right-hand side, which reads

$$G \in C^1(\mathbb{R}^2),\ G(0, 0) = 0, \qquad \partial_+ G \overset{def}{=} \frac{\partial G}{\partial f^+} \geq 0, \qquad \partial_- G \overset{def}{=} \frac{\partial G}{\partial f^-} < 0. \tag{1.3}$$

[*] Received by the editors December 10, 2001; accepted for publication (in revised form) October 21, 2002; published electronically May 6, 2003. This research was partially supported by the European Union "Asymptotic methods in kinetic theory" TMR project ERBFMRXCT970157.

http://www.siam.org/journals/sinum/41-2/39939.html

[†] Istituto per le Applicazioni del Calcolo (sezione di Bari), Via G. Amendola 122-I, 70126 Bari, Italy (l.gosse@area.ba.cnr.it).

[‡] Dipartimento di Matematica, Università degli Studi di Pavia, Via Ferrata 1, 27100 Pavia, Italy (toscani@dimat.unipv.it).

This matches essentially the standard hypotheses encountered in [8, 29, 30], with the notable exception of [31], in which compactness results are established by means of a different methodology. Our present objective is then to develop and study robust numerical processes for (1.1), (3.4), stable and reliable on the whole range $0 \leq \varepsilon \leq 1$. We aim also at establishing rigorous compactness properties for these schemes in the spirit of the former articles [8, 29].

Part of this program has already been reached by the authors of [17, 21, 22], who introduced the notion of *asymptotic preserving (AP)* schemes relying on former works devoted to semilinear hyperbolic relaxation; in the present context, we refer to [30] and the quotations therein. Although the aforementioned schemes are computationally competitive, they still lack rigorous stability results as the stiffness strengthens in the diffusive limit; see, however, [19, 23]. On the other hand, the discretizations proposed in [2, section 3.1] do not fully address the issue of stability as $\varepsilon \to 0$.

Thus we propose to work out this twofold objective making use of the so-called *well-balanced (WB)* schemes [12, 15] (see also [4, 18, 26, 34]), built on a localization procedure already used in [1, 10, 11]. It relies on a rather simple idea: to reformulate the collision terms as a *nonconservative (NC) product* [25], in order to be able to solve *exactly* self-similar Riemann problems inside a classical Godunov procedure [9].

The concentration process for the right-hand side is carried out within section 2 by means of uniform BV estimates and representation of weak-$\star$ limits of measures. Then in section 3, we deduce in a rather straightforward way a Godunov scheme whose building blocks are nonconservative Riemann problems. Such a scheme turns out to be WB in the rarefied regime (see section 3.1) and AP in the diffusive limit $\varepsilon \to 0$ (see sections 3.2, 3.3). We give several compactness results under very reasonable CFL conditions of the type $\Delta t \simeq \max(\varepsilon h, h^2)$ (where $h, \Delta t$ stand for the space/time steps) by means of uniform BV bounds. Finally, we display some numerical results to illustrate various estimates in section 4 on widely used discrete kinetic models such as Carleman's [6] or Ruijgrok and Wu's [37].

2. A nonconservative reformulation for the kinetic model.

In the present section, we are about to follow the canvas of [10] in order to reformulate (1.1) as a *homogeneous* but *nonconservative* weakly coupled 2×2 system.

2.1. Uniform BV estimates.

We consider the Cauchy problem for $1 \geq \epsilon > 0$:

$$(2.1) \quad \partial_t f^{\pm} \pm \partial_x f^{\pm} = \mp G(f^+, f^-)\partial_x a^{\epsilon}, \qquad 0 \leq f^{\pm}(0,x) = f_0^{\pm}(x) \in L^1 \cap BV(\mathbb{R}).$$

We assume that a^ϵ is Lipschitz continuous for $\epsilon > 0$; more precisely,

$$a^\epsilon(x) = \begin{cases} jh & \text{for} \quad x \in \left] jh, \left(j + \frac{1}{2} - \frac{\epsilon}{2}\right) h \right], \\ \frac{x}{\epsilon} + \left(j + \frac{1}{2}\right) h \left(1 - \frac{1}{\epsilon}\right) & \text{for} \quad x \in \left] \left(j + \frac{1}{2} - \frac{\epsilon}{2}\right) h, \left(j + \frac{1}{2} + \frac{\epsilon}{2}\right) h \right], \\ (j+1)h & \text{for} \quad x \in \left] \left(j + \frac{1}{2} + \frac{\epsilon}{2}\right) h, (j+1)h \right]. \end{cases}$$

(2.2)

This means that $a^{\epsilon=1}(x) = x$, $a^\epsilon \in BV_{loc}(\mathbb{R})$ uniformly in ϵ, $\partial_x a^\epsilon \geq 0$, and moreover, ($\mathbf{1}_A$ stands for the characteristic function of a set A)

$$a^\epsilon \overset{\epsilon \to 0}{\longrightarrow} \sum_{j \in \mathbb{Z}} jh \mathbf{1}_{](j-\frac{1}{2})h, (j+\frac{1}{2})h]}, \qquad h > 0.$$

From [24, 33], one deduces that the Cauchy problem for (2.1) is well posed for any $\epsilon > 0$ but becomes ambiguous in the limit $\epsilon \to 0$ because a so-called *nonconservative product* [25] appears on the right-hand side as $\partial_x a^\epsilon$ concentrates into a Dirac comb.

LEMMA 1. *Let $f_0^\pm \in L^1 \cap BV(\mathbb{R})$ have compact support; then the weak solutions to* (2.1) *$f^\pm$ belong to $BV_{loc}(\mathbb{R}_*^+ \times \mathbb{R})$ uniformly in ϵ.*

Proof. We split the proof into several steps for the sake of clarity.

(i) From [33] and the fact that G is quasi-monotone and $\partial_x a^\epsilon \geq 0$, one deduces a $L^1(\mathbb{R})$ contraction principle for any value $\epsilon > 0$: if $\tilde{f}_0^\pm \in L^1 \cap BV(\mathbb{R})$,

$$(2.3) \qquad \forall t > 0, \qquad \partial_t \int_{\mathbb{R}} (|f^+(t,x) - \tilde{f}^+(t,x)| + |f^-(t,x) - \tilde{f}^-(t,x)|)dx \leq 0.$$

As $G(0,0) = 0$, the null solution trivially satisfies (2.1), and this ensures the L^1 stability and the positivity-preserving property. As we assumed that either $-\partial_- G$ (or $\partial_+ G$) is strictly positive, we can use the implicit function theorem to deduce that the equation $G(u,v) = 0$ admits as a unique solution a smooth curve $v = M(u)$, $M' \geq 0$, called the *Maxwellian distribution*. Therefore, using these curves as comparison functions inside (2.3) gives a maximum principle. More precisely, as in [38], the following domain,

$$[0, \|f_0^+\|_{L^\infty(\mathbb{R})}] \times [0, M(\|f_0^+\|_{L^\infty(\mathbb{R})})],$$

is positively invariant for (2.1). But since (2.1) isn't translation invariant, (2.3) doesn't guarantee the uniform $BV(\mathbb{R})$ stability.

(ii) Differentiating each equation in (2.1) with respect to time, multiplying by $(\mathrm{sgn}(\partial_t f^+), \mathrm{sgn}(\partial_t f^-))^T$, and integrating on $x \in \mathbb{R}$, the same way one reaches

$$(2.4) \qquad \forall t > 0, \qquad \partial_t \int_{\mathbb{R}} (|\partial_t f^+(t,x)| + |\partial_t f^-(t,x)|)dx \leq 0.$$

This implies that $\partial_t f^\pm(t,.)$ are bounded measures on $\mathbb{R}$, and the same holds true for $G(f^+, f^-)\partial_x a^\epsilon$ by the L^∞ stability and (2.2); but by their very definition, one has also

$$(2.5) \qquad |\partial_x f^\pm| - |G(f^+, f^-)\partial_x a^\epsilon| \leq |\partial_t f^\pm| \leq |\partial_x f^\pm| + |G(f^+, f^-)\partial_x a^\epsilon|.$$

So inside any interval $a < 0 < b$ large enough, one gets out of (2.4), (2.5) for any $t > 0$,

$$\int_a^b |\partial_x f^+(t,x)| + |\partial_x f^-(t,x)| dx \leq \int_{\mathbb{R}} |\partial_x f_0^+| + |\partial_x f_0^-| + 4\|G(f^+, f^-)\|_{L^\infty} \int_a^b |\partial_x a^\epsilon|,$$

and this ensures the $BV_{loc}(\mathbb{R})$ stability for $t > 0$.

(iii) It remains to check the L^1 modulus of continuity in the time variable. Thanks to the BV_{loc} bound, we deduce from (2.1) that on the same interval there holds for $t > s \geq 0$ (TV stands for the total variation in space):

$$\int_a^b |f^\pm(t,x) - f^\pm(s,x)| dx \leq |t-s| \left[TV(f_0^+) + TV(f_0^-) + 6\|G(f^+, f^-)\|_{L^\infty} \int_a^b |\partial_x a^\epsilon| \right].$$

And this is enough to conclude the proof. □

Of course, by the classical Helly's compactness principle, we deduce that the sequence of weak solutions to (2.1) is relatively compact in the strong topology of $L^1_{loc}(\mathbb{R}_*^+ \times \mathbb{R})$ as $\epsilon \to 0$.

2.2. Limiting values of the right-hand side. In order to shed complete light on the limit system emanating from (2.1) as $\epsilon \to 0$, we must give a precise meaning to the ambiguous product appearing on its right-hand side. This can be done within the recent theory of *nonconservative products* [25], which can be applied thanks to the uniform BV-bound established in the preceding section.

In order to reveal the nature of the limit for $G(f^+, f^-)\partial_x a^\epsilon$ in the weak-$\star$ topology of measures, we pick up a test function $\psi \in C^0_c(\mathbb{R}^+_* \times \mathbb{R})$ that is continuous and compactly supported and look at the behavior of the sequence

$$\mathcal{I}^\epsilon = \int_{\mathbb{R}^+_* \times \mathbb{R}} G(f^+, f^-)\partial_x a^\epsilon \psi(t,x)dtdx, \qquad \epsilon \to 0.$$

PROPOSITION 1. *Under the assumptions of Lemma* 1, *there holds as* $\epsilon \to 0$,

$$G(f^+, f^-)\partial_x a^\epsilon \overset{weak-\star}{\rightharpoonup} \mathcal{M}$$
$$\sum_{j\in\mathbb{Z}} h \left(\int_0^1 G(\bar{f}^+_{j+\frac{1}{2}}, \bar{f}^-_{j+\frac{1}{2}})(t,\xi)d\xi \right) \delta\left(x - \left(j + \frac{1}{2}\right)h\right),$$

where δ stands for the Dirac mass in $x = 0$ and the "microscopic profiles" $\bar{f}^\pm_{j+\frac{1}{2}}$ satisfy the ordinary differential system

$$\partial_\xi \bar{f}^\pm_{j+\frac{1}{2}} = -hG(\bar{f}^+_{j+\frac{1}{2}}, \bar{f}^-_{j+\frac{1}{2}}), \qquad \xi \in [0,1], \tag{2.6}$$

with the initial data for $t \in \mathbb{R}^+$ and $x = \left(j + \frac{1}{2}\right)h$, $j \in \mathbb{Z}$:

$$\bar{f}^+_{j+\frac{1}{2}}(t, \xi = 0) = f^+(t, x-0), \qquad \bar{f}^-_{j+\frac{1}{2}}(t, \xi = 1) = f^+(t, x+0). \tag{2.7}$$

We stress that the left/right values of $f^\pm(t,.)$ in (2.7) make sense, thanks to the uniform BV-regularity.

Proof. We notice at once that, thanks to the definition (2.2) of a^ϵ, we have

$$\mathcal{I}^\epsilon = \int_{\mathbb{R}^+_*} \sum_{j\in\mathbb{Z}} \int_{(j+\frac{1}{2}-\frac{\epsilon}{2})h}^{(j+\frac{1}{2}+\frac{\epsilon}{2})h} \frac{G(f^+, f^-)}{\epsilon} \psi(t,x)dxdt.$$

Thus it is convenient to perform a rescaling of the space variable:

$$[0,1] \ni \xi = \frac{1}{h\epsilon}\left(x - \left(j + \frac{1}{2} - \frac{\epsilon}{2}\right)h\right), \quad x = \left(j + \frac{1}{2} - \frac{\epsilon}{2}\right)h + \xi h\epsilon \overset{\epsilon\to 0}{\to} \left(j + \frac{1}{2}\right)h.$$

Inside any stripe $\left]\left(j + \frac{1}{2} - \frac{\epsilon}{2}\right)h, \left(j + \frac{1}{2} + \frac{\epsilon}{2}\right)h\right]$, the unknowns $f^\pm$ satisfy the following semilinear boundary value problem for $\xi \in [0,1]$:

$$\begin{cases} \epsilon h \partial_t f^\pm \pm \partial_\xi f^\pm = \mp h.G(f^+, f^-), \qquad t > 0, \\ f^+(t, \xi = 0) = f^+\left(t, \left(j + \frac{1}{2} - \frac{\epsilon}{2}\right)h\right), \\ f^-(t, \xi = 1) = f^-\left(t, \left(j + \frac{1}{2} + \frac{\epsilon}{2}\right)h\right), \\ f^\pm(t = 0, \xi) = f^\pm_0(\xi). \end{cases}$$

Its solution can be computed by the method of characteristics for $t \in [\tau_0, \tau_0 + \epsilon h]$,

$$\dot{\xi}^\pm_{\tau_0} = \tfrac{\pm 1}{\epsilon h}, \qquad \xi^+_{\tau_0}(\tau_0) = 0,\ \xi^-_{\tau_0}(\tau_0) = 1;$$
$$\dot{f}^\pm = \tfrac{\mp 1}{\epsilon} G(f^+, f^-), \qquad f^\pm(t) = f^\pm(t, \xi^\pm_{\tau_0}(t)).$$

One sees, therefore, that along $\xi^{\pm}_{\tau_0}$, there holds

$$\partial_\xi f^{\pm} = \dot{f}^{\pm}\left(\dot{\xi}^{\pm}_{\tau_0}\right)^{-1} = -h.G(f^+, f^-).$$

Since the system is semilinear, $\xi^{\pm}_{\tau_0}$ realizes a diffeomophism from $[\tau_0, \tau_0 + \epsilon h]$ onto $[0,1]$ and has an inverse we note $\tau^{\pm}$; it satisfies

$$\forall \xi \in [0,1], \qquad \tau^+(0) = \tau^-(1) = \tau_0, \qquad \frac{d\tau^{\pm}}{d\xi} = \epsilon h \overset{\epsilon \to 0}{\to} 0.$$

This means in particular that for $\xi \in [0,1]$,

$$f^{\pm}(t) = f^{\pm}(t, \xi^{\pm}_{\tau_0}(t)) = f^{\pm}(\tau^{\pm}(\xi), \xi) \overset{\epsilon \to 0}{\to} f^{\pm}(\tau_0, \xi)$$

and keeps on satisfying the differential equation. It remains to rewrite

$$\mathcal{I}^{\epsilon} = \int_{\mathbb{R}^+_*} \sum_{j \in \mathbb{Z}} \int_0^1 h.G(f^+, f^-)(t, \xi)\psi\left(\left(j + \frac{1}{2} - \frac{\epsilon}{2}\right) h + \xi h \epsilon\right) d\xi dt.$$

We can invoke Lebesgue's dominated convergence theorem in order to pass to the limit $\epsilon \to 0$ in $\mathcal{I}^{\epsilon}$, and we are done. □

From now on, the meaning of the "distributions product" in

$$\partial_t f^{\pm} \pm \partial_x f^{\pm} = \mp \sum_{j \in \mathbb{Z}} h.G(f^+, f^-)\delta\left(x - \left(j - \frac{1}{2}\right) h\right) \tag{2.8}$$

is to be *always* understood as following from Proposition 1. In particular, this last result provides a unique way to solve the Riemann problem for (2.8) with three simple waves, two of them moving with velocity ±1 associated with the convection process, and the static one rendering the action of the localized collision term. More precisely, if we supply four constant states at time $t = 0$ separated by a discontinuity in $x = \left(j - \frac{1}{2}\right) h$ $f^{\pm}_{L/R}$, the self-similar solution to (2.8) is given by

$$\begin{cases} (f^+_L, f^-_L) & \text{for} \quad x - \left(j - \frac{1}{2}\right) h < -t, \\ (f^+_L, \tilde{f}^-) & \text{for} \quad -t < x - \left(j - \frac{1}{2}\right) h < 0, \\ (\tilde{f}^+, f^-_R) & \text{for} \quad 0 < x - \left(j - \frac{1}{2}\right) h < t, \\ (f^+_R, f^-_R) & \text{for} \quad x - \left(j - \frac{1}{2}\right) h > t, \end{cases} \tag{2.9}$$

where, according to the notation of Proposition 1,

$$\tilde{f}^+ = \bar{f}^+_{j-\frac{1}{2}}(t, \xi = 1) \quad \text{and} \quad \tilde{f}^- = \bar{f}^-_{j-\frac{1}{2}}(t, \xi = 0).$$

Such a construction has already been successfully used inside a numerical processing of the so-called *hyperbolic heat equations* in [13]. We also stress that it is but a particular case of the "*h*-Riemann solvers" introduced in [1] within a different context.

2.3. Uniqueness via $L^1(\mathbb{R})$ contraction. Thanks to the dissipative structure of (2.1), it is straightforward to establish uniqueness for the singular problem (2.8) as soon as it has been given a rigorous sense within the theory of distributions.

Proposition 2. *Under the assumptions of Lemma* 1*, let $f_0^\pm, \tilde{f}_0^\pm$ be two sets of initial data for* (2.8)*. The following holds for all $t > 0$:*

$$\|f^+(t,.)-\tilde{f}^+(t,.)\|_{L^1(\mathbb{R})}+\|f^-(t,.)-\tilde{f}^-(t,.)\|_{L^1(\mathbb{R})} \leq \|f_0^+-\tilde{f}_0^+\|_{L^1(\mathbb{R})}+\|f_0^--\tilde{f}_0^-\|_{L^1(\mathbb{R})}.$$

In particular, the weak solution in the sense of Proposition 1 *to* (2.8)*, $f_0^\pm \in L^1 \cap BV(\mathbb{R})$ with compact support, is unique and belongs to $L^\infty(\mathbb{R}^+; BV(\mathbb{R})) \cap Lip(\mathbb{R}^+; L^1(\mathbb{R}))$.*

Proof. We proceed by approximation and come back to (2.1), for which the contraction property (2.3) holds uniformly in ϵ. It remains to integrate it in time and make use of the compactness results to conclude. □

Remark 1. *We would like to mention at this level that the present construction provides an alternative route to the compactness results of* [10] *concerning the convex scalar balance law whose right-hand side concentrates like* (2.1), (2.2):

$$\partial_t u + \partial_x f(u) = k(x) g(u) \partial_x a^\epsilon, \qquad 0 \leq u(t=0,x) = u_0(x) \in L^1 \cap BV(\mathbb{R}).$$

Indeed, if it is assumed that $0 \leq k \in C_c^0(\mathbb{R})$, $g' \leq 0$, and $g(\bar{u}) = 0$ for a $\bar{u} > 0$, then the $L^1(\mathbb{R})$ contraction principle [24] *holds uniformly in ϵ, and the interval*

$$[0, \max(\bar{u}, \|u_0\|_{L^\infty(\mathbb{R})})]$$

is positively invariant since its endpoints give rise as initial data to sub-/supersolutions of the associated homogeneous conservation law. Therefore, it is easy to follow the lines of Lemma 1 *to derive*

$$\partial_t \left(\int_{\mathbb{R}} |\partial_t u|(t,x) dx \right) \leq 0,$$

which gives in turn, as long as k has compact support,

$$\int_{\mathbb{R}} |\partial_x f(u)|(t,x) dx \leq \int_{\mathbb{R}} |\partial_x f(u_0)| + 2\|k(x)g(u)\|_{L^\infty} \int_{Supp(k)} |\partial_x a^\epsilon|,$$

together with a time Lipschitz-modulus of continuity. If, moreover, a nonresonance assumption $f' \geq c > 0$ holds, it turns out that this is enough to establish a uniform $BV_{loc}(\mathbb{R}_*^+ \times \mathbb{R})$ bound for u as $\epsilon \to 0$. Therefore we recover (part of) the conclusion of Lemma 7 *in* [10].

We close this section by introducing $\mathbf{S}$, the *solution operator* for (2.8) as follows.

Proposition 3. *There exists a unique "nonconservative contraction semigroup" $\mathbf{S}$, whose domain is $L^1 \cap BV(\mathbb{R})$ and such that any trajectory $0 < t \mapsto \mathbf{S}(t) f_0^\pm$ coincides with the unique weak solution to* (2.8)*, $f^\pm(t=0,.) = f_0^\pm$, in the sense of distributions.*

3. Derivation and convergence of well-balanced schemes.

Roughly speaking, we are about to develop and study Godunov schemes [9] for (1.1), relying on solving elementary Riemann problems for (2.8) whose self-similar solution is given by (2.9). More precisely, given a time-step $\Delta t > 0$ and a mesh-size $h > 0$, we can define a computational Cartesian grid. The cells read for all $j, n \in \mathbb{Z} \times \mathbb{N}$,

$$C_j = \left] \left(j - \frac{1}{2}\right) h, \left(j + \frac{1}{2}\right) h \right], \qquad I^n = [n\Delta t, (n+1)\Delta t[.$$

Let $\mathcal{P}^h$ be the standard L^2 projector on piecewise-constant functions:

$$\begin{array}{rccc} \mathcal{P}^h: & L^1 \cap BV(\mathbb{R}) & \to & L^1 \cap BV(\mathbb{R}), \\ & \varphi & \mapsto & \left(\int_{C_j} \frac{\varphi(x)}{h} dx \right)_{j \in \mathbb{Z}}. \end{array}$$

It remains to discretize the initial data as follows: we define a piecewise constant approximation $f_h^{\pm}(t=0,.)$ by taking the pointwise values

$$\forall j \in \mathbb{Z}, \qquad f_{j,0}^{\pm} = f_0^{\pm}(jh),$$

and this makes sense thanks to the BV regularity of the considered functions. Our well-balanced Godunov scheme therefore reads as

$$f_h^{\pm}(t,.) = \mathbf{S}(t - n\Delta t) \circ (\mathcal{P}^h \circ \mathbf{S}(\Delta t))^n f_h^{\pm}(t=0,.), \tag{3.1}$$

where n stands for the integer part of $t/\Delta t$. Therefore Riemann problems for (2.8) are to be solved at the endpoints of each C_j, $j \in \mathbb{Z}$.

3.1. The rarefied regime. We focus first on (1.1) in its hyperbolic scaling. Using the divergence theorem, one sees that the Godunov scheme (3.1) generates the following values as $n \in \mathbb{N}$, $j \in \mathbb{Z}$:

$$f_{j,n+1}^{+} = f_{j,n}^{+} - \frac{\Delta t}{h}(f_{j,n}^{+} - f_{j-\frac{1}{2},n}^{+}), \qquad f_{j,n+1}^{-} = f_{j,n}^{-} + \frac{\Delta t}{h}(f_{j+\frac{1}{2},n}^{-} - f_{j,n}^{-}). \tag{3.2}$$

The values at the borders of each cell C_j are given by the generalized jump relations (2.6), (2.7). Thus the upwind scheme (3.2) can be rewritten as

$$\begin{aligned} f_{j,n+1}^{+} &= f_{j,n}^{+} - \frac{\Delta t}{h}(f_{j,n}^{+} - f_{j-1,n}^{+}) - \Delta t \int_0^1 G(\bar{f}_{j-\frac{1}{2}}^{+}, \bar{f}_{j-\frac{1}{2}}^{-})(n\Delta t, \xi) d\xi, \\ f_{j,n+1}^{-} &= f_{j,n}^{-} + \frac{\Delta t}{h}(f_{j+1,n}^{-} - f_{j,n}^{-}) + \Delta t \int_0^1 G(\bar{f}_{j+\frac{1}{2}}^{+}, \bar{f}_{j+\frac{1}{2}}^{-})(n\Delta t, \xi) d\xi. \end{aligned} \tag{3.3}$$

This highlights the ability of this scheme to preserve *exactly* the steady-state curves of (1.1) since, by their very definition, they satisfy the following (cf. (2.6), (2.7)) for all $j \in \mathbb{Z}$:

$$\begin{aligned} f_{j,0}^{+} - f_{j-1,0}^{+} &= -h \int_0^1 G(\bar{f}_{j-\frac{1}{2}}^{+}, \bar{f}_{j-\frac{1}{2}}^{-})(0, \xi) d\xi, \\ f_{j+1,0}^{-} - f_{j,0}^{-} &= -h \int_0^1 G(\bar{f}_{j+\frac{1}{2}}^{+}, \bar{f}_{j+\frac{1}{2}}^{-})(0, \xi) d\xi. \end{aligned}$$

The following stability result is easily established.

LEMMA 2. *Let $f_0^{\pm} \in L^1 \cap BV(\mathbb{R})$; under the hyperbolic CFL condition $\Delta t \leq h$, the approximate solutions $f_h^{\pm}$ obtained from* (3.1), (3.2) *satisfy*

$$TV(f_h^{+}(t,.)) + TV(f_h^{-}(t,.)) \leq \exp\left(2\frac{t}{h}(\exp(Lip(G)h) - 1)\right)[TV(f_0^{+}) + TV(f_0^{-})],$$

where $Lip(G)$ stands for the Lipschitz constant of G.

Proof. We start from (3.3) and follow the proof of Lemma 10 in [10]: by the classical theory of differential equations and a linearization of G, one gets the following inequalities:

$$\begin{aligned}
|f^+_{j,n+1} - f^+_{j-1,n+1}| \leq {} & |f^+_{j,n} - f^+_{j-1,n}| \left(1 - \frac{\Delta t}{h}\right) \\
& + \frac{\Delta t}{h}(1 + (\exp(Lip(G)h) - 1))|f^+_{j-1,n} - f^+_{j-2,n}| \\
& + \frac{\Delta t}{h}(\exp(Lip(G)h) - 1)|f^-_{j+1,n} - f^-_{j,n}|, \\
|f^-_{j+1,n+1} - f^-_{j,n+1}| \leq {} & |f^-_{j+1,n} - f^-_{j,n}| \left(1 - \frac{\Delta t}{h}\right) \\
& + \frac{\Delta t}{h}(1 + (\exp(Lip(G)h) - 1))|f^-_{j+2,n} - f^-_{j+1,n}| \\
& + \frac{\Delta t}{h}(\exp(Lip(G)h) - 1)|f^+_{j,n} - f^+_{j-1,n}|.
\end{aligned}$$

It remains to add these two inequalities, to sum on $j \in \mathbb{Z}$ to derive

$$\begin{aligned}
& \sum_{j\in\mathbb{Z}}(|f^+_{j,n+1} - f^+_{j-1,n+1}| + |f^-_{j+1,n+1} - f^-_{j,n+1}|) \\
\leq {} & \left(1 + \frac{2\Delta t}{h}(\exp(Lip(G)h) - 1)\right) \sum_{j\in\mathbb{Z}}(|f^+_{j,n} - f^+_{j-1,n}| + |f^-_{j+1,n} - f^-_{j,n}|).
\end{aligned}$$

And this is enough to conclude, since $BV(\mathbb{R}) \subset L^\infty(\mathbb{R})$. □

REMARK 2. *This proof is a direct adaptation to* (1.1) *of the one given for Lemma* 10 *in* [10]*. We mention here that there exists, however, a much quicker way to establish this former compactness result, relying on* [27]*. Indeed, one can consider any nonhomogeneous scalar balance law endowed with a localized right-hand side,*

$$\partial_t u + \partial_x f(u) - g(u)\partial_x a = 0, \qquad \partial_t a = 0,$$

as an elementary but nonconservative 2×2 *Temple system whose wave curves are the level sets of the strong Riemann invariants (in the notation of* [10]*):*

$$a \text{ and } w(u, a) = \phi^{-1} \circ (\phi(u) - a), \qquad \phi'(u) = \frac{f'(u)}{g(u)}.$$

Therefore Lemmas 3.1 *and* 3.2 *in* [27] *imply that the Godunov scheme decreases the total variation of* $w(t, .)$*. This entails control on the* $u(t, .)$ *variable in the case in which*

$$0 < c \leq \partial_u w(u, a) \leq C < +\infty,$$

and this is a consequence of the nonresonance assumption $f'(u) \neq 0$*. Thus strong* L^1_{loc} *compactness for the scalar "well-balanced" scheme follows. It does not seem that the same shortcut applies here in order to shrink the proof of Lemma* 2*.*

By standard arguments, we can establish strong L^1_{loc} compactness for $f^\pm_h$ as $h \to 0$, relying on the bound stated in Lemma 2.

3.2. The diffusive regime: BV stability. We move now to the study of numerical approximations to (1.1) in its *diffusive scaling*, that is to say,

$$\partial_t f^{\pm} \pm \frac{1}{\varepsilon}\partial_x f^{\pm} = \mp\frac{1}{\varepsilon^2}G(f^+,f^-), \qquad 0<\varepsilon<1, \qquad x\in\mathbb{R}, \qquad t>0. \tag{3.4}$$

In this perspective, the so-called *hyperbolic heat equations* treated in [13] correspond to the special case $G(f^+,f^-)=f^+-f^-$. We assume also that the Maxwellian distribution is given by $M(f)=f$, that is to say,

$$f^+=f^- \qquad \Rightarrow \qquad G(f^+,f^-)=0.$$

In this setting and for any $\varepsilon>0$, the previous technique relying on a localization of the source term onto a Dirac comb still applies, but the differential system (2.6), (2.7) in Proposition 1 has to be rescaled,

$$\forall j\in\mathbb{Z}, \qquad \partial_\xi \bar f^{\pm}_{j+\frac12} = \frac{-h}{\varepsilon}G(\bar f^+_{j+\frac12}, \bar f^-_{j+\frac12}), \tag{3.5}$$

and the stability result given in Lemma 2 becomes obsolete because of both the unrealistic restriction $\Delta t\le\varepsilon h$ and the fact that $Lip(G)/\varepsilon$ can be made arbitrarily big. It is therefore of interest to consider the *macroscopic variables*, which read

$$\rho=f^++f^-, \qquad J=\frac{f^+-f^-}{\varepsilon}, \tag{3.6}$$

and within which the system (3.5) can be rewritten as

$$\partial_\xi J=0, \qquad \partial_\xi\rho=-\frac{2h}{\varepsilon}G\left(\frac12(\rho+\varepsilon J),\frac12(\rho-\varepsilon J)\right)\overset{def}{=}-2hA(\rho,J,\varepsilon J). \tag{3.7}$$

Indeed, the precise form of A can be revealed, relying on the mean-value theorem:

$$A(\rho,J,\varepsilon J)=\frac1\varepsilon\underbrace{G\left(\frac\rho2,\frac\rho2\right)}_{=0}+\frac J2(\partial_+G-\partial_-G)\left(\frac{\rho+\theta\varepsilon J}{2},\frac{\rho-\theta\varepsilon J}{2}\right)$$

for some $\theta\in[0,1]$. As ρ and J also realize the first two moments of the discrete kinetic model (1.1), it can be expected that they satisfy the semilinear hyperbolic system

$$\partial_t\rho+\partial_xJ=0, \qquad \varepsilon^2\partial_tJ+\partial_x\rho=-2A(\rho,J,\varepsilon J), \tag{3.8}$$

which has been shown recently to exhibit diffusive asymptotics as $\varepsilon\to0$. More precisely, by the implicit function theorem, one goes formally from (3.8) to

$$J=-B(\rho,\partial_x\rho), \qquad \partial_t\rho=\partial_x(B(\rho,\partial_x\rho)),$$

and this limiting behavior holds rigorously, for instance, if

- $A(\rho,J,\varepsilon J)=A(\rho,J)=\rho^\alpha J$ and $B(\rho,\partial_x\rho)=\frac12\rho^{-\alpha}\partial_x\rho$, $\alpha<-1$ (see [29]),
- $A(\rho,J,\varepsilon J)=J-\frac12(\rho^2+(\varepsilon J)^2)$ and $B(\rho,\partial_x\rho)=-\frac12(\rho^2-\partial_x\rho)$ (see [8]).

Some other results are available in different contexts; see, for instance, [5, 31].

The main point of the AP schemes [17, 21] is to capture these features numerically as $\varepsilon\to0$ with a fixed (and reasonable!) $h>0$.

Following Proposition 1, we integrate (3.7) on $\xi \in [0,1]$ and, inverting (3.6), we find within the notation of (2.9) and with $\bar\rho = \bar f^+ + \bar f^-$

$$\Phi(J; f_L^+, f_R^-) \overset{def}{=} (2f_R^- + \varepsilon J) - (2f_L^+ - \varepsilon J) + 2h \int_0^1 A(\bar\rho, J, \varepsilon J) d\xi = 0. \tag{3.9}$$

We plan to apply the implicit function theorem to Φ since

$$\partial_J \Phi = 2\varepsilon + 2h \int_0^1 \partial_J A(\bar\rho, J, \varepsilon J) d\xi, \qquad \partial_J A = \frac{1}{2}(\partial_+ G - \partial_- G) > 0.$$

Therefore, the solution of (3.9) in J is given by a smooth *flux function*:

$$\begin{array}{rccc} F: & (\mathbb{R}_*^+)^2 & \to & \mathbb{R}, \\ & (f_L^+, f_R^-) & \mapsto & J = F(f_L^+, f_R^-). \end{array} \tag{3.10}$$

Moreover, we know that

$$\nabla F = \left(\frac{-1}{\partial_J \Phi} \right) \nabla \Phi.$$

Therefore we propose the following definition.

DEFINITION 1. *We say that the flux function F* (3.10) *is monotone if $F(0,0) = 0$, and it is increasing (resp., decreasing) with respect to its first (resp., second) variable:*

$$\partial_+ F \geq 0, \qquad \partial_- F \leq 0.$$

We also aim at treating (3.4) by means of the modified (partly implicit) numerical scheme one gets out of (3.1), (3.2), (3.3):

$$\begin{aligned} f_{j,n+1}^+ &= f_{j,n}^+ - \frac{\Delta t}{\varepsilon h}(f_{j,n+1}^+ - f_{j,n+1}^-) + \frac{\Delta t}{h} F(f_{j-1,n}^+, f_{j,n}^-), \\ f_{j,n+1}^- &= f_{j,n}^- + \frac{\Delta t}{\varepsilon h}(f_{j,n+1}^+ - f_{j,n+1}^-) - \frac{\Delta t}{h} F(f_{j,n}^+, f_{j+1,n}^-). \end{aligned} \tag{3.11}$$

In sharp contrast with (3.3), this emphasizes the consistency of such a discretization with a diffusive asymptotic behavior for $\rho_{j,n} = f_{j,n}^+ + f_{j,n}^-$. Of course, we keep on using the notation $f_h^\pm$ for the piecewise constant numerical approximations to (3.4) generated by (3.11).

LEMMA 3. *Let $0 \leq f_0^\pm \in L^1 \cap BV(\mathbb{R})$; if the flux function F is monotone and under the parabolic CFL condition $(\Delta t + \varepsilon h) Lip(F) \leq h$, one has for all $t > 0$, $\varepsilon > 0$*

- $\|f_h^+(.,t)\|_{L^1(\mathbb{R})} + \|f_h^-(.,t)\|_{L^1(\mathbb{R})} \leq \|f_0^+\|_{L^1(\mathbb{R})} + \|f_0^-\|_{L^1(\mathbb{R})}$,
- $TV(f_h^+(.,t)) + TV(f_h^-(.,t)) \leq TV(f_0^+) + TV(f_0^-)$,

and the scheme (3.11) *is positivity preserving.*

The aforementioned CFL condition is said to be *parabolic* because $Lip(F)$ is $O(h^{-1})$ and doesn't blow up as $\varepsilon \to 0$. It means also in most cases that $\varepsilon \leq O(h)$; in this sense, it completes the picture with (3.2), which is stable in the complementary range of parameters.

REMARK 3. *We stress that there exist many cases of interest for which the monotonicity of F can be established rigorously. For instance if $A(\rho, J) = k(\rho)J$, $k > 0$ (see* [29]*), one may use a different functional:*

$$\tilde\Phi(J; f_L^+, f_R^-) \overset{def}{=} \phi(2f_R^- + \varepsilon J) - \phi(2f_L^+ - \varepsilon J) + 2hJ = 0, \qquad \phi'(\rho) = \frac{1}{k(\rho)}. \tag{3.12}$$

Proof. For ease of writing, we denote $a = 1 + \frac{\Delta t}{\varepsilon h}$, $b = \frac{\Delta t}{\varepsilon h}$. The system (3.11) can be explicitly solved, and this is a desirable feature according to [17]:

$$\begin{aligned} f^+_{j,n+1} &= \frac{a}{a+b}\left(f^+_{j,n} + \frac{\Delta t}{h}F(f^+_{j-1,n}, f^-_{j,n})\right) + \frac{b}{a+b}\left(f^-_{j,n} - \frac{\Delta t}{h}F(f^+_{j,n}, f^-_{j+1,n})\right), \\ f^-_{j,n+1} &= \frac{b}{a+b}\left(f^+_{j,n} + \frac{\Delta t}{h}F(f^+_{j-1,n}, f^-_{j,n})\right) + \frac{a}{a+b}\left(f^-_{j,n} - \frac{\Delta t}{h}F(f^+_{j,n}, f^-_{j+1,n})\right). \end{aligned}$$

In order to control the L^1 norm, we linearize F around $(0,0)$:

$$F(f^+_{j-1,n}, f^-_{j,n}) = \partial_+F(\xi_{j-\frac{1}{2},n})f^+_{j-1,n} + \partial_-F(\xi_{j-\frac{1}{2},n})f^-_{j,n}.$$

The monotonicity property of the flux function together with the CFL restriction give

$$\begin{aligned} |f^+_{j,n+1}| \le \frac{1}{a+b}\Bigg[&|f^+_{j,n}|\left(a - \frac{b\Delta t}{h}\partial_+F(\xi_{j+\frac{1}{2},n})\right) + |f^+_{j-1,n}|\frac{a\Delta t}{h}\partial_+F(\xi_{j-\frac{1}{2},n}) \\ &+ |f^-_{j,n}|\left(b + \frac{a\Delta t}{h}\partial_-F(\xi_{j-\frac{1}{2},n})\right) - |f^-_{j+1,n}|\frac{b\Delta t}{h}\partial_-F(\xi_{j+\frac{1}{2},n})\Bigg], \\ |f^-_{j,n+1}| \le \frac{1}{a+b}\Bigg[&|f^+_{j,n}|\left(b - \frac{a\Delta t}{h}\partial_+F(\xi_{j+\frac{1}{2},n})\right) + |f^+_{j-1,n}|\frac{b\Delta t}{h}\partial_+F(\xi_{j-\frac{1}{2},n}) \\ &+ |f^-_{j,n}|\left(a + \frac{b\Delta t}{h}\partial_-F(\xi_{j-\frac{1}{2},n})\right) - |f^-_{j+1,n}|\frac{a\Delta t}{h}\partial_-F(\xi_{j+\frac{1}{2},n})\Bigg]. \end{aligned}$$

Such a convex combination ensures the positivity-preserving property for (3.11). Adding the two inequalities leads to

$$\sum_{j\in\mathbb{Z}} h(|f^+_{j,n+1}| + |f^-_{j,n+1}|) \le \sum_{j\in\mathbb{Z}} h(|f^+_{j,n}| + |f^-_{j,n}|).$$

The decay in time of the total variation in space is shown by similar arguments. □

This stability result is already enough to ensure strong compactness for the numerical approximations $f^\pm_h$ generated by (3.11) as $h \to 0$, as long as the relaxation parameter ε remains strictly positive.

3.3. The diffusive regime: Limiting behavior. We are now interested in the behavior of (3.11) as $\varepsilon \to 0$. The forthcoming result completes Lemma 3.

LEMMA 4. *Under the hypotheses of Lemma* 3, *we suppose* $\varepsilon\|\tilde{F}\|_{L^\infty} < 1$ *in the following decomposition, which holds for* F, *uniformly in* $h \ge 0$:

$$\begin{gathered} F(f^+_h, f^-_h) = (f^+_h - f^-_h)\tilde{F}(f^+_h, f^-_h) + g(f^+_h, f^-_h), \\ \tilde{F} \in C^1(\mathbb{R}^2), \quad \tilde{F}(f^+_h, f^-_h) \in L^\infty(\mathbb{R}), \quad g(f^+_h, f^-_h) \in L^1(\mathbb{R}). \end{gathered} \tag{3.13}$$

Then one has the estimates uniformly in $\varepsilon \ge 0$ *(where* C *is an absolute constant):*

- $\|f^+(t,.) - f^-(t,.)\|_{L^1(\mathbb{R})} \le \|f^+_0 - f^-_0\|_{L^1(\mathbb{R})}$
  $+ C\varepsilon(\|g\|_{L^1(\mathbb{R})} + hLip(F)(TV(f^+_0) + TV(f^-_0))),$
- $\|f^+(t,.) - f^+(s,.)\|_{L^1(\mathbb{R})} + \|f^-(t,.) - f^-(s,.)\|_{L^1(\mathbb{R})} \le \sqrt{|t-s|}$
  $\times \left[\frac{2}{\varepsilon}\|f^+_0 - f^-_0\|_{L^1(\mathbb{R})} + hLip(F)(1+C)(TV(f^+_0) + TV(f^-_0)) + C\|g\|_{L^1(\mathbb{R})}\right].$

The condition $\varepsilon\|\tilde{F}\|_{L^\infty} < 1$ roughly means that $\varepsilon < O(h)$; this kind of restriction has already been encountered in [14] in order to show compactness in the context of another relaxation problem. The technical assumption (3.13) will be checked in the numerical examples later on: it expresses the fact that the Maxwellian distribution for (3.4), (1.1) should be given by $f^+ = f^-$. (In the case of the Goldstein–Taylor model, the decomposition is trivial, $\tilde{F} \equiv \frac{1}{h+\varepsilon}$, $g \equiv 0$; see [13].)

Proof. We keep on using the same notations; from (3.11), we get

$$(1+2b)(f^+_{j,n+1} - f^-_{j,n+1}) = (f^+_{j,n} - f^-_{j,n}) + \frac{2\Delta t}{h} F(f^+_{j,n}, f^-_{j,n})$$
$$+ \frac{\Delta t}{h}[F(f^+_{j-1,n}, f^-_{j,n}) + F(f^+_{j,n}, f^-_{j+1,n}) - 2F(f^+_{j,n}, f^-_{j,n})].$$

We now use the decomposition (3.13) in order to get

$$(1+2b)|f^+_{j,n+1} - f^-_{j,n+1}| \leq |f^+_{j,n} - f^-_{j,n}| \left(1 + \frac{2\Delta t}{h}\|\tilde{F}\|_{L^\infty}\right)$$
$$+ \frac{2\Delta t}{h}|g(f^+_{j,n}, f^-_{j,n})| + \frac{\Delta t}{h} Lip(F)[|f^+_{j-1,n} - f^+_{j,n}| + |f^-_{j+1,n} - f^-_{j,n}|].$$

Thus an elementary computation shows that

$$\alpha \overset{def}{=} \frac{1 + \frac{2\Delta t}{h}\|\tilde{F}\|_{L^\infty}}{1 + \frac{2\Delta t}{\varepsilon h}} < 1 \qquad \Leftrightarrow \qquad \varepsilon\|\tilde{F}\|_{L^\infty} < 1.$$

This implies that

$$\|f^+(t,.) - f^-(t,.)\|_{L^1(\mathbb{R})} \leq \|f^+_0 - f^-_0\|_{L^1(\mathbb{R})}$$
$$+ \frac{\varepsilon}{1-\alpha}(\|g\|_{L^1(\mathbb{R})} + h.Lip(F)[TV(f^+_0) + TV(f^-_0)]),$$

and the control on the Maxwellian distribution follows.

Concerning the L^1-modulus of continuity in time, following [13], we rewrite the first equation in (3.11) as

$$f^+_{j,n+1} - f^+_{j,n} + \frac{\Delta t}{\varepsilon h}(f^+_{j,n+1} - f^+_{j,n}) = \frac{\Delta t}{\varepsilon h}(f^-_{j,n+1} - f^-_{j,n}) - \frac{\Delta t}{\varepsilon h}(f^+_{j,n} - f^-_{j,n}) + \frac{\Delta t}{h}F(f^+_{j-1,n}, f^-_{j,n}),$$

in order to derive for all $j \in \mathbb{Z}$, $n \in \mathbb{N}$,

$$|f^+_{j,n+1} - f^+_{j,n}|\left(1 + \frac{\Delta t}{\varepsilon h}\right) - \frac{\Delta t}{\varepsilon h}|f^-_{j,n+1} - f^-_{j,n}| \leq \frac{\Delta t}{\varepsilon h}|f^+_{j,n} - f^-_{j,n}| + \frac{\Delta t}{h}|F(f^+_{j-1,n}, f^-_{j,n})|,$$

together with a similar expression for the f^- variable. Therefore, we arrive at

$$|f^+_{j,n+1} - f^+_{j,n}| + |f^-_{j,n+1} - f^-_{j,n}| \leq \frac{2\Delta t}{\varepsilon h}|f^+_{j,n} - f^-_{j,n}|$$
$$+ \frac{\Delta t}{h} Lip(F)[|f^+_{j-1,n} - f^+_{j,n}| + |f^-_{j+1,n} - f^-_{j,n}|],$$

and we are done with the second inequality of Lemma 4, just summing on $j \in \mathbb{Z}$ and noticing that under a parabolic CFL condition, $\frac{\Delta t}{h} = O(h) = O(\sqrt{\Delta t})$. □

As a consequence of Lemma 4, strong compactness as $\varepsilon \to 0$, $h > 0$ fixed follows as soon as one provides a so-called well-prepared initial datum, that is to say,

$$\|f_0^+ - f_0^-\|_{L^1(\mathbb{R})} = O(\varepsilon). \tag{3.14}$$

Such an initialization for (3.4) cancels any kind of *initial layer*, which would destroy the Hölder time regularity of the process (see also [35] for a similar remark) in $t = 0+$.

With obvious notation, we can easily deduce the asymptotic behavior of the scheme (3.11) as $\varepsilon \to 0$; adding the two equations, we derive

$$\rho_{j,n+1} = \rho_{j,n} + \frac{\Delta t}{h}\left(F\left(\frac{\rho_{j-1,n}}{2}, \frac{\rho_{j,n}}{2}\right) - F\left(\frac{\rho_{j,n}}{2}, \frac{\rho_{j+1,n}}{2}\right)\right) + \mathcal{R}_{j,n}, \tag{3.15}$$

and the estimates from Lemma 4 ensure that the remaining term is of the order of ε in L^1 as a consequence of the smoothness of F. This has already been evidenced by explicit computations with the Goldstein–Taylor model in [13].

Remark 4. *At this point, we underline that, relying on the monotonicity of F,* (3.13) *can be easily fulfilled using the mean-value theorem:*

$$F(f^+, f^-) = (f^+ - f^-)\partial_+ F(\xi) + \operatorname{div}(F)(\xi) f^-, \qquad \xi \in (\mathbb{R}^+)^2. \tag{3.16}$$

Since $f^\pm(t,.) \in L^1(\mathbb{R})$, what we need is just $\operatorname{div}(F) = O(1)$ *uniformly in $h \geq 0$. It turns out that this corresponds to a splitting of the flux function F between a diffusive and a convective part when looking at the scheme* (3.15) *since it can be rewritten as*

$$\rho_{j,n+1} = \rho_{j,n} + \frac{\Delta t}{2h}\partial_+ F(\xi_{j,n})(\rho_{j+1,n} - 2\rho_{j,n} + \rho_{j-1,n}) - \frac{\Delta t}{2h}\operatorname{div}(F)(\xi_{j,n})(\rho_{j+1,n} - \rho_{j,n}).$$

One readily checks that this scheme is L^∞-stable; its L^1 and BV stability are just particular cases of Lemma 3.

As a byproduct of Lemmas 3 and 4, we can state the following theorem, which deals with the cases investigated by Lions and Toscani in [29].

Theorem 1. *Assume that $0 \leq f_0^\pm \in L^1 \cap BV(\mathbb{R})$ are initial data for* (3.4) *with $G(f^+, f^-) = (f^+ + f^-)^\alpha (f^+ - f^-)$, $\alpha \leq 0$, and satisfy* (3.14). *Then as $h, \varepsilon \to 0$ under the prescribed CFL conditions, the sequence $f_h^\pm$ generated by the scheme* (3.11) *is relatively compact in the strong topology of $L^1_{loc}(\mathbb{R}^+_* \times \mathbb{R})$. In particular, $\rho_h = f_h^+ + f_h^-$ converges towards the unique solution in the sense of distributions to*

$$\partial_t \rho = \frac{1}{2}\partial_x(\rho^{-\alpha}\partial_x \rho), \qquad \rho(t=0,.) = 2f_0^+ = 2f_0^-.$$

Proof. We are precisely in position to use the modified functional $\tilde{\Phi}$ in (3.12) since in the notation of Remark 3, $k(\rho) = \rho^\alpha \geq 0$. Thus the flux function F exists and is monotone in the sense of Definition 1. Moreover, the Maxwellian distribution is $f^+ = f^-$, and we can take $g \equiv 0$ in (3.13) since by the mean-value theorem we get some expression for the flux function from (3.12):

$$F(f_L^+, f_R^-) = \frac{f_L^+ - f_R^-}{h.k(\zeta) + \varepsilon}, \qquad \zeta \in \mathbb{R}^+. \tag{3.17}$$

The conclusion thus follows from Lemma 4 and (3.15). □

We left behind the cases $0 < \alpha \leq 1$ (fast diffusion equations) as $\tilde{\Phi}$ can become singular if $\rho = 0$; this isn't an issue in practical computations (see section 4.3).

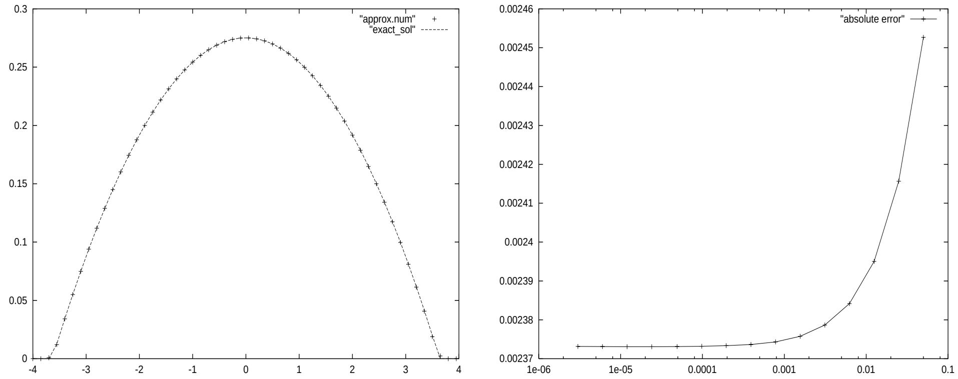


FIG. 4.1. *Numerical results for* (3.11) *on Barenblatt's problem in* $T = 3$*:* $3.10^{-6} \leq \varepsilon \leq 5.10^{-2}$.

4. Numerical experiments.

We illustrate in this section the results previously stated by means of some numerical runs of increasing difficulty inspired by those in [13, 19].

4.1. The porous media equation.

We simulate here the so-called Barenblatt's problem [3], which consists of finding a particular solution to the porous media equation whose analytical expression is known. More precisely, we select $A(\rho, J) = \frac{J}{4\rho}$ in (3.8), which gives $\alpha = -1$ in Theorem 1, and we look for

$$\rho(t,x) = \frac{1}{r(t)} \left(1 - \left(\frac{x}{r(t)} \right)^2 \right) \mathbf{1}_{|x| \leq r(t)}, \qquad r(t) = \big(12(1+t)\big)^{\frac{1}{3}}.$$

In this case, (3.12) is solved explicitly, and the flux function is given by

$$F(f_L^+, f_R^-) = \frac{(f_L^+)^2 - (f_R^-)^2}{(h/4) + \varepsilon(f_L^+ + f_R^-)}.$$

The partial derivatives do not change signs, whatever values ε, h take, since the scheme (3.11) preserves positivity and

$$\partial_+ F(f_L^+, f_R^-) = \frac{2hf_L^+ + 4\varepsilon(f_L^+ + f_R^-)^2}{\big(h + 4\varepsilon(f_L^+ + f_R^-)\big)^2}, \qquad \partial_- F(f_L^+, f_R^-) = -\frac{2hf_R^- + 4\varepsilon(f_L^+ + f_R^-)^2}{\big(h + 4\varepsilon(f_L^+ + f_R^-)\big)^2}.$$

Notice that, since we deal with a quadratic nonlinearity, the value of ζ in (3.17) is simply given by an arithmetic average. Numerical results are shown in Figure 4.1 in time $T = 3$, with the parameters $h = 0.15$ and $\Delta t = 0.01$. On the right, the absolute error between (3.11) and the exact solution is displayed as a function of ε; it stalls below a certain value as the value of h becomes a limiting factor.

4.2. The advection-diffusion equation.

We move on now to another equation which has been investigated from the relaxation point of view in [20, 5, 2]:

$$(4.1) \qquad \partial_t \rho + \partial_x \rho = \frac{1}{2} \partial_{xx} \rho, \qquad x \in \mathbb{R}, \qquad t > 0.$$

Here, we apply the same program, relying on (3.8) together with the right-hand side $A(\rho, J) = J - \rho$. Once again, we are able to solve (3.7), and the flux function F is

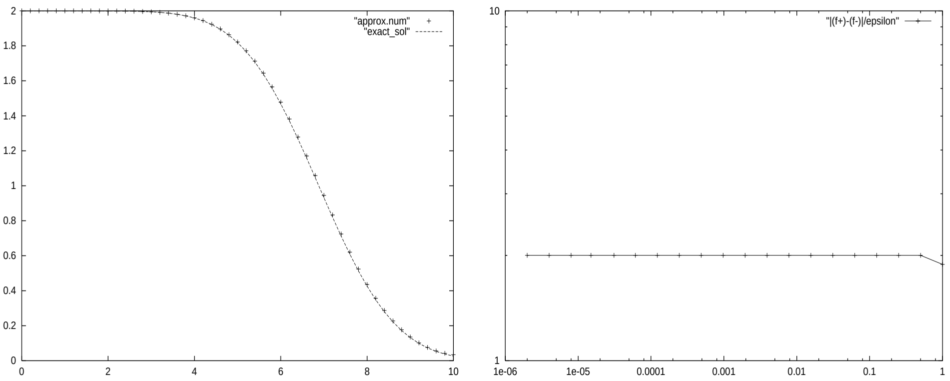


FIG. 4.2. *Numerical results for* (3.11), (4.1) *on a Riemann problem:* $2.10^{-6} \leq \varepsilon \leq 1$.

monotone since it comes out to be

$$F(f_L^+, f_R^-) = 2\frac{\exp(2h)f_L^+ - f_R^-}{\exp(2h) - 1 + \varepsilon(1 + \exp(2h))}.$$

By means of Lemma 3, compactness holds in $h \geq 0$ under the CFL condition

$$\frac{2(\Delta t + \varepsilon h)}{1 - \exp(-2h)} \leq h \qquad \Rightarrow \qquad 0 \leq \Delta t = O(h^2),\ \varepsilon < h,$$

and one can check therefore that for (3.16),

$$\operatorname{div}(F) = \frac{2}{1 + \varepsilon\left(\frac{\exp(2h)+1}{\exp(2h)-1}\right)} = O(1), \qquad 0 \leq h \leq 1.$$

Moreover, concerning (3.13) and Lemma 4, we can choose $\tilde{F} = \partial_+ F$ and we get

$$\varepsilon\|\tilde{F}\|_{L^\infty} < 1 \quad \Leftrightarrow \quad \varepsilon < h < 1.$$

Therefore, gathering inside Lemmas 3, 4 and Remark 4, we obtain an analogue of Theorem 1 for (4.1).

THEOREM 2. *Assume that* $0 \leq f_0^\pm \in L^1 \cap BV(\mathbb{R})$ *are initial data for* (3.6), (3.8) *with* $A(\rho, J) = J - \rho$ *and satisfy* (3.14). *Then as* $h, \varepsilon \to 0$ *under the prescribed conditions, the sequence* $f_h^\pm$ *generated by the scheme* (3.11) *is relatively compact in the strong topology of* $L^1_{loc}(\mathbb{R}_*^+ \times \mathbb{R})$. *In particular,* $\rho_h = f_h^+ + f_h^-$ *converges towards the unique solution to* (4.1), $\rho(t = 0, .) = 2f_0^+ = 2f_0^-$.

We close this paragraph by presenting some numerical results which illustrate our statements. We choose some Maxwellian initial data $f_0^\pm(x) = \mathbf{1}_{x<5}$ to compute the solution of (3.8) in $x \in [0, 10]$ for $T = 2$. Of course, the exact solution of (4.1) is given by $\rho(t, x) = 1 - \operatorname{erf}((x - t - 5)/\sqrt{2t})$. We took $h = 0.2$ and $\Delta t = 0.02$; see Figure 4.2.

4.3. The Carleman model. We consider the so-called Carleman's model [6], which corresponds to the choice $A(\rho, J) = \rho J$, that is to say, $\alpha = 1$ in Theorem 1. This presents a difficulty as the asymptotic behavior is singular if $\rho = 0$; anyway, we succeeded in simulating the following *initial-boundary value problem* by means of the

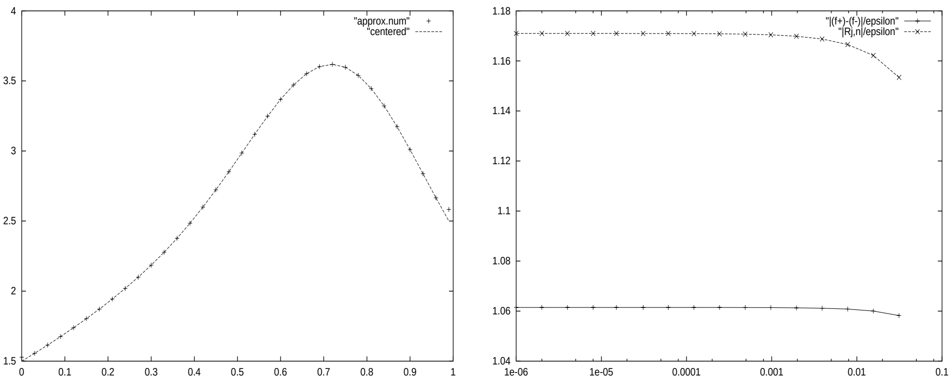


FIG. 4.3. *Transient regime (*$T = 0.01$*) for Carleman's model on the IBVP (initial boundary value problem)* (4.2): $10^{-6} \leq \varepsilon \leq 3.10^{-2}$.

scheme (3.11):

$$\begin{aligned} &\partial_t \rho = \frac{1}{2}\partial_{xx}(\ln(\rho)), \qquad \rho(0,.) = 1 + \mathbf{1}_{x>0.5}; \\ &\rho(.,x=0) = 1.5, \qquad \rho(.,x=1) = 2.5. \end{aligned} \tag{4.2}$$

The computational domain is $x \in [0,1]$, and we took $h = 0.03$, $\Delta t = 0.001$ in order to produce the results displayed in Figure 4.3. In this case, the flux function cannot be known analytically, but it turns out from (3.12) that the equation

$$F(f_L^+, f_R^-) = J = \frac{1}{2h}\ln\left(\frac{2f_L^+ - \varepsilon J}{2f_R^- + \varepsilon J}\right)$$

can be easily solved by means of a fixed point algorithm if ε is small enough. We compared our results with those generated by a standard second order centered scheme for (4.2): the absolute error between them is of the order of ε, as announced in (3.15).

4.4. The Ruijgrok–Wu model. We close this section devoted to numerical tests with the Ruijgrok and Wu model of the Boltzmann equation [8, 29, 30], which relaxes under a diffusive scaling of variables towards the viscous Burgers equation [16]. More precisely, we aim at reproducing with (3.11) the smooth solution of the initial-boundary value problem

$$\begin{aligned} &\partial_t \rho + \rho\partial_x \rho = \frac{1}{2}\partial_{xx}\rho, \qquad \rho(0,.) = 2\big(2 - \mathbf{1}_{x>0}\big); \\ &\rho(.,x=-1) = 4, \qquad \rho(.,x=1) = 2, \end{aligned} \tag{4.3}$$

by means of (3.8) with the right-hand side $A(\rho, J, \varepsilon J) = J - \frac{\varepsilon^2}{2}J^2 - \frac{1}{2}\rho^2$.

Several features make this difficult: first is that, according to [8, 30], this 2×2 system is not unconditionally quasi-monotone. Indeed, this property holds only in case the initial data satisfy $f_0^- \leq 1/2\varepsilon$; this has to be taken into account when starting the simulation. Moreover, the differential equation (3.7) cannot always be solved analytically, and we integrated it approximately through the midpoint rule. Thus the flux function we used comes out of a fixed point algorithm on the following equation, which is deduced from (3.9):

$$J = \frac{1}{2h}\left[(2f_L^+ - \varepsilon J)\left(1 + \frac{h}{2}(2f_L^+ - \varepsilon J)\right) - (2f_R^- + \varepsilon J)\left(1 - \frac{h}{2}(2f_R^- + \varepsilon J)\right) + \varepsilon h J^2\right].$$

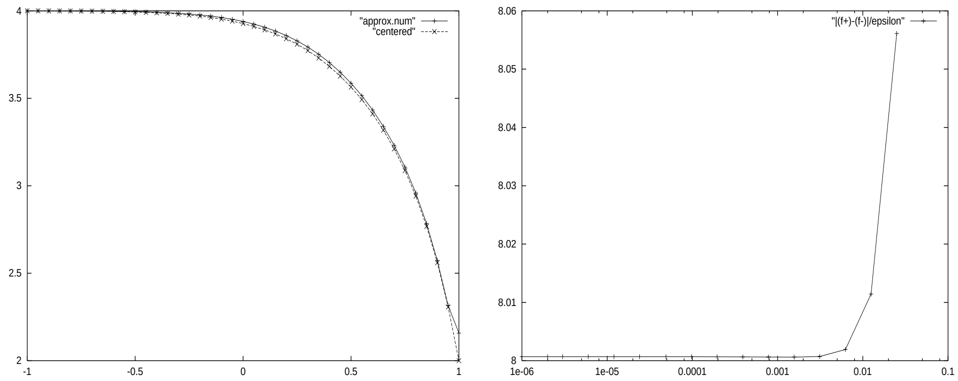


FIG. 4.4. *Transient regime ($T = 0.3$) for Ruijgrok and Wu's model on* (4.3): $10^{-6} \le \varepsilon \le 3.10^{-2}$.

This choice led to the results displayed in Figure 4.4 with the parameters $h = 0.05$, $\Delta t = 0.001$. This demonstrates the fact that an approximate treatment of the jump relation (3.7) still gives a good outcome on a nonlinear problem, and the Maxwellian estimate of Lemma 4 is kept.

5. Conclusion. We presented in this paper a study of a well-balanced scheme for discrete 2×2 kinetic models. This scheme is endowed with several interesting properties as it preserves steady-state curves in the rarefied regime (1.1) and is asymptotic preserving in the sense of [17] as $\varepsilon \to 0$ in (3.4). Moreover, these statements can be rigorously established in many significant situations like, for instance, the cases investigated in [29]. From this point of view, this work can be seen as an extension of [10, 11] to relaxation problems with a diffusive behavior. The present approach may be further developed in several directions; first, multidimensional problems could be considered, then more complex asymptotics could also be tackled in the spirit of [31]. In any case, intermediate scalings considered in [8, 32] (the so-called *incompressible Euler limits*) relaxing to conservation laws can be handled by a suitable modification of (3.5) inside the scheme (3.11).

REFERENCES

[1] D. Amadori, L. Gosse, and G. Guerra, *Global BV entropy solutions and uniqueness for hyperbolic systems of balance laws*, Arch. Ration. Mech. Anal., 162 (2002), pp. 327–366.

[2] D. Aregba-Driollet, R. Natalini, and S.Q. Tang, *Diffusive kinetic explicit schemes for nonlinear degenerate parabolic systems*, Math. Comp., to appear.

[3] G.I. Barenblatt, *On some steady motion of a liquid or a gas in a porous medium*, Prikl. Mat. Mekh., 16 (1952), pp. 67–78.

[4] R. Botschorijvili, B. Perthame, and A. Vasseur, *Equilibrium schemes for scalar conservation laws with stiff source terms*, Math. Comp., 72 (2003), pp. 131–157.

[5] F. Bouchut, F.R. Guarguaglini, and R. Natalini, *Diffusive BGK approximations for nonlinear multidimensional parabolic equations*, Indiana Univ. Math. J., 49 (2000), pp. 723–749.

[6] T. Carleman, *Problèmes mathématiques de la théorie cinétique des gaz*, Almqvist and Wiksell, Uppsala, Sweden, 1957.

[7] C. Cercignani, R. Illner, and M. Pulvirenti, *The Mathematical Theory of Dilute Gases*, Appl. Math. Sci. 106, Springer-Verlag, New York, 1994.

[8] E. Gabetta and B. Perthame, *Scaling limits for the Ruijgrok-Wu model of the Boltzmann equation*, Math. Methods Appl. Sci., 24 (2001), pp. 949–967.

[9] S.K. Godunov, *Finite difference schemes for numerical computation of solutions of the equations of fluid dynamics*, Mat. USSR Sbornik, 47 (1959), pp. 271–306.

[10] L. Gosse, *Localization effects and measure source terms in numerical schemes for balance laws*, Math. Comp., 71 (2002), pp. 553–582.

[11] L. Gosse, *Time-splitting schemes and measure source terms for a quasilinear relaxing system*, Math. Models Methods Appl. Sci., to appear.

[12] L. Gosse, *A well-balanced scheme using non-conservative products designed for hyperbolic systems of conservation laws with source terms*, Math. Models Methods Appl. Sci., 11 (2001), pp. 339–365.

[13] L. Gosse and G. Toscani, *An asymptotic-preserving well-balanced scheme for the hyperbolic heat equations*, C.R. Math. Acad. Sci. Paris Sér. I Math., 334 (2002), pp. 337–342.

[14] L. Gosse and A.E. Tzavaras, *Convergence of relaxation schemes to the equations of elastodynamics*, Math. Comp., 70 (2001), pp. 555–577.

[15] J.M. Greenberg and A.Y. LeRoux, *A well-balanced scheme for the numerical processing of source terms in hyperbolic equations*, SIAM J. Numer. Anal., 33 (1996), pp. 1–16.

[16] E. Hopf, *The partial differential equation* $u_t + uu_x = \mu u_{xx}$, Comm. Pure Appl. Math., 3 (1950), pp. 201–230.

[17] S. Jin, *Efficient asymptotic-preserving (AP) schemes for some multiscale kinetic equations*, SIAM J. Sci. Comput., 21 (1999), pp. 441–454.

[18] S. Jin, *A steady-state capturing method for hyperbolic systems with geometrical source terms*, M2AN Math. Model. Numer. Anal., 35 (2001), pp. 631–645.

[19] F. Golse, S. Jin, and C.D. Levermore, *The convergence of numerical transfer schemes in diffusive regimes* I: *Discrete-ordinate method*, SIAM J. Numer. Anal., 36 (1999), pp. 1333–1369.

[20] S. Jin and H. Liu, *Diffusion limit of a hyperbolic system with relaxation*, Methods Appl. Anal., 5 (1998), pp. 317–334.

[21] S. Jin, L. Pareschi, and G. Toscani, *Diffusive relaxation schemes for multiscale discrete-velocity kinetic equations*, SIAM J. Numer. Anal., 35 (1998), pp. 2405–2439.

[22] A. Klar, *An asymptotic-induced scheme for nonstationary transport equations in the diffusive limit*, SIAM J. Numer. Anal., 35 (1998), pp. 1073–1094.

[23] A. Klar and A. Unterreiter, *Uniform stability of a finite difference scheme for transport equations in diffusive regimes*, SIAM J. Numer. Anal., 40 (2002), pp. 891–913.

[24] S.N. Kružkov, *First order quasilinear equations in several independent space variables*, Mat. USSR Sbornik, 81 (1970), pp. 228–255.

[25] Ph. G. LeFloch and A.E. Tzavaras, *Representation of weak limits and definition of non-conservative products*, SIAM J. Math. Anal., 30 (1999), pp. 1309–1342.

[26] R.J. LeVeque, *Balancing source terms and flux gradients in high resolution Godunov methods: The quasi steady wave propagation algorithm*, J. Comput. Phys., 146 (1998), pp. 346–365.

[27] R.J. LeVeque and B. Temple, *Stability of Godunov's method for a class of* 2×2 *systems of conservation laws*, Trans. Amer. Math. Soc., 288 (1985), pp. 115–123.

[28] P.L. Lions, *Mathematical topics in fluid mechanics* I, *Incompressible models*, Oxford University Press, London, 1996.

[29] P.L. Lions and G. Toscani, *Diffusive limit for finite velocity Boltzmann kinetic models*, Rev. Mat. Iberoamericana, 13 (1997), pp. 473–513.

[30] H. Liu, J. Wang, and G. Warnecke, *Convergence of a splitting scheme applied to the Ruijgrok–Wu model of the Boltzmann equation*, J. Comput. Appl. Math., 134 (2001), pp. 343–367.

[31] P. Marcati and A. Milani, *The one-dimensional Darcy's law as the limit of a compressible Euler flow*, J. Differential Equations, 84 (1990), pp. 129–147.

[32] G. Naldi and L. Pareschi, *Numerical schemes for hyperbolic systems of conservation laws with stiff diffusive relaxation*, SIAM J. Numer. Anal., 37 (2000), pp. 1246–1270.

[33] R. Natalini and B. Hanouzet, *Weakly coupled systems of quasilinear hyperbolic equations*, Differential Integral Equations, 9 (1997), pp. 1279–1292.

[34] B. Perthame and C. Simeoni, *A kinetic scheme for the Saint-Venant system with a source term*, Calcolo, 38 (2001), pp. 201–231.

[35] B. Perthame and E. Tadmor, *A kinetic equation with kinetic entropy functions for scalar conservation laws*, Comm. Math. Phys., 136 (1991), pp. 501–517.

[36] T. Platkowski and R. Illner, *Discrete velocity models of the Boltzmann equation: A survey on the mathematical aspects of the theory*, SIAM Rev., 30 (1988), pp. 213–255.

[37] W. Ruijgrok and T.T. Wu, *A completely solvable model of the nonlinear Boltzmann equation*, Phys. A, 113 (1982), pp. 401–416.

[38] A.E. Tzavaras, *On the mathematical theory of fluid dynamic limits to conservation laws*, in Advances in Mathematical Fluid Mechanics, J. Malek, J. Nečas, and M. Rokyta, eds., Springer, New York, 2000, pp. 192–222.

[39] A.I. Vol'Pert, *Spaces BV and quasilinear equations*, Mat. USSR Sbornik, 2 (1967), pp. 225–267.

SIAM J. Numer. Anal.
Vol. 41, No. 2, pp. 659–672

OPTIMAL ERROR ESTIMATES OF THE CHEBYSHEV–LEGENDRE SPECTRAL METHOD FOR SOLVING THE GENERALIZED BURGERS EQUATION*

HUA WU†, HEPING MA†, AND HUIYUAN LI†‡

Abstract. In this paper the Chebyshev–Legendre collocation method is applied to the generalized Burgers equation. Optimal error estimate of the method is proved for the problem with the Dirichlet boundary conditions. Also, a Legendre–Galerkin–Chebyshev collocation method is given for the generalized Burgers equation. The scheme is basically formulated in the Legendre spectral form but with the nonlinear term being treated by the Chebyshev collocation method so that the scheme can be implemented at Chebyshev–Gauss–Lobatto points efficiently. Optimal order convergence is also obtained through coupling estimates in the L^2-norm and the H^1-norm.

Key words. optimal error estimate, Chebyshev–Legendre method, generalized Burgers equation

AMS subject classifications. 65M12, 65M70

PII. S0036142901399781

1. Introduction.

The spectral method is used widely in seeking numerical solutions of partial differential equations because of its "infinite order" of convergence [4, 2, 9]. It is well known that the Chebyshev method is easier to implement than the Legendre method, but it is more difficult to justify in numerical analysis than the latter. In 1994, Don and Gottlieb first introduced the Chebyshev–Legendre (CL) method [6], where the Legendre method was implemented on Chebyshev points. The boundary conditions were imposed via a penalty technique, and the scheme was in a collocation form. It was shown that the method was stable in the unweighted L^2-norm, and error estimates were given for linear problems [6, 9]. In 1989, Reyna [21] analyzed an L^2-estimate of a modified Chebyshev collocation method for a linear system of equations. Shen [24, 25] recommended the CL method for elliptic problems with simple schemes and efficient implementation. Ma [15, 16] applied CL viscosity methods to the nonlinear conservation laws. Error analysis of a similar CL method but using the Chebyshev–Gauss (CG) points, which are not as common as the Chebyshev–Gauss–Lobatto (CGL) points, can be found in [13].

In this paper, we consider CL methods for the following generalized Burgers equation:

$$
(1.1)\qquad \begin{cases} \partial_t U(x,t) + \partial_x F(U(x,t)) - \nu \partial_x^2 U(x,t) = 0, & (x,\,t) \in (-1,\,1) \times (0,\,T), \\ \alpha U(1,\,t) + \beta \partial_x U(1,\,t) = g^+(t), & t \in (0,\,T), \\ \gamma U(-1,\,t) + \delta \partial_x U(-1,\,t) = g^-(t), & t \in (0,\,T), \\ U(x,0) = U_0(x), & x \in (-1,\,1), \end{cases}
$$

where $F(z)$ is a smooth function of z, the parameters α, γ, β are nonnegative, and δ is nonpositive. Much work has been done on numerical analysis of spectral methods

*Received by the editors December 17, 2001; accepted for publication (in revised form) November 13, 2002; published electronically May 6, 2003. This work was supported by the Special Funds for State Major Basic Research Projects of China N.G1999032804.

http://www.siam.org/journals/sinum/41-2/39978.html

†Department of Mathematics, Shanghai University, Shanghai 200436, People's Republic of China (hwu@mail.shu.edu.cn, hpma@mail.shu.edu.cn).

‡Institute of Software, Chinese Academy of Sciences, Beijing 100080, People's Republic of China (hynli@mail.rdcps.ac.cn).

for Burgers' equation [18, 19, 3, 17, 7], where error estimates of the Legendre spectral/pseudospectral method and the Chebyshev spectral/pseudospectral method for Burgers' equation have been established. The key to the numerical analysis of the Chebyshev method is the coercive property of the nonsymmetric bilinear form, which is true if the function vanishes on the boundaries. This becomes unavailable at the interface points when the method is used with domain decompositions. In order to overcome this difficulty, a modified Chebyshev pseudospectral method is given in [8] which introduced linear polynomials so that the new unknown vanishes at interface points. The CL methods enjoy both advantages of easy implementation of the Chebyshev method and good stability of the Legendre method for the nonlinear problems and can be applied to multidomain cases without such a difficulty. The aim of this paper is to give optimal error estimates of CL methods for solving the generalized Burgers equation. Besides the Chebyshev–Legendre collocation (CLC) method in [6], a Legendre–Galerkin–Chebyshev collocation (LGCC) scheme is presented. The scheme is basically formulated in the Legendre–Galerkin form but with the nonlinear term being treated with the Chebyshev collocation method. Here the CGL points are adopted. By combining Galerkin and collocation methods, the scheme seems more flexible and easier to be generalized to multidomain approaches. In numerical analysis of such methods, we need to consider the stability and approximation properties of the Chebyshev interpolation operator in the L^2-norm rather than in the Chebyshev weighted norm. Also, due to the property of the Chebyshev interpolation operator, we find that it is difficult to get the desired L^2-estimate directly for our fully discrete scheme. This is why an H^1-estimate is involved in analysis. Optimal convergence rate of the methods is obtained through combining L^2- and H^1-estimates.

Given the history of success with the preconditioning methods discussed in [20, 5, 10, 11, 12], such methods will probably be successful in the CLC case, which may be more preferable for high dimensional problems. The LGCC scheme can be solved by efficient direct solvers developed in [22, 23].

2. Schemes. Let $I = (-1, 1)$. Denote by $(\cdot,\cdot)$ and $\|\cdot\|$ the inner product and norm of the space $L^2(I)$, respectively. For $\sigma > 0$, let $H^\sigma(I)$ be the classical Sobolev space equipped with the norm $\|\cdot\|_\sigma$ and the seminorm $|\cdot|_\sigma$. Let $\mathbb{P}_N$ be the set of all algebraic polynomials of degree at most N.

We first apply the CLC method in [6] to the following problem (1.1) [9]: Find $u \in \mathbb{P}_N$ such that for $0 \le j \le N$,

$$\begin{cases} \partial_t u(x_j, t) + (\partial_x I_N^C F(u))(x_j, t) - \nu\partial_x^2 u(x_j, t) = -R(x_j, t), \\ u(x_j, 0) = U_0(x_j), \end{cases} \tag{2.1}$$

where $x_j = \cos(\frac{\pi j}{N})$ $(0 \le j \le N)$ are the CGL points, I_N^C is the Chebyshev interpolation operator at the CGL points, and [6]

$$R(x, t) = \tau_0 Q^+(x)[B^+(t) - g^+(t)] + \tau_N Q^-(x)[B^-(t) - g^-(t)]$$

with

$$Q^+(x) = \frac{(1+x)L_N'(x)}{2L_N'(1)}, \qquad B^+(t) = \alpha u(1, t) + \beta\partial_x u(1, t),$$

$$Q^-(x) = \frac{(1-x)L_N'(x)}{2L_N'(-1)}, \qquad B^-(t) = \gamma u(-1, t) + \delta\partial_x u(-1, t),$$

where $L_N(x)$ is the Legendre polynomial of degree N. For the time advance, we adopt the second order Crank–Nicolson/leapfrog (CNLF) scheme. Let τ be the mesh size in time, and let $S_t = \{k\tau \ : \ k = 1, 2, \ldots, n_t, \ t = n_t\tau\}$. The notations $v_{\hat{t}}(t)$ and $\hat{v}(t)$ are used as

$$v_{\hat{t}}(t) = \frac{v(t+\tau) - v(t-\tau)}{2\tau}, \qquad \hat{v}(t) = \frac{1}{2}[v(t+\tau) + v(t-\tau)].$$

The fully discrete CL scheme for (1.1) is to find $u \in \mathbb{P}_N$ such that for $0 \le j \le N$,

$$\text{(2.2)} \qquad \begin{cases} u_{\hat{t}}(x_j, t) + (\partial_x I_N^C F(u))(x_j, t) - \nu\partial_x^2 \hat{u}(x_j, t) = -\hat{R}(x_j, t), \quad t \in S_{T-\tau}, \\ u(x_j, \tau) = U_0(x_j) + \tau\partial_t U(x_j, 0), \\ u(x_j, 0) = U_0(x_j), \end{cases}$$

where $\partial_t U(x, 0) = \nu\partial_x^2 U_0(x) - \partial_x F(U_0(x))$ is computed via (1.1).

We next give an LGCC scheme. For simplicity, we consider only the homogenous Dirichlet boundary condition. Other kinds of cases can be treated as in [22]. Define the approximation space

$$V_N^0 = \mathbb{P}_N \cap H_0^1(I), \qquad H_0^1(I) = \left\{v \in H^1(I) \ : \ v(-1) = v(1) = 0\right\}.$$

The fully discrete LGCC scheme for (1.1) is to find $u(t) \in V_N^0$ such that

$$\text{(2.3)} \qquad \begin{cases} (u_{\hat{t}},\, v) + (\partial_x I_N^C F(u),\, v) + \nu(\partial_x \hat{u},\, \partial_x v) = 0 \quad \forall v \in V_N^0, \quad t \in S_{T-\tau}, \\ u(\tau) = I_N^C(U_0 + \tau\partial_t U(0)), \\ u(0) = I_N^C U_0, \end{cases}$$

where $\partial_t U(0) = \nu\partial_x^2 U_0 - \partial_x F(U_0)$. We choose appropriate base functions of V_N^0 as in [22] to set up the corresponding system of equations. For $0 \le n \le N-2$, let $c_n = 1/(2n+1)$ and $\phi_n(x) = c_{n+1}[L_n(x) - L_{n+2}(x)]$, where $\{L_n(x)\}$ are the Legendre polynomials, so that $\partial_x\phi_n(x) = -L_{n+1}(x)$. Expanding $u(x,t) = \sum_{n=0}^{N-2} a_n(t)\phi_n(x)$ and taking $v = \phi_m$ in (2.3) lead to

$$\text{(2.4)} \quad \begin{aligned} \sum_{n=0}^{N-2} [(\phi_n, \phi_m) + 2c_{n+1}\nu\tau\delta_{mn}]a_n(t+\tau) &= -2\tau(I_N^C F(u(t)), L_{m+1}) \\ &\quad + \sum_{n=0}^{N-2} [(\phi_n, \phi_m) - 2c_{n+1}\nu\tau\delta_{mn}]a_n(t-\tau), \quad 0 \le m \le N-2, \ t \in S_{T-\tau}. \end{aligned}$$

The matrix of above system is pentadiagonal [22]. We note that the nonlinear term in (2.4) can be computed by the fast Legendre transform (FLT) [1] between the coefficients of the Legendre series and its values at the CGL points, such as

$$\{a_n\} \xrightarrow{FLT} \{u(x_j)\} \to \{F(u(x_j)\} \xrightarrow{FLT} \{(\widehat{I_N^C F(u)})_n^L\},$$

where $(\widehat{I_N^C F(u)})_n^L$ are the Legendre expansion coefficients of $I_N^C F(u)$.

3. Preliminary. In this section, we introduce a suitable comparison function and give some lemmas needed in error analysis. We shall denote by C a generic positive constant independent of N or any function. Define $P_{1,N} : H^1(I) \to \mathbb{P}_N$ such that

$$P_{1,N}u(x) = u(-1) + \int_{-1}^{x} P_{N-1}^{L} \partial_y u \, dy,$$

where $P_N^L : L^2(I) \to \mathbb{P}_N$ denotes the Legendre orthogonal projection operator. We have from the definition of $P_{1,N}$ immediately that

$$(\partial_x P_{1,N}u, \partial_x v) = (\partial_x u, \partial_x v) \qquad \forall\, v \in \mathbb{P}_N. \tag{3.1}$$

Also, it is easy to see that $P_{1,N}u - u \in H_0^1(I)$ and

$$\begin{aligned}(P_{1,N}u - u, v) &= (P_{1,N}u - u, \partial_x \partial_x^{-1} v) \\ &= (P_{N-1}^{L}\partial_x u - \partial_x u, \partial_x^{-1} v) = 0 \qquad \forall\, v \in \mathbb{P}_{N-2},\end{aligned} \tag{3.2}$$

where $\partial_x^{-1} v := \int_{-1}^{x} v(y)\, dy \in \mathbb{P}_{N-1}$. We will need an estimate in the following negative norm:

$$\|u\|_{-1} := \sup_{v \in H_0^1(I), v \neq 0} \frac{|(u,v)|}{\|v\|_1}.$$

Lemma 3.1 (see [4, 15]). *If $u \in H^\sigma(I)\,(\sigma \geq 1)$, then*

$$\|P_{1,N}u - u\|_l \leq CN^{l-\sigma}\, \|u\|_\sigma, \qquad -1 \leq l \leq 1. \tag{3.3}$$

Proof. We prove only (3.3) with $l = -1$. The other cases can be found in [4, 15]. For any $v \in H_0^1(I)$, on the use of (3.2) and the result (3.3) with $l = 0$, we get

$$(P_{1,N}u - u, v) = (P_{1,N}u - u, v - P_{1,N-2}v) \leq CN^{-\sigma-1}\|u\|_\sigma \|v\|_1,$$

which gives the desired result. □

In general, the discrete inner product and norm are defined as follows:

$$(u,v)_N = \sum_{j=0}^{N} u(y_j)v(y_j)\omega_j, \qquad \|u\|_N = \sqrt{(u,u)_N},$$

where y_j and ω_j $(j = 0, \dots, N)$ are the Legendre–Gauss–Lobatto points and the corresponding quadrature weights. Associating with this quadrature rule, we denote by I_N^L the Legendre interpolation operator.

Lemma 3.2 (see [4, 9]). *If $u \in H^\sigma(I)\,(\sigma \geq 1)$, then*

$$\|I_N^L u - u\|_l \leq CN^{l-\sigma}\|u\|_\sigma, \qquad 0 \leq l \leq 1, \tag{3.4}$$

$$|(u,v) - (u,v)_N| \leq CN^{-\sigma}\|u\|_\sigma\, \|v\| \qquad \forall v \in \mathbb{P}_N. \tag{3.5}$$

Further, if $u \in \mathbb{P}_N$, then

$$\|u\| \leq \|u\|_N \leq \sqrt{2 + \frac{1}{N}}\, \|u\|, \tag{3.6}$$

$$\|u\|_{L^\infty(I)} \leq \frac{N+1}{\sqrt{2}}\|u\|. \tag{3.7}$$

Lemma 3.3 (see [15]). *If $u \in H^1(I)$, then*

$$N\|I_N^C u - u\| + |I_N^C u|_1 \le C\|u\|_1. \tag{3.8}$$

Moreover, if $u \in H^\sigma(I)\,(\sigma \ge 1)$, then

$$\|I_N^C u - u\|_l \le CN^{l-\sigma}\|u\|_\sigma, \qquad 0 \le l \le 1. \tag{3.9}$$

We note that although I_N^C is the Chebyshev interpolation operator, the norms in (3.8) and (3.9) are in the Legendre form rather than in the weighted Chebyshev form. These results are useful for numerical analysis of CL spectral methods.

4. The stability and convergence of the CLC method. In this section, we first consider the stability of the semidiscrete scheme (2.1) with the Dirichlet boundary conditions ($\alpha = \gamma = 1, \beta = \delta = 0$) and then give the proof of its convergence. Suppose that u and the term on the right-hand side in (2.1) have the errors $\tilde{u}$ and $\tilde{f}$, respectively. Then by (2.1) we have

$$(\partial_t \tilde{u}, v)_N + (\partial_x I_N^C \tilde{F}, v) - \nu(\partial_x^2 \tilde{u}, v) + (\tilde{R}, v)_N = (\tilde{f}, v)_N \quad \forall v \in \mathbb{P}_N, t \in (0,T), \tag{4.1}$$

where $\tilde{F} := F(u+\tilde{u}) - F(u)$ and

$$\tilde{R}(x,t) := \tau_0 Q^+(x)\tilde{u}(1,t) + \tau_N Q^-(x)\tilde{u}(-1,t).$$

Taking $v = \tilde{u}$ in (4.1), we get

$$\begin{aligned} &\frac{d}{dt}\|\tilde{u}\|_N^2 + \nu|\tilde{u}|_1^2 + \tau_0\omega_0\tilde{u}^2(1) + \tau_N\omega_N\tilde{u}^2(-1) \\ &\qquad = (\tilde{f}, \tilde{u})_N - (\partial_x I_N^C \tilde{F}, \tilde{u}) + \nu\tilde{u}(1)\partial_x\tilde{u}(1) - \nu\tilde{u}(-1)\partial_x\tilde{u}(-1). \end{aligned} \tag{4.2}$$

We bound the terms on the right-hand side of the above equation. First,

$$|(\tilde{f}, \tilde{u})_N| \le \|\tilde{f}\|_N^2 + \|\tilde{u}\|_N^2.$$

Next, let

$$\tau_0, \tau_N \ge \frac{1}{4}\max\{4\nu, 1\}\, N^2(N+1)^2. \tag{4.3}$$

Then it can be seen that

$$\begin{aligned} &|\nu\tilde{u}(1)\partial_x\tilde{u}(1) - \nu\tilde{u}(-1)\partial_x\tilde{u}(-1)| \\ &\qquad \le \frac{\nu}{4}(\omega_0|\partial_x\tilde{u}(1)|^2 + \omega_N|\partial_x\tilde{u}(-1)|^2) + \frac{\nu}{\omega_0}|\tilde{u}(1)|^2 + \frac{\nu}{\omega_N}|\tilde{u}(-1)|^2 \\ &\qquad \le \frac{\nu}{4}|\tilde{u}|_1^2 + \frac{1}{4}\tau_0\omega_0|\tilde{u}(1)|^2 + \frac{1}{4}\tau_N\omega_N|\tilde{u}(-1)|^2. \end{aligned}$$

As for the nonlinear term, by integrating by parts and noting that $\omega_0 = \omega_N = \frac{2}{N(N+1)}$, we get from (3.7) and (3.8) that

$$\begin{aligned} &\left|(\partial_x I_N^C \tilde{F}, \tilde{u})\right| \\ &\le \left|(I_N^C\tilde{F} - \tilde{F}, \partial_x\tilde{u}) + (\tilde{F}, \partial_x\tilde{u})\right| + \left|I_N^C\tilde{F}(1)\tilde{u}(1) - I_N^C\tilde{F}(-1)\tilde{u}(-1)\right| \\ &\le \frac{\nu}{4}|\tilde{u}|_1^2 + 2\|I_N^C\tilde{F} - \tilde{F}\|^2 + 2\|\tilde{F}\|^2 + \omega_0|\tilde{F}(1)|^2 + \omega_N|\tilde{F}(-1)|^2 + \frac{\tilde{u}^2(1)}{4\omega_0} + \frac{\tilde{u}^2(-1)}{4\omega_N} \\ &\le \frac{\nu}{4}|\tilde{u}|_1^2 + 2\|I_N^C\tilde{F} - \tilde{F}\|^2 + 5\|\tilde{F}\|^2 + 3\|I_N^L\tilde{F} - \tilde{F}\|^2 + \frac{1}{4}\tau_0\omega_0\tilde{u}^2(1) + \frac{1}{4}\tau_N\omega_N\tilde{u}^2(-1) \\ &\le \frac{\nu}{4}|\tilde{u}|_1^2 + CN^{-2}|\tilde{F}|_1^2 + C\|\tilde{F}\|^2 + \frac{1}{4}\tau_0\omega_0\tilde{u}^2(1) + \frac{1}{4}\tau_N\omega_N\tilde{u}^2(-1). \end{aligned}$$

Let C_0 be a positive constant and

$$\begin{aligned} u_M &= \max_{0\le s\le T}\left\{\|u(s)\|_{L^\infty(I)} + N^{-1}\|\partial_x u(s)\|_{L^\infty(I)}\right\}, \\ C_F(z_1,z_2) &= \max_{|z|\le|z_1|+|z_2|}|F'(z)| + (|z_1|+|z_2|)\max_{|z|\le|z_1|+|z_2|}|F''(z)|. \end{aligned} \tag{4.4}$$

For any given $t\in(0,T)$, if

$$\|\tilde{u}(s)\|_{L^\infty(I)} \le C_0 \qquad \forall\, s\in(0,t), \tag{4.5}$$

then

$$\begin{aligned} &\|\tilde{F}\| + N^{-1}|\tilde{F}|_1 \\ &= \left\|\int_0^1 F'(u+\theta\tilde{u})\tilde{u}d\theta\right\| + N^{-1}\left\|\int_0^1 (F''(u+\theta\tilde{u})(\partial_x u + \theta\partial_x\tilde{u})\tilde{u} + F'(u+\theta\tilde{u})\partial_x\tilde{u})d\theta\right\| \\ &\le C_F(u_M,C_0)\|\tilde{u}\| + \frac{1}{4}\nu|\tilde{u}|_1 \qquad \forall\, s\in(0,t). \end{aligned}$$

Therefore, integrating (4.2) in time leads to

$$E(\tilde{u},t) \le \rho(\tilde{u},\tilde{f},t) + C^*\int_0^t E(\tilde{u},s)ds, \tag{4.6}$$

where C^* is a positive constant depending on $C_F(u_M,C_0)$ and

$$\begin{aligned} E(\tilde{u},t) &= \|\tilde{u}(t)\|_N^2 + \int_0^t \{\nu|\tilde{u}(s)|_1^2 + \tau_0\omega_0\tilde{u}^2(1,s) + \tau_N\omega_N\tilde{u}^2(-1,s)\}ds, \\ \rho(\tilde{u},\tilde{f},t) &= \|\tilde{u}(0)\|_N^2 + \int_0^t \|\tilde{f}(s)\|_N^2 ds. \end{aligned}$$

We have the following stability result.

Theorem 4.1. *If $\rho(\tilde{u},\tilde{f},T)\le 2C_0^2e^{-C^*T}(N+1)^{-2}$, then*

$$E(\tilde{u},t) \le \rho(\tilde{u},\tilde{f},t)e^{C^*t}. \tag{4.7}$$

Proof. Following the line in [14], we first prove that

$$\max_{0\le s\le T}\|\tilde{u}(s)\|_{L^\infty(I)} \le C_0. \tag{4.8}$$

Otherwise, there must exist $t_1 < T$ such that

$$\max_{0\le s\le t_1}\|\tilde{u}(s)\|_{L^\infty(I)} \le C_0, \quad \|\tilde{u}(t_1)\|_{L^\infty(I)} = C_0, \tag{4.9}$$

while by (4.6) and the Gronwall inequality we have

$$E(\tilde{u},t_1) \le \rho(\tilde{u},\tilde{f},t_1)e^{C^*t_1} < \rho(\tilde{u},\tilde{f},T)e^{C^*T} \le 2C_0^2(N+1)^{-2}.$$

Thus, from Lemma 3.2,

$$\begin{aligned} \|\tilde{u}(t_1)\|_{L^\infty(I)} &\le \frac{N+1}{\sqrt{2}}\|\tilde{u}(t_1)\| \le \frac{N+1}{\sqrt{2}}\|\tilde{u}(t_1)\|_N \\ &\le (N+1)\sqrt{\frac{E(\tilde{u},t_1)}{2}} < C_0, \end{aligned}$$

which is contradictory with (4.9). Thus (4.8) holds, and we derive (4.7) from the Gronwall inequality. □

Next we consider the convergence of the scheme (2.1) with the Dirichlet boundary conditions $\alpha = \gamma = 1$, $\beta = \delta = 0$. Setting $w = P_{1,N}U$ and $\eta = u - w$, we get from (1.1) and (2.1) that

$$(4.10) \qquad (\partial_t \eta, v)_N + (\partial_x I_N^C \tilde{G}, v) - \nu(\partial_x^2 \eta, v) + (R(\eta, t), v)_N = g(v) \qquad \forall\, v \in \mathbb{P}_N,$$

where $\tilde{G} = F(w + \eta) - F(w)$ and

$$(4.11) \quad \begin{aligned} g(v) &:= [(\partial_t U, v) - (\partial_t w, v)_N] + (\partial_x (F(U) - I_N^C F(w)), v) - \nu(\partial_x^2 (U - w), v), \\ R(\eta, t) &= \tau_0 Q^+(x)\eta(1, t) + \tau_N Q^-(x)\eta(-1, t). \end{aligned}$$

Similar to (4.4), we let

$$U_M = \max_{0 \le s \le T} \{\|w(s)\|_{L^\infty(I)} + N^{-1}\|\partial_x w(s)\|_{L^\infty(I)}\}.$$

Then, as in the analysis of stability, we need to estimate $g(\eta)$ in (4.11). We separate it into

$$\begin{aligned} g(\eta) &= [(\partial_t U, \eta) - (\partial_t w, \eta)_N] + (\partial_x (F(U) - I_N^C F(w)), \eta) - \nu(\partial_x^2 (U - w), \eta) \\ &:= \sum_{i=1}^{3} J_i. \end{aligned}$$

First, by (3.5), (3.6), (3.4), and (3.3),

$$(4.12) \qquad \begin{aligned} |J_1| &\le |(\partial_t U, \eta) - (\partial_t U, \eta)_N| + |(\partial_t I_N^L U, \eta)_N - (\partial_t w, \eta)_N| \\ &\le C N^{-\sigma} \|\partial_t U\|_\sigma \|\eta\|_N + 2\|\partial_t I_N^L U - \partial_t w\| \, \|\eta\|_N \\ &\le C N^{-\sigma} \|\partial_t U\|_\sigma \|\eta\|_N. \end{aligned}$$

Next, by (3.9) we have

$$\begin{aligned} |J_2| &= |(F(U) - I_N^C F(w), \partial_x \eta)| \le C\|F(U) - I_N^C F(w)\|^2 + \frac{\nu}{8}|\eta|_1^2 \\ &\le C(\|(I - I_N^C)(F(w) - F(U))\|^2 + \|(I - I_N^C)F(U)\|^2 + \|F(U) - F(w)\|^2) + \frac{\nu}{8}\nu|\eta|_1^2 \\ &\le C(N^{-2}|F(w) - F(U)|_1^2 + \|F(U) - F(w)\|^2) + C N^{-2\sigma}\|F(U)\|_\sigma^2 + \frac{\nu}{8}|\eta|_1^2 \\ &\le C C_F'\Big(\|U\|_{L^\infty(I)}, \|w\|_{L^\infty(I)}\Big)\Big(1 + N^{-1}(\|\partial_x U\|_{L^\infty(I)} + \|\partial_x w\|_{L^\infty(I)})\Big) N^{-2\sigma}\|U\|_\sigma^2 \\ &\quad + \frac{\nu}{8}|\eta|_1^2 + C N^{-2\sigma}\|F(U)\|_\sigma^2, \end{aligned}$$

where

$$C_F'(z_1, z_2) = \max_{|z| \le \max\{z_1, z_2\}} \{|F'(z)|, |F''(z)|\}.$$

Finally, let $\tau_0, \tau_N \ge C N^6$. Then, from integrating by parts, (3.1), (3.3), and (3.4), we also have

$$\begin{aligned}|J_3| &= |\nu\{\partial_x(U-w)\eta\}|_{-1}^{1}|\\ &\le \frac{1}{8}\tau_0\omega_0\eta^2(1)+\frac{1}{8}\tau_N\omega_N\eta^2(-1)+\frac{2\nu^2}{\tau_0\omega_0}|\partial_x(U-w)(1)|^2+\frac{2\nu^2}{\tau_N\omega_N}|\partial_x(U-w)(-1)|^2\\ &\le \frac{1}{8}\tau_0\omega_0\eta^2(1)+\frac{1}{8}\tau_N\omega_N\eta^2(-1)+CN^{-2}\|\partial_x(U-w)\|_N^2\\ &\le \frac{1}{8}\tau_0\omega_0\eta^2(1)+\frac{1}{8}\tau_N\omega_N\eta^2(1)+CN^{-2\sigma}\|U\|_\sigma^2.\end{aligned}$$

For the initial error, we have from (3.3) and (3.9) that

$$\begin{aligned}\|\eta(0)\|_N &\le C\|\eta(0)\| \le C\|(I_N^C-P_{1,N})U_0\|^2\\ &\le C\|(I_N^C-I)U_0\|^2+C\|(I-P_{1,N})U_0\|^2\le CN^{-2\sigma}\|U_0\|_\sigma^2.\end{aligned} \tag{4.13}$$

We end this section with the following convergence theorem.

THEOREM 4.2. *Let U and u be the solutions of* (1.1) *and* (2.1)*, respectively. Assume that $\sigma \ge 2$, $U \in H^1(0,T;H^\sigma(I))$, $F(z)\in C^\sigma(\mathbb{R})$, and $\tau_0,\tau_N \ge CN^6$. Then there exists a positive constant C depending on ν^{-1} and the regularities of U and F such that*

$$\|u(t)-U(t)\| \le CN^{-\sigma} \qquad \forall\, t\in(0,T).$$

5. The stability and convergence of LGCC method. In this section, we consider the stability and convergence of the fully discrete scheme (2.3). We assume that all functions below are valued at time s unless otherwise specified. Suppose u and the term on the right-hand side in (2.3) have the error $\tilde{u}$ and $\tilde{f}$, respectively. Then by (2.3) we have

$$(\tilde{u}_{\hat{t}},v)+(\partial_x I_N^C\tilde{F},v)+\nu(\partial_x\hat{\tilde{u}},\partial_x v)=(\tilde{f},v) \qquad \forall\, v\in V_N^0,\quad t\in S_{t-\tau}. \tag{5.1}$$

Taking $v=\hat{\tilde{u}}$ in (5.1), we get

$$\frac{1}{2}\|\tilde{u}\|_{\hat{t}}^2+\nu|\hat{\tilde{u}}|_1^2=(\tilde{f},\hat{\tilde{u}})+(I_N^C\tilde{F},\partial_x\hat{\tilde{u}}). \tag{5.2}$$

We need to bound the term $\|I_N^C\tilde{F}\|$, which would be easy to do, provided that I_N^C is replaced by the Legendre–Galerkin/collocation operator or the norm is in the weighted Chebyshev one. But here we cannot deal with this directly. So we turn to the stability property of I_N^C given in (3.8). Taking $v=\tilde{u}_{\hat{t}}$ in (5.1), we have

$$\|\tilde{u}_{\hat{t}}\|^2+\frac{1}{2}\nu(|\tilde{u}|_1^2)_{\hat{t}}=(\tilde{f},\tilde{u}_{\hat{t}})-(\partial_x I_N^C\tilde{F},\tilde{u}_{\hat{t}}). \tag{5.3}$$

Combining (5.2) and (5.3) through the factor N^{-2}, we arrive at

$$\begin{aligned}&(\|\tilde{u}\|^2+N^{-2}\nu|\tilde{u}|_1^2)_{\hat{t}}+2(\nu|\hat{\tilde{u}}|_1^2+N^{-2}\|\tilde{u}_{\hat{t}}\|^2)\\ &\quad=2(\tilde{f},\hat{\tilde{u}})+2(I_N^C\tilde{F},\partial_x\hat{\tilde{u}})+2N^{-2}((\tilde{f},\tilde{u}_{\hat{t}})-(\partial_x I_N^C\tilde{F},\tilde{u}_{\hat{t}}))\\ &\quad\le C(\|\tilde{f}\|_{-1}^2+\|I_N^C\tilde{F}\|^2+N^{-2}\|\tilde{f}\|^2+N^{-2}|I_N^C\tilde{F}|_1^2)+\nu|\hat{\tilde{u}}|_1^2+N^{-2}\|\tilde{u}_{\hat{t}}\|^2,\end{aligned} \tag{5.4}$$

where C is a positive constant dependent on ν^{-1}. Summing (5.4) for $s\in S_{t-\tau}$ gives

$$E(\tilde{u},t)\le\rho(\tilde{u},\tilde{f},t)+C\tau\sum_{s\in S_{t-\tau}}(\|I_N^C\tilde{F}(s)\|^2+N^{-2}|I_N^C\tilde{F}(s)|_1^2), \tag{5.5}$$

where

$$\begin{aligned} E(\tilde u, t) &= \|\tilde u(t)\|^2 + N^{-2}\nu|\tilde u(t)|_1^2 + 2\tau \sum_{s\in S_{t-\tau}} (\nu|\hat{\tilde u}(s)|_1^2 + N^{-2}\|\tilde u_{\hat t}(s)\|^2),\\ \rho(\tilde u, \tilde f, t) &= \|\tilde u(0)\|^2 + N^{-2}\nu|\tilde u(0)|_1^2 + \|\tilde u(\tau)\|^2 + N^{-2}\nu|\tilde u(\tau)|_1^2\\ &\quad + C\tau \sum_{s\in S_{t-\tau}} (\|\tilde f(s)\|_{-1}^2 + N^{-2}\|\tilde f(s)\|^2). \end{aligned}$$

For any given $t \in S_T$, if

$$\|\tilde u(s)\|_{L^\infty(I)} \le C_0 \qquad \forall\, s \in S_{t-\tau}, \tag{5.6}$$

then by (3.8) and (3.9)

$$\begin{aligned} &\|I_N^C\tilde F\| + N^{-1}|I_N^C\tilde F|_1 \le \|\tilde F\| + \|I_N^C\tilde F - \tilde F\| + N^{-1}|I_N^C\tilde F|_1 \le \|\tilde F\| + CN^{-1}|\tilde F|_1\\ &= \left\|\int_0^1 F'(u+\theta\tilde u)\tilde u d\theta\right\| + CN^{-1}\left\|\int_0^1 (F''(u+\theta\tilde u)(\partial_x u + \theta\partial_x\tilde u)\tilde u + F'(u+\theta\tilde u)\partial_x\tilde u)d\theta\right\|\\ &\le C_F(u_M, C_0)(\|\tilde u\| + N^{-1}|\tilde u|_1) \qquad \forall\, s \in S_{t-\tau}. \end{aligned}$$

Thus, we have shown that for any $t \in S_T$, if (5.6) holds, then

$$E(\tilde u, t) \le \rho(\tilde u, \tilde f, t) + C^*\tau \sum_{s\in S_{t-\tau}} E(\tilde u, s), \tag{5.7}$$

where C^* is a positive constant dependent on $C_F(u_M, C_0)$ and ν^{-1}.

THEOREM 5.1. *Let τ be suitably small. If $\rho(\tilde u, \tilde f, T) \le 2C_0^2 e^{-C^*T}/(N+1)^2$, then*

$$E(\tilde u, t) \le \rho(\tilde u, \tilde f, t)e^{C^*t} \qquad \forall\, t \in S_T. \tag{5.8}$$

Proof. We prove the result by induction over $t \in S_T$. It is easy to see that the result (5.8) is true for $t = \tau$. Assume that it is true for all $s \in S_{t-\tau}$:

$$E(\tilde u, s) \le \rho(\tilde u, \tilde f, s)e^{C^*s}. \tag{5.9}$$

Then, from the inverse inequality (3.7), we have

$$\|\tilde u(s)\|_{L^\infty(I)}^2 \le \frac{(N+1)^2}{2}\|\tilde u(s)\|^2 \le \frac{(N+1)^2}{2}\rho(\tilde u, \tilde f, s)e^{C^*s} \le C_0^2,$$

which means (5.6) holds. Therefore, we have from (4.6) and (5.9) that

$$\begin{aligned} E(\tilde u, t) &\le \rho(\tilde u, \tilde f, t) + C^*\tau \sum_{s\in S_{t-\tau}} E(\tilde u, s) \le \rho(\tilde u, \tilde f, t) + C^*\tau \sum_{s\in S_{t-\tau}} \rho(\tilde u, \tilde f, s)e^{C^*s}\\ &\le \rho(\tilde u, \tilde f, t)\left(1 + C^*\tau \sum_{s\in S_{t-\tau}} e^{C^*s}\right) \le \rho(\tilde u, \tilde f, t)e^{C^*t}. \end{aligned}$$

Thus the proof is completed. □

Next we consider the convergence of the scheme (2.3). Setting $w = P_{1,N}U$ and $\eta = u - w$, we get from (1.1) and (2.3) that

$$(\eta_{\hat t}, v) + (\partial_x I_N^C\tilde G, v) + \nu(\partial_x\hat\eta, \partial_x v) = (\tilde g, v) \qquad \forall\, v \in V_N^0, \tag{5.10}$$

where

$$\tilde{g} = \partial_t \hat{U} - U_{\hat{t}} + U_{\hat{t}} - w_{\hat{t}} + \partial_x(F(U) - I_N^C F(w) + \frac{\tau^2}{2} F(U)_{t\bar{t}}), \tag{5.11}$$

$$v_{t\bar{t}}(t) := \frac{1}{\tau^2}(v(t+\tau) - 2v(t) + v(t-\tau)). \tag{5.12}$$

Let C_* be a positive constant dependent on $C_F(U_M, C_0)$ and ν^{-1}. Then, as in Theorem 5.1, if $\rho(\eta, \tilde{g}, T) \leq 2C_0 e^{-C_* T}/(N+1)^2$, we have

$$E(\eta, t) \leq \rho(\eta, \tilde{g}, t) e^{C_* t} \qquad \forall\, t \in S_T. \tag{5.13}$$

We arrive at the following convergence result.

Theorem 5.2. *Let U and u be the solution of* (1.1) *and* (2.3), *respectively. Assume that $\sigma \geq 2$ and $U \in C(0,T; H_0^1(I) \cap H^\sigma(I)) \cap H^1(0,T; H_0^1(I) \cap H^{\sigma-1}(I)) \cap H^3(0,T; L^2(I)) \cap H^2(0,T; H^1(I))$, $\partial_t U(0) \in H^{\sigma/2}(I)$, and $F(z) \in C^{\max\{3,\sigma\}}(\mathbb{R})$. Then there exists a positive constant C depending on ν^{-1} and the regularities of U and F such that if $\tau\sqrt{N} \leq c_0$ being suitably small,*

$$\|u(t) - U(t)\| \leq C(\tau^2 + N^{-\sigma}) \qquad \forall\, t \in S_T.$$

Proof. We need only to estimate $\rho(\eta, \tilde{g}, t)$ in (5.13). We separate $\tilde{g}$ in (5.11) into

$$\begin{aligned}\tilde{g} &= (\partial_t \hat{U} - U_{\hat{t}}) + (U_{\hat{t}} - w_{\hat{t}}) + \partial_x(F(U) - I_N^C F(U)) \\ &\quad + \partial_x I_N^C(F(U) - F(w)) + \frac{\tau^2}{2}\partial_x F(U)_{t\bar{t}} := \sum_{j=1}^{5} \tilde{g}_j.\end{aligned}$$

A simple calculation and (3.3) give

$$\begin{aligned}&\tau \sum_{s \in S_{t-\tau}} (\|\tilde{g}_1\|_{-1}^2 + N^{-2}\|\tilde{g}_1\|^2) \leq C\tau^4(\|\partial_t^3 U\|_{L^2(0,T;H^{-1}(I))}^2 + N^{-2}\|\partial_t^3 U\|_{L^2(0,T;L^2(I))}^2), \\ &\tau \sum_{s \in S_{t-\tau}} (\|\tilde{g}_2\|_{-1}^2 + N^{-2}\|\tilde{g}_2\|^2) \leq CN^{-2\sigma}\tau \sum_{s \in S_{t-\tau}} \|U_{\hat{t}}\|_{\sigma-1}^2 \\ &\qquad \leq CN^{-2\sigma}\|\partial_t U\|_{L^2(0,T;H^{\sigma-1}(I))}^2.\end{aligned}$$

From (3.9) and (3.3),

$$\begin{aligned}&\|\tilde{g}_3\|_{-1} + N^{-1}\|\tilde{g}_3\| \\ &\quad \leq \|F(U) - I_N^C F(U)\| + N^{-1}|F(U) - I_N^C F(U)|_1 \leq CN^{-\sigma}\|F(U)\|_\sigma, \\ &\|\tilde{g}_4\|_{-1} + N^{-1}\|\tilde{g}_4\| \\ &\quad \leq \|F(U) - F(w)\| + \|(I_N^C - I)(F(U) - F(w))\| + N^{-1}|I_N^C(F(U) - F(w))|_1 \\ &\quad \leq \|F(U) - F(w)\| + CN^{-1}|F(U) - F(w)|_1 \\ &\quad \leq CC_F'(\|U\|_{L^\infty(I)}, \|w\|_{L^\infty(I)})(1 + N^{-1}(\|\partial_x U\|_{L^\infty(I)} + \|\partial_x w\|_{L^\infty(I)}))N^{-\sigma}\|U\|_\sigma.\end{aligned}$$

Also,

$$\begin{aligned}\tau \sum_{s \in S_{t-\tau}} (\|\tilde{g}_5\|_{-1}^2 + N^{-2}\|\tilde{g}_5\|^2) &\leq \tau^5 \sum_{s \in S_{t-\tau}} (\|F(U)_{t\bar{t}}\|^2 + N^{-2}|F(U)_{t\bar{t}}|_1^2) \\ &\leq C\tau^4(\|\partial_t^2 F(U)\|_{L^2(0,T;L^2(I))}^2 + N^{-2}\|\partial_t^2 F(U)\|_{L^2(0,T;H^1(I))}^2) \\ &\leq CC_F''\tau^4(\|U\|_{H^2(0,T;L^2(I))}^2 + \|\partial_t U\|_{L^4(I\times(0,T))}^4 + N^{-2}\|U\|_{H^2(0,T;H^1(I))}^2) \\ &\leq CC_F''\tau^4\|U\|_{H^2(0,T;H^1(I))}^2,\end{aligned}$$

where

$$C_F'' = \max_{|z| \le \|U\|_{C(\bar{I}\times[0,T])}} \left\{ |\partial_z^l F(z)|^2 \ : \ l = 1,2,3 \right\}.$$

For the initial errors we have

$$\|\eta(0)\| + N^{-1}|\eta(0)|_1 = \|(I_N^C - P_{1,N})U_0\| + N^{-1}|(I_N^C - P_{1,N})U_0|_1 \le CN^{-\sigma}\|U_0\|_\sigma,$$

and from Taylor's formula

$$\begin{aligned}
&\|\eta(\tau)\| + N^{-1}|\eta(\tau)|_1\\
&\le \|(I_N^C - I)U_0\| + N^{-1}|(I_N^C - I)U_0|_1 + \tau(\|(I_N^C - I)U(\tau)\| + N^{-1}|(I_N^C - I)U(\tau)|_1)\\
&\quad + \tau^2(\|\partial_t^2 U\|_{C(0,\tau;L^2(I))} + N^{-1}\|\partial_t^2 U\|_{C(0,\tau;H^1(I))})\\
&\quad + \|(I - P_{1,N})U(\tau)\| + N^{-1}|(I - P_{1,N})U(\tau)|_1\\
&\le CN^{-\sigma}(\|U_0\|_\sigma + \|U(\tau)\|_\sigma) + C\tau N^{-\sigma/2}\|\partial_t U(0)\|_{\sigma/2} + C\tau^2\|\partial_t^2 U\|_{C(0,\tau;H^1(I))}.
\end{aligned}$$

Thus the proof is completed by (5.13) and (3.3) with the triangle inequality. □

Remark 5.1. If we consider only the semidiscrete scheme, the stability and optimal error estimate similar to Theorems 4.1–4.2 can be established for (2.1) with all three kinds of boundary conditions by combining the arguments given in [6, 13], provided that

$$\tau_0 = \begin{cases} \nu\beta^{-1}\omega_0^{-1}, & \beta \ne 0, \\ CN^6, & \beta = 0, \end{cases} \qquad \tau_N = \begin{cases} -\nu\delta^{-1}\omega_N^{-1}, & \delta \ne 0, \\ CN^6, & \delta = 0. \end{cases} \tag{5.14}$$

For the fully discrete CNLF scheme, it seems difficult to get desired results in that way. The difference is that in the semidiscrete case one has the term $|\tilde{u}(t)|_1$ which can be used to control the nonlinear term, while in the fully discrete case one has only the term $|\tilde{u}(t+\tau) + \tilde{u}(t-\tau)|_1$ (but the nonlinear term is set at t for ease of computation). However, an H^1-estimate can be obtained for the fully discrete scheme of the problem with the Neumann or Robin boundary condition. Under the condition (5.14) with $\beta \ne 0$ and $\delta \ne 0$, which is also applied in [6] to the problem with the Neumann boundary condition, the scheme (2.1) reads as follows: For any $v \in \mathbb{P}_N$,

(5.15)
$$\begin{cases} (\partial_t u(t), v)_N + (\partial_x I_N^C F(u(t)), v) + \nu(\partial_x u(t), \partial_x v) \\ \qquad + \nu\beta^{-1}(\alpha u(1,t) - g^+(t))v(1) + \nu|\delta|^{-1}(\gamma u(-1,t) - g^-(t))v(-1) = 0, \\ (u(0), v)_N = (I_N^C U_0, v)_N. \end{cases}$$

This is just a pseudospectral scheme with the boundary condition being treated in a natural way. Then the fully discrete CNLF scheme is the following: For any $v \in \mathbb{P}_N$,

(5.16)
$$\begin{cases} (u_{\hat{t}}(t),\, v)_N + (\partial_x I_N^C F(u(t)),\, v) + \nu(\partial_x \hat{u}(t),\, \partial_x v) \\ \qquad + \nu\beta^{-1}(\alpha\hat{u}(1,t) - \hat{g}^+(t))v(1) + \nu|\delta|^{-1}(\gamma\hat{u}(-1,t) - \hat{g}^-(t))v(-1) = 0, \end{cases}$$

with the initial values as in (2.2). The LGCC scheme can be set by simply replacing $(\cdot,\cdot)_N$ with $(\cdot,\cdot)$ in (5.15) and (5.16), respectively. In both cases, the stability and

Table 1

Maximum error at $t = 12$ for Example 6.1 with $m = 1$, $\nu = 0.1$, $\omega = 0.5$, $I = [-5, 5]$, $x_0 = -3$.

		LGCC		LC	
τ	N	L^∞ error	Time	L^∞ error	Time
1.e-1	16	5.4400e-02	0.09	5.4894e-02	0.14
1.e-2	32	4.5000e-03	0.89	4.5000e-03	0.83
1.e-3	64	2.9282e-05	10.72	2.7465e-05	10.73
1.e-4	128	3.4310e-09	166.21	3.4576e-09	533.19
1.e-5	256	2.0935e-11	4583.69	2.0800e-11	13872.49

Table 2

Maximum error at $t = 1$ for Example 6.1 with $m = 2$, $\nu = 7$, $\omega = 10$, $I = [-10, 10]$, $x_0 = -6$.

		LGCC		LC	
τ	N	L^∞ error	Time	L^∞ error	Time
2.5e-3	32	3.2600e-2	0.39	3.4200e-2	0.32
8.0e-4	64	8.2341e-4	1.27	8.2723e-4	1.24
5.0e-5	128	4.4544e-7	29.11	3.9143e-7	92.64
5.0e-6	256	4.3200e-9	800.72	4.5900e-9	2280.66

optimal error estimates in the H^1-norm can be derived for the fully CNLF scheme by an argument similar to the one given in the proofs of Theorems 5.1–5.2. When an H^1-estimate is considered, the nonlinear term can be estimated directly by using the stability property (3.8) of the Chebyshev interpolation operator.

6. Numerical results. In this section, we will give some numerical results of the LGCC method and the Legendre collocation (LC) method to make a comparison.

Example 6.1. *Consider the generalized Burgers equation*

$$\partial_t U + U^m \partial_x U - \nu \partial_x^2 U = 0, \tag{6.1}$$

with the following analytical solution:

$$U(x,t) = \left\{ \frac{(m+1)\omega}{2} \left[1 - \tanh\left(\frac{m\omega}{2\nu}(x - \omega t - x_0) \right) \right] \right\}^{1/m}, \qquad \omega > 0.$$

It is computed by the LGCC method and the LC method with $m = 1$, $\nu = 0.1$, $\omega = 0.5$, $I = [-5, 5]$, $x_0 = -3$, $t \in [0, 12]$, and $m = 2$, $\nu = 7$, $\omega = 10$, $I = [-10, 10]$, $x_0 = -6$, $t \in [0, 1]$. The maximum errors are reported in Table 1 at $t = 12$ and Table 2 at $t = 1$, respectively.

Example 6.2. *Consider the Burgers equation*

$$\partial_t U + U^2 \partial_x U - \nu \partial_x^2 U = g, \tag{6.2}$$

with a soliton-like solution

$$U(x,t) = \operatorname{sech}^2(ax - bt - c).$$

Numerical results of the LGCC method and the LC method with $a = b = 1$, $c = -4$, $\nu = 1$, and $I = [-10, 10]$ are reported in Table 3 at $t = 8$. From Tables 1–3, we can

Table 3
Maximum error at $t = 8$ *for Example* 6.2 *with* $a = b = 1$, $c = -4$, $\nu = 1$, $I = [-10, 10]$.

		LGCC		LC	
τ	N	L^∞ error	Time	L^∞ error	Time
1.e-1	16	1.5660e-01	0.79	2.3770e-01	0.16
1.e-2	32	2.9800e-02	0.89	3.2700e-02	0.81
1.e-3	64	3.1580e-04	9.79	3.0556e-04	11.00
1.e-4	128	1.7111e-08	149.21	1.5117e-08	279.72
1.e-5	256	8.2430e-12	3629.76	8.3099e-12	9937.84

see that the LGCC method obtained nearly the same precision as the LC method. But the LGCC method spends less time than the LC method, especially when the number of collocation points increases.

Acknowledgment. The authors thank the referees for their valuable suggestions and comments.

REFERENCES

[1] B. K. Alpert and V. Rokhlin, *A fast algorithm for the evaluation of Legendre expansions*, SIAM J. Sci. Statist. Comput., 12 (1991), pp. 158–179.

[2] C. Bernardi and Y. Maday, *Spectral methods*, in Handbook of Numerical Analysis, Vol. V, Techniques of Scientific Computing (Part 2), P. G. Ciarlet and J. L. Lions, eds., North-Holland, Amsterdam, 1997, pp. 209–486.

[3] N. Bressan and A. Quarteroni, *An implicit/explicit spectral method for Burgers' equation*, Calcolo, 23 (1986), pp. 265–284.

[4] C. Canuto, M. Y. Hussaini, A. Quarteroni, and T. A. Zang, *Spectral Methods in Fluid Dynamics*, Springer–Verlag, Berlin, 1988.

[5] E. Coutsias, T. Hagstrom, J. Hesthaven, and D. Torres, *Integration preconditioners for differential operators in spectral tau-methods*, in Proceedings of the Third International Conference on Spectral and High Order Methods, A. V. Ilin and L. R. Scott, eds., Houston, TX, 1996, pp. 21–38.

[6] W. S. Don and D. Gottlieb, *The Chebyshev–Legendre method: Implementing Legendre methods on Chebyshev points*, SIAM J. Numer. Anal., 31 (1994), pp. 1519–1534.

[7] W. E, *Convergence of spectral methods for Burgers' equation*, SIAM J. Numer. Anal., 29 (1992), pp. 1520–1541.

[8] D. Funaro, *Domain decomposition methods for pseudospectral approximations part* I*, second order equations in one dimension*, Numer. Math., 52 (1988), pp. 329–344.

[9] B.-Y. Guo, *Spectral Methods and Their Applications*, World Scientific, Singapore, 1998.

[10] J. S. Hesthaven, *Integration preconditioning of pseudospectral operators.* I*. Basic linear operators*, SIAM J. Numer. Anal., 35 (1998), pp. 1571–1593.

[11] S. D. Kim and S. V. Parter, *Preconditioning Chebyshev spectral collocation method for elliptic partial differential equations*, SIAM J. Numer. Anal., 33 (1996), pp. 2375–2400.

[12] S. D. Kim and S. V. Parter, *Preconditioning Chebyshev spectral collocation by finite-difference operators*, SIAM J. Numer. Anal., 34 (1997), pp. 939–958.

[13] H.-Y. Li, H. Wu, and H.-P. Ma, *Legendre Galerkin–Chebyshev collocation method for Burgers-like equation*, IMA J. Numer. Anal., 23 (2003), pp. 109–124.

[14] H.-P. Ma, *A three-level Fourier pseudospectral scheme for Burgers' equation*, Chinese J. Numer. Math. Appl., 10 (1988), pp. 11–18.

[15] H. Ma, *Chebyshev–Legendre spectral viscosity method for nonlinear conservation laws*, SIAM J. Numer. Anal., 35 (1998), pp. 869–892.

[16] H. Ma, *Chebyshev–Legendre super spectral viscosity method for nonlinear conservation laws*, SIAM J. Numer. Anal., 35 (1998), pp. 893–908.

[17] H.-P. Ma and B.-Y. Guo, *The Chebyshev spectral method for Burgers-like equations*, J. Comput. Math., 6 (1988), pp. 48–53.

[18] Y. Maday and A. Quarteroni, *Legendre and Chebyshev spectral approximations of Burgers' equation*, Numer. Math., 37 (1981), pp. 321–332.

[19] Y. Maday and A. Quarteroni, *Approximations of Burgers' equation by pseudo-spectral methods*, RAIRO Anal. Numér., 16 (1982), pp. 375–404.

[20] A. Quarteroni and E. Zampieri, *Finite element preconditioning for Legendre spectral collocation approximations to elliptic equations and systems*, SIAM J. Numer. Anal., 29 (1992), pp. 917–936.

[21] L. G. Reyna, *L^2-estimates for Chebyshev collocation*, J. Sci. Comput., 3 (1988), pp. 1–24.

[22] J. Shen, *Efficient spectral-Galerkin method* I. *Direct solvers of second- and fourth-order equations using Legendre polynomials*, SIAM J. Sci. Comput., 15 (1994), pp. 1489–1505.

[23] J. Shen, *Efficient spectral-Galerkin method* II. *Direct solvers of second- and fourth-order equations using Chebyshev polynomials*, SIAM J. Sci. Comput., 16 (1995), pp. 74–87.

[24] J. Shen, *Efficient Chebyshev–Legendre Galerkin methods for elliptic problems*, in Proceedings of the International Conference on Spectral and High Order Methods, Houston, TX, 1995, pp. 233–239.

[25] J. Shen, *Efficient spectral-Galerkin methods* III: *Polar and cylindrical geometries*, SIAM J. Sci. Comput., 18 (1997), pp. 1583–1604.

SIAM J. Numer. Anal.
Vol. 41, No. 2, pp. 673–694

FAST SWEEPING ALGORITHMS FOR A CLASS OF HAMILTON–JACOBI EQUATIONS[*]

YEN-HSI RICHARD TSAI[†], LI-TIEN CHENG[‡], STANLEY OSHER[§], AND HONG-KAI ZHAO[¶]

Abstract. We derive a Godunov-type numerical flux for the class of strictly convex, homogeneous Hamiltonians that includes $H(p,q)=\sqrt{ap^2+bq^2-2cpq}$, $c^2<ab$. We combine our Godunov numerical fluxes with simple Gauss–Seidel-type iterations for solving the corresponding Hamilton–Jacobi (HJ) equations. The resulting algorithm is fast since it does not require a sorting strategy as found, e.g., in the fast marching method. In addition, it provides a way to compute solutions to a class of HJ equations for which the conventional fast marching method is not applicable. Our experiments indicate convergence after a few iterations, even in rather difficult cases.

Key words. Hamilton–Jacobi equations, fast marching, fast sweeping, upwind finite differencing, eikonal equations

AMS subject classifications. 35, 65

PII. S0036142901396533

1. Introduction.

Hamilton–Jacobi (HJ) equations have a rich pool of applications, ranging from those of optimal control theory and geometrical optics, to essentially any problem that needs the (weighted) distance function [14]. Examples include crystal growth, ray tracing, etching, robotic motion planning, and computer vision. Solutions of these types of equations usually develop singularities in their derivatives, and thus the unique viscosity solution [6] is sought.

In this article, we focus on the class of time independent HJ equations with Dirichlet boundary condition

$$H(\mathbf{x},\nabla u)=r(\mathbf{x}),\qquad u|_\Gamma=0;$$

$H(\mathbf{x},\mathbf{p})$ are strictly convex nonnegative, and $\lim_{\lambda\to 0}H(\mathbf{x},\lambda\mathbf{p})=0$. We explain our method using the following important model equation:

$$H(\phi_x,\phi_y)=\sqrt{a\phi_x^2+b\phi_y^2-2c\phi_x\phi_y}=r, \tag{1.1}$$

where $\phi:\mathbb{R}^2\mapsto\mathbb{R}$ is continuous and a, b, c, and r can be either constants or scalar functions; in the latter case, H depends also on x, defined on $\mathbb{R}^2$, satisfying $ab>c^2$, $a,b,r>0$. With $a=b=1$ and $c=0$, we have the standard eikonal equation for

[*] Received by the editors October 15, 2001; accepted for publication (in revised form) October 17, 2002; published electronically May 6, 2003.

http://www.siam.org/journals/sinum/41-2/39653.html

[†] Department of Mathematics and Program in Applied and Computational Mathematics, Princeton University, Princeton, NJ 08544 (ytsai@math.princeton.edu). The research of this author was supported by ONR N00014-97-1-0027, DARPA/NSF VIP grant NSF DMS 9615854, and ARO DAAG 55-98-1-0323.

[‡] Department of Mathematics, University of California at San Diego, La Jolla, CA 92093-0112 (lcheng@math.ucsd.edu). The research of this author was supported by NSF grants 0112413 and 0208449.

[§] Department of Mathematics, University of California at Los Angeles, Los Angeles, CA 90095 (sjo@math.ucla.edu). The research of this author was supported by ONR N00014-97-1-0027, DARPA/NSF VIP grant NSF DMS 9615854, and ARO DAAG 55-98-1-0323.

[¶] Department of Mathematics, University of California at Irvine, Irvine, CA 92697-3875 (zhao@math.ucla.edu).

which many numerical methods have been developed. This equation has the essential features of HJ equations with convex Hamiltonians, so that we can easily explain our algorithm, and is general enough that fast marching is not applicable.

In the following subsections, we will review some of the solution methods for the eikonal equation since it forms the motivation of our work. We then present a fast Gauss–Seidel-type iteration method for (1.1) which utilizes a monotone upwind Godunov flux for the Hamiltonian. We show numerically that this algorithm can be applied directly to equations of the above type with variable coefficients.

1.1. Solving eikonal equations. In geometrical optics [10], the eikonal equation

$$\sqrt{\phi_x^2 + \phi_y^2} = r(x, y) \tag{1.2}$$

is derived from the leading term in an asymptotic expansion

$$e^{i\omega(\phi(x,y)-t)} \sum_{j=0}^{\infty} A_j(x, y, t)(i\omega)^{-j}$$

of the wave equation

$$w_{tt} - c^2(x, y)(w_{xx} + w_{yy}) = 0,$$

where $r(x, y) = 1/|c(x, y)|$ is the function of slowness. The level sets of the solution ϕ can thus be interpreted as the first arrival time of the wave front that is initially Γ. It can also be interpreted as the "distance" function to Γ.

We first restrict our attention for now to the case in which $r = 1$. Let Γ be a closed subset of $\mathbb{R}^2$. It can be shown easily that the distance function defined by

$$d(\mathbf{x}) = \text{dist }(\mathbf{x}, \Gamma) := \min_{p\in\Gamma} ||\mathbf{x} - p||, \qquad \mathbf{x} = (x, y) \in \mathbb{R}^2,$$

is the viscosity solution to (1.2) with the boundary condition

$$\phi(x, y) = 0 \qquad \text{for } (x, y) \in \Gamma.$$

Rouy and Tourin [20] proved the convergence to the viscosity solution of an iterative method solving (1.2) with the Godunov Hamiltonian approximating $||\nabla\phi||$. The Godunov Hamiltonian function can be written in the following form:

$$H_G(p_-, p_+, q_-, q_+) = \sqrt{\max\{p_-^+, p_+^-\}^2 + \max\{q_-^+, q_+^-\}^2}, \tag{1.3}$$

where $p_\pm = D_\pm^x\phi_{i,j}$, $q_\pm = D_\pm^y\phi_{i,j}$, $D_\pm^x\phi_{i,j} = \pm(\phi_{i\pm1,j} - \phi_{i,j})/h$ and accordingly for $D_\pm^y\phi_{i,j}$, and $x^+ = \max(x, 0)$, $x^- = -\min(x, 0)$.

Osher [13] provided a link to time dependent eikonal equations by proving that the t-level set of $\phi(x, y)$ is the zero level set of the viscosity solution of the evolution equation at time t,

$$\psi_t = ||\nabla\psi|| = 0,$$

with appropriate initial conditions. In fact, the same is true for a very general class of HJ equations (see [13]). As a consequence, one can try to solve the time dependent equation by the level set formulation [17] with high order approximations on the partial derivatives [9, 18]. Crandall and Lions proved that the discrete solution obtained with a consistent, monotone numerical Hamiltonian converges to the desired viscosity solution [5].

Tsitsiklis [25] combined heap sort with a variant of the classical Dijkstra algorithm to solve the steady state equation of the more general problem

$$||\nabla\phi|| = r(\mathbf{x}).$$

This was later rederived in [23] and also reported in [8]. It has become known as the fast marching method, whose complexity is $\mathcal{O}(N\log(N))$, where N is the number of grid points. Osher and Helmsen [15] have extended the fast marching-type method to somewhat more general HJ equations. We will comment on this in a following section.

1.2. Anisotropic eikonal equation. We return to the Hamiltonian in question: $H(p,q) = \sqrt{ap^2+bq^2-2cpq}$. Writing the quadratic form as

$$ap^2+bq^2-2cpq = \begin{pmatrix} p & q \end{pmatrix}\begin{pmatrix} a & -c \\ -c & b \end{pmatrix}\begin{pmatrix} p \\ q \end{pmatrix},$$

it is easy to see that we can diagonalize the symmetric matrix in the middle of the equation for our previously noted choices of a, b, c and find a coordinate system ξ-η such that, after rescaling, the Hamiltonian becomes

$$H(\tilde{p},\tilde{q}) = \sqrt{\tilde{p}^2+\tilde{q}^2}.$$

The eigensystem of the above matrix defines the anisotropy. Indeed, the authors in [15] proposed to solve the constant coefficient equation (1.1) by first transforming it to (1.2) in the ξ-η coordinate system.

This anisotropy occurs in fields such as ray tracing in special media, e.g., crystals, in which there are "preferred" directions. Furthermore, we will see that it can be a result of considering the geodesic distance function on a manifold M that is defined as the graph of a smooth function f.

Let ϕ be the distance function such that

$$\phi(x,y) = \min_{\gamma\subset M}\int_\gamma ds$$

and γ connects the point (x,y) with the set $\Gamma\subset M$. The minimizing curve is called the geodesic, and ϕ the distance function to Γ on M. Moreover, ϕ solves

$$||P_{\nabla\psi}\nabla\phi||^2 = 1, \qquad \phi|_\Gamma = 0, \tag{1.4}$$

where $\psi(x,y,z) = f(x,y) - z$, and the projection operator [4]

$$P_{\nabla\psi} = I - \frac{\nabla\psi\bigotimes\nabla\psi}{||\nabla\psi||^2},$$

which projects a vector onto a plane whose normal is parallel to $\nabla\psi$. Using the fact that $P_{\nabla\psi}$ is a projection operator, a simple calculation shows that

(1.5)

$$||P_{\nabla\psi}\nabla\phi||^2 = \left(1-\frac{f_x^2}{f_x^2+f_y^2+1}\right)\phi_x^2 + \left(1-\frac{f_y^2}{f_x^2+f_y^2+1}\right)\phi_y^2 - 2\frac{f_xf_y}{f_x^2+f_y^2+1}\phi_x\phi_y.$$

This is clearly of the form of Hamiltonians that we are interested in. We will apply our algorithm to compute the geodesic distance later in this paper.

There are other approaches that are designed to compute distances on manifolds. For example, [11] provided an algorithm to compute the geodesic distance on triangulated manifolds. Barth [2] uses the discontinuous Galerkin method to find distance on graphs of functions that are represented by spline functions. In [4], the authors embed the manifold as the zero level set of a Lipschitz continuous function and solve the corresponding time dependent eikonal equation (1.4) in the embedding space. As we have mentioned in the previous subsection, the zero level set of the time dependent eikonal equation at time t_1 is the t_1-level set of the solution to the stationary eikonal equations (see [13]). In [12], the authors adopted the standard fast marching method to solve the isotropic eikonal equation in a thin band of thickness ϵ, which encloses the manifold M, and proved that the restriction of the solution to M converges to the geodesic distance as ϵ goes to 0. In [21, 22], the authors provide an ordered upwind method to solve a general class of static HJ equations. We will comment on their method in a later subsection.

1.3. Osher's fast marching criteria. Since the fast marching method is by now well known, we will not give much detail on its implementation in this paper. In general, this involves a sorting procedure and the solution of

$$H_G(D_-^x\phi_{i,j}, D_+^x\phi_{i,j}, D_-^y\phi_{i,j}, D_+^y\phi_{i,j}) = 1 \tag{1.6}$$

for ϕ_{ij} in terms of its four neighboring values. More precisely, the heap sort strategy of the fast marching method requires a monotone update sequence. The updated value of a grid node has to be greater than or equal to those of the grid nodes used to form the finite difference stencil. This amounts to the condition

$$pH_p + qH_q \geq 0,$$

which dictates that the solution be nondecreasing along the characteristics. However, if we use one-sided upwind finite difference approximations for partial derivatives of ϕ on a Cartesian grid, it is equivalent to demanding that the partial derivatives of ϕ (i.e., p and q) and their corresponding components of the characteristics directions (i.e., dx/dt and dy/dt) have the same sign. Since $dx/dt = H_p$ and $dy/dt = H_q$, we have the stricter Osher's fast marching criterion

$$pH_p \geq 0, \quad qH_q \geq 0. \tag{1.7}$$

It does not matter whether the Hamiltonian is convex or not; as long as criterion (1.7) is satisfied, a simple fast marching algorithm can be applied. But if the criterion is not satisfied, fast marching cannot be applied to the problem on a Cartesian grid. Of course there are Hamiltonians that do not satisfy (1.7). In the class of Hamiltonians that we consider, as long as $c \neq 0$, it is likely that the values of p and q differ to the extent that the above criterion is no longer satisfied. In light of criterion (1.7), we have also tried to find directions $\xi(x, y)$ and $\eta(x, y)$ locally in which $\tilde{p}H_{\tilde{p}} \geq 0, \tilde{q}H_{\tilde{q}} \geq 0$. However, if one insists on using Cartesian grids, the implementation of this approach might be a bit hairy. We are interested especially in tackling, over a Cartesian grid, problems where the solution is nondecreasing along characteristics but where Osher's fast marching criterion is not satisfied.

1.4. The sweeping idea. Danielsson [7] proposed an algorithm to compute Euclidean distance to a subset of grid points on a two dimensional grid by visiting each grid node in some predefined order. In [3], Boué and Dupuis suggest a similar "sweeping" approach to solve the steady state equation which, by experience, results in an $\mathcal{O}(N)$ algorithm for the problem at hand. This "sweeping" approach has recently been used in [24] and [27] to compute the distance function to an arbitrary data set in computer vision. In [26], the author provides some theoretical evidence indicating that sweeping converges to an approximate Euclidean distance function, i.e., to an approximate viscosity solution of $|\nabla\phi| = 1$ in $2d$ predetermined iterations. We will talk about these iterations in a later section. Using this "sweeping" approach, the complexity of the algorithms drops from $\mathcal{O}(N \log N)$ in fast marching to $\mathcal{O}(N)$, and the implementation of the algorithms becomes a bit easier than the fast marching method that requires heap sort.

This sweeping idea is best illustrated by solving the eikonal equation in $[0, 1]$:

$$|u_x| = 1, \qquad u(0) = u(1) = 0.$$

Let $u_i = u(x_i)$ denote the grid values associated with the uniform grid composed of the gridpoints $0 = x_0 < x_1 < \cdots < x_n = 1$. We then solve the discretized nonlinear system

$$\sqrt{\max(\max(D_- u_i, 0)^2, \min(D_+ u_i, 0)^2)} = 1, \qquad u_0 = u_n = 0, \tag{1.8}$$

by our sweeping approach. We initially set $u_i^{(0)} = \infty$, $i = 1, \ldots, n-1$. In practice, ∞ can be replaced by some number K, which is larger than $\max_{x\in[0,1]} u$. Let us begin by sweeping from 0 to 1; i.e., we update u_i from $i = 1$ increasing to $i = n-1$. This is "equivalent" to following the characteristics emanating from x_0. Let $u_i^{(1)}$ denote the grid values after this sweep. We then have

$$u_i^{(1)} = \begin{cases} \frac{i}{n} \text{ if } i = 1, \ldots, n-2, \\ \frac{1}{n} \text{ if } i = n-1. \end{cases}$$

Notice that at $i = n-1$, we actually use the upwind information from the neighboring right boundary point. Furthermore, notice that $u_i^{(1)}$ already has the correct desired values for $i \le n/2$ since the sweep goes from left to right, the desired upwind direction for these i. In the second sweep, we update u_i from $i = n-1$ decreasing to 1, starting with $u_i^{(1)}$. During this sweep, we follow the characteristics emanating from x_n. The use of (1.8) is essential, since it determines what happens when two characteristics cross each other. It is then not hard to see that, after the second sweep,

$$u_i = \begin{cases} \frac{i}{n} \text{ if } i \le \frac{n}{2}, \\ \frac{(n-i)}{n} \text{ otherwise.} \end{cases}$$

Notice that the correct values at $i \le n/2$ derived after the first sweep are unchanged, and new and correct values for $i > n/2$ are created. In summary, this simple iterative algorithm can be described as follows: at the kth iteration, solve

$$\max\left(\max\left(\frac{u_i^{(k)} - u_{i-1}^{(k-1)}}{\Delta x}, 0\right), \min\left(\frac{u_{i+1}^{(k-1)} - u_i^{(k)}}{\Delta x}, 0\right)\right) = 1$$

for $u_i^{(k)}$ for each i going from 1 to $n-1$ in the first iteration ($k = 1$), and from $n-1$ to 1 for the second iteration ($k = 2$). However, for more complicated equations and boundary conditions, it is not so easy to write down the equivalent explicit solution.

In this paper, we will extend this sweeping approach to a class of HJ equations that cannot be solved by the fast marching algorithm, by first deriving a Godunov Hamiltonian.

In [21, 22], the authors proposed a one-pass method that is based on a control-theoretic viewpoint. In principle, they solve the HJB equation

$$\max_{\mathbf{a}} \nabla u \cdot \mathbf{a} f(\mathbf{a}, \mathbf{x}) = 1, \tag{1.9}$$

where $\mathbf{p} = (p, q)$ and the function $f(\mathbf{a}, \mathbf{x})$ is the speed of motion. This formula is the second Legendre transform taken on the sphere; see, e.g., [16, 19].

The idea is still to follow the characteristics and update the grid value in a monotone sequence. In a notation similar to the two dimensional setting of [21, 22], we let u_o be the grid value we are updating. To update u_o, we have to look for two other grid values u_r and u_s, which are not necessarily the immediate grid neighbors of u_o. For example, if u_o is the grid value $u_{i,j}$, the immediate neighbors of u_o are then $u_{i+1,j}$,$u_{i,j+1}$,$u_{i-1,j}$, and $u_{i,j-1}$. As we indicated in the previous subsection, it is possible that u_o is less than all its immediate neighboring values. We then need to find two other grid values, here denoted as u_s and u_r, to form an upwinding stencil. Then u_o is found by minimizing a nonlinear expression derived from (1.9), using the values of u_r, u_s, and f. The heap sort data structures are used in order to find u_r and u_s; therefore, the complexity is $N \log N$, where N is the total number of grid points. Also, since u_r and u_s may not lie on the immediate neighbors, this algorithm may need a larger region around the initial wave front to get started.

As one will see in the following section, our proposed method is also based on following the characteristics. To update u_o, our method uses only the immediate neighboring grid values and does not need the heap sort data structure. More importantly, our algorithm follows the characteristics with certain directions simultaneously, in a parallel way, instead of a sequential way as in the fast marching method. The Godunov flux is essential in our algorithm, since it determines what neighboring grid values should be used to update u on a given grid node o. At least in the examples presented, we need only to solve a simple quadratic equation and run some simple tests to determine the value to be updated. This simple procedure is performed in each sweep, and the solution is obtained after a few sweeps. Our code is not much more than what is presented in section 3.2. We also point out the ease of implementing our proposed algorithm and its extension to more dimensions; this will be described in a sequel paper.

2. A Godunov flux for strictly convex Hamiltonians.

By solving the Riemann problem for HJ equations (Godunov's procedure), Bardi and Osher [1] proved rigorously that

$$H_G(p_-, p_+; q_-, q_+) = \underset{p \in I[p_-, p_+]}{\text{ext}} \; \underset{q \in I[q_-, q_+]}{\text{ext}} H(p, q), \tag{2.1}$$

where

$$\underset{p \in I[a,b]}{\text{ext}} = \min_{p \in [a,b]} \quad \text{if } a \le b,$$

$$\underset{p \in I[a,b]}{\text{ext}} = \max_{p \in [b,a]} \quad \text{if } a > b,$$

$$H_G(D_-^x \phi_{ij}, D_+^x \phi_{ij}, D_-^y \phi_{ij}, D_+^y \phi_{ij}) = H_G(p_-, p_+; q_-, q_+),$$

and $I[a,b]$ denotes the closed interval bounded by a and b. This is a monotone upwind flux function, which implies convergence. Godunov's scheme (1.3) for the eikonal equation $\sqrt{\phi_x^2+\phi_y^2}=1$ can be derived from the above formula. It is one of the central topics of this paper to derive an explicit formula for the class of strictly convex Hamiltonians in question. Especially, we will demonstrate our numerical methods on $H=\sqrt{a\phi_x^2+b\phi_y^2-2c\phi_x\phi_y}$, $c^2<ab$.

Note that, in general, if we reverse the order on p and q in our ext-ext decision, the result might be different, although they both give convergent monotone methods. However, in the convex Hamiltonian at hand, the results are order independent.

For convenience, we will also use $H_G(\phi_{i,j},\phi_{i\pm1,j},\phi_{i,j\pm1})$ to denote the evaluation of our Godunov Hamiltonian $H_G(D_-^x\phi_{ij},D_+^x\phi_{ij},D_-^y\phi_{ij},D_+^y\phi_{ij})$.

2.1. Derivation of the flux. In order to derive a compact expression that satisfies (2.1), we need to study the extremum of the Hamiltonian on $I_p\times I_q\subset\mathbb{R}^2$, where I_p is a shorthand for $I[p_-,p_+]$.

The extremum may occur on either the critical points of H or the boundary of $I_p\times I_q$. Let us first look at the partial derivatives of H, i.e., H_p and H_q, and their zeros. Fix a q_0; the extremum of $H(p,q_0)$ occurs at either the critical point of $H(p,q_0)$ (i.e., where $H_p=0$) or the boundary of $I[p_-,p_+]$. We denote the critical point by $p_\sigma(q_0)$. Similarly, given p_0, we obtain the critical point $q_\sigma(p_0)$. For convenience, we shall denote $p_\sigma(q_0)$ by p_σ when q_0 can be determined from the context, and (p_σ,q_σ) is the critical point of H such that $H_p(p_\sigma,q_\sigma)=H_q(p_\sigma,q_\sigma)=0$. *Therefore, we consider separately* $H(p_\sigma,q_\sigma)$, $H(p_-,q_\sigma(p_-))$, $H(p_+,q_\sigma(p_+))$, $H(p_\sigma(q_-),q_-)$, $H(p_\sigma(q_+),q_+)$, *and* $H(p_\pm,q_\pm)$ *as possible evaluations of* (2.1).

For fixed p, we have

$$H_G(p,q_-,q_+)=H(p,\operatorname{sgn}\max\{(q_--q_\sigma)^+,(q_+-q_\sigma)^-\}+q_\sigma\,), \tag{2.2}$$

where

$$\begin{aligned}\operatorname{sgn}\max(x,y)&=x^+\quad\text{if}\quad\max\{x^+,y^-\}=x^+,\\ \operatorname{sgn}\max(x,y)&=-y^-\quad\text{if}\quad\max\{x^+,y^-\}=y^-.\end{aligned}$$

The expression for fixed q is a direct analogy to (2.2). It is easy to see that $H_G(\cdot,\cdot;q_-,q_+)$ is increasing in q_- and decreasing in q_+. By symmetry, $H_G(p_-,p_+;\cdot,\cdot)$ is increasing in p_- and decreasing in p_+.

Details of the derivation of the above expression are provided in the appendix.

The following proposition will be of use in analyzing this introduced Godunov flux.

PROPOSITION 1. *If* $H_{pp}>0,H_{qq}>0$, $pH_p\geq0,qH_q\geq0$, *and* $p_\sigma(0)=q_\sigma(0)=0$, *then* $p_\sigma(q)\equiv0\ \forall q$, *and* $q_\sigma(p)\equiv0\ \forall p$.

Proof. $p_\sigma(q)$, by definition, is the zero of $H_p(p_\sigma(q),q)=0$. We will write p_σ in place of $p_\sigma(q)$ for brevity. This proposition is then proved by simple manipulation of the definitions:

$$\begin{aligned}\frac{d}{dq}p_\sigma H(p_\sigma,q)&=p_\sigma'H_p(p_\sigma,q)+p_\sigma(H_{pp}(p_\sigma,q)p_\sigma'+H_{pq}(p_\sigma,q))\\ &=p_\sigma p_\sigma'H_{pp}(p_\sigma,q)+\frac{\partial}{\partial q}H_p(p_\sigma,q)\\ &=p_\sigma p_\sigma'H_{pp}(p_\sigma,q)\\ &=0.\end{aligned}$$

The hypothesis $H_{pp} > 0$ implies that

$$p_\sigma(q) = 0 \ \forall q \quad \text{or} \quad p'_\sigma(q) = 0 \ \forall q.$$

Again, by the hypothesis that $p_\sigma(0) = q_\sigma(0) = 0$, we can conclude that $p_\sigma(q) \equiv 0 \ \forall q$.

Similarly, $q_\sigma(p) \equiv 0 \ \forall p$. □

Notice that if the Hamiltonian is $\sqrt{p^2+q^2}$, our upwinding expression in (2.2) is identical to the conventional expression $\max(p_-^+, p_+^-)$. (In this case, the sign of the second argument does not matter since we are really evaluating its square product in the eikonal equation.) In fact, we have the following corollary, which is a direct consequence of Proposition 1.

Corollary 1. *If $H_{pp} > 0, H_{qq} > 0$, $pH_p \geq 0, qH_q \geq 0$, $p_\sigma(0) = q_\sigma(0) = 0$, and $H(p,q) = H(|p|,|q|)$, then the Godunov flux can be simplified to*

$$H_G(p_-, p_+; q_-, q_+) = H(\max\{p_-^+, p_+^-\}, \max\{q_-^+, q_+^-\}).$$

3. The sweeping algorithms.

We will use the model equation (1.1) as a concrete example for the exposition of our algorithm. We stress here again that the scheme described below is valid for a general class of convex, homogeneous HJ equations.

From the assumption that the solution is nondecreasing along the characteristics, i.e.,

$$pH_p + qH_q \geq 0,$$

we can easily deduce that the solution is nondecreasing at least in either the x- or y- direction; i.e., either $pH_p \geq 0$ or $qH_q \geq 0$. Since we approximate the derivatives $\phi_x(x_{i,j})$ by finite differencing using the neighbors of $\phi_{i,j}$, the above monotonicity property translates to the following requirement in the solution $\phi_{i,j}$.

Definition 1. *Let $\phi_{i,j}$ be the solution of $H_G(\phi, \phi_{i\pm1,j}, \phi_{i,j\pm j}) = r_{i,j}$. We say that ϕ satisfies the monotonicity requirement if*

$$\phi_{i,j} \geq \min\{\phi_{i\pm1,j}, \phi_{i,j\pm1}\}.$$

3.1. Derivation of the scheme.

Without loss of generality, we assume that $r(x,y) = 1$. Let us reexamine the equation to be solved:

$$H(p,q) = 1, \tag{3.1}$$

where

$$H : \mathbb{R} \times \mathbb{R} \to \mathbb{R}.$$

Equation (3.1) dictates a level set relation; namely, the solution is the 1-level set of H in the p-q plane (denoted here as Λ). Correspondingly, the solutions of the HJ equation with the Godunov Hamiltonian

$$H_G(p_+, p_-; q_+, q_-) = \underset{p \in I[p_-,p_+]}{\text{ext}} \ \underset{q \in I[q_-,q_+]}{\text{ext}} H(p,q) = r \tag{3.2}$$

satisfy the following two properties:

- they are the intersections of Λ and the set $I[p_-,p_+] \times I[q_-,q_+]$;
- they are either the critical points of H or the boundary points of the set $I[p_-,p_+] \times I[q_-,q_+]$.

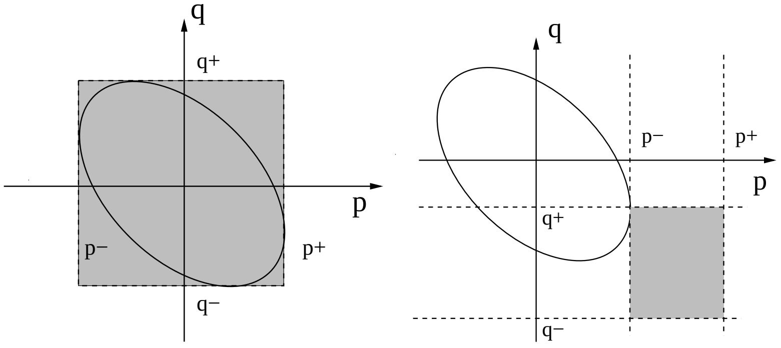


FIG. 1. *The* 1*-level set of* H *and the box* $I[p-,p+]xI[q-,q+]$.

Figure 1 demonstrates two possible configurations of the intervals. So what our algorithm should do is find a suitable value of ϕ on each grid node so that the divided forward and backward differences of ϕ at that grid node satisfy (3.2).

Suppose we are on the grid node (i,j), and it is determined that

$$H_G(p_+,p_-;q_+,q_-) = H(p_-,q_+) = 1.$$

Correspondingly, for our model equation (1.1) we have to solve the following quadratic equation:

$$a\left(\frac{\phi_{i,j}-\phi_{i-1,j}}{\Delta x}\right)^2 + b\left(\frac{\phi_{i,j+1}-\phi_{i,j}}{\Delta y}\right)^2 - 2c\left(\frac{\phi_{i,j}-\phi_{i-1,j}}{\Delta x}\right)\left(\frac{\phi_{i,j+1}-\phi_{i,j}}{\Delta y}\right) = 1. \tag{3.3}$$

The solution $\phi_{i,j}$ not only has to satisfy the above equation, but ultimately has to be a solution to (3.2), given its four neighbors $\phi_{i-1,j}$, $\phi_{i+1,j}$, $\phi_{i,j-1}$, and $\phi_{i,j+1}$. The subfigure on the right in Figure 1 shows one such possible configuration; i.e.,

$$\frac{\phi_{i,j}-\phi_{i-1,j}}{\Delta x} < \frac{\phi_{i+1,j}-\phi_{i,j}}{\Delta x} \quad \text{and} \quad \frac{\phi_{i,j}-\phi_{i,j-1}}{\Delta y} < \frac{\phi_{i,j+1}-\phi_{i,j}}{\Delta y}$$

such that

$$\operatorname*{ext}_{p\in I[p_-,p_+]}\ \operatorname*{ext}_{q\in I[q_-,q_+]} H(p,q) = \min_{p\in I[p_-,p_+]}\ \min_{q\in I[q_-,q_+]} H(p,q) = 1.$$

One can, of course, implement a tree of all the probable cases from the complete listing of that of the Godunov Hamiltonian (2.1). However, we have a more straightforward approach that utilizes the compact expressions for the Godunov Hamiltonian (2.2) that we obtained from the previous section.

Instead, we solve the equation with the following reduced formulas for the original Godunov Hamiltonian:

$$H_G(p_+,p_-;q_+,q_-) = \operatorname*{ext}_{q\in I[q_-,q_+]} H(p_-,q), \tag{3.4}$$

$$H_G(p_+,p_-;q_+,q_-) = \operatorname*{ext}_{q\in I[q_-,q_+]} H(p_+,q), \tag{3.5}$$

$$H_G(p_+, p_-; q_+, q_-) = \underset{p \in I[p_-, p_+]}{\text{ext}} H(p, q_-), \tag{3.6}$$

$$H_G(p_+, p_-; q_+, q_-) = \underset{p \in I[p_-, p_+]}{\text{ext}} H(p, q_+), \tag{3.7}$$

$$H_G(p_+, p_-; q_+, q_-) = H(p_\sigma, q_\sigma). \tag{3.8}$$

For example, in the first case, the flux is equivalent to

$$H(p_-, \operatorname{sgn}\max\{(q_- - q_\sigma)^+, (q_+ - q_\sigma)^-\} + q_\sigma) = 1.$$

The possible evaluations of $\operatorname{sgn}\max\{(q_- - q_\sigma)^+, (q_+ - q_\sigma)^-\} + q_\sigma$ are q_-, q_+, $q_\sigma(p_-)$, and 0. We thus end up solving the HJ equation with all possible arguments for the Hamiltonian.

Suppose we algebraically solve $H(p_-, q_+) = 1$ for $\phi_{i,j}$ and call the solution ϕ^{can}. We then compute the divided differences $p_\pm$ and $q_\pm$ using this ϕ^{can} in place of $\phi_{i,j}$. We call ϕ^{can} valid if both

$$H(p_-, \operatorname{sgn}\max\{(q_- - q_\sigma)^+, (q_+ - q_\sigma)^-\} + q_\sigma) = 1,$$

$$H(\operatorname{sgn}\max\{(p_- - p_\sigma)^+, (p_+ - p_\sigma)^-\} + p_\sigma\, , q_+) = 1,$$

and ϕ^{can} satisfies the monotonicity requirement (Definition 1).

Finally, we set $\phi_{i,j}$ to be the minimum of those in the set of all valid candidate solutions ϕ^{can} obtained from using all the possible combinations of the arguments of H. This is motivated by the first arrival time interpretation of the function ϕ.

In essence, we are solving for the central value in the Godunov Hamiltonian in terms of its four neighbors. It is well known and easy to show that any monotone Hamiltonian, let alone Godunov's, is a monotone function of this value. For these Hamiltonians, this value goes from $-\infty$ to $+\infty$. Thus there is always a unique solution.

Definition 2 (sweeping iteration). *A compact way of writing this sweeping iterations in C/C++ is the following:*

```
for(s1=-1;s1<=1;s1+=2)
for(s2=-1;s2<=1;s2+=2)
for(i=(s1<0?nx:0);(s1<0?i>=0:i<=nx);i+=s1)
for(j=(s2<0?ny:0);(s2<0?j>=0:j<=ny);j+=s2)
  update $\phi_{i,j}$.
```

3.2. The algorithm. For the brevity of the algorithm, we define respectively

$$h_{G1}(p, q_-, q_+) := \operatorname{sgn}\max\{(q_- - q_\sigma(p))^+, (q_+ - q_\sigma(p))^-\} + q_\sigma(p),$$

$$h_{G2}(p_-, p_+, q) := \operatorname{sgn}\max\{(p_- - p_\sigma(q))^+, (p_+ - p_\sigma(q))^-\} + p_\sigma(q),$$

where $q_\sigma(p) = pc/b$ and $p_\sigma(q) = qc/a$.

Algorithm. We assume that $\phi(i,j)$ is given the exact values in a small neighborhood of Γ. We denote this neighborhood $\text{Nbd}(\Gamma)$. We initialize ϕ by setting $\phi(i,j) = \phi^{(0)}_{i,j}$ to ∞.[1] We begin by computing $\phi^{(n)}_{i,j}$ for $n = 1$.

Do the following steps while $||\phi^{(n)} - \phi^{(n-1)}|| > \delta$ ($\delta > 0$ is the given tolerance):

[1] Notice that we only need to use a large value in actual implementation.

1. For each grid point (i, j) visited in the sweeping iteration, if $x_{i,j} \neq \mathrm{Nbd}(\Gamma)$, do the following:
   - (a) For $(s_x, s_y) = (-1, 1)$, $(-1, -1)$, $(1, -1)$, and $(1, 1)$
     - i. Solve

       $$H\left(\tfrac{s_x \cdot (\phi_{tmp}(s_x,s_y) - \phi^{(n)}(i-s_x,j))}{dx}, \tfrac{s_y \cdot (\phi_{tmp}(s_x,s_y) - \phi^{(n)}(i,j-s_y))}{dy}\right) = r(i, j)$$

       for $\phi_{tmp}(s_x, s_y)$.
     - ii. Let

       $$p(s_x, s_y) = \frac{s_x \cdot (\phi_{tmp}(s_x, s_y) - \phi^{(n)}(i - s_x, j))}{dx}$$

       and

       $$q(s_x, s_y) = \frac{s_y \cdot (\phi_{tmp}(s_x, s_y) - \phi^{(n)}(i, j - s_y))}{dy}.$$
     - iii. Let $T_{G1}(s_x, s_y)$ be the logical evaluation of the equality

       $$H(p(s_x, s_y), h_{G1}(p(s_x, s_y), q(s_x, 1), q(s_x, -1))) = r(i, j),$$

       and $T_{G2}(s_x, s_y)$ be that of

       $$H(h_{G2}(p(1, s_y), p(-1, s_y), q(s_x, s_y)), q(s_x, s_y)) = r(i, j).$$
     - iv. Let $M(s_x, s_y) = \phi_{tmp}(s_x, s_y) - \min(\phi^{(n)}(i - s_x, j), \phi^{(n)}(i, j - s_y))$.
     - v. If $T_{G1}(s_x, s_y)$, $T_{G2}(s_x, s_y)$ are true and $M(s_x, s_y) \geq 0$, add $\phi_{tmp}(s_x, s_y)$ to the list `phi_candidate.`
   - (b) For $(s_x, s_y) = (1, 0)$, $(-1, 0)$
     - i. Solve

       $$H\left(\tfrac{s_x \cdot (\phi_{tmp}(s_x,0) - \phi^{(n)}(i-s_x,j))}{dx}, \tfrac{s_x \cdot (\phi_{tmp}(s_x,0) - \phi^{(n)}(i-s_x,j))}{dx} \tfrac{c}{b}\right) = r(i, j)$$

       for $\phi_{tmp}(s_x, s_y)$.
     - ii. Compute $p(s_x, s_y)$ and $q(s_x, s_y)$, following the definition.
     - iii. Evaluate $T_{G1}(s_x, s_y)$.
     - iv. If $T_{G1}(s_x, s_y)$ is true and $M(s_x, s_y) \geq 0$, add $\phi_{tmp}(s_x, s_y)$ to the list `phi_candidate.`
   - (c) For $(s_x, s_y) = (0, 1)$ and $(0, -1)$
     - i. Solve

       $$H\left(\tfrac{s_y \cdot (\phi_{tmp}(0,s_y) - \phi^{(n)}(i,j-s_y))}{dy} \tfrac{c}{a}, \tfrac{s_y \cdot (\phi_{tmp}(0,s_y) - \phi^{(n)}(i,j-s_y))}{dy}\right) = r(i, j)$$

       for $\phi_{tmp}(s_x, s_y)$.
     - ii. Compute $p(s_x, s_y)$ and $q(s_x, s_y)$, following the definition.
     - iii. Evaluate $T_{G2}(s_x, s_y)$.
     - iv. If $T_{G2}(s_x, s_y)$ is true and $M(s_x, s_y) \geq 0$, add $\phi_{tmp}(s_x, s_y)$ to the list `phi_candidate.`
   - (d) Let ϕ_{min} be the minimum element of `phi_candidate`.

     $$\phi^{(n)}(i, j) = \min(\phi^{(n)}(i, j), \phi_{min}).$$

(e) Clear `phi_candidate`.

2. Set $n = n + 1$; go back to step 1.

As described in the previous section, we have to solve the HJ equation with all possible arguments for the Hamiltonian and take the minimum of those in the set of all valid candidate solutions. The possible arguments of the Hamiltonian consist of the forward/backward differences of ϕ and the critical points centered at each grid node. In the above algorithm, this set of all possible arguments is indexed by $\{-1, 0, 1\}^2$. Therefore, by $X(-1, 1)$ we denote the quantity X that is computed using $H(p_-, q_+)$. The number 0 encodes the cases of critical points. For example, $\phi_{tmp}(-1, 1)$ denotes the roots of the quadratic equation formed by $H(p_-, q_+) = r$; $\phi_{tmp}(1, 0)$ denotes that of $H(p_+, q_\sigma(p_+))$.

We remark that in the case of $c = 0$, our algorithm is equivalent to what is used in the fast marching method under the Rouy–Tourin formula (1.3). Secondly, in our numerical implementation, we put a threshold value in the evaluations T_{G1} and T_{G2} for numerical accuracy reasons.

4. Examples.

Proposition 1 and Corollary 1 show the equivalence of the Godunov flux derived in this paper to the one commonly used in the fast marching applications. The use of this sweeping approach with the Godunov flux (1.3) has been reported in [24, 26] for eikonal equations; we will not repeat those examples in this paper. Instead, we present results of our algorithm applied to our model equation.

4.1. Quadratic Hamiltonians $\sqrt{ap^2 + bq^2 - 2cpq}$, $ab > c^2$, $a, b > 0$.

In each of the following examples, we compute the difference in the approximations in each successive iteration, i.e., $||\phi^{n+1} - \phi^n||_{L_1}$, and say that the iterations have converged if this distance is less than $\varepsilon\Delta x$, where $\varepsilon > 0$ and Δx is the grid size. In the examples presented in this paper, we simply set the threshold to be 10^{-10}. Notice also that the set Γ, on which $\phi = 0$, is either a rectangle, an L-shaped piecewise linear object, or a set of isolated points. The reader can identify their location easily from the figures.

We started out by testing our algorithm on constant coefficient cases. In the case

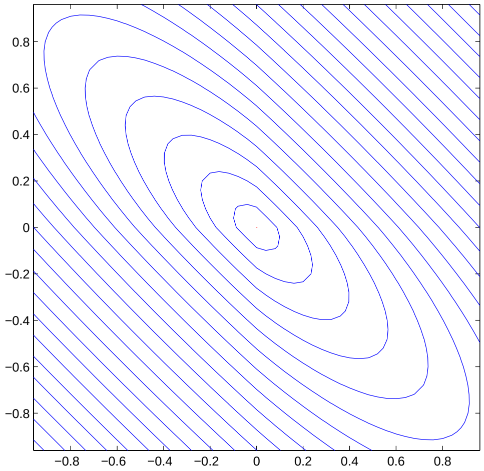


FIG. 2. *A sweeping result after* 2 *sweeping iterations on a* 50×50 *grid. The initial boundary is a single point in the center.* $a = 1.0$, $b = 1.0$, $c = 0.9$.

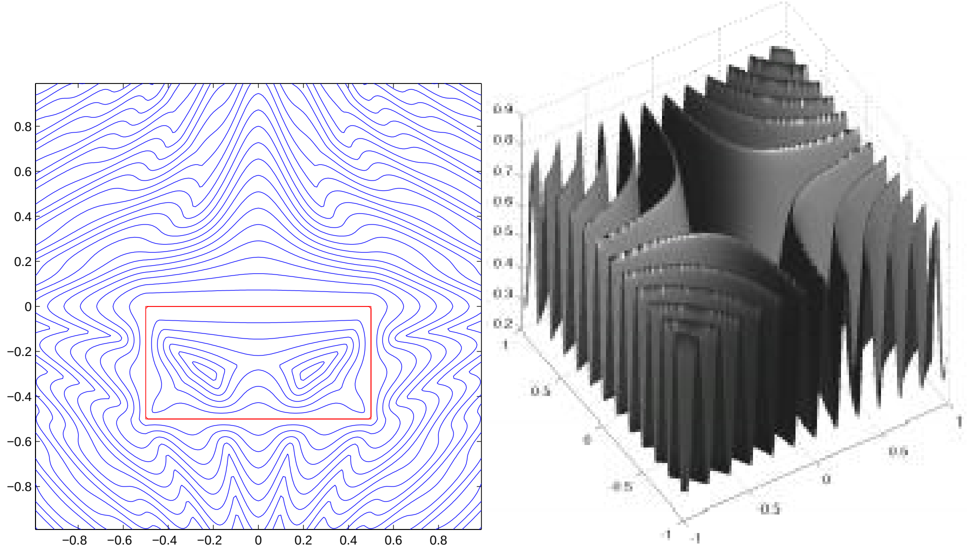


Fig. 3. *$a = 1$, $b = 1$, $c = 0$, with a more oscillatory $r(x) = 2.1 - \cos(4\pi^2 xy)$, on a 200×200 grid; convergence is reached in 7 sweeping iterations. The subplot on the left is the contour of the solution started with the square in the center. On the right is the graph of $r(x)$. Level curves with step 0.02 are plotted.*

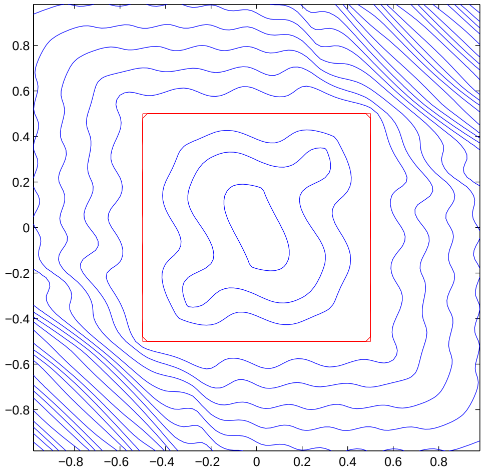


Fig. 4. *(A very degenerate case) $a = 0.375$, $b = 0.25$, $c = 0.29$, with a more oscillatory $r(x) = (2.1 - \cos(4\pi^2 xy))/4.0$, on a 100×100 grid. Notice that, in this case, $ab = 0.0938$ is barely greater than $c^2 = 0.0841$. The contour of the solution is plotted. Convergence is reached at 43 sweeping iterations.*

of $a = b$, $c = 0$, we have solutions that match the fast marching solutions. Figure 2 shows a result of a computation of the anisotropic case in which $a = b = 1$, $c = 0.9$. This is our first example in which the fast marching method is not applicable.

Next we apply the sweeping algorithm directly to cases in which the coefficients of the quadratic Hamiltonian or the right-hand sides are not constant. Figure 3 shows a computational result on a constant coefficient isotropic Hamiltonian and rather oscillatory forcing function. The rectangle in the middle is the set Γ. Figure 4

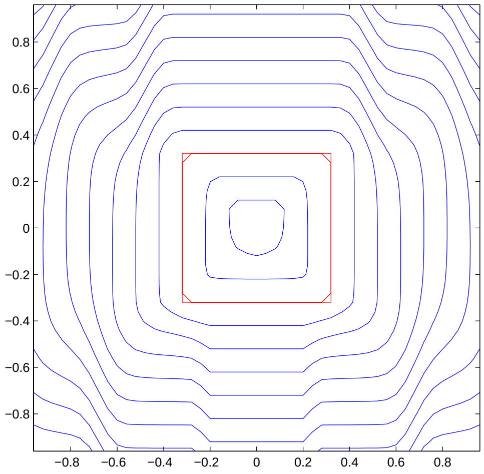


Fig. 5. *$a = 1$, $b = 1$, $c(x, y) = 0.9 \sin(5\pi x)$, and $r(x, y) = 1$, on a 50×50 grid. Convergence occurs after 10 iterations.*

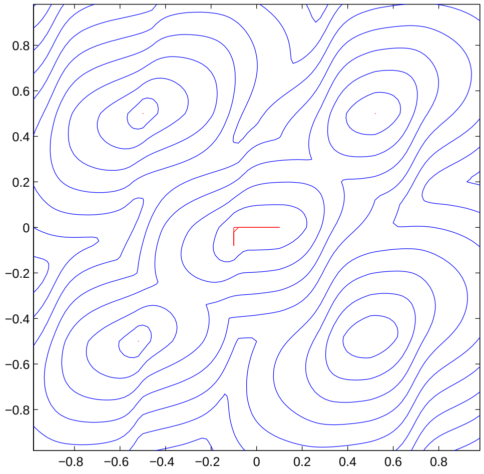


Fig. 6. *$a = 1.5 + \sin(5\pi x)$, $b = 1$, $c = -0.6$, on a 50×50 grid. Convergence occurs after 10 iterations.*

shows a computational result for a very anisotropic case. We notice that the number of iterations needed for convergence seems to depend on the anisotropy of the Hamiltonian and also on how oscillatory the forcing term is. Figures 5, 6, and 7 show results obtained from variable coefficient Hamiltonians with constant and variable forcing function $r(x, y)$.

4.2. Examples of distance on manifolds. We now apply our sweeping algorithm to compute the geodesic distance on manifolds that are the graphs of certain functions. Given a function $f(x, y)$, with graph $z = f(x, y)$, we compute the coefficients $a(x, y)$, $b(x, y)$, and $c(x, y)$ according to (1.5) and apply our algorithm directly to the corresponding HJ equation. We first test the algorithm on a half-sphere with radius one. Figures 8 and 9 show the equidistance lines to one and two seed points,

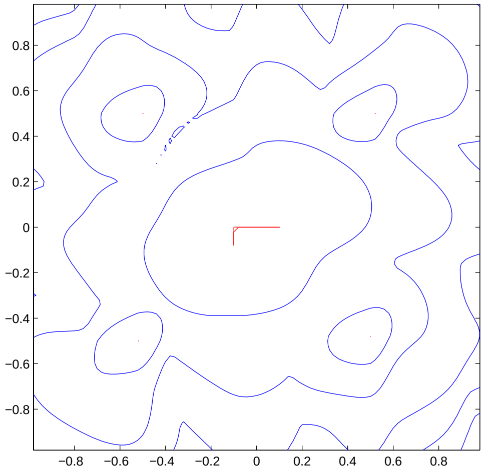


FIG. 7. *$a = 1.5 + \sin(5\pi x)$, $b = 1$, $c = -0.6$, and $r(x, y) = 2.1 + \cos(4\pi xy)$, on a 100×100 grid. Convergence occurs after 10 iterations.*

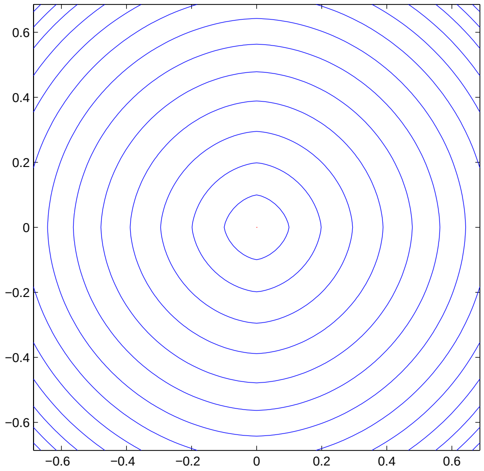


FIG. 8. *This is an example of the distance on a half-sphere. The sweeping algorithm was applied to the graph of $f(x, y) = \sqrt{1.0 - (x^2 + y^2)}$, with $\phi(0, 0) = 0$ as boundary condition, on a 100×100 grid.*

respectively. Figures 10, 11, and 12 show similar computation results applied to somewhat more oscillatory manifolds. As we expected, more sweeping iterations are required for convergence.

4.3. Grid effects. We first perform a rotation of the coordinate system. We represent this by

$$(x, y) \mapsto (\tilde{x}, \tilde{y})$$

and let

$$(a, b, c) \mapsto (\tilde{a}, \tilde{b}, \tilde{c}).$$

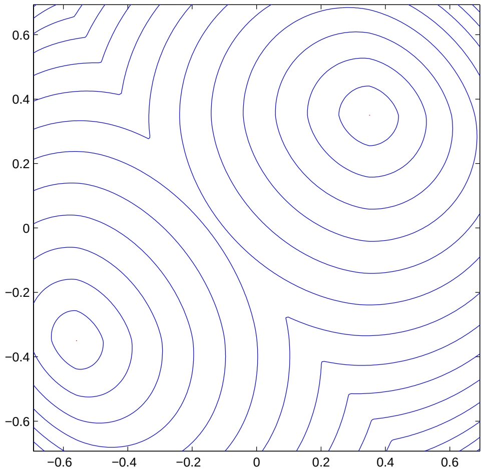


FIG. 9. *This is an example of the distance on a half-sphere. The sweeping algorithm was applied to the graph of* $f(x, y) = \sqrt{1.0 - (x^2 + y^2)}$, *with two seed points. Convergence is reached after* 2 *sweeping iterations.*

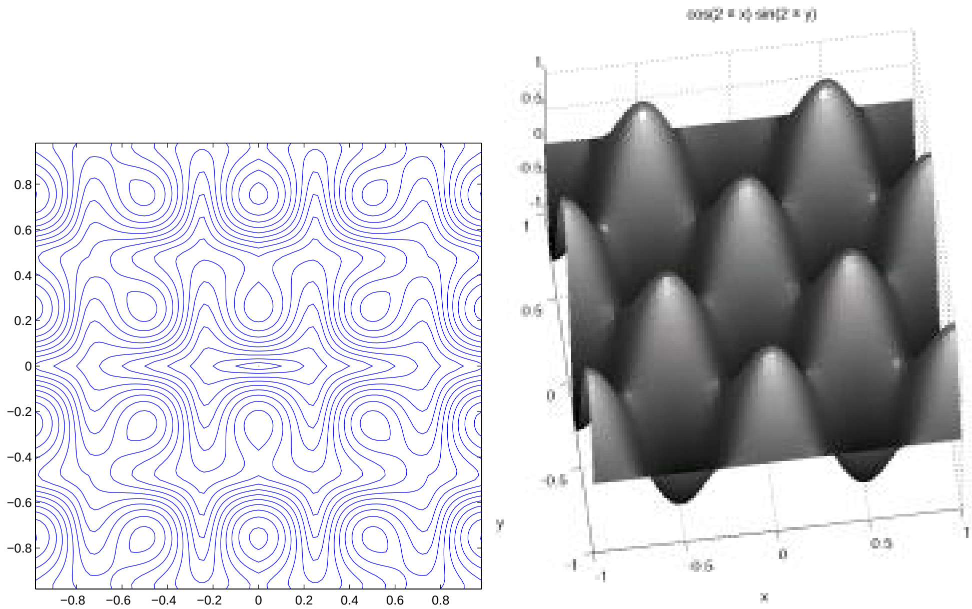


FIG. 10. *The distance contour from the seed point* $(0, 0)$ *on the graph of* $f(x, y) = \cos(2\pi x)\sin(2\pi y)$, *on a* 100×100 *grid. Convergence occurs after* 9 *iterations.*

To study the grid effects of our sweeping algorithm, we set $u = 0$ on a rotated square whose sides do not align with the grid lines. Comparing the results, shown in Figure 13, we see that the second picture, concentrating especially on the diamond-shaped contour in the middle, indeed shows grid effects compared to the first picture. However, with further grid refinement, as shown in the third picture, grid effects become unnoticeable, and the solution from our sweeping algorithm accurately ap-

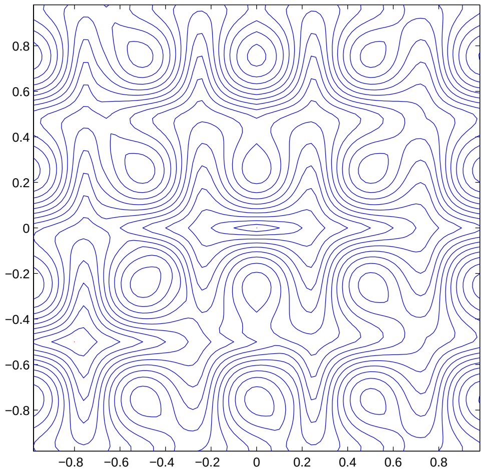


Fig. 11. *The distance contour from the seed point* $(0,0)$ *and* $(-0.8,-0.5)$ *on the graph of* $f(x,y)=\cos(2\pi x)\sin(2\pi y)$, *on a* 100×100 *grid. Convergence occurs after* 11 *iterations.*

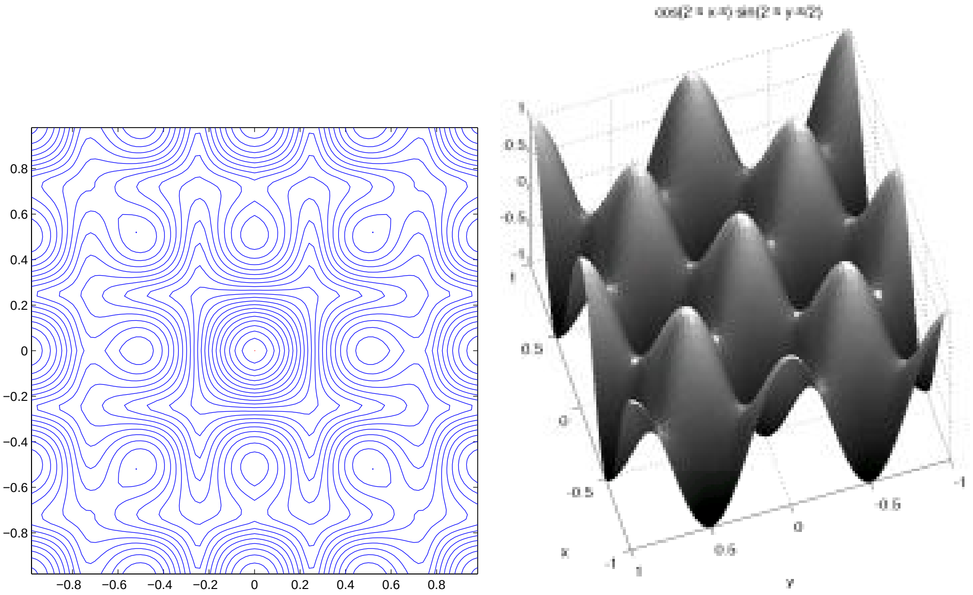


Fig. 12. *The distance contour from the seed point* $(0,0)$ *on the graph of* $f(x,y)=\cos(2\pi x-\pi)\sin(2\pi y-\pi/2)$, *on a* 100×100 *grid. Convergence occurs after* 9 *iterations.*

proximates the exact solution.

4.4. Comparison with the time marching solutions. We use the first order Runge–Kutta–Lax–Friedrichs method [18] to discretize the following equation and march to steady state:

$$\tilde{\phi}_t+\operatorname{sgn}(\phi(x,y))(H(x,y,\tilde{\phi}_x,\tilde{\phi}_y)-r(x,y))=0, \tag{4.1}$$

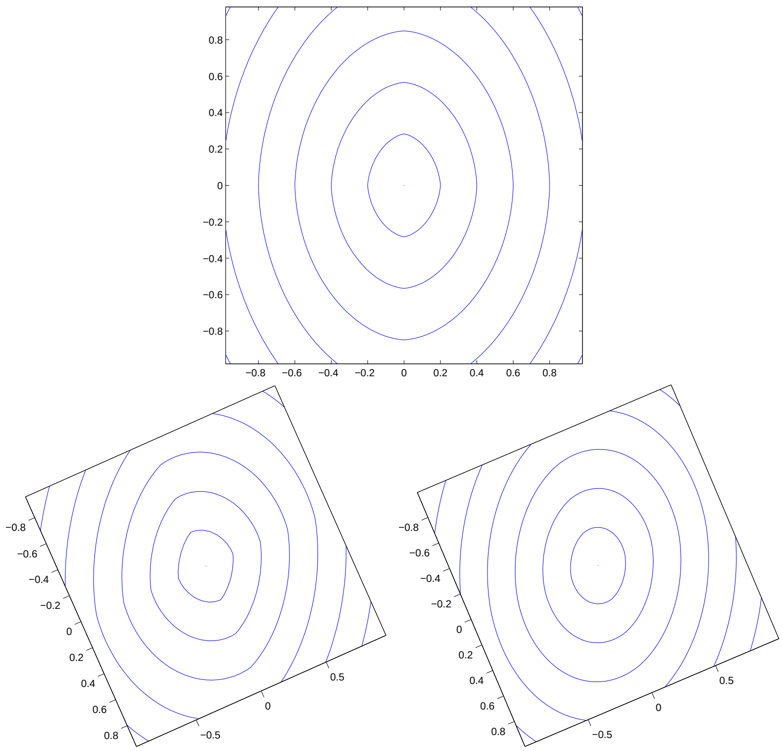


Fig. 13. *Anisotropic case with a point source at* $(0,0)$. $a = 1, b = 1, c = 0.9$ *and* $\tilde{a} = 1.70365$ $\tilde{b} = 0.296352$, *and* $\tilde{c} = -0.561141$, *on* 50×50 *and* 100×100 *grids. Convergence occurs after* 2 *iterations.*

Table 1

Comparison of the time marching and sweeping solutions to the example shown in Figure 12.

	$dx = 2/50$	2/100	2/200	2/400	2/800
$\|\|\phi - \tilde{\phi}\|\|_{L_1}$	2.85423	1.83377	1.04008	0.56206	0.295738
$\|\|\phi - \tilde{\phi}\|\|_\infty$	1.03825	0.708986	0.436469	0.246439	0.133858

where $\tilde{\phi}(x, y, t = 0) = \phi(x, y) = 0$ for $(x, y) \in \Gamma$ and ϕ is the solution obtained from the sweeping algorithm.

We remark that solving (4.1) is by no means a practical method for solving the steady state equation. Thousands of iterations are required for steady state, even if we take ϕ as the initial Cauchy data. We use it only to verify the validity of our algorithm. Secondly, the solutions of (4.1) suffer from excessive smearing due to the numerical viscosity introduced by the Lax–Friedrichs method. As a consequence, $\tilde{\phi}$ does not match well with ϕ on coarse grids. The reader can compare Figure 14 with Figure 12, for example. However, we do see that $||\phi - \tilde{\phi}||$ decreases with the refinement of the grid size; see Table 1 and Figure 14. We remark that higher order approximation schemes such as RK3-WENO5 will greatly reduce the numerical viscosity; the reader is referred to [18]. Our purpose here is only to show that the sweeping approximations converge to the viscosity solution.

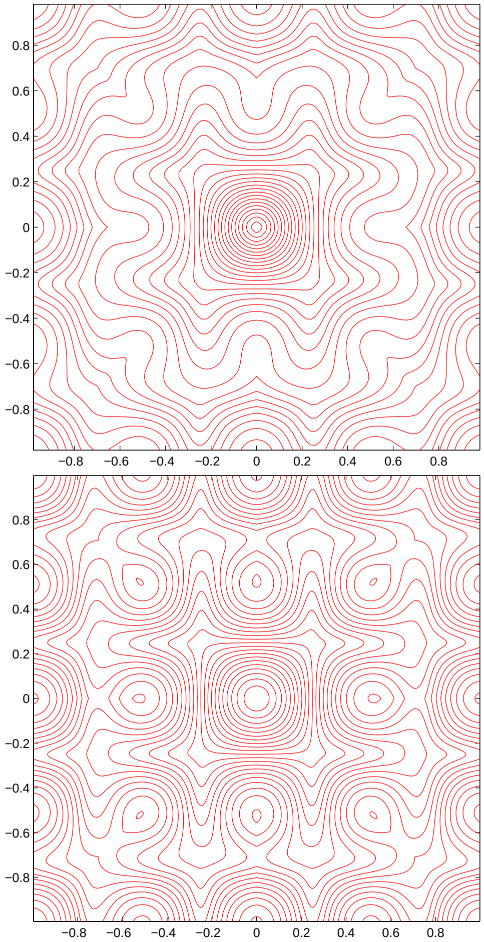


Fig. 14. *Steady state of the time marching on a* 100×100 *and* 800×800 *grid.*

Table 2
A numerical convergence study of the sweeping algorithm applied to the graph of $f(x, y) = \sqrt{1.0 - (x^2 + y^2)}$, *with* $\phi(0, 0) = 0$ *as boundary condition on the domain* $[-0.7, 0.7] \times [-0.7, 0.7]$.

	$dx = 1.4/200$	1.4/400	1.4/800	1.4/1600
$\|\|\phi - \tilde{\phi}\|\|_{L_1}$	0.0138803	0.0079927	0.00453004	0.00253513
rate		0.796	0.819	0.84

4.5. Numerical convergence. Since we can easily compute the geodesic distance on a sphere, we will use it as an example to show numerical convergence of our algorithm. A distance contour plot is shown in Figure 8. Table 2 shows a numerical convergence of order 1. We have also noticed that the number of iterations needed for the L_1 difference of the approximations in each successive iteration to decrease below the given tolerance seems to be bounded independently of the grid size. This number seems to depend on the anisotropy (c^2/ab), the forcing function r, and the configuration of the interface Γ.

5. Conclusion.

In this article, we studied a fast method for solving a class of time independent HJ equations with Dirichlet boundary conditions. The Hamiltonians of interest are homogeneous and convex. This fast method combines the idea of tracing the characteristics with Godunov construction and Gauss–Seidel iterations with smart choices of different updating sequences. In particular, we discussed some important properties of the Hamiltonian $H = \sqrt{ap^2 + bq^2 - 2cpq}$, $c^2 < ab$, and the corresponding HJ equations. By the simple structure of the convexity, we derived a compact expression for the Godunov Hamiltonian that involves taking extrema of the Hamiltonian in relation to the evaluations of the derivatives of the solution. With our compact Godunov flux, the complexity of evaluating the Godunov Hamiltonian is reduced to only eight cases in two space dimensions. We then incorporated the expression into a simple Gauss–Seidel-type iteration procedure. We have produced some computational results using this algorithm. In particular, we have applied our algorithm to compute geodesic distances on graphs of functions. This is of some importance since people are interested in finding the geodesics on terrain-like manifolds.

We also remark that this Godunov-flux sweeping approach can be extended to higher dimensional cases. We are currently preparing another paper on this subject.

Our experience shows that the number of iterations needed depends on the amount of anisotropy and the nature of the forcing function. Under normal nondegenerate circumstances, experience shows an $\mathcal{O}(N)$ complexity for convergence, where N is the number of grid points. Recently, in [26], the author provided some theoretical evidence on the bound of the number of iterations for isotropic, homogeneous eikonal equations. This points out a future research direction of bounding the number of sweeping iterations needed for convergence in relation to the anisotropy.

6. Appendix.

6.1. Derivation of the flux for homogeneous convex Hamiltonians.

To obtain the formula used earlier in this paper, we simply verify its equivalence to the following cases, which rely only on the convexity of H:

$p_- < p_+$, and $q_- < q_+$:

$$H_G = \min_{p\in[p_-,p_+]} \min_{q\in[q_-,q_+]} H(p,q).$$

- If $q_\sigma \in [q_-, q_+]$,
  - $p_\sigma < p_- < p_+$, $H(p_-, q_\sigma)$,
  - $p_- < p_+ < p_\sigma$, $H(p_+, q_\sigma)$,
  - $p_- < p_\sigma < p_+$, $H(p_\sigma, q_\sigma)$.
- If $q_\sigma < q_-$,
  - $p_\sigma < p_- < p_+$, $H(p_-, q_-)$,
  - $p_- < p_+ < p_\sigma$, $H(p_+, q_-)$,
  - $p_- < p_\sigma < p_+$, $H(p_\sigma, q_-)$.
- If $q_\sigma > q_+$,
  - $p_\sigma < p_- < p_+$, $H(p_-, q_+)$,
  - $p_- < p_+ < p_\sigma$, $H(p_+, q_+)$,
  - $p_- < p_\sigma < p_+$, $H(p_\sigma, q_+)$.

$p_- < p_+$, and $q_- > q_+$:

$$H_G = \min_{p\in[p_-,p_+]} \max_{q\in[q_+,q_-]} H(p,q) = \min_{p\in[p_-,p_+]} \max\{H(p,q_-), H(p,q_+)\}.$$

- If $q_\sigma < q_+$,
  - $p_\sigma < p_- < p_+$, $H(p_-, q_-)$,
  - $p_- < p_+ < p_\sigma$, $H(p_+, q_-)$,
  - $p_- < p_\sigma < p_+$, $H(p_\sigma, q_-)$.
- If $q_\sigma > q_-$,
  - $p_\sigma < p_- < p_+$, $H(p_-, q_+)$,
  - $p_- < p_+ < p_\sigma$, $H(p_+, q_+)$,
  - $p_- < p_\sigma < p_+$, $H(p_\sigma, q_+)$.
- If $q_+ < q_\sigma < q_-$,
  - $(q_\sigma - q_+) > (q_- - q_\sigma)$, $H(\cdot, q_+)$,
  - $(q_\sigma - q_+) \leq (q_- - q_\sigma)$, $H(\cdot, q_-)$.

$p_- > p_+$, and $q_- > q_+$:

$$H_G = \max_{p \in [p_+, p_-]} \max_{q \in [q_+, q_-]} H(p, q).$$

- If $q_\sigma > q_-$,
  - $p_\sigma > p_-$, $H(p_+, q_+)$,
  - $p_\sigma < p_+$, $H(p_-, q_+)$.
- If $q_\sigma < q_+$,
  - $p_\sigma > p_-$, $H(p_+, q_-)$,
  - $p_\sigma < p_+$, $H(p_-, q_-)$.
- If $q_+ < q_\sigma < q_-$
  - $(q_\sigma - q_+) > (q_- - q_\sigma)$, $H(\cdot, q_+)$,
  - $(q_\sigma - q_+) \leq (q_- - q_\sigma)$, $H(\cdot, q_-)$.

$p_- > p_+$, and $q_- < q_+$:

$$H_G = \max_{p \in [p_+, p_-]} \min_{q \in [q_-, q_+]} H(p, q).$$

- If $q_\sigma \in [q_-, q_+]$,
  - $p_\sigma > p_-$, $H(p_+, q_\sigma)$,
  - $p_\sigma < p_+$, $H(p_-, q_\sigma)$.
- If $q_\sigma < q_-$,
  - $p_\sigma > p_-$, $H(p_+, q_-)$,
  - $p_\sigma < p_+$, $H(p_-, q_-)$.
- If $q_\sigma > q_+$,
  - $p_\sigma > p_-$, $H(p_+, q_+)$,
  - $p_\sigma < p_+$, $H(p_-, q_+)$.

Acknowledgments. YT thanks Paul Burchard and C.Y. Kao for their useful conversations on this topic. The authors also thank Prof. Giovanni Russo for his helpful suggestions to improve this paper.

REFERENCES

[1] M. Bardi and S. Osher, *The nonconvex multi-dimensional Riemann problem for Hamilton–Jacobi equations*, SIAM J. Math. Anal., 22 (1991), pp. 344–351.

[2] T. J. Barth, *On the Marchability of Interior Stabilized Discontinuous Galerkin Approximations of the Eikonal and Related PDEs with Non-Divergence Structure*, NASA Technical report, NAS-01-010, NASA Ames Research Center, Moffett Field, CA, 2001.

[3] M. Boué and P. Dupuis, *Markov chain approximations for deterministic control problems with affine dynamics and quadratic cost in the control*, SIAM J. Numer. Anal., 36 (1999), pp. 667–695.

[4] L.-T. Cheng, P. Burchard, B. Merriman, and S. Osher, *Motion of curves constrained on surfaces using a level set approach*, J. Comput. Phys., 175 (2002), pp. 604–644.

[5] M.G. Crandall and P.L. Lions, *Two approximations of solutions of Hamilton–Jacobi equations*, Math. Comp., 43 (1984), pp. 1–19.

[6] M. G. Crandall and P.-L. Lions, *Viscosity solutions of Hamilton–Jacobi equations*, Trans. Amer. Math. Soc., 277 (1983), pp. 1–42.

[7] P.-E. Danielsson, *Euclidean distance mapping*, Computer Graphics and Image Processing, 14 (1980), pp. 227–248.

[8] J. Helmsen, E. Puckett, P. Colella, and M. Dorr, *Two new methods for simulating photolithography development in* 3*d*, in SPIE 2726, Bellingham, WA, 1996, pp. 253–261.

[9] G.-S. Jiang and D. Peng, *Weighted ENO schemes for Hamilton–Jacobi equations*, SIAM J. Sci. Comput., 21 (2000), pp. 2126–2143.

[10] J. B. Keller, *Geometrical theory of diffraction*, J. Opt. Soc. Amer., 52 (1962), pp. 116–130.

[11] R. Kimmel and J. A. Sethian, *Computing geodesic paths on manifolds*, Proc. Natl. Acad. Sci. USA, 95 (1998), pp. 8431–8435.

[12] F. Memoli and G. Sapiro, *Fast computation of weighted distance functions and geodesics on implicit hyper-surfaces*, J. Comput. Phys., 173 (2001), pp. 730–764.

[13] S. Osher, *A level set formulation for the solution of the Dirichlet problem for Hamilton–Jacobi equations*, SIAM J. Math. Anal., 24 (1993), pp. 1145–1152.

[14] S. Osher and R. P. Fedkiw, *Level set methods: An overview and some recent results*, J. Comput. Phys., 169 (2001), pp. 463–502.

[15] S. Osher and J. Helmsen, *A Generalized Fast Algorithm with Applications to Ion Etching*, manuscript.

[16] S. Osher and B. Merriman, *The Wulff shape as the asymptotic limit of a growing crystalline interface*, Asian J. Math., 1 (1997), pp. 560–571.

[17] S. Osher and J. A. Sethian, *Fronts propagating with curvature-dependent speed: Algorithms based on Hamilton–Jacobi formulations*, J. Comput. Phys., 79 (1988), pp. 12–49.

[18] S. Osher and C.-W. Shu, *High-order essentially nonoscillatory schemes for Hamilton–Jacobi equations*, SIAM J. Numer. Anal., 28 (1991), pp. 907–922.

[19] D. Peng, S. Osher, B. Merriman, and H.-K. Zhao, *The geometry of Wulff crystal shapes and its relations with Riemann problems*, in Nonlinear Partial Differential Equations (Proceedings of the Nonlinear PDE Emphasis Year meeting in Evanston, IL, 1998), G.-Q. Chen and E. DiBenedetto, eds., AMS, Providence, RI, 1999, pp. 251–303.

[20] E. Rouy and A. Tourin, *A viscosity solutions approach to shape-from-shading*, SIAM J. Numer. Anal., 29 (1992), pp. 867–884.

[21] J. A. Sethian and A. Vladimirsky, *Fast methods for the eikonal and related Hamilton–Jacobi equations on unstructured meshes*, Proc. Natl. Acad. Sci. USA, 97 (2000), pp. 5699–5703.

[22] J. A. Sethian and A. Vladimirsky, *Ordered upwind methods for static Hamilton–Jacobi equations*, Proc. Natl. Acad. Sci. USA, 98 (2001), pp. 11069–11074.

[23] J. A. Sethian, *Fast marching level set methods for three dimensional photolithography development*, in SPIE 2726, Bellingham, WA, 1996, pp. 261–272.

[24] Y.-H. R. Tsai, *Rapid and accurate computation of the distance function using grids*, J. Comput. Phys., 178 (2002), pp. 175–195.

[25] J. Tsitsiklis, *Efficient algorithms for globally optimal trajectories*, IEEE Trans. Automat. Control, 40 (1995), pp. 1528–1538.

[26] H.-K. Zhao, *Fast sweeping method for eikonal equations* I: *Distance function*, SIAM J. Numer. Anal., submitted; also available at www.math.uci.edu/~zhao, 2002.

[27] H.-K. Zhao, S. Osher, B. Merriman, and M. Kang, *Implicit and non-parametric shape reconstruction from unorganized points using variational level set method*, Computer Vision and Image Understanding, 80 (2000), pp. 295–319.

SIAM J. Numer. Anal.
Vol. 41, No. 2, pp. 695–714

THE POSTPROCESSING GALERKIN AND NONLINEAR GALERKIN METHODS—A TRUNCATION ANALYSIS POINT OF VIEW[*]

LEN G. MARGOLIN[†], EDRISS S. TITI[‡], AND SHANNON WYNNE[§]

Abstract. We revisit the postprocessing algorithm and give a justification from a classical truncation analysis point of view. We assume a perturbation expansion for the high frequency mode component of solutions to the underlying equation. Keeping terms to certain orders, we then generate approximate systems which correspond to numerical schemes. We show that the first two leading order methods are in fact the postprocessed Galerkin and postprocessed nonlinear Galerkin methods, respectively. Hence postprocessed Galerkin is a natural leading order method, more natural than the standard Galerkin method, for approximating solutions of parabolic dissipative PDEs. The analysis is presented in the framework of the two-dimensional Navier–Stokes equation (NSE); however, similar analysis may be done for any parabolic, dissipative nonlinear PDE.

The truncation analysis is based on asymptotic estimates (in time) for the low and high mode components. We also introduce and investigate an alternative postprocessing scheme, which we call the dynamic postprocessing method, for the case in which the asymptotic estimates (in time) do not hold (i.e., in the situation of long transients, nonsmooth initial data, or highly oscillatory time-dependent solutions).

Key words. dissipative equations, spectral methods, approximate inertial manifolds, nonlinear Galerkin methods, postprocessing algorithm, multigrid

AMS subject classification. 65P25

PII. S0036142901390500

1. Introduction.

We revisit the postprocessing algorithm for the Galerkin and nonlinear Galerkin methods. Postprocessing methods first evolved from the theory of approximate inertial manifolds (AIMs) (see, e.g., [3], [5], [6], [7], [15], [19], and [21]) and take advantage of the observation that, for dissipative evolution equations, the Galerkin and nonlinear Galerkin methods do better approximating the low modes of the exact solution u than approximating the solution itself. AIMs are used to "post-process" the low modes in order to obtain a more accurate approximation for the high modes. For a variety of applications, the postprocessed Galerkin has been shown to be a very efficient algorithm for improving the accuracy of Galerkin/nonlinear Galerkin methods with very little extra computational cost (see, for example, [8], [10], [11], [12], and [17]). However, postprocessing is not simply a technique for improving efficiency. In this paper we show that postprocessing methods arise in a very natural way through a classical truncation analysis of the dissipative evolution equation. More specifically, we will show that, to leading order, the correct approximative scheme is

[*] Received by the editors June 17, 2001; accepted for publication (in revised form) August 20, 2002; published electronically May 6, 2003. This work was supported in part by National Science Foundation grants DMS-9706964 and DMS-9704632, by the Department of Energy under contract W-7405-ENG-36, and by the U.S. Air Force Office of Scientific Research under grants AFOSR-F49620-98-1-0180 and AFOSR-49620-01-1-0026.

http://www.siam.org/journals/sinum/41-2/39050.html

[†] Institute for Geophysics and Planetary Physics and Center of Nonlinear Studies, Los Alamos National Laboratory, Los Alamos, NM 87545 (len@lanl.gov).

[‡] Department of Mathematics and Department of Mechanical and Aerospace Engineering, University of California, Irvine, CA 92697-3875 and Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot 76100, Israel (etiti@math.uci.edu).

[§] Center for Research in Scientific Computing, North Carolina State University, Raleigh, NC 27695-8205 (snwynne@unity.ncsu.edu).

actually the postprocessed Galerkin method, and not the standard Galerkin method as is commonly believed.

We present this work in the context of the two-dimensional Navier–Stokes equations (NSE) in Ω, an open bounded set of $\mathbb{R}^2$, with smooth boundary $\partial\Omega$,

$$\begin{aligned} \frac{\partial u}{\partial t} - \nu\Delta u + (u\cdot\nabla)u + \nabla\pi &= f, \\ \nabla\cdot u &= 0, \\ u(0,x) &= u_0(x), \end{aligned} \tag{1.1}$$

where the unknowns are the vector velocity u and the scalar pressure π; $f(x,t)$ is a given body forcing, and $\nu > 0$ is the kinematic constant viscosity. The equations are subject to either nonslip Dirichlet boundary conditions for $\partial\Omega$ smooth enough, or periodic boundary conditions when Ω is a square. To this end, we define the Hilbert space H as

$$H = \{u \in L^2(\Omega)^2, \nabla\cdot u = 0, u\cdot\vec{n} = 0 \text{ on } \partial\Omega\}$$

in the case of nonslip Dirichlet boundary conditions, where $\vec{n}$ denotes the outward normal unit vector to $\partial\Omega$, or

$$H = \left\{u \in L^2_{\text{per}}(\Omega)^2, \nabla\cdot u = 0, \int_\Omega u\ dx = 0\right\}$$

in the case of periodic boundary conditions. The space H is a closed subspace of $L^2(\Omega)^2$ and is endowed with the scalar product and norm from $L^2(\Omega)^2$, denoted by $(\cdot,\cdot)$ and $\|\cdot\|$, respectively. We also define the Hilbert space V as $V = \{u \in H^1_0(\Omega)^2, \nabla\cdot u = 0\}$ or $V = \{u \in H^1_{\text{per}}(\Omega)^2, \nabla\cdot u = 0, \int_\Omega u\ dx = 0\}$, depending on the boundary conditions. Let P be the Leray orthogonal projection from $L^2(\Omega)^2$ onto H. Then (1.1) projected onto H may be written as an abstract functional differential equation of the form

$$\begin{aligned} \frac{du}{dt} + \nu Au + B(u,u) &= f, \\ u(0) &= u_0; \end{aligned} \tag{1.2}$$

see, e.g., [2] or [20].

The Stokes operator A is defined as $-P\Delta$ with the appropriate boundary conditions. The domain of A in H, denoted $D(A)$, is either $H^2(\Omega)^2 \cap V$ or $H^2_{\text{per}}(\Omega)^2 \cap V$, depending on the boundary conditions. The nonlinear term is $B(u,u)$ and is defined in general as $B(u,v) = P[(u\cdot\nabla)v]$. Finally, f (or $f = Pf$) is the forcing term and is assumed to be at least in H. The operator A is a positive, self-adjoint, densely defined, unbounded operator with compact inverse. The eigenfunctions of A, $\{\omega_1, \omega_2, \ldots\}$, form a complete orthonormal basis for the space H. The associated eigenvalues $\{\lambda_1, \lambda_2, \ldots\}$ satisfy $0 < \lambda_1 \le \lambda_2 \le \cdots$ and the asymptotic formula $\lambda_j \sim j$. Properties of the spaces H, $V = D(A^{1/2})$, and $D(A)$ may be found in [2], [18], or [20].

We decompose the solution u into low mode and high mode components by letting $H_N = \text{span}\{\omega_1, \omega_2, \ldots, \omega_N\}$, the span of the first N eigenfunctions of the Stokes operator A. Let P_N be the orthogonal projection of H onto H_N, and $Q_N = I - P_N$ be the projection onto the orthogonal complement space $H_N^\perp$. Then, for any $u \in H$, we can uniquely decompose $u = p + q$, where $p = P_N u$ and $q = Q_N u$. Projecting (1.2) onto H_N and $H_N^{\perp}$, we get an equivalent system for the NSE

$$\frac{dp}{dt} + \nu Ap + P_N\left[B(p,p) + B(p,q) + B(q,p) + B(q,q)\right] = P_N f, \tag{1.3}$$

$$\frac{dq}{dt} + \nu Aq + Q_N\left[B(p,p) + B(p,q) + B(q,p) + B(q,q)\right] = Q_N f, \tag{1.4}$$

$$p(0) = P_N u_0 \quad \text{and} \quad q(0) = Q_N u_0.$$

The truncation analysis is accomplished by using estimates for the low modes p and high modes q of a solution u. We present the truncation analysis for the two-dimensional NSE; however, similar analysis may be done for general nonlinear parabolic evolution or elliptic equations, such as reaction-diffusion systems, the Bénard convection problem, etc. The key to the analysis is understanding the interaction of the low and high modes, and estimating the nonlinear term.

The truncation analysis is based on asymptotic (in time) estimates for the low and high mode components, as was done in [10] and [11] when developing the postprocessing algorithm. These asymptotic estimates hold when solutions of the NSE are on, or near, the attractor, i.e., for the case of autonomous systems (f time-independent) and provided that t is large enough. However, these estimates may not hold, for example, in the case of nonsmooth initial data, long transients, or nonautonomous systems with highly oscillatory (in time) forcing. In [15] the authors showed that, for a highly oscillatory time-dependent forcing function, the dominant balance in (1.4) is between the dq/dt term and the forcing term; hence the dq/dt term should not be dropped in the AIM construction. This case leads to and justifies an alternate/reform postprocessing method, proposed in [24] for integrating along transients, which we call here *dynamic postprocessing*.

Let us emphasize again that there is a basic difference between the nonlinear Galerkin methods and the postprocessing Galerkin method. Specifically, unlike the usual multigrid (in this case two-grid) and the nonlinear Galerkin methods, in the postprocessing Galerkin methods the evolution/integration on the coarse mesh, i.e., low frequencies, does not use at all the information on the fine mesh (small scales or high frequencies). Only at the end of the calculations does one use the solution on the coarse mesh to refine the solution. On the other hand, in standard two-grid methods, including the nonlinear Galerkin methods and their variants, one uses cycles in which one has to compute the solution on the fine mesh in order to update the time step integration on the coarse mesh and vice versa. In fact, this occasional updating of the solutions on the fine mesh is the major source of computational disadvantage of the nonlinear Galerkin method in comparison to the Galerkin method, as was demonstrated computationally in, for instance, [10] and [11].

In this paper we first present a classical truncation analysis of the NSE using established asymptotic (in time) estimates. In section 2 we present several approximate systems of varying orders of accuracy based on the truncation analysis results. We introduce a more general postprocessing algorithm in section 3 for the case in which the asymptotic estimates no longer hold (i.e., in the presence of long transients, nonsmooth initial data, or highly oscillatory forcing). In section 4 we analyze the accuracy of the various postprocessing methods for the case of a highly time-oscillatory solution. In section 5 we present some numerical experiments to support the analysis of sections 2, 3, and 4 and compare the computational efficiency of the standard and the more general postprocessing methods. Finally, we give some concluding remarks in section 6. Preliminary results of this study were reported in [23].

2. Near-attractor truncation analysis. We first present the truncation analysis based on asymptotic estimates for u, p, and q. It is well known (see, e.g., [2] or [18]) that for $f \in H$ and independent of time, (1.2) is dissipative in the spaces H, V, and $D(A)$. This means that any solution $u(t)$ of (1.2) will, after a certain time, enter and remain in a ball in H centered at 0 with radius ρ_0. The same is true for a ball in V of radius ρ_1, and a ball in $D(A)$ of radius ρ_2. The radii ρ_0, ρ_1, and ρ_2 depend on $\|f\|$, ν, and λ_1. Therefore, we will assume that for $t \geq T_0$, for some positive T_0 that depends on ν, $\|f\|$, λ_1, and the initial data $\|u_0\|$, we have

$$\|u(t)\| \leq \rho_0, \qquad \|A^{1/2}u(t)\| \leq \rho_1, \qquad \|Au(t)\| \leq \rho_2. \tag{2.1}$$

Notice that the global attractor for (1.2) is contained in these balls. For solutions on the attractor, $T_0 = 0$ and the uniform bounds apply for all time $t \in \mathbb{R}$, since the global attractor is invariant (see, e.g., [2] and [18]).

From the above bounds for u, we have that q is also bounded in H, $D(A^{1/2})$, and $D(A)$ for $t > T_0$. Using the bound $\|Aq\| \leq \rho_2$ and the fact that $\|A^{\alpha}q\| \leq \lambda_{N+1}^{-\alpha}\|q\|$, we quickly obtain estimates for q in terms of λ_{N+1}. We denote $\epsilon = (\lambda_1/\lambda_{N+1})^{1/2}$. Then for $t > T_0$ the following estimates for q and dq/dt are at hand:

$$\begin{gathered}\|q\| \leq \lambda_{N+1}^{-1}\|Aq\| \leq \lambda_{N+1}^{-1}\rho_2 = O(\epsilon^2), \\ \|A^{1/2}q\| \leq \lambda_{N+1}^{-1/2}\|Aq\| \leq \lambda_{N+1}^{-1/2}\rho_2 = O(\epsilon), \\ \|Aq\| \leq \|Au\| \leq \rho_2 = O(1)\end{gathered} \tag{2.2}$$

as $\epsilon \to 0$. Using the fact that the solutions are analytic in time (see, e.g., [2] and [20]), one can apply the Cauchy formula for the derivatives of complex analytic functions to obtain an estimate for $\|dq/dt\|$ of the same order as $\|q\|$ (again, see, e.g., [2], [5], and [20]). We have

$$\left\|\frac{dq}{dt}\right\| = O(\epsilon^2) \quad \text{as } \epsilon \to 0. \tag{2.3}$$

Let us stress that the constant ρ_2, which depends on the physical parameters but not on N, is quite large in comparison with the constants ρ_0 or ρ_1 for small values of the viscosity ν or large values of $\|f\|$. It is preferable to avoid using the ρ_2 bound and to derive more delicate estimates for $\|q\|$, $\|dq/dt\|$, and $\|A^{1/2}q\|$ of the same orders as above involving only ρ_0 and ρ_1. Indeed, the authors of [5] derive bounds of the type given in (2.2) and (2.3) involving ρ_0 and ρ_1 but not ρ_2. However, this is done at the expense of adding a term of the order $|\log \epsilon|$. In practice, this is a more reasonable bound, since the best available bound for ρ_2 is many orders of magnitude larger than those for ρ_0 and ρ_1. Moreover, for practical computations, $|\log \epsilon|$ will be of order 1 even if ϵ is very small.

For the low mode component, we have only that p is bounded in H, $D(A^{1/2})$, and $D(A)$ for $t > T_0$. Hence, we set

$$\|p\|,\ \|A^{1/2}p\|,\ \|Ap\| = O(1) \quad \text{as } \epsilon \to 0. \tag{2.4}$$

For the truncation analysis we consider a perturbation expansion for q of the form

$$q = q_1 + q_2 + q_3 + q_4 + \cdots. \tag{2.5}$$

To leading order, we have estimates (2.2) and (2.3) for q. Hence, the corresponding estimates for the first expansion term q_1 are as follows:

$$\begin{aligned} \|q_1\|,\ \left\|\frac{dq_1}{dt}\right\| &= O(\epsilon^2), \\ \|A^{1/2}q_1\| &= O(\epsilon), \\ \|Aq_1\| &= O(1) \quad \text{as } \epsilon \to 0. \end{aligned} \tag{2.6}$$

Each successive term q_j is assumed to be of higher order in ϵ, i.e., $\|dq_j/dt\|, \|q_j\| = O(\epsilon^{j+1})$, $\|A^{1/2}q_j\| = O(\epsilon^j)$, and $\|Aq_j\| = O(\epsilon^{j-1})$, for $j = 1, 2, \ldots$. In principle, the initial value u_0 should also be decomposed accordingly, i.e., $u_0 = P_N u_0 + Q_N u_0$, with $Q_N u_0 = q_1^0 + q_2^0 + \cdots$ such that $O(q_j^0) = \epsilon O(q_{j-1}^0)$. In particular, for solutions on or near the attractor, we should have $\|q_j^0\| = O(\epsilon^{j+1})$, $\|A^{1/2}q_j^0\| = O(\epsilon^j)$, and $\|Aq_j^0\| = O(\epsilon^{j-1})$, for $j = 1, 2, \ldots$.

We substitute expansion (2.5) into system (1.3)–(1.4) above and estimate the order of each term in the system. By keeping terms up to order $\epsilon^{1/2}$, $\epsilon^{3/2}$, and so on, we generate approximate systems for NSEs of increasing orders of accuracy. The challenge comes with estimating the nonlinear terms. Substituting expansion (2.5) into the nonlinear terms results in the following:

$$\begin{aligned} B(p,q) = {} & B(p, q_1 + q_2 + q_3 + \cdots) \\ = {} & B(p,q_1) + B(p,q_2) + B(p,q_3) + \cdots, \\ B(q,p) = {} & B(q_1 + q_2 + q_3 + \cdots, p) \\ = {} & B(q_1,p) + B(q_2,p) + B(q_3,p) + \cdots, \\ B(q,q) = {} & B(q_1 + q_2 + q_3 + \cdots, q_1 + q_2 + q_3 + \cdots) \\ = {} & B(q_1,q_1) + B(q_1,q_2) + B(q_1,q_3) + \cdots \\ & + B(q_2,q_1) + B(q_2,q_2) + B(q_2,q_3) + \cdots \\ & + B(q_3,q_1) + B(q_3,q_2) + B(q_3,q_3) + \cdots \\ & + \cdots. \end{aligned}$$

We majorize each term using inequalities for the nonlinear term given, for instance, in [2], [20], or [22] for inequalities (2.9) and (2.10). For convenience we recall the two-dimensional version of these inequalities. For any $u, v \in D(A)$,

$$\|B(u,v)\| \le c_1 \|u\|^{1/2} \|A^{1/2}u\|^{1/2} \|A^{1/2}v\|^{1/2} \|Av\|^{1/2} \tag{2.7}$$

$$\le c_2 \|u\|^{1/2} \|Au\|^{1/2} \|A^{1/2}v\| \tag{2.8}$$

$$\le c_3 \|A^{1/2}u\| \|A^{1/2}v\| \left(1 + \log \frac{\|Au\|^2}{\lambda_1 \|A^{1/2}u\|^2}\right)^{1/2} \tag{2.9}$$

and

$$\|B(u,v)\| \le c_4 \|u\|\, \|Av\| \left(1 + \log \frac{\|A^{3/2}v\|^2}{\lambda_1 \|Av\|^2}\right)^{1/2} \tag{2.10}$$

for $u \in D(A)$ and $v \in D(A^{3/2})$. The constants c_1–c_4 are independent of u, v, and the size of Ω, but might depend on its shape. For each nonlinear term we choose the inequality that results in the highest order of ϵ. Using estimates (2.4) for p and (2.6) for q_1, we obtain $\|B(p,p)\| = O(1)$, $\|B(p,q_1)\| = O(\epsilon)$, and $\|B(q_1,q_1)\| = O(\epsilon^2)$. For the $B(q_1,p)$ term, inequality (2.7) gives $\|B(q_1,p)\| = O(\epsilon^{3/2})$, and inequality (2.10) gives $\|B(q_1, p)\| = O(\epsilon^2 L_\epsilon^{1/2})$, where $L_\epsilon = (1 + 2|\log \epsilon|)$. The $O(\epsilon^2 L_\epsilon^{1/2})$ estimate is "closer" to being of the order $O(\epsilon^2)$ than $O(\epsilon)$. However, in either case, the term is definitely of the order $O(\epsilon^{3/2})$. For simplicity of ordering the various terms, we will consider this term to be $O(\epsilon^{3/2})$.

2.1. Near-attractor approximate systems. To produce approximate schemes for the Navier–Stokes system, we keep only terms in (1.3)–(1.4) to certain orders in ϵ. Below, we list the approximate systems produced by keeping terms to order $\epsilon^{1/2}$ and $\epsilon^{3/2}$. We will set nonlinear terms of the order $O(\epsilon^2 L_\epsilon^{1/2})$ and $O(\epsilon^3 L_\epsilon^{1/2})$ to be of the order $O(\epsilon^{3/2})$ and $O(\epsilon^{5/2})$, respectively.

$\boxed{O(\epsilon^{1/2}):}$

$$\frac{dp}{dt} + \nu Ap + P_N\left[B(p,p)\right] \approx P_N f, \tag{2.11}$$

$$\nu Aq_1 + Q_N\left[B(p,p)\right] \approx Q_N f, \tag{2.12}$$

$$p(0) = P_N u_0. \tag{2.13}$$

Equation (2.11) is an evolution equation for the low mode component p. Equation (2.12) is coupled to (2.11); it defines q_1, the leading order approximation term of the high modes, in terms of the low modes and is therefore a postprocessing step. From (2.12) one can verify that $\|Aq_1\| = O(1)$, which is consistent with our assumptions.

$\boxed{O(\epsilon^{3/2}):}$

$$\frac{dp}{dt} + \nu Ap + P_N\left[B(p,p) + B(p,q_1) + B(q_1,p)\right] \approx P_N f, \tag{2.14}$$

$$\nu A(q_1 + q_2) + Q_N\left[B(p,p) + B(p,q_1) + B(q_1,p)\right] \approx Q_N f, \tag{2.15}$$

$$\nu Aq_2 + Q_N\left[B(p,q_1) + B(q_1,p)\right] \approx 0, \tag{2.16}$$

$$p(0) = P_N u_0. \tag{2.17}$$

Equation (2.14) is the evolution equation for p with q_1 in the nonlinear term defined by (2.12); it is a nonlinear Galerkin method as defined in [5], [13], [14], and [16]. Equation (2.15) defines $q_1 + q_2$, which is a higher order approximation of the high modes. Equation (2.16) defines q_2; it is derived from (2.15) and the definition of q_1 given in (2.12). From (2.16) one can show that $\|Aq_2\| \leq \|B(p,q_1)\| + \|B(q_1,p)\| = O(\epsilon)$. Hence $\|q_2\| = O(\epsilon^2)$, which is consistent with our assumptions.

Similarly, we may obtain an approximate system for the NSE valid to order $O(\epsilon^{5/2})$, $O(\epsilon^{7/2})$, and in general, valid to order $O(\epsilon^{j+1/2})$. In the general case, the low mode equation is evolved with linear combinations of q_1 through q_j in the nonlinear term; the high mode equation involves linear combinations of q_1 through q_{j+1}, where $(q_1 + q_2 + \cdots + q_{j+1})$ is used to approximate q. Thus the high modes of the solution u should be approximated to one order higher in ϵ than the order of the high mode terms used in the low mode equation. The term q_{j+1} is not used to evolve the low modes; it needs to be evaluated only once, at some final time T, and may therefore be considered a postprocessing step. The approximate systems above, produced with a classical truncation analysis, demonstrate that the postprocessing step is a very natural and significant part of approximating the original system.

2.2. Standard (near-attractor) postprocessing schemes. For each approximate system from the previous section, we may generate a postprocessed Galerkin or nonlinear Galerkin scheme of increasing order of accuracy. From the truncation analysis we know that, to approximate the low and high modes of a solution u to the same order in ϵ, we must include the postprocessing step. In general the solution of the evolution equation is sought as an approximation of the low modes of the exact solution u, and the solution of the high mode equation is sought as an approximation of the high modes of the solution u. The goal for each $\epsilon^{j+1/2}$ postprocessing scheme is to produce a more accurate approximation of the low and high modes as j increases.

From system (2.11)–(2.12), i.e., keeping terms to order $O(\epsilon^{1/2})$, we obtain the postprocessed standard Galerkin method

$$\frac{du_N}{dt} + \nu A u_N + P_N\left[B(u_N, u_N)\right] = P_N f, \tag{2.18}$$

$$\nu A \phi_1 + Q_N\left[B(u_N, u_N)\right] = Q_N f, \tag{2.19}$$

$$u_N(0) = P_N u_0, \tag{2.20}$$

where $u_N \in H_N$ is the solution of the evolution equation and is an approximation of the low modes p, and $\phi_1 \in Q_N H$ is an approximation of the high modes q (i.e., $\phi_1 \approx q_1$). Note that $\phi_1(t) = \Phi^1(u_N(t))$, where Φ^1 is exactly the Foias–Manley–Temam (FMT) AIM first introduced in [5]. This is the same postprocessed Galerkin method originally defined in [10] and [11]. Solving for $u_N(t)$ does not depend on ϕ_1, and hence one does not need to evaluate $\phi_1(t) = \Phi^1(u_N(t))$ at all times, but only when an approximate solution is needed. This is typically done once at some final time T. It is therefore a postprocessing step. The approximate solution at time T is then $u_N(T) + \phi_1(T) = u_N(T) + \Phi^1(u_N(T))$, and not $u_N(T)$ as is traditionally used with the standard Galerkin method. This scheme indicates that, to leading order in ϵ, the correct approximation method is the postprocessed Galerkin method.

The approximation properties of the postprocessed Galerkin method ($j = 0$) are well understood. We know, for instance, that Φ_1 is Lipschitz continuous, $\|\phi_1(t)\| = O(\epsilon^2)$, $\|q(t) - \phi_1(t)\| = O(\epsilon^3)$, and $\|p(t) - u_N(t)\| = O(\epsilon^3)$. Proofs of the first three properties may be found, for example, in [4] and [5]. The fourth property is proven in [11], specifically for the two-dimensional NSE.

In general, keeping terms of order $\epsilon^{j+1/2}$ for $j \geq 1$, we obtain a postprocessed nonlinear Galerkin scheme with successively more accurate approximation properties as j increases. In particular, for the scheme generated from system (2.14)–(2.16), i.e., the case $j = 1$, we can show that the low and high mode approximation errors are of the order $O(\epsilon^4)$ using the techniques as in [11] and [17]. Though more accurate, the $O(\epsilon^{j+1/2})$ systems are not computationally competitive for $j \geq 1$. We know from numerical experiments presented in [10], [11], and [17] that the more computationally efficient schemes are the postprocessed Galerkin method (2.18)–(2.19) and variants thereof, such as the postprocessed filtered Galerkin method. Hence, for the purposes of this paper, we will concentrate on the postprocessed Galerkin method, system (2.18)–(2.20).

3. A more general truncation analysis.

The standard postprocessing scheme and systems in the previous section were generated based on asymptotic (in time) estimates for the low and high mode components. These estimates hold for autonomous systems when solutions of the NSE are on or near the attractor, i.e., for t large enough. However, these estimates may no longer hold, for instance, in the case of nonsmooth initial data, long transients, or nonautonomous systems with highly oscillatory time-dependent forcing. For these cases, the leading order approximation for system (1.3)–(1.4) is no longer clear. In particular, the dq/dt term may no longer be small in comparison with the other terms in (1.4). For instance, in the case of a highly oscillatory time-dependent force, the authors of [15] presented an analytic example showing that the dominant balance in (1.4) is between the dq/dt term and the forcing term, and not between the dissipative term and the forcing and nonlinear terms. In this case they concluded that the dq/dt term should not be dropped in the AIM construction. In this section we consider the special case in which the forcing f is a highly oscillatory time-dependent function.

We start with the nonautonomous Navier–Stokes system (1.3)–(1.4) with highly oscillatory forcing. We assume that the force remains bounded (i.e., $f \in L^\infty((0,\infty);H)$ but oscillatory in time (defined later in Theorem 4.4). Furthermore, we assume that the solution $u(t)$ is bounded in $D(A)$ for $t \geq 0$ and that the initial condition is smooth, i.e., $u_0 \in D(A)$. As before, we observe that $\|Au(t)\| = O(1)$ since $\|Au(t)\|$ is bounded uniformly. Then for all $t \geq 0$ we have

$$\|p(t)\|, \|A^{1/2}p(t)\|, \|Ap(t)\| = O(1) \qquad \text{and} \qquad \|Aq(t)\| = O(1).$$

Again using the fact that $\|q\| \leq \lambda_{N+1}^{-\alpha}\|A^\alpha q\|$, we obtain that $\|q(t)\| = O(\epsilon^2)$ and $\|A^{1/2}q(t)\| = O(\epsilon)$ as before. Since the forcing is highly oscillatory in time, we cannot assume that the time derivative of the solution u, and hence q, is necessarily small. In this situation we will suppose that $\|dq/dt\| = O(1)$, the same order as the $\|Aq(t)\|$ term or larger. We have the following bounds for q,

$$\begin{aligned} \|q(t)\| &= O(\epsilon^2), \\ \|A^{1/2}q(t)\| &= O(\epsilon), \\ \|Aq(t)\|, \|dq/dt\| &= O(1). \end{aligned} \tag{3.1}$$

Without assuming that the forcing term is real analytic in time with values in H, one could not show that the solution $u(t)$ is real analytic in time with values in $D(A)$. Therefore, it would not be possible to employ the techniques used in [5] to get tight estimates on the constants involved in the bounds given in (3.1).

3.1. More general approximate systems. For the truncation analysis, we again assume a perturbation expansion for q of the form $q = (q_1 + q_2 + q_3 + \cdots)$. Since q_1 is the leading order approximation for q, the above estimates hold for q_1 as well. We then substitute the perturbation expansion for q into system (1.3)–(1.4) and estimate the orders of the various terms as before. The only differences are the orders of the $d(q_1 + q_2 + \cdots)/dt$ terms.

Keeping terms up to order $\epsilon^{1/2}$, we have the following leading order approximate system:

$$\frac{dp}{dt} + \nu Ap + P_N\left[B(p,p)\right] \approx P_N f, \tag{3.2}$$

$$\frac{dq_1}{dt} + \nu Aq_1 + Q_N\left[B(p,p)\right] \approx Q_N f, \tag{3.3}$$

$$p(0) = P_N u_0, \tag{3.4}$$

$$q_1(0) = Q_N u_0. \tag{3.5}$$

Equation (3.2) is the usual evolution equation for the low mode component; it is the standard Galerkin method. Equation (3.3) is used to define q_1, the leading order approximation of the high modes, only now it is an evolution equation. Here q_1 is not needed for the evolution of the low modes; hence (3.3) may be considered a postprocessing step. This is the same postprocessing step introduced in [24] for the case of nonsmooth initial data and long transients, and justifies the postprocessing method given therein. It is worth noting that one can think about the above system (3.2)–(3.5) as a two-level multigrid method, where one integrates (3.2) on the coarse mesh and then postprocesses on the fine mesh using (3.3).

Keeping terms up to order $\epsilon^{3/2}$, we have the following approximate system:

$$\frac{dp}{dt} + \nu Ap + P_N\left[B(p, p+q_1) + B(q_1, p)\right] \approx P_N f, \tag{3.6}$$

$$\frac{d(q_1+q_2)}{dt} + \nu A(q_1+q_2) + Q_N\left[B(p, p+q_1) + B(q_1, p)\right] \approx Q_N f, \tag{3.7}$$

$$p(0) = P_N u_0, \tag{3.8}$$

$$(q_1+q_2)(0) = Q_N u_0. \tag{3.9}$$

Equation (3.6) is the same evolution equation for p as in (2.14), but with q_1 now defined by (3.3). It is a nonlinear Galerkin method. Equation (3.7) defines q_1+q_2, the high mode approximation. From (3.3), (3.6), and (3.7) one concludes

$$\frac{dp}{dt} + \nu Ap + P_N\left[B(p, p+q_1) + B(q_1, p)\right] \approx P_N f, \tag{3.10}$$

$$\frac{dq_1}{dt} + \nu Aq_1 + Q_N\left[B(p, p)\right] \approx Q_N f, \tag{3.11}$$

$$\frac{dq_2}{dt} + \nu Aq_2 + Q_N\left[B(p, q_1) + B(q_1, p)\right] \approx 0, \tag{3.12}$$

$$p(0) = P_N u_0, \tag{3.13}$$

$$q_1(0) = q_1^0, \tag{3.14}$$

$$q_2(0) = q_2^0. \tag{3.15}$$

Equation (3.12) is a postprocessing step since q_2 is not used in the evolution equation for the low mode component p. Here again one can think about the above scheme as a two-level multigrid method.

We may continue this process as before, keeping terms to higher and higher orders in ϵ to generate a general postprocessing scheme. However, for computational efficiency, we are interested only in the leading order postprocessing algorithms.

3.2. A dynamic postprocessing scheme.

Motivated by the approximate system (3.2)–(3.5), we introduce the *dynamic postprocessing scheme*

$$\frac{du_N}{dt} + \nu Au_N + P_N\left[B(u_N, u_N)\right] = P_N f, \tag{3.16}$$

$$\frac{d\tilde{\phi}_1}{dt} + \nu A\tilde{\phi}_1 + Q_N\left[B(u_N, u_N)\right] = Q_N f, \tag{3.17}$$

$$u_N(0) = P_N u_0, \tag{3.18}$$

$$\tilde{\phi}_1(0) = Q_N u_0, \tag{3.19}$$

where the approximation for the high modes $\tilde{\phi}_1$ is obtained as the solution of evolution equation (3.17). Notice that $\tilde{\phi}_1 = \tilde{\phi}_1(t; u_N(t))$.

4. Error analysis.

In the following, we will compare the accuracy of the standard postprocessing method, system (2.18)–(2.20), with the dynamic postprocessing method, system (3.16)–(3.19), in the case of a highly oscillatory forcing function.

Since the approximation for the low mode component is exactly the same in each case, namely the Galerkin approximation, we will compare only the postprocessing approximation of the high mode component.

For comparison purposes we will use the uniform bounds

$$\|Au(t)\| \leq \rho_2, \qquad \|Au_N(t)\| \leq \rho_2^*, \qquad t \geq 0, \tag{4.1}$$

where ρ_2 and ρ_2^* are constants which depend on the data of the problem (i.e., ν, f, $\|u_0\|$, and λ_1) but are independent of N. Let us observe that usually $\rho_2^* = \rho_2$. We will also utilize a low mode accuracy estimate, which we restate below without proof (see [11], Theorem 1).

THEOREM 4.1. *Let $T > 0$ be fixed. Let $u = p + q$ be the solution of* (1.2) *on $[0, T]$ such that the bounds in* (3.1) *and* (4.1) *hold. Then, there exists a constant $C = C(T, \rho_1, \rho_2)$ such that for any $t \in [0, T]$ the solution $u_N(t)$ of* (2.18) *and* (2.20) *satisfies*

$$\|p(t) - u_N(t)\| \leq C \frac{L_\epsilon^2}{\lambda_{N+1}^{3/2}} = O(\epsilon^3 L_\epsilon^2), \tag{4.2}$$

where $L_\epsilon = 1 + 2|\log \epsilon| = 1 + \log(\lambda_{N+1}/\lambda_1)$.

The theorem is proven in the case of f time-independent. However (see [11, Remark 2]), f plays no role in the estimates, and hence the result is valid for $f = f(t)$ as well.

We first work with the leading order postprocessing method presented in section 2, whose corresponding scheme is given by (2.18)–(2.19). Here u_N is the Galerkin low mode approximation. The high mode approximation is given by $\phi_1 = \Phi^1(u_N)$, where Φ^1 is the FMT AIM introduced in [5] and is defined in general as

$$\Phi^1(v) = (\nu A)^{-1}(Q_N f - Q_N B(v, v)), \qquad v \in H_N.$$

A common approach for estimating the error $\|q(t) - \Phi^1(u_N(t))\|$ is to first bound $\|q - \Phi^1(p)\|$ using asymptotic estimates for p, q, and dq/dt, where $u(t) = p(t) + q(t)$ is the exact solution. Since we no longer assume that $\|dq/dt\|$ is small, we first reexamine the $\|q - \Phi^1(p)\|$ estimate in the case of a highly oscillatory forcing function. We have the following theorem.

THEOREM 4.2. *Let $f(t) \in L^\infty((0, \infty); H)$ and $u_0 \in D(A)$. Then for any solution $u(t) = p(t) + q(t)$ of* (1.2), *and $u_N(t)$ the solution of* (2.18) *and* (2.20) *such that estimates* (3.1) *and* (4.1) *hold, we have*

$$\|q(t) - \Phi^1(p(t))\| \leq \frac{C}{\nu \lambda_{N+1}} \left(\left\| \frac{dq}{dt} \right\| + \lambda_{N+1}^{-1/2} \|Aq\| \right), \tag{4.3}$$

$$\|q(t) - \Phi^1(u_N(t))\| \leq L\|p(t) - u_N(t)\| + \frac{C}{\nu \lambda_{N+1}} \left(\left\| \frac{dq}{dt} \right\| + \lambda_{N+1}^{-1/2} \|Aq\| \right) \tag{4.4}$$

for every $t \geq 0$. Here L is the Lipschitz constant for Φ^1, which is known to be of the order $o(1)$ as $\lambda_{N+1} \to \infty$ (see, e.g., [4], [5], *and* [21]*).*

Proof. Let $t \geq 0$. Subtracting (2.11) from (1.4) and taking the L^2 norm, we obtain

$$\|\nu A(q(t) - \Phi^1(p(t)))\| \leq \left\| \frac{dq}{dt} \right\| + \|B(p, q)\| + \|B(q, p)\| + \|B(q, q)\|.$$

Using inequalities (2.7)–(2.10) to bound the nonlinear terms, we have

$$\|\nu A(q-\Phi^1(p))\| \leq \left\|\frac{dq}{dt}\right\| + \frac{c_1\|Ap\|}{\lambda_{N+1}^{1/2}}\|Aq\| + \frac{c_1\|Ap\|}{\lambda_{N+1}^{3/4}}\|Aq\| + \frac{c_1}{\lambda_{N+1}}\|Aq\|^2,$$

and hence

$$\|q-\Phi^1(p)\| \leq (\nu\lambda_{N+1})^{-1}\left(\left\|\frac{dq}{dt}\right\| + \frac{C}{\lambda_{N+1}^{1/2}}\|Aq\|\right), \tag{4.5}$$

where $C = C(c_1, \rho_1, \rho_2)$. This proves estimate (4.3). To obtain estimate (4.4), first apply the triangle inequality,

$$\|q(t)-\Phi^1(u_N(t))\| \leq \|q(t)-\Phi^1(p(t))\| + \|\Phi^1(p(t))-\Phi^1(u_N(t))\|.$$

Then use the Lipschitz continuity of Φ^1 (see [5]) and estimate (4.3). □

In this section we assume that $\|dq/dt\|, \|Aq\| = O(1)$, and thus $\|q-\Phi^1(p)\| = O(\epsilon^2)$, rather than of the order $O(\epsilon^3) = O(\lambda_{N+1}^{-3/2})$ for the solutions on or near the attractor in the case of autonomous systems. Different asymptotic estimates for the q and dq/dt terms result in different accuracy estimates for $\|q-\Phi^1(p)\|$ and for the total accuracy estimate of the standard postprocessing algorithm. In particular, using Theorem 4.1 to bound the $\|p(t)-u_N(t)\|$ term, the bounds for $\|dq/dt\|$ and $\|Aq\|$ dominate the error in estimate (4.4). The accuracy estimate for the high mode approximation using the standard postprocessing method is given below.

COROLLARY 4.3. *Let $f(t) \in L^\infty((0,\infty); H)$, $u_0 \in D(A)$, and $T > 0$. Then for any solution $u(t) = p(t) + q(t)$ of* (1.2)*, and $u_N(t)$ the solution of* (2.18) *and* (2.20) *such that the estimates* (3.1) *and* (4.1) *hold, we have*

$$\|q(t)-\Phi^1(u_N(t))\| = O(\epsilon^2) \tag{4.6}$$

for $t \geq 0$.

Note that, without a better estimate for the $\|dq/dt\|$ term, we could easily have obtained the same estimate for $\|q-\Phi^1(p)\|$ by first applying the triangle inequality to get

$$\|q-\Phi^1(p)\| \leq \|q\| + \|\Phi^1(p)\|.$$

Then, under the assumptions of Theorem 4.2, one can show that $\|\Phi^1(p)\| = O(\lambda_{N+1}^{-1}) = O(\epsilon^2)$. Since $\|q\| = O(\epsilon^2)$ as well, we obtain $\|q-\Phi^1(p)\| = O(\epsilon^2)$. Hence, Corollary 4.3 only indicates that q and $\Phi^1(p)$ are of the same order.

With additional assumptions on $f(t)$, the above estimate for $\|q-\Phi^1(p)\|$ may be improved. In particular, in [15] the authors show that $\|q-\Phi^1(p)\| = O(\epsilon^{1+2\theta})$ for f Hölder continuous in time (with exponent θ), with values in H for N large enough. For convenience we restate the theorem.

THEOREM 4.4. *Let $f(t)$ be Hölder continuous (i.e., $\|f(t_1)-f(t_2)\| \leq L_1|t_1 - t_2|^\theta$) and satisfy $\sup_{t\geq 0}|f(t)| \leq f_\infty < \infty$. Furthermore, impose sufficient conditions on $f(t)$ so that $\|Au\|$ is uniformly bounded and, hence, the solution $p(t)$ of* (1.3) *is uniformly Lipschitz in time, (i.e., $\|p(t_1)-p(t_2)\| \leq L_2|t_2-t_1|$, where L_2 depends on ν, f_∞, and λ_1). Let $\|q(0)-\Phi^1(p(0))\| = O(\lambda_{N+1}^{(1/2+\theta)})$. Then, for N sufficiently large and $t \geq 0$, any solution $u(t) = p(t)+q(t)$ of* (1.2) *satisfies*

$$\|q(t)-\Phi^1(p(t))\| \leq \frac{4\alpha_5}{\lambda_{N+1}^{1/2+\theta}},$$

where $\alpha_5 = \alpha_4(1+(1+e)^{-1})$, $\alpha_4 = \lambda_{N+1}^{-1/2} 2\nu^{-1}(\alpha_3 L_2 \lambda_N^{1/2} + L_1) + \alpha_2 L_2 + \nu^{-1} L_1 \lambda_{N+1}^{-1/2}$, L_1 *is the Hölder constant for* f, *and* L_2 *is the Lipschitz constant for* p.

Proof. We refer the reader to [15, Theorem 5.11] for specific conditions on N, definitions of α_2, α_3, α_4, and the proof of the above theorem. □

If $\theta > 1/2$, then Theorem 4.4 represents an improvement from the previous $O(\epsilon^2)$ estimate for $\|q - \Phi^1(p)\|$, where f was only assumed to be bounded. Using the Lipschitz continuity of Φ^1 and Theorem 4.1, we have an improved estimate for $\|q - \Phi^1(u_N)\|$ in the case in which f is Hölder continuous.

COROLLARY 4.5. *Let* $f(t)$ *satisfy the conditions of Theorem* 4.4, $u_0 \in D(A)$, $T > 0$, *and* N *sufficiently large. Then for any solution* $u(t) = p(t) + q(t)$ *of* (1.2), *and* $u_N(t)$ *the solution of* (2.18) *and* (2.20) *such that estimates* (3.1) *and* (4.1) *hold, we have*

$$\|q(t) - \Phi^1(u_N(t))\| \leq L\|p(t) - u_N(t)\| + \frac{4\alpha_5}{\lambda_{N+1}^{1/2+\theta}} = O(\epsilon^{2+2(\theta-1/2)}) \tag{4.7}$$

for $t \in [0,T]$. *Here* $L = o(1)$ *is the Lipschitz constant for* Φ^1.

We now examine the dynamic postprocessing method, system (3.16)–(3.19), and obtain an estimate for $\|q(t) - \tilde{\phi}_1(t; u_N(t))\|$.

THEOREM 4.6. *Let* $f(t) \in L^\infty((0,\infty); H)$, $u_0 \in D(A)$, *and* $T > 0$. *Let* $u(t) = p(t) + q(t)$ *be a solution of* (1.2), *and* $u_N(t)$ *and* $\tilde{\phi}_1(t; u_N(t))$ *be a solution of system* (3.16)–(3.19) *such that estimates* (3.1) *and* (4.1) *hold. Then, for* $t \in [0,T]$, *we have*

$$\|q(t) - \tilde{\phi}_1(t; u_N(t))\| \leq \frac{C}{\nu\lambda_{N+1}^{1/2}} \left(\max_{s\in[0,T]} \|p(s) - u_N(s)\| + \lambda_{N+1}^{-1} \max_{s\in[0,T]} \|Aq(s)\| \right), \tag{4.8}$$

where $C = C(\rho_2, \rho_2^*)$; ρ_2 *and* ρ_2^* *are defined in* (2.1) *and* (4.1), *respectively.*

Proof. We subtract (3.17) from (1.4). Letting $\Delta(t) = q(t) - \tilde{\phi}_1(t; u_N(t))$, we have

$$\frac{d\Delta}{dt} + \nu A\Delta = Q_N \left[B(p+q,q) + B(q,p) + B(p, p-u_N) + B(p-u_N, u_N) \right].$$

Taking the inner product of Δ and the above equation, we obtain

$$\begin{aligned} \frac{1}{2}\frac{d}{dt}\|\Delta\|^2 + \nu\|A^{1/2}\Delta\|^2 \leq{}& |(B(p+q,q),\Delta)| + |(B(q,p),\Delta)| \\ &+ |(B(p,p-u_N),\Delta)| + |(B(p-u_N,u_N),\Delta)|. \end{aligned}$$

We apply the Cauchy–Schwarz inequality and Young's inequality to get a factor of $\nu\|A^{1/2}\Delta\|$ from each term on the right-hand side. Combining all $\|A^{1/2}\Delta\|$ terms with the left-hand side of the inequality and using the fact that $\lambda_{N+1}^{1/2}\|\Delta\| \leq \|A^{1/2}\Delta\|$, we have

$$\begin{aligned} \frac{d}{dt}\|\Delta\|^2 + \nu\lambda_{N+1}\|\Delta\|^2 \leq{}& \frac{5}{\nu\lambda_{N+1}} \left(\|B(p,q)\|^2 + \|B(q,p)\|^2 + \|B(q,q)\|^2 \right. \\ &\left. + \|B(p,p-u_N)\|^2 + \|B(p-u_N,u_N)\|^2 \right). \end{aligned}$$

Estimating the nonlinear terms using estimates (2.7)–(2.10) as before,

$$
\begin{aligned}
\frac{d}{dt}\|\Delta\|^2 + \nu\lambda_{N+1}\|\Delta\|^2 &\le \frac{5}{\nu\lambda_{N+1}}\Big(c_2^2\|p\|\|Ap\|\|A^{1/2}q\|^2 \\
&\quad + c_4^2(1+\log(\lambda_N/\lambda_1))\|Ap\|^2\|q\|^2 + c_2^2\|q\|\|A^{1/2}q\|^2\|Aq\| \\
&\quad + c_2^2\|p\|\|Ap\|\|A^{1/2}(p-u_N)\|^2 + c_4^2(1+\log(\lambda_N/\lambda_1))\|Au_N\|^2\|p-u_N\|^2\Big) \\
&\le \frac{5}{\nu\lambda_{N+1}}\left(\frac{c_2^2\|Ap\|^2}{\lambda_{N+1}}\|Aq\|^2 + \frac{c_4^2L_\epsilon\|Ap\|^2}{\lambda_{N+1}^2}\|Aq\|^2 + \frac{c_2^2\|Aq\|^2}{\lambda_{N+1}^2}\|Aq\|^2\right. \\
&\qquad\left. + \lambda_N c_2^2\|p\|\|Ap\|\|p-u_N\|^2 + c_4^2L_\epsilon\|Au_N\|^2\|p-u_N\|^2\right) \\
&\le \frac{C}{\nu}\left(\|p-u_N\|^2 + \lambda_{N+1}^{-2}\|Aq\|^2\right),
\end{aligned}
$$

where $C = C(c_2, c_3, c_4, \rho_2, \rho_2^*)$. We then apply Gronwall's inequality to obtain

$$
\|\Delta(t)\|^2 \le \|\Delta(0)\|^2 e^{-\nu\lambda_{N+1}t} + \frac{C}{\nu^2\lambda_{N+1}}\left(\max_{[0,T]}\|p-u_N\|^2 + \lambda_{N+1}^{-2}\max_{[0,T]}\|Aq\|^2\right)
$$

for $t \in [0,T]$. Finally, by initializing $\tilde{\phi}_1(0;u_N) = Q_N u_0 = q(0)$, we have estimate (4.8). □

COROLLARY 4.7. *Let $f(t) \in L^\infty((0,\infty);H)$, $u_0 \in D(A)$, and $T > 0$. Let $u(t) = p(t) + q(t)$ be a solution of* (1.2), *and $u_N(t)$ and $\tilde{\phi}_1(t;u_N(t))$ be a solution of system* (3.16)–(3.19) *such that estimates* (3.1) *and* (4.1) *hold. Then, for $t \in [0,T]$, we have*

$$
\|q(t) - \tilde{\phi}_1(t;u_N(t))\| = O(\epsilon^3). \tag{4.9}
$$

Proof. Since $\|Aq\| = O(1)$ and $\|p(t) - u_N(t)\| = O(\epsilon^3)$ from Theorem 4.1, the $\|Aq\|$ term dominates the right-hand side of (4.8). Thus $\|q - \tilde{\phi}_1(t;u_N(t))\| = O(\epsilon^3)$ as $\epsilon \to 0$. □

In this case, the dynamic postprocessing method produces a more accurate high mode approximation. In particular, the dynamic postprocessing method produces a high mode approximation of the same order as the low mode approximation.

5. Numerical experiments.

In this section we present some numerical experiments to support the above accuracy analysis and compare the efficiency of the two leading order methods. Opting for a one-dimensional calculation, we integrated Burgers equation with homogeneous Dirichlet boundary conditions on the interval $[0,\pi]$. That is, we used the equation

$$
\begin{aligned}
\frac{\partial u}{\partial t} - \nu\frac{\partial^2 u}{\partial x^2} + u\frac{\partial u}{\partial x} &= f(x,t), \\
u(0,t) = u(\pi,t) &= 0, \\
u(x,0) &= u_0(x).
\end{aligned}
$$

Using notation similar to the NSE, the above equation is equivalent to the functional differential equation

$$
\frac{du}{dt} + \nu Au + B(u,u) = f,
$$

where, in this case, $A = -\frac{\partial^2}{\partial x^2}$ with domain $D(A) = H^2(0,\pi) \cap H_0^1(0,\pi)$. The eigenfunctions of A are $\omega_k = \sqrt{2/\pi}\sin(kx)$, with corresponding eigenvalues $\lambda_k = k^2$, for $k = 1, 2, \ldots$. The bilinear term $B(u,u)$ is defined by $B(u,v) = \frac{2}{3}uv_x + \frac{1}{3}u_x v$ for every u, $v \in H_0^1(0,\pi)$. In particular, we have $B(u,u) = uu_x$ for every $u \in H_0^1(0,\pi)$.

We chose an exact solution $u_e(x,t)$ and then computed the "highly oscillatory" time-dependent forcing term from the exact solution. In this way we checked errors without computing a large Galerkin approximation as an "exact" solution. We chose $u_e(x,t)$ as follows,

$$(5.1) \quad u_e(x,t) = \sum_{k=1}^{\infty} \frac{a_k(t)}{k^3} \sin kx, \qquad a_k(t) = \begin{cases} 1 + \gamma \sin k^2 t, & 1 \geq k \geq 100, \\ 1, & k > 100, \end{cases}$$

and then calculated the forcing function as $f(x,t) := du_e/dt + \nu A u_e + B(u_e, u_e)$. The exact solution u_e is in $D(A)$ for $t \geq 0$ as assumed above in section 3. Actually, we can compute sharper estimates for $\|Aq\|$ and $\|dq/dt\|$ using expression (5.1) to obtain $\|Aq\| \leq \sqrt{2}/\lambda_N^{1/4} = O(\epsilon^{1/2})$, and similarly $\|dq/dt\| \leq \sqrt{\gamma}/\lambda_N^{1/4} = \gamma O(\epsilon^{1/2})$. Note that the $\|Aq\|$ and $\|dq/dt\|$ terms are of the same order, depending on the magnitude of γ. We then obtain $\|q\| = O(\epsilon^{5/2})$ and $\|A^{1/2}q\| = O(\epsilon^{3/2})$. These estimates are of slightly higher order in ϵ than those assumed in the theoretical section. However, using these estimates for the truncation analysis and keeping terms up to order ϵ, we obtain the same leading order approximate system as in system (3.2)–(3.3).

The experiments in this section were run on a Sun Ultra 5. The time integrator used was the VODE code [1] with computed diagonal Jacobians (VODE option MF=23). This code is a reliable and efficient tool for the time integration of systems of ODEs, especially for stiff problems like those arising from the spatial discretizations of dissipative PDEs. VODE consists of a backward differentiation formula (BDF) implemented with variable time step and variable order. Specifically, in the algorithm the time levels are unevenly spaced and the step sizes are produced by the code as the integration proceeds. Also, formulas of different orders (up to order six) are used, the order of the formula being selected by the code at every time step. For problems similar to those in this section, the superior efficiency of codes like VODE with respect to other frequently used time integrators was experimentally checked in [9].

For each value of N we sought the Galerkin approximation, the standard postprocessed Galerkin approximation (standard PP) from system (2.18)–(2.20), and the dynamic postprocessed Galerkin approximation (dynamic PP) from system (3.16)–(3.19). Each experiment was carried out with decreasing values of the time-integration tolerance (an input parameter to VODE) until additional reduction did not improve the accuracy of the solution any further. This means that the time discretization error is negligible in comparison with the spatial error that we are interested in examining.

Figure 5.1 shows the total errors for each of the approximations at time $t = 2.0$ units, with $\gamma = 0.1$ and $\nu = 1$. The initial condition was the projection of $u_e(x,0)$. The solid line represents errors from the Galerkin method, (2.18) and (2.20); dashed lines represent errors from the standard postprocessing method, system (2.18)–(2.20); and dotted lines represent errors from the dynamic postprocessing method, system (3.16)–(3.19). It is clear from Figure 5.1 that the dynamic postprocessing method achieves the best rate of convergence as indicated by the most negative slope. The improvement of the standard postprocessing method over the Galerkin method is only algebraic, and not a significant improvement in the rate of convergence. Thus, the addition of the dq_1/dt term in the high mode equation is beneficial in this case, at least in terms of accuracy. The low mode errors for each of the three methods are essentially the same. The high mode errors are very similar to the total errors shown in Figure 5.1, since the high mode error dominates the total Galerkin and standard postprocessing error. Only the dynamic postprocessing method produces high mode errors (and rate of convergence) that are approximately of the same order as those for the low mode errors. In the case being studied, i.e., highly oscillatory solutions, the dynamic postprocessing method is the most accurate.

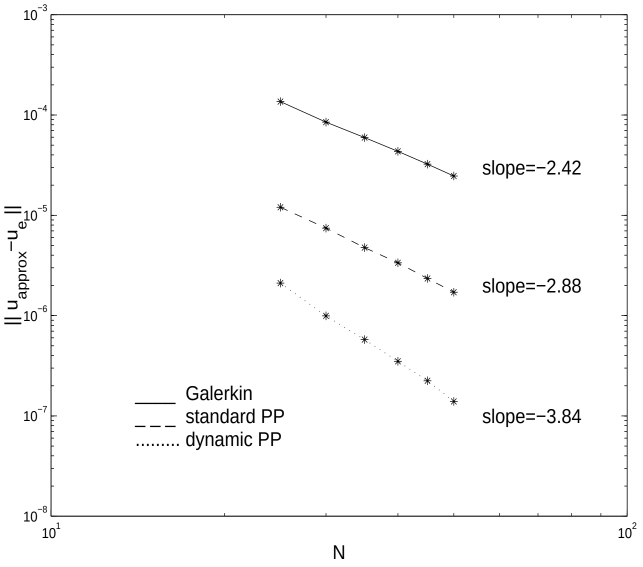


FIG. 5.1. *Total errors* $\|u_{approx} - u_e\|$.

The dynamic postprocessing method requires a numerical integration to obtain the high modes, rather than evaluating the high modes once at some final time, as with the standard postprocessing method [8], [10], [11], [17]. We next looked at the efficiency of the dynamic postprocessing method in the case of highly time-oscillatory solutions to determine whether the error improvement justifies any additional computational cost (CPU).

We again integrated Burgers equation with homogeneous Dirichlet boundary conditions on the interval $[0, \pi]$ as in section 5. However, this time we provided the forcing function defined by

$$f(x,t) = \sum_{k=1}^{\infty} \left(\frac{\dot{a}_k(t)}{k^3} + \frac{a_k(t)}{k} \right) \sin kx, \quad a_k(t) = \left\{ \begin{array}{ll} 1 + \gamma \sin k^2 t, & 1 \geq k \geq 100, \\ 1, & k > 100, \end{array} \right. \tag{5.2}$$

and used a large Galerkin run as an “exact” solution for computing errors. This way we did not accumulate the cost of computing the forcing function from an exact solution at each time step. The experiments in this section were run on an SGI Origin 2000 with $\gamma = 0.1$, $\nu = 1$, and initial condition $u_0(x) = \sum k^{-3} \sin kx$.

We first verified that we obtain accuracy results with this forcing function similar to those in the previous numerical experiment. Figure 5.2 shows the total error estimates at $t = 2.0$ units using the VODE time integrator. The rates of convergence are similar to those in Figure 5.1; the dynamic postprocessing method is again the most accurate method. In Figure 5.3 we plot the total errors $\|u_{approx} - u_e\|$ from Figure 5.2 versus the amount of computing time (in seconds) needed by each method to achieve those errors. We added results from larger mode standard postprocessing runs to better indicate any overlap. A horizontal line across the plot indicates, for a particular error, the CPU time needed by each method. Again, the solid line represents errors from the Galerkin method, dashed lines represent errors from the standard postprocessing method, and dotted lines represent errors from the dynamic

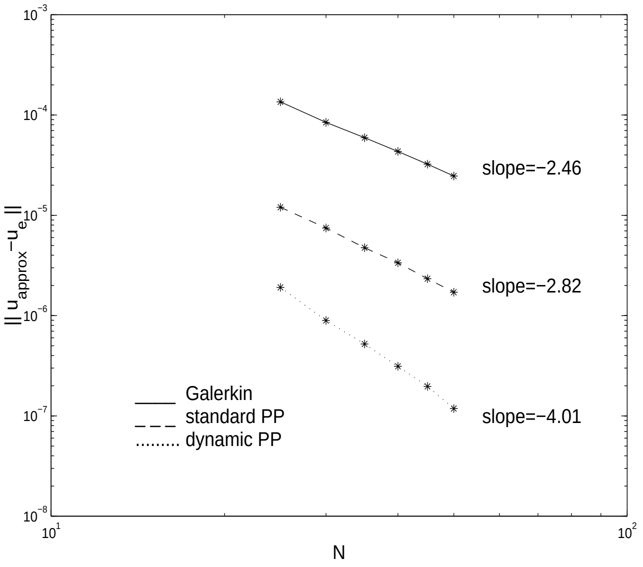


FIG. 5.2. *Total errors* $\|u_{approx} - u_e\|$.

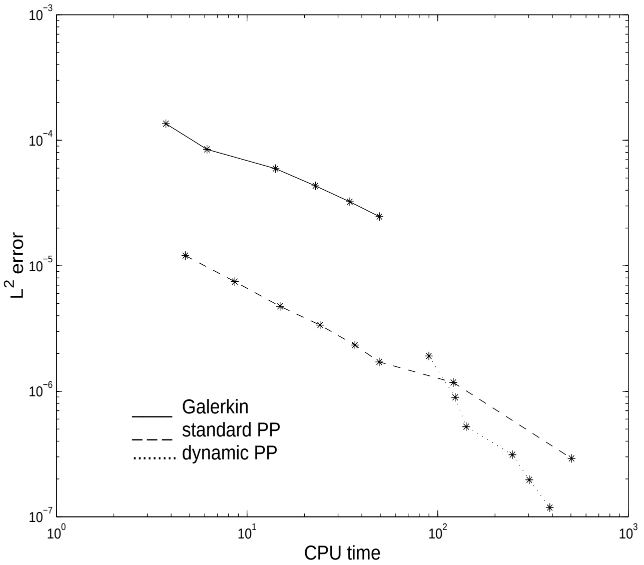


FIG. 5.3. *Error vs. CPU: VODE.*

postprocessing method. For the larger mode runs, the dynamic postprocessing method is slightly more efficient than the standard postprocessing method.

We also sought to take advantage of using a larger time step to integrate the high modes, subcycling the low mode integration within the high mode integration. For this experiment we used a semi-implicit backward Euler scheme. The low modes were integrated using the scheme

$$p^{n+1} - p^n + \Delta t \left(Ap^{n+1} + P_N B(p^n, p^n) \right) = \Delta t P_N f^n,$$

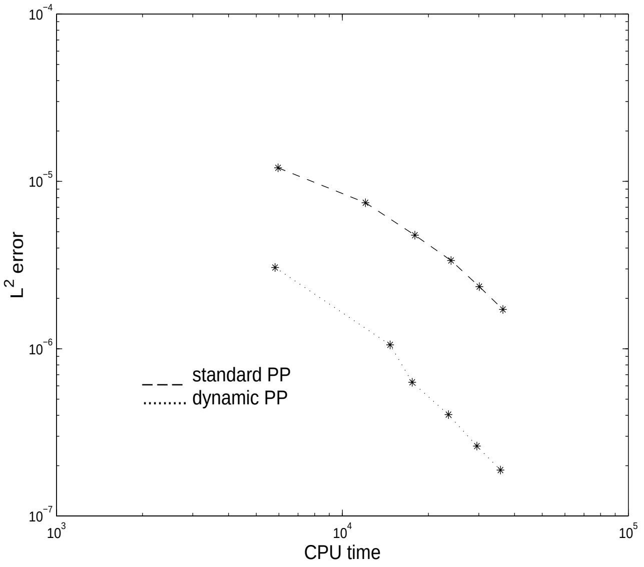


FIG. 5.4. *Error vs. CPU: backward Euler.*

and the high modes were integrated, in the dynamic postprocessing method, using

$$q^{n+1} - q^n + \Delta t \left(Aq^{n+1} + Q_N B(p^n, p^n) \right) = \Delta t Q_N f^n.$$

Otherwise, the experimental setup was the same as with the VODE time integrator, i.e., we computed the errors at time $t = 2$ units with $\nu = 1$, $\gamma = 0.1$, and initial condition $u_0(x) = \sum k^{-3} \sin kx$. In each experiment we set Δt small enough so that the low mode error ($\|p - u_N\|$) was equivalent to the low mode errors from the VODE experiments. The effect of subcycling was to slightly increase the error with little improvement in CPU time. This is because the cost of evaluating the high modes at each time step is minimal compared to the cost of evaluating the nonlinear term within each subcycle in order to integrate the low modes. The CPU comparison for the semi-implicit backward Euler experiments without subcycling is given in Figure 5.4. We plot error versus CPU time for the standard and dynamic postprocessing methods. Again the dynamic postprocessing method is the more efficient method for the case of a *highly oscillatory* (in time) solution due to a highly oscillatory (in time) forcing function.

Performing the same experiments with a less time-oscillatory forcing function, we obtained different accuracy and CPU comparisons. In this final set of experiments, we used the forcing function

$$f(x,t) = \sum_{k=1}^{\infty} \left(\frac{\dot{a}_k(t)}{k^3} + \frac{a_k(t)}{k} \right) \sin kx, \quad a_k(t) = \begin{cases} 1 + \gamma \sin t, & 1 \geq k \geq 100, \\ 1, & k > 100. \end{cases} \tag{5.3}$$

In this case we expect the solution to be less oscillatory, and hence the $\|dq/dt\|$ term should be of a smaller order, and the standard postprocessing method should be as accurate as the dynamic postprocessing method. Figure 5.5 shows the total errors for the Galerkin, standard postprocessing, and dynamic postprocessing approximations using the VODE time integrator at time $t = 3.0$ units, again with $\gamma = 0.1$, $\nu = 1.0$, and $u_0(x) = \sum k^{-3} \sin kx$. Note that the standard and dynamic postprocessing methods

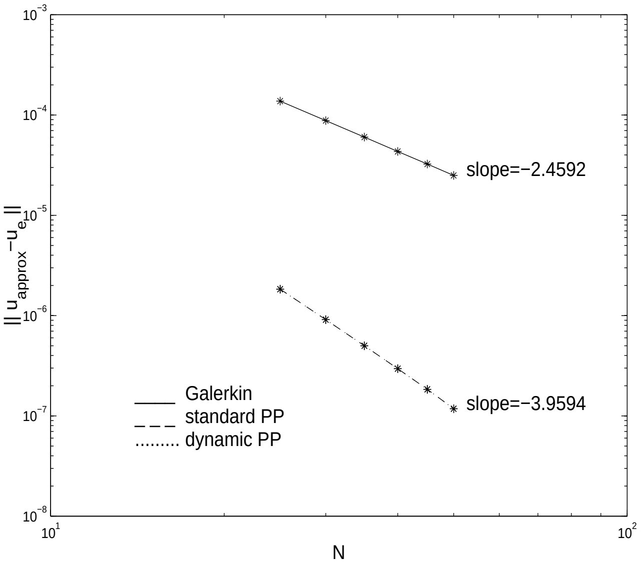


FIG. 5.5. *Total errors: f slowly oscillating in time.*

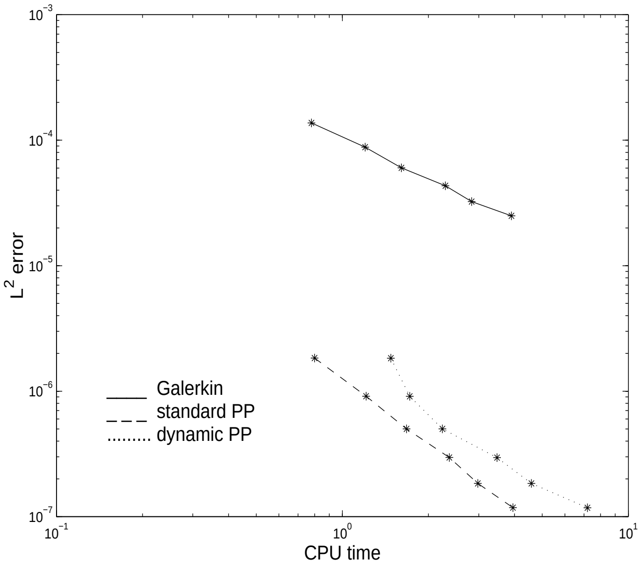


FIG. 5.6. *Error vs. CPU: VODE and f slowly oscillating in time.*

have the same rate of convergence, and there is no improvement with the dynamic postprocessing method.

In Figure 5.6 we plot the total error versus CPU time for the case of the slowly oscillating (in time) forcing function when using the VODE time integrator; the standard postprocessing method proves to be more efficient. Results using the backward Euler scheme are similar, though the differences are less pronounced.

6. Concluding remarks. Through a classical truncation analysis we have shown that postprocessing appears as a natural part of approximate systems and corresponding schemes for numerically integrating the two-dimensional NSE. For autonomous systems, we generated a family of approximate systems (and schemes) of increasing orders of approximation using the asymptotic (in time) estimates from [5] for $\|Aq\|$ and $\|dq/dt\|$. Each system included a postprocessing step that resulted in a high mode approximation of the same order as the low mode approximation. We found that the leading order system is the standard Galerkin method with postprocessing as introduced in [11]. The standard Galerkin method alone uses a less accurate approximation for the high modes than for the low modes. Hence, the accuracy of the high mode approximation, or lack thereof, dominates the error.

By assuming different asymptotic (in time) estimates for $\|Aq\|$ and $\|dq/dt\|$, we obtained the *dynamic postprocessing* method as the leading order method. This was done for the case of a highly oscillatory (in time) solution; the algorithm applies to the case of nonsmooth initial data as well [24]. In the case of a highly oscillatory solution, the dynamic postprocessing method was more accurate and efficient than the standard postprocessing method. For nonautonomous systems with solutions that are not so oscillatory, both methods obtained the same accuracy; however, the standard postprocessing method was more efficient in this case.

The method of using truncation analysis with asymptotic estimates for the low and high modes may easily be extended to general nonlinear parabolic evolution or elliptic equations to obtain postprocessing systems and schemes which approximate the low and high modes to the same order of accuracy.

7. Acknowledgments. We would like to thank Bosco Garcı́a-Archilla and Julia Novo for the stimulating and inspiring discussions. This work was initiated when E. S. T. and S. W. were visiting the Institute of Geophysical and Planetary Physics (IGPP) at the Los Alamos National Laboratory, while E. S. T. was the Orson Anderson Scholar from 1997–1998.

REFERENCES

[1] P. N. Brown, G. D. Byrne, and A. C. Hindmarsh, *VODE: A variable-coefficient ODE solver*, SIAM J. Sci. Statist. Comput., 10 (1989), pp. 1038–1051.

[2] P. Constantin and C. Foias, *Navier–Stokes Equations,* University of Chicago Press, Chicago, 1988.

[3] C. Devulder and M. Marion, *A class of numerical algorithms for large time integration: The nonlinear Galerkin methods*, SIAM J. Numer. Anal., 29 (1992), pp. 462–483.

[4] C. Devulder, M. Marion, and E. S. Titi, *On the rate of convergence of the nonlinear Galerkin methods*, Math. Comp., 60 (1993) pp. 495–514.

[5] C. Foias, O. Manley, and R. Temam, *Modelling of the interaction of small and large eddies in two dimensional turbulent flow*, RAIRO Modél. Math. Anal. Numér., 22 (1988), pp. 93–118.

[6] C. Foias, G. Sell, and R. Temam, *Inertial manifolds for nonlinear evolutionary equations*, J. Differential Equations, 73 (1988), pp. 309–353.

[7] C. Foias, G. Sell, and E. S. Titi, *Exponential tracking and approximation of inertial manifolds for dissipative nonlinear equations*, J. Dynam. Differential Equations, 1 (1989), pp. 199–244.

[8] J. de Frutos, B. García-Archilla, and J. Novo, *A postprocessed Galerkin method with Chebyshev and Legendre polynomials*, Numer. Math., 86 (2000), pp. 377–417.

[9] B. García-Archilla, *Some practical experience with the time integration of dissipative equations*, J. Comput. Phys., 122 (1995), pp. 25–29.

[10] B. Garciá-Archilla, J. Novo, and E. S. Titi, *Postprocessing the Galerkin method: A novel approach to approximate inertial manifolds*, SIAM J. Numer. Anal., 35 (1998), pp. 941–972.

[11] B. Garciá-Archilla, J. Novo, and E. S. Titi, *An approximate inertial manifolds approach to postprocessing the Galerkin method for the Navier–Stokes equations*, Math. Comp., 68 (1999), pp. 893–911.

[12] B. García-Archilla and E. S. Titi, *Postprocessing the Galerkin method: The finite-element case*, SIAM J. Numer. Anal., 37 (2000), pp. 470–499.

[13] M. S. Jolly, I. G. Kevrekidis, and E. S. Titi, *Approximate inertial manifolds for the Kuramoto–Sivashinsky equation: Analysis and computations*, Phys. D, 44 (1990), pp. 38–60.

[14] M. S. Jolly, I. G. Kevrekidis, and E. S. Titi, *Preserving dissipation in approximate inertial forms for the Kuramoto–Sivashinsky equation*, J. Dynam. Differential Equations, 3 (1991), pp. 179–197.

[15] D. A. Jones and E. S. Titi, *A Remark on quasi-stationary approximate inertial manifolds for the D Navier–Stokes equations*, SIAM J. Math. Anal., 25 (1994), pp. 894–914.

[16] M. Marion and R. Temam, *Nonlinear Galerkin methods*, SIAM J. Numer. Anal., 26 (1989), pp. 1139–1157.

[17] J. Novo, E. S. Titi, and S. Wynne, *Efficient methods using high accuracy approximate inertial manifolds*, Numer. Math., 87 (2001), pp. 555–595.

[18] R. Temam, *Infinite-Dimensional Dynamical Systems in Mechanics and Physics*, Appl. Math. Sci. 68, Springer-Verlag, New York, 1988.

[19] R. Temam, *Attractors for the Navier-Stokes equations, Localization and approximation*, J. Fac. Sci. Univ. Tokyo, Sect. IA Math., 36 (1989), pp. 629–647.

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

[21] E. S. Titi, *On approximate inertial manifolds to the Navier-Stokes equations*, J. Math. Anal. Appl., 149 (1990), pp. 540–557.

[22] E. S. Titi, *On a criterion for locating stable stationary solutions to the Navier-Stokes equations*, Nonlinear Anal., 9 (1987), pp. 1085–1102.

[23] S. Wynne, *Efficient Numerical Algorithms for Simulating Evolution Equations*, Ph.D. Thesis, Department of Mathematics, University of California, Irvine, 1999.

[24] He Yinnian and R. M. M. Mattheij, *Stability and convergence for the reform postprocessing Galerkin method*, Nonlinear Anal. Real World Appl., 4 (2000), pp. 517–533.

SIAM J. Numer. Anal.
Vol. 41, No. 2, pp. 715–730

FIRST-ORDER SYSTEM LEAST SQUARES FOR THE STRESS-DISPLACEMENT FORMULATION: LINEAR ELASTICITY[*]

ZHIQIANG CAI[†] AND GERHARD STARKE[‡]

Abstract. This paper develops a least-squares finite element method for linear elasticity in both two and three dimensions. The least-squares functional is based on the stress-displacement formulation with the symmetry condition of the stress tensor imposed in the first-order system. For the respective displacement and stress, using the Crouzeix–Raviart and Raviart–Thomas finite element spaces, our least-squares finite element method is shown to be optimal in the (broken) H^1 and $H(\mathrm{div})$ norms uniform in the incompressible limit.

Key words. least-squares finite element method, linear elasticity, incompressible limit

AMS subject classifications. 65M60, 65M15

PII. S003614290139696X

1. Introduction.

The practical need of the stress tensor has motivated extensive studies of mixed finite element methods in the stress-displacement formulation (see [1, 4, 2, 3, 5, 11, 14, 20]). Unlike mixed methods for second-order scalar elliptic boundary value problems, stress-displacement finite elements are extremely difficult to construct. This is due to the fact that the stress tensor is symmetric. A beautiful finite element space had not been constructed until recently by Arnold and Winther [5]. Their space is a natural extension of the Raviart–Thomas space of $H(\mathrm{div})$. The minimum degree of freedom on each triangle of Arnold and Winther space for the symmetric stress tensor in two dimensions is 24, which is very expensive. Previous works impose the symmetry condition weakly via a Lagrange multiplier (see [1, 2, 20]). Like scalar elliptic problems, mixed methods lead to saddle-point problems, and mixed finite elements are subject to the inf-sup condition. Many solution methods which work well for symmetric positive definite problems cannot be applied directly. Although substantial progress in solution methods for saddle-point problems has been achieved, these problems may still be difficult and expensive to solve.

Finite element methods of least-squares type have been the object of many studies recently (see, e.g., the survey [7] and the monograph [18]). Least-squares finite element methods have also been applied to first-order system formulations of linear elasticity, for example, in [13], where displacement gradients are used as additional degrees of freedom. Recently, a displacement-stress-rotation least-squares formulation has been investigated in [19] (see also the references therein for some other least-squares approaches in the engineering literature). Our aim is to present a least-squares formulation that computes approximations for the stress and displacement only. These are

[*] Received by the editors October 25, 2001; accepted for publication (in revised form) November 4, 2002; published electronically May 6, 2003. This work was sponsored in part by National Science Foundation grants ASC-9720257 and INT-9910010.

http://www.siam.org/journals/sinum/41-2/39696.html

[†] Department of Mathematics, Purdue University, 1395 Mathematical Sciences Building, West Lafayette, IN 47907-1395 (zcai@math.purdue.edu). The research of this author was performed in part under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under contract W-7405-Eng-48.

[‡] Institut für Angewandte Mathematik, Universität Hannover, Welfengarten 1, 30167 Hannover, Germany (starke@ifam.uni-hannover.de). The research of this author was supported in part by the Deutscher Akademischer Austauschdienst (DAAD) through a bilateral travel grant.

the quantities of interest in many practical applications including coupling of elastic deformation with fluid flow models. The least-squares formulation presented in this paper also has some advantages for the extension to geometrically nonlinear elasticity computations, as will be considered in a companion paper.

The purpose of this paper is to develop a least-squares finite element method based on the stress-displacement formulation. To circumvent the numerical difficulty on the symmetry of the stress tensor, we impose such a symmetry condition in the first-order system and then apply the least-squares principle to this overdetermined, but consistent, system. The least-squares functional uses the L^2 norm, and it is shown that the homogeneous functional is equivalent to the energy norm involving the Lamé constant for the displacement and the standard $H(\mathrm{div})$ norm for the stress. This implies that our least-squares finite element method using the respective Crouzeix–Raviart and Raviart–Thomas spaces for the displacement and stress yields optimal error estimates uniformly in the incompressible limit. The algebraic system resulting in this discretization may be efficiently solved by multigrid methods, which will be considered in a forthcoming paper. Additionally, we consider an inverse norm least-squares functional and show that its homogeneous form is equivalent to the energy norm for the displacement and the L^2 norm for the stress. This functional can be used to develop a discrete inverse norm least-squares method (see, e.g., [9]).

An outline of the paper is as follows. The linear elasticity system is introduced in section 2, along with some notations. Section 3 develops the least-squares functionals based on the extended first-order system of the stress and displacement and establishes their ellipticity and continuity. Section 4 discusses the finite element approximation. Finally, section 5 establishes an inequality in the stress tensor space, used in section 3, through a Helmholtz decomposition.

2. Linear elasticity and preliminaries.

We consider an isotropic elastic material in the configuration space $\Omega \subset \Re^d$ ($d = 2$ or 3). Assume that Ω is a bounded, open, connected domain with Lipschitz boundary $\partial\Omega$. Let $\mathbf{u} = (u_1, \dots, u_d)^t$ be the displacement and $\mathbf{f} = (f_1, \dots, f_d)^t$ be the body force. The constituent law expresses a linear relation between the stress tensor $\boldsymbol{\sigma}(\mathbf{u}) = (\sigma_{ij}(\mathbf{u}))_{d\times d}$ and the linearized strain tensor $\boldsymbol{\epsilon}(\mathbf{u}) = (\epsilon_{ij}(\mathbf{u}))_{d\times d}$, with $\epsilon_{ij}(\mathbf{u}) = \frac{1}{2}(\partial_j u_i + \partial_i u_j)$:

$$\sigma_{ij}(\mathbf{u}) = \lambda \mathrm{tr}\,\big(\boldsymbol{\epsilon}(\mathbf{u})\big)\delta_{ij} + 2\mu\epsilon_{ij}(\mathbf{u}), \tag{2.1}$$

where tr stands for the trace operator (i.e., $\mathrm{tr}\,(\boldsymbol{\epsilon}(\mathbf{u})) = \sum_{j=1}^d \epsilon_{jj}(\mathbf{u}) = \nabla\cdot\mathbf{u}$), δ_{ij} is the Kronecker tensor, and the positive constants λ and μ are the Lamé constants such that $\mu \in [\mu_1, \mu_2]$ with $0 < \mu_1 < \mu_2$ and $\lambda \in (0, \infty)$. We have the equilibrium equation

$$\sum_{i=1}^d \frac{\partial\sigma_{ij}(\mathbf{u})}{\partial x_i} + f_j = 0 \quad \text{for } j = 1, \dots, d. \tag{2.2}$$

Let Γ_D and Γ_N be a partition of the boundary of Ω such that $\partial\Omega = \bar{\Gamma}_D \cup \bar{\Gamma}_N$ and $\Gamma_D \cap \Gamma_N = \emptyset$. Let $\mathbf{n} = (n_1, \dots, n_d)^t$ be the outward unit vector normal to the boundary. We impose the homogeneous displacement and traction boundary conditions

$$\left\{\begin{aligned} \mathbf{u} = \mathbf{0} \ \text{ on } \ \Gamma_D,& \\ \sum_{i=1}^d \sigma_{ij}(\mathbf{u})n_i = 0 \ \text{ on } \ \Gamma_N& \quad \text{for } j = 1, \dots, d. \end{aligned}\right. \tag{2.3}$$

For simplicity, we assume that Γ_D is not empty (i.e., $\operatorname{mes}(\Gamma_D) \neq 0$). For the pure traction problem ($\Gamma_D = \emptyset$), our approach may be easily extended to the space of infinitesimal rigid motions.

We use the standard notation and definition for the Sobolev spaces $H^s(\Omega)$ for $s \geq 0$, the associated inner products are denoted by $(\cdot,\, \cdot)_{s,\,\Omega}$, and their norms by $\|\cdot\|_{s,\,\Omega}$. (We will omit Ω from the inner product and norm designation when there is no risk of confusion.) For $s = 0$, $H^s(\Omega)$ coincides with $L^2(\Omega)$. In this case, the norm and inner product will be denoted by $\|\cdot\|$ and $(\cdot,\, \cdot)$, respectively. Let

$$H_D^1(\Omega) = \{v \in H^1(\Omega) \,:\, v = 0 \text{ on } \Gamma_D\} \quad \text{and} \quad H_N^1(\Omega) = \{v \in H^1(\Omega) \,:\, v = 0 \text{ on } \Gamma_N\}.$$

We use $H_D^{-1}(\Omega)$ to denote the dual of $H_D^1(\Omega)$ with the norm defined by

$$\|\phi\|_{-1,\,D} = \sup_{0 \neq \psi \in H_D^1(\Omega)} \frac{(\phi,\, \psi)}{\|\psi\|_1}$$

(see [6, section 6.2]). Let

$$H(\operatorname{div};\Omega) = \{\mathbf{q} \in L^2(\Omega)^d \,:\, \nabla\cdot\mathbf{q} \in L^2(\Omega)\}$$

and

$$H(\mathbf{curl};\, \Omega) = \{\mathbf{q} \in L^2(\Omega)^d \,:\, \nabla\times\mathbf{q} \in L^2(\Omega)^{2d-3}\},$$

which are Hilbert spaces under the respective norms

$$\|\mathbf{q}\|_{H(\operatorname{div};\,\Omega)} = \left(\|\mathbf{q}\|^2 + \|\nabla\cdot\mathbf{q}\|^2\right)^{\frac{1}{2}} \quad \text{and} \quad \|\mathbf{q}\|_{H(\mathbf{curl};\,\Omega)} = \left(\|\mathbf{q}\|^2 + \|\nabla\times\mathbf{q}\|^2\right)^{\frac{1}{2}}.$$

Define the subspaces

$$H_N(\operatorname{div};\, \Omega) = \{\mathbf{q} \in H(\operatorname{div};\, \Omega) \,:\, \mathbf{n}\cdot\mathbf{q} = 0 \ \text{ on } \ \Gamma_N\}$$

and

$$H_D(\mathbf{curl};\, \Omega) = \{\mathbf{q} \in H(\mathbf{curl};\, \Omega) \,:\, \mathbf{n}\times\mathbf{q} = \mathbf{0} \ \text{ on } \ \Gamma_D\}.$$

Finally, define the product spaces

$$H_D^{-1}(\Omega)^d = \prod_{i=1}^d H_D^{-1}(\Omega)\,, \quad H_N(\operatorname{div};\Omega)^d = \prod_{i=1}^d H_N(\operatorname{div};\Omega),$$
$$\text{and} \quad H_D(\mathbf{curl};\Omega)^d = \prod_{i=1}^d H_D(\mathbf{curl};\Omega)$$

with standard product norms. We also use the notations

$$\boldsymbol{\sigma} : \boldsymbol{\tau} = \sum_{i,j=1}^d \sigma_{ij}\tau_{ij} \quad \text{and} \quad |\boldsymbol{\tau}| = \sqrt{\boldsymbol{\tau} : \boldsymbol{\tau}}.$$

The weak form of boundary value problem for the displacement in (2.2) and (2.3) has a unique solution $\mathbf{u} \in H_D^1(\Omega)^d$ for every $\mathbf{f} \in H_D^{-1}(\Omega)^d$. Moreover, the solution $\mathbf{u}$ satisfies the following H^1 regularity estimate:

$$\|\mathbf{u}\|_1 + \lambda\|\nabla\cdot\mathbf{u}\| \leq C\,\|\mathbf{f}\|_{-1}. \tag{2.4}$$

If the domain Ω is convex or its boundary is $C^{1,1}$, then the H^2 regularity estimate holds:

$$\|\mathbf{u}\|_2 + \lambda \|\nabla \cdot \mathbf{u}\|_1 \le C\,\|\mathbf{f}\| \tag{2.5}$$

for the pure displacement or pure traction problems (see, e.g., [10]). We use C with or without subscripts to denote a generic positive constant, possibly different at different occurrences, which is independent of the Lamé constant λ and the mesh size h introduced in the subsequent section but may depend on the Lamé constant μ and the domain Ω. We will frequently use the term *uniform* in reference to a relation to mean that it holds independent of λ and h.

3. First-order system least squares.

Let $\mathcal{C} = \lambda \mathbf{b}\mathbf{b}^t + 2\mu I$ be a $d^2 \times d^2$ matrix, where

$$\mathbf{b} = \begin{cases} (1,\,0,\,0,\,1)^t, & d = 2, \\ (1,\,0,\,0,\,0,\,1,\,0,\,0,\,0,\,1)^t, & d = 3. \end{cases}$$

It is easy to see that $\mathcal{C}$ is symmetric and positive definite and that its inverse has the form of

$$\mathcal{C}^{-1} = \frac{1}{2\mu}\left(I - \frac{\lambda}{d\lambda + 2\mu}\mathbf{b}\mathbf{b}^t \right).$$

It is convenient to view $d \times d$-matrices as d^2-vectors, e.g., $(\sigma_{ij})_{d\times d}$ as $(\boldsymbol{\sigma}_1, \ldots, \boldsymbol{\sigma}_d)^t$, where $\boldsymbol{\sigma}_j = (\sigma_{1j}, \ldots, \sigma_{dj})^t$ is the jth column of $(\sigma_{ij})_{d\times d}$ for $j = 1, \ldots, d$. Thus,

$$\operatorname{tr} \boldsymbol{\sigma} = \operatorname{tr}\,(\sigma_{ij})_{d\times d} = \sum_{i=1}^{d} \sigma_{ii} = \mathbf{b}^t \begin{pmatrix} \boldsymbol{\sigma}_1 \\ \vdots \\ \boldsymbol{\sigma}_d \end{pmatrix} = \mathbf{b}^t \boldsymbol{\sigma}.$$

Now, the constituent law may be rewritten in terms of the matrix $\mathcal{C}$:

$$\boldsymbol{\sigma}(\mathbf{u}) = \mathcal{C}\boldsymbol{\epsilon}(\mathbf{u}). \tag{3.1}$$

By treating the stress tensor as independent variables, we then have the following first-order system:

$$\begin{cases} \boldsymbol{\sigma} - \mathcal{C}\boldsymbol{\epsilon}(\mathbf{u}) & = & \mathbf{0} & \text{in} & \Omega, \\ \nabla \cdot \boldsymbol{\sigma} + \mathbf{f} & = & \mathbf{0} & \text{in} & \Omega, \end{cases} \tag{3.2}$$

with boundary conditions

$$\mathbf{u} = \mathbf{0} \text{ on } \Gamma_D \quad \text{and} \quad \mathbf{n} \cdot \boldsymbol{\sigma} = \mathbf{0} \text{ on } \Gamma_N. \tag{3.3}$$

Here, the respective divergence and normal operators $\nabla\cdot$ and $\mathbf{n}\cdot$ (and other operators encountered in the subsequent section) are extended componentwise:

$$\nabla \cdot \boldsymbol{\sigma} = \begin{pmatrix} \nabla \cdot \boldsymbol{\sigma}_1 \\ \vdots \\ \nabla \cdot \boldsymbol{\sigma}_d \end{pmatrix} \quad \text{and} \quad \mathbf{n} \cdot \boldsymbol{\sigma} = \begin{pmatrix} \mathbf{n} \cdot \boldsymbol{\sigma}_1 \\ \vdots \\ \mathbf{n} \cdot \boldsymbol{\sigma}_d \end{pmatrix}.$$

Note that the stress tensor is symmetric; that is,

$$\boldsymbol{\sigma} = \boldsymbol{\sigma}^t \quad \text{in } \Omega. \tag{3.4}$$

(Here, $\boldsymbol{\sigma}^t$ denotes the transpose of $\boldsymbol{\sigma}$ as a $d \times d$ matrix.) One can impose such symmetry in the solution space as in [5]. By doing so, it complicates the construction and increases the dimension of the finite element space. The construction of a piecewise linear $H(\mathrm{div})$-conforming finite element space for the stress field would necessarily be of the form

$$\boldsymbol{\sigma}|_T = \begin{pmatrix} \alpha_T + \gamma_T x_1 & \beta_T + \gamma_T x_2 \\ \rho_T + \delta_T x_1 & \sigma_T + \delta_T x_2 \end{pmatrix}$$

with $\alpha_T, \beta_T, \gamma_T, \delta_T, \rho_T, \sigma_T \in \Re$. The symmetry condition would imply $\gamma_T = \delta_T = 0$, leaving us with nothing but constants and therefore with $\operatorname{div} \boldsymbol{\sigma} = \mathbf{0}$. This does certainly not lead to an acceptable approximation property in the $H(\mathrm{div})$ norm, and therefore, piecewise linear finite element spaces are not admissible in this context. Instead of using higher-order polynomials, we choose to impose the symmetry condition in the system. To this end, an equivalent extended system for (3.2) is

$$\text{(3.5)} \qquad \left\{ \begin{array}{rcll} \mathcal{C}^{-\frac{1}{2}}\boldsymbol{\sigma} - \mathcal{C}^{\frac{1}{2}}\boldsymbol{\epsilon}(\mathbf{u}) & = & \mathbf{0} & \text{in} \quad \Omega, \\ \nabla \cdot \boldsymbol{\sigma} + \mathbf{f} & = & \mathbf{0} & \text{in} \quad \Omega, \\ \frac{1}{2}\left(\boldsymbol{\sigma} - \boldsymbol{\sigma}^t\right) & = & \mathbf{0} & \text{in} \quad \Omega. \end{array} \right.$$

Applying the trace operator to (3.1) gives

$$\text{(3.6)} \qquad \operatorname{tr} \boldsymbol{\sigma} = \operatorname{tr} \mathcal{C}\boldsymbol{\epsilon}(\mathbf{u}) = (d\lambda + 2\mu)\nabla \cdot \mathbf{u} \quad \text{in} \quad \Omega.$$

If $\Gamma_N = \emptyset$, then $\int_\Omega \nabla \cdot \mathbf{u}\, dx = \int_{\partial\Omega} \mathbf{n} \cdot \mathbf{u}\, ds = 0$, which implies $\int_\Omega \operatorname{tr} \boldsymbol{\sigma}\, dx = 0$. Therefore, we are at liberty to impose such a condition for $\boldsymbol{\sigma}$. Let $\mathbf{X}$ denote $H_N(\mathrm{div};\, \Omega)^d$ if $\Gamma_N \neq \emptyset$, and its subspace $\{\boldsymbol{\tau} \in H_N(\mathrm{div};\, \Omega)^d : \int_\Omega \operatorname{tr} \boldsymbol{\tau}\, dx = 0\}$ otherwise. For $\mathbf{f} \in L^2(\Omega)^d$, we define the following least-squares functionals:

$$\text{(3.7)} \qquad G_{-1}(\mathbf{u},\, \boldsymbol{\sigma};\, \mathbf{f}) = \|\mathcal{C}^{-\frac{1}{2}}\boldsymbol{\sigma} - \mathcal{C}^{\frac{1}{2}}\boldsymbol{\epsilon}(\mathbf{u})\|^2 + \|\nabla \cdot \boldsymbol{\sigma} + \mathbf{f}\|^2_{-1,D} + \left\| \frac{1}{2}\left(\boldsymbol{\sigma} - \boldsymbol{\sigma}^t\right) \right\|^2$$

and

$$\text{(3.8)} \qquad G(\mathbf{u},\, \boldsymbol{\sigma};\, \mathbf{f}) = \|\mathcal{C}^{-\frac{1}{2}}\boldsymbol{\sigma} - \mathcal{C}^{\frac{1}{2}}\boldsymbol{\epsilon}(\mathbf{u})\|^2 + \|\nabla \cdot \boldsymbol{\sigma} + \mathbf{f}\|^2 + \left\| \frac{1}{2}\left(\boldsymbol{\sigma} - \boldsymbol{\sigma}^t\right) \right\|^2$$

for $(\mathbf{u},\, \boldsymbol{\sigma}) \in \mathbf{H} \equiv H^1_D(\Omega)^d \times \mathbf{X}$. We first establish uniform boundedness and ellipticity (i.e., equivalence) of the homogeneous functionals $G_{-1}(\mathbf{v},\, \boldsymbol{\tau};\, \mathbf{0})$ and $G(\mathbf{v},\, \boldsymbol{\tau};\, \mathbf{0})$ in terms of the respective functionals $M_{-1}(\mathbf{v},\, \boldsymbol{\tau})$ and $M(\mathbf{v},\, \boldsymbol{\tau})$ defined on $\mathbf{H}$ by

$$M_{-1}(\mathbf{v},\, \boldsymbol{\tau}) = \|\mathcal{C}^{\frac{1}{2}}\boldsymbol{\epsilon}(\mathbf{v})\|^2 + \|\mathcal{C}^{-\frac{1}{2}}\boldsymbol{\tau}\|^2 + \|\nabla \cdot \boldsymbol{\tau}\|^2_{-1,D}$$

and

$$M(\mathbf{v},\, \boldsymbol{\tau}) = \|\mathcal{C}^{\frac{1}{2}}\boldsymbol{\epsilon}(\mathbf{v})\|^2 + \|\mathcal{C}^{-\frac{1}{2}}\boldsymbol{\tau}\|^2 + \|\nabla \cdot \boldsymbol{\tau}\|^2.$$

THEOREM 3.1. *There exist positive constants C_1 and C_2, independent of λ, such that*

$$\text{(3.9)} \qquad \frac{1}{C_1} M_{-1}(\mathbf{v},\, \boldsymbol{\tau}) \le G_{-1}(\mathbf{v},\, \boldsymbol{\tau};\, \mathbf{0}) \le C_1 M_{-1}(\mathbf{v},\, \boldsymbol{\tau})$$

and that

$$\frac{1}{C_2} M(\mathbf{v},\, \boldsymbol{\tau}) \le G(\mathbf{v},\, \boldsymbol{\tau};\, \mathbf{0}) \le C_2 M(\mathbf{v},\, \boldsymbol{\tau}) \tag{3.10}$$

hold for all $(\mathbf{v},\, \boldsymbol{\tau}) \in H_D^1(\Omega)^d \times H_N(\mathrm{div};\, \Omega)^d$.

Proof. Decomposing the tensor $\boldsymbol{\tau}$ into symmetric and skew-symmetric parts

$$\boldsymbol{\tau} = \frac{\boldsymbol{\tau} + \boldsymbol{\tau}^t}{2} + \frac{\boldsymbol{\tau} - \boldsymbol{\tau}^t}{2},$$

we then have

$$\mathcal{C}^{-1}\boldsymbol{\tau} = \mathcal{C}^{-1}\left(\frac{\boldsymbol{\tau} + \boldsymbol{\tau}^t}{2}\right) + \frac{1}{2\mu}\frac{\boldsymbol{\tau} - \boldsymbol{\tau}^t}{2}.$$

Note that $A : B = 0$ if A and B are symmetric and skew-symmetric tensors, respectively. Hence,

$$\|\mathcal{C}^{-\frac{1}{2}}\boldsymbol{\tau}\|^2 = \left\|\mathcal{C}^{-\frac{1}{2}}\frac{\boldsymbol{\tau} + \boldsymbol{\tau}^t}{2}\right\|^2 + \left\|\mathcal{C}^{-\frac{1}{2}}\frac{\boldsymbol{\tau} - \boldsymbol{\tau}^t}{2}\right\|^2 = \left\|\mathcal{C}^{-\frac{1}{2}}\frac{\boldsymbol{\tau} + \boldsymbol{\tau}^t}{2}\right\|^2 + \frac{1}{2\mu}\left\|\frac{\boldsymbol{\tau} - \boldsymbol{\tau}^t}{2}\right\|^2.$$

Now, the upper bounds in both (3.9) and (3.10) follow from the triangle inequality. To show the validity of the lower bound in (3.9), note first that $\boldsymbol{\epsilon}(\mathbf{v}) = \frac{1}{2}(\nabla\mathbf{v} + (\nabla\mathbf{v})^t)$ is the symmetric part of the gradient, and hence, using integration by parts,

$$\begin{aligned}(\boldsymbol{\tau},\, \boldsymbol{\epsilon}(\mathbf{v})) &= \left(\frac{\boldsymbol{\tau} + \boldsymbol{\tau}^t}{2},\, \boldsymbol{\epsilon}(\mathbf{v})\right) = \left(\frac{\boldsymbol{\tau} + \boldsymbol{\tau}^t}{2},\, \nabla\mathbf{v}\right) \\ &= (\boldsymbol{\tau},\, \nabla\mathbf{v}) - \left(\frac{\boldsymbol{\tau} - \boldsymbol{\tau}^t}{2},\, \nabla\mathbf{v}\right) = -(\nabla\cdot\boldsymbol{\tau},\, \mathbf{v}) - \left(\frac{\boldsymbol{\tau} - \boldsymbol{\tau}^t}{2},\, \nabla\mathbf{v}\right). \end{aligned} \tag{3.11}$$

Using the Cauchy–Schwarz and Korn inequalities, we then have that

$$\begin{aligned}\|\mathcal{C}^{1/2}\boldsymbol{\epsilon}(\mathbf{v})\|^2 &= (\mathcal{C}\boldsymbol{\epsilon}(\mathbf{v}),\, \boldsymbol{\epsilon}(\mathbf{v})) = (\mathcal{C}\boldsymbol{\epsilon}(\mathbf{v}) - \boldsymbol{\tau},\, \boldsymbol{\epsilon}(\mathbf{v})) + (\boldsymbol{\tau}, \boldsymbol{\epsilon}(\mathbf{v})) \\ &\le \|\mathcal{C}^{-1/2}\boldsymbol{\tau} - \mathcal{C}^{1/2}\boldsymbol{\epsilon}(\mathbf{v})\|\, \|\mathcal{C}^{1/2}\boldsymbol{\epsilon}(\mathbf{v})\| + \|\nabla\cdot\boldsymbol{\tau}\|_{-1,D}\, \|\mathbf{v}\| + \left\|\frac{\boldsymbol{\tau} - \boldsymbol{\tau}^t}{2}\right\|\, \|\nabla\mathbf{v}\| \\ &\le C\left(\|\mathcal{C}^{-1/2}\boldsymbol{\tau} - \mathcal{C}^{1/2}\boldsymbol{\epsilon}(\mathbf{v})\| + \|\nabla\cdot\boldsymbol{\tau}\|_{-1,D} + \left\|\frac{\boldsymbol{\tau} - \boldsymbol{\tau}^t}{2}\right\|\right)\|\mathcal{C}^{1/2}\boldsymbol{\epsilon}(\mathbf{v})\|\,, \end{aligned} \tag{3.12}$$

which implies that

$$\|\mathcal{C}^{\frac{1}{2}}\boldsymbol{\epsilon}(\mathbf{v})\|^2 \le C\left(\|\mathcal{C}^{-\frac{1}{2}}\boldsymbol{\tau} - \mathcal{C}^{\frac{1}{2}}\boldsymbol{\epsilon}(\mathbf{v})\| + \|\nabla\cdot\boldsymbol{\tau}\|_{-1,D} + \left\|\frac{\boldsymbol{\tau} - \boldsymbol{\tau}^t}{2}\right\|\right)^2 \le C\, G_{-1}(\mathbf{v},\, \boldsymbol{\tau};\, \mathbf{0}).$$

Together with the triangle inequality, it is easy to see that $\|\mathcal{C}^{-\frac{1}{2}}\boldsymbol{\tau}\|^2$ is also bounded above by the homogeneous functional. This completes the proof of the lower bound in (3.9). Since $G_{-1}(\mathbf{v},\, \boldsymbol{\tau};\, \mathbf{0}) \le G(\mathbf{v},\, \boldsymbol{\tau};\, \mathbf{0})$ and $\|\nabla\cdot\boldsymbol{\tau}\|^2 \le G(\mathbf{v},\, \boldsymbol{\tau};\, \mathbf{0})$, the lower bound in (3.10) follows from that in (3.9). The proof of the theorem is therefore finished. □

Note that

$$\|\mathcal{C}^{\frac{1}{2}}\boldsymbol{\epsilon}(\mathbf{v})\|^2 = 2\mu\|\boldsymbol{\epsilon}(\mathbf{v})\|^2 + \lambda\|\nabla\cdot\mathbf{v}\|^2.$$

Hence, Korn's inequality (see, e.g., Braess [8, section VI.3]),

$$\|\mathbf{v}\|_1^2 \le C\,\|\boldsymbol{\epsilon}(\mathbf{v})\|^2 \quad \forall\ \mathbf{v}\in H^1_D(\Omega)^d,$$

implies the uniform equivalence of $\|\mathcal{C}^{\frac{1}{2}}\boldsymbol{\epsilon}(\mathbf{v})\|^2$ and

$$|||\mathbf{v}||| \equiv \|\mathbf{v}\|_1^2 + \lambda\|\nabla\cdot\mathbf{v}\|^2;$$

i.e., there exists a positive constant C independent of λ such that

$$\frac{1}{C}\left(\|\mathbf{v}\|_1^2 + \lambda\|\nabla\cdot\mathbf{v}\|^2\right) \le \|\mathcal{C}^{\frac{1}{2}}\boldsymbol{\epsilon}(\mathbf{v})\|^2 \le C\,\left(\|\mathbf{v}\|_1^2 + \lambda\|\nabla\cdot\mathbf{v}\|^2\right) \tag{3.13}$$

holds for all $\mathbf{v}\in H^1_D(\Omega)^d$. It is easy to see that

$$\|\mathcal{C}^{-1/2}\boldsymbol{\tau}\|^2 = \frac{1}{2\mu}\left(\|\boldsymbol{\tau}\|^2 - \frac{\lambda}{d\lambda+2\mu}\|\mathrm{tr}\,\boldsymbol{\tau}\|^2\right).$$

We may split $\mathcal{C}^{-1}$ into its deviatoric and volumetric parts as

$$\mathcal{C}^{-1}\boldsymbol{\tau} = \frac{1}{2\mu}\left(I - \frac{1}{d}\mathbf{b}\mathbf{b}^t\right)\boldsymbol{\tau} + \frac{1}{d(d\lambda+2\mu)}\mathbf{b}\mathbf{b}^t\boldsymbol{\tau} = \frac{1}{2\mu}\mathbf{dev}\,\boldsymbol{\tau} + \frac{1}{d(d\lambda+2\mu)}\mathrm{tr}\,\boldsymbol{\tau}\,I,$$

which implies

$$\|\mathcal{C}^{-1/2}\boldsymbol{\tau}\|^2 = \frac{1}{2\mu}\|\mathbf{dev}\,\boldsymbol{\tau}\|^2 + \frac{1}{d(d\lambda+2\mu)}\|\mathrm{tr}\,\boldsymbol{\tau}\|^2\,. \tag{3.14}$$

This means that the nondeviatoric part of the stress is unweighted in the incompressible limit. Particularly, in two dimensions one has

$$\|\mathcal{C}^{-1/2}\boldsymbol{\tau}\|^2 = \frac{1}{2\mu}\|\tau_{12}\|^2 + \frac{1}{2\mu}\|\tau_{21}\|^2 + \frac{1}{4\mu}\|\tau_{11}-\tau_{22}\|^2 + \frac{1}{4(\lambda+\mu)}\|\mathrm{tr}\,\boldsymbol{\tau}\|^2\,. \tag{3.15}$$

LEMMA 3.2. *For any $\boldsymbol{\tau}\in\mathbf{X}$, there exists a positive constant C independent of λ such that*

$$\|\boldsymbol{\tau}\| \le C\,(\|\mathcal{C}^{-1/2}\boldsymbol{\tau}\| + \|\nabla\cdot\boldsymbol{\tau}\|_{-1,D}). \tag{3.16}$$

Proof. The validity of (3.16) follows from Lemmas 5.3 and 5.4 (see section 5) and the fact that

$$\|\boldsymbol{\tau}\|^2 = 2\mu\|\mathcal{C}^{-1/2}\boldsymbol{\tau}\|^2 + \frac{\lambda}{d\lambda+2\mu}\|\mathrm{tr}\,\boldsymbol{\tau}\|^2 \le 2\mu\|\mathcal{C}^{-1/2}\boldsymbol{\tau}\|^2 + \frac{1}{d}\|\mathrm{tr}\,\boldsymbol{\tau}\|^2.$$

This completes the proof of the lemma. □

Since, for all $\boldsymbol{\tau}\in\mathbf{X}$,

$$\|\nabla\cdot\boldsymbol{\tau}\|_{-1,D} \le \|\boldsymbol{\tau}\| \quad \text{and} \quad \|\nabla\cdot\boldsymbol{\tau}\|_{-1,D} \le \|\nabla\cdot\boldsymbol{\tau}\|\,,$$

it is then easy to see that there exist positive constants C_1 and C_2 such that

$$\frac{1}{C_1}\|\boldsymbol{\tau}\|^2 \le \|\mathcal{C}^{-1/2}\boldsymbol{\tau}\|^2 + \|\nabla\cdot\boldsymbol{\tau}\|^2_{-1,D} \le C_1\|\boldsymbol{\tau}\|^2 \tag{3.17}$$

and that

$$\frac{1}{C_2}\|\boldsymbol{\tau}\|^2_{H(\mathrm{div};\Omega)} \le \|\mathcal{C}^{-1/2}\boldsymbol{\tau}\|^2 + \|\nabla\cdot\boldsymbol{\tau}\|^2 \le C_2\|\boldsymbol{\tau}\|^2_{H(\mathrm{div};\Omega)}. \tag{3.18}$$

THEOREM 3.3. *There exist positive constants C_1 and C_2, independent of λ, such that*

$$\frac{1}{C_1}\left(|||\mathbf{v}|||^2 + \|\boldsymbol{\tau}\|^2\right) \le G_{-1}(\mathbf{v},\,\boldsymbol{\tau};\,\mathbf{0}) \le C_1\left(|||\mathbf{v}|||^2 + \|\boldsymbol{\tau}\|^2\right) \tag{3.19}$$

and that

$$\frac{1}{C_2}(|||\mathbf{v}|||^2 + \|\boldsymbol{\tau}\|^2_{H(\mathrm{div};\Omega)}) \le G(\mathbf{v},\,\boldsymbol{\tau};\,\mathbf{0}) \le C_2(|||\mathbf{v}|||^2 + \|\boldsymbol{\tau}\|^2_{H(\mathrm{div};\Omega)}) \tag{3.20}$$

hold for all $(\mathbf{v},\,\boldsymbol{\tau}) \in H^1_D(\Omega)^d \times \mathbf{X}$.

Proof. The theorem is a direct consequence of Theorem 3.1, (3.13), (3.17), and (3.18). □

4. Finite element approximation. For the finite element approximation of the system (3.5), the least-squares functional in (3.8) is minimized with respect to appropriate finite-dimensional spaces. For the stress approximation, the standard $H(\mathrm{div};\Omega)$-conforming Raviart–Thomas elements may be used. Due to the special structure of $\mathcal{C}^{-1}$, we have proved the uniform equivalence of $M(0,\boldsymbol{\tau})$ and the $H(\mathrm{div};\Omega)$ norm in (3.18). Therefore, [11, Proposition 3.9] gives approximation properties which are uniform in λ with respect to $M(0,\,\cdot)$. However, the situation is more complicated for the displacement approximation. In order to get approximation properties with respect to

$$\|\mathbf{v}\|_1^2 + \lambda\|\nabla\cdot\mathbf{v}\|^2\;,$$

standard continuous piecewise polynomial elements are not sufficient. Following [11, section VI.3] we may use nonconforming finite element spaces; see also [10, section 9.4] for the case of Crouzeix–Raviart elements.

To this end, let $\mathcal{T}_h$ be a regular triangulation of the domain Ω with elements of size $O(h)$ (see [14]). The minimization is then carried out for the discrete least-squares functional

$$G_h(\mathbf{u}_h,\boldsymbol{\sigma}_h;\mathbf{f}) = \sum_{K\in\mathcal{T}_h}\|\mathcal{C}^{-\frac12}\boldsymbol{\sigma}_h - \mathcal{C}^{\frac12}\boldsymbol{\epsilon}(\mathbf{u}_h)\|^2_{0,K} + \|\nabla\cdot\boldsymbol{\sigma}_h + \mathbf{f}\|^2 + \left\|\frac12\left(\boldsymbol{\sigma}_h - \boldsymbol{\sigma}_h^t\right)\right\|^2 \tag{4.1}$$

over a finite dimensional space $\mathbf{V}_h \times \mathbf{X}_h$. If we define the associated bilinear form

$$\begin{aligned}\mathcal{B}_h(\mathbf{u},\boldsymbol{\sigma};\mathbf{v},\boldsymbol{\tau}) = \sum_{K\in\mathcal{T}_h}&(\mathcal{C}^{-\frac12}\boldsymbol{\sigma} - \mathcal{C}^{\frac12}\boldsymbol{\epsilon}(\mathbf{u}),\mathcal{C}^{-\frac12}\boldsymbol{\tau} - \mathcal{C}^{\frac12}\boldsymbol{\epsilon}(\mathbf{v}))_{0,K}\\ &+ (\nabla\cdot\boldsymbol{\sigma},\nabla\cdot\boldsymbol{\tau}) + \frac14(\boldsymbol{\sigma}-\boldsymbol{\sigma}^t,\boldsymbol{\tau}-\boldsymbol{\tau}^t)\;,\end{aligned}$$

then the minimum $(\mathbf{u}_h,\boldsymbol{\sigma}_h) \in \mathbf{V}_h\times\mathbf{X}_h$ of the least-squares functional in (4.1) satisfies

$$\mathcal{B}_h(\mathbf{u}_h,\boldsymbol{\sigma}_h;\mathbf{v},\boldsymbol{\tau}) = -(f,\nabla\cdot\boldsymbol{\tau}) \tag{4.2}$$

for all $(\mathbf{v},\boldsymbol{\tau}) \in \mathbf{V}_h \times \mathbf{X}_h$.

For simplicity, we restrict ourselves to triangular elements in two dimensions. Specifically, for $k \geq 1$,

$$\begin{aligned}\mathbf{V}_h = \{&\mathbf{v} \in L^2(\Omega)^2 \,:\, \mathbf{v}|_T \text{ is a polynomial of degree } k \text{ for each } K \in \mathcal{T}_h\,,\\ &\text{such that } \mathbf{v} \text{ is continuous at the } k \text{ Gauss points on interior edges}\,,\\ &\text{and } \mathbf{v} = \mathbf{0} \text{ at the } k \text{ Gauss points of edges in } \Gamma_D\}\end{aligned}$$

and

$$\begin{aligned}\mathbf{X}_h = \{&\boldsymbol{\tau}_h \subset \mathbf{X} \,:\, \mathbf{v}|_T \text{ is a polynomial of degree } k \text{ for each } K \in \mathcal{T}_h\,,\\ &\text{such that } \mathbf{n}\cdot\boldsymbol{\tau}_h \text{ is a polynomial of degree } k-1 \text{ along edges}\}\,.\end{aligned}$$

In order to establish approximation properties for this approach, we need to modify the result of Theorem 3.1 for the discrete least-squares functional in (4.1). To this end, we define a discrete norm by

$$|||(\mathbf{v},\boldsymbol{\tau})|||_h \equiv \left(\sum_{K\in\mathcal{T}_h} \|\mathcal{C}^{\frac{1}{2}}\boldsymbol{\epsilon}(\mathbf{v})\|_{0,K}^2 + \|\mathcal{C}^{-\frac{1}{2}}\boldsymbol{\tau}\|^2 + \|\nabla\cdot\boldsymbol{\tau}\|^2\right)^{\frac{1}{2}} \tag{4.3}$$

and show its equivalence with respect to the discrete least-squares functional.

THEOREM 4.1. *There exist positive constants C_E and C_C, independent of λ, such that*

$$\begin{aligned}&G_h(\mathbf{v},\boldsymbol{\tau};\mathbf{0}) \geq C_E|||(\mathbf{v},\boldsymbol{\tau})|||_h^2\\ &\qquad\qquad \forall\,(\mathbf{v},\boldsymbol{\tau}) \in \mathbf{V}_h \times \mathbf{X}_h\,,\\ &G_h(\mathbf{v},\boldsymbol{\tau};\mathbf{0}) \leq C_C|||(\mathbf{v},\boldsymbol{\tau})|||_h^2\\ &\qquad\qquad \forall\,(\mathbf{v},\boldsymbol{\tau}) \in (H_D^1(\Omega) + \mathbf{V}_h) \times H_N(\mathrm{div};\Omega).\end{aligned} \tag{4.4}$$

Proof. We proceed similarly to the proof of Theorem 3.1. As in (3.11), we obtain

$$\begin{aligned}\sum_{K\in\mathcal{T}_h} (\boldsymbol{\tau},\boldsymbol{\epsilon}(\mathbf{v}))_{0,K} &= \sum_{K\in\mathcal{T}_h}\left[(\boldsymbol{\tau},\nabla\mathbf{v})_{0,K} - \left(\frac{\boldsymbol{\tau}-\boldsymbol{\tau}^t}{2},\nabla\mathbf{v}\right)_{0,K}\right]\\ &= \sum_{K\in\mathcal{T}_h} (\mathbf{n}\cdot\boldsymbol{\tau},\mathbf{v})_{0,\partial K} - \sum_{K\in\mathcal{T}_h} (\nabla\cdot\boldsymbol{\tau},\mathbf{v})_{0,K} - \sum_{K\in\mathcal{T}_h}\left(\frac{\boldsymbol{\tau}-\boldsymbol{\tau}^t}{2},\nabla\mathbf{v}\right)_{0,K}.\end{aligned}$$

The first sum on the right-hand side can be written as a sum over all edges

$$\sum_{\mathcal{E}_h\ni E\subseteq\Gamma_N} (\mathbf{n}\cdot\boldsymbol{\tau},\mathbf{v})_{0,E} + \sum_{\mathcal{E}_h\ni E\subseteq\Gamma_D} (\mathbf{n}\cdot\boldsymbol{\tau},\mathbf{v})_{0,E} + \sum_{\mathcal{E}_h\ni E\not\subseteq\partial\Omega} (\mathbf{n}\cdot\boldsymbol{\tau},[\mathbf{v}])_{0,E}\,, \tag{4.5}$$

where $\mathcal{E}_h$ is the collection of all edges of the triangulation $\mathcal{T}_h$, and $[\mathbf{v}]$ denotes the jump of $\mathbf{v}$ on E. For $(\mathbf{v},\boldsymbol{\tau}) \in \mathbf{V}_h \times \mathbf{X}_h$, the first term above vanishes since $\mathbf{n}\cdot\boldsymbol{\tau} = \mathbf{0}$ on Γ_N. For the remaining two terms, we see that $\mathbf{n}\cdot\boldsymbol{\tau}$ is a polynomial of degree $k-1$, and $\mathbf{v}$ or $[\mathbf{v}]$, respectively, is a polynomial of degree k which vanishes at the Gauss points. In both cases, the integrand is therefore a polynomial of degree $2k-1$, which is zero at the k Gauss points, implying that the second and third terms in (4.5) also vanish. We therefore have in analogy to (3.12)

$$\sum_{K\in\mathcal{T}_h}(\boldsymbol{\tau},\boldsymbol{\epsilon}(\mathbf{v}))_{0,K}=-(\nabla\cdot\boldsymbol{\tau},\mathbf{v})-\sum_{K\in\mathcal{T}_h}\left(\frac{\boldsymbol{\tau}-\boldsymbol{\tau}^t}{2},\nabla\mathbf{v}\right)_{0,K}. \tag{4.6}$$

The rest of the proof is completely analogous to that of Theorem 3.1. □

Remark. Theorem 4.1 is also valid if nonconforming elements of degree k for the displacement are combined with Raviart–Thomas elements of lower degree for the stress. For example, quadratic nonconforming elements may be combined with the lowest-order Raviart–Thomas spaces.

The quasioptimality of the least-squares finite element approximation follows from the coercivity result in Theorem 4.1 in the usual way.

COROLLARY 4.2. *Let* $(\mathbf{u},\boldsymbol{\sigma})$ *be the solution of* (3.5) *with boundary conditions* (3.3)*, and let* $(\mathbf{u}_h,\boldsymbol{\sigma}_h)\in\mathbf{V}_h\times\mathbf{X}_h$ *be the solution of* (4.2)*. Then*

$$|||(\mathbf{u}-\mathbf{u}_h,\boldsymbol{\sigma}-\boldsymbol{\sigma}_h)|||_h\le C\inf_{(\mathbf{v}_h,\boldsymbol{\tau}_h)\in\mathbf{V}_h\times\mathbf{X}_h}|||(\mathbf{u}-\mathbf{v}_h,\boldsymbol{\sigma}-\boldsymbol{\tau}_h)|||_h\,. \tag{4.7}$$

Proof. The triangle inequality and the first inequality in (4.4) give

$$\begin{aligned}|||(\mathbf{u}-\mathbf{u}_h,\boldsymbol{\sigma}-\boldsymbol{\sigma}_h)|||_h&\le|||(\mathbf{u}-\mathbf{v}_h,\boldsymbol{\sigma}-\boldsymbol{\tau}_h)|||_h+|||(\mathbf{u}_h-\mathbf{v}_h,\boldsymbol{\sigma}_h-\boldsymbol{\tau}_h)|||_h\\&\le|||(\mathbf{u}-\mathbf{v}_h,\boldsymbol{\sigma}-\boldsymbol{\tau}_h)|||_h+C_E^{-1/2}G_h(\mathbf{u}_h-\mathbf{v}_h,\boldsymbol{\sigma}_h-\boldsymbol{\tau}_h;0)^{1/2}\end{aligned}$$

for all $(\mathbf{v}_h,\boldsymbol{\tau}_h)\in\mathbf{V}_h\times\mathbf{X}_h$. The following orthogonality property is the consequence of (3.5) and (4.2):

$$\mathcal{B}_h(\mathbf{u}-\mathbf{u}_h,\boldsymbol{\sigma}-\boldsymbol{\sigma}_h;\mathbf{u}_h-\mathbf{v}_h,\boldsymbol{\sigma}_h-\boldsymbol{\tau}_h)=0.$$

Hence,

$$\begin{aligned}G_h(\mathbf{u}_h-\mathbf{v}_h,\boldsymbol{\sigma}_h-\boldsymbol{\tau}_h;0)&=\mathcal{B}_h(\mathbf{u}_h-\mathbf{v}_h,\boldsymbol{\sigma}_h-\boldsymbol{\tau}_h;\mathbf{u}_h-\mathbf{v}_h,\boldsymbol{\sigma}_h-\boldsymbol{\tau}_h)\\&=\mathcal{B}_h(\mathbf{u}-\mathbf{v}_h,\boldsymbol{\sigma}-\boldsymbol{\tau}_h;\mathbf{u}_h-\mathbf{v}_h,\boldsymbol{\sigma}_h-\boldsymbol{\tau}_h)\\&\le G_h(\mathbf{u}-\mathbf{v}_h,\boldsymbol{\sigma}-\boldsymbol{\tau}_h;0)^{1/2}G_h(\mathbf{u}_h-\mathbf{v}_h,\boldsymbol{\sigma}_h-\boldsymbol{\tau}_h;0)^{1/2}\,,\end{aligned}$$

which, combined with the second inequality in (4.4), implies

$$G_h(\mathbf{u}_h-\mathbf{v}_h,\boldsymbol{\sigma}_h-\boldsymbol{\tau}_h;0)\le G_h(\mathbf{u}-\mathbf{v}_h,\boldsymbol{\sigma}-\boldsymbol{\tau}_h;0)\le C_C|||(\mathbf{u}-\mathbf{v}_h,\boldsymbol{\sigma}-\boldsymbol{\tau}_h)|||_h^2\,.$$

We have therefore proved

$$|||(\mathbf{u}-\mathbf{u}_h,\boldsymbol{\sigma}-\boldsymbol{\sigma}_h)|||_h\le\left(1+\left(\frac{C_C}{C_E}\right)^{1/2}\right)|||(\mathbf{u}-\mathbf{v}_h,\boldsymbol{\sigma}-\boldsymbol{\tau}_h)|||_h \tag{4.8}$$

for all $(\mathbf{v}_h,\boldsymbol{\tau}_h)\in\mathbf{V}_h\times\mathbf{X}_h$. □

THEOREM 4.3. *Assume that* $\mathbf{f}\in L^2(\Omega)^2$ *and that the regularity estimate in* (2.5) *holds. Then, for* $k=1$*, i.e., for* $\mathbf{V}_h$ *the Crouzeix–Raviart elements and* $\mathbf{Q}_h$ *the lowest-order Raviart–Thomas elements, we have the error estimate*

$$|||(\mathbf{u}-\mathbf{u}_h,\boldsymbol{\sigma}-\boldsymbol{\sigma}_h)|||_h\le C\,h\,\|\mathbf{f}\|\,. \tag{4.9}$$

Proof. The definition of the discrete norm in (4.3) implies that it is sufficient to bound the two terms

$$\left(\sum_{K\in\mathcal{T}_h} \|\mathcal{C}^{\frac{1}{2}}\boldsymbol{\epsilon}(\mathbf{u}-\mathbf{v}_h)\|_{0,K}^2\right)^{1/2} \qquad \text{and} \qquad \left(\|\mathcal{C}^{-1/2}(\boldsymbol{\sigma}-\boldsymbol{\tau}_h)\|^2+\|\nabla\cdot(\boldsymbol{\sigma}-\boldsymbol{\tau}_h)\|^2\right)^{1/2}$$

separately. For the first term we conclude in analogy to [10, section 9.4] that there is a mapping $\mathcal{I}_h : H_D^1(\Omega)^2 \to \mathbf{V}_h$ such that

$$\begin{aligned}\left(\sum_{K\in\mathcal{T}_h} \|\mathcal{C}^{1/2}\boldsymbol{\epsilon}(\mathbf{u}-\mathcal{I}_h\mathbf{u})\|_{0,K}^2\right)^{1/2} &\\ = \left(\sum_{K\in\mathcal{T}_h}\left(2\mu\|\boldsymbol{\epsilon}(\mathbf{u}-\mathcal{I}_h\mathbf{u})\|_{0,K}^2+\lambda\|\nabla\cdot(\mathbf{u}-\mathcal{I}_h\mathbf{u})\|_{0,K}^2\right)\right)^{1/2} &\\ \le\ C\,h\ (\|\mathbf{u}\|_2+\lambda\|\nabla\cdot\mathbf{u}\|_1)&\end{aligned}$$

uniformly as $\lambda\to\infty$. For the second term we know that there exists a projection $\mathcal{R}_h : H_N(\text{div};\Omega)^2 \to \mathbf{X}_h$ such that

$$\|\mathcal{C}^{-1/2}(\boldsymbol{\sigma}-\mathcal{R}_h\boldsymbol{\sigma})\| \le \frac{1}{2\mu}\|\boldsymbol{\sigma}-\mathcal{R}_h\boldsymbol{\sigma}\| \le C\,h\ (\|\boldsymbol{\sigma}\|_1+\|\nabla\cdot\boldsymbol{\sigma}\|_1)\,,$$

$$\|\nabla\cdot(\boldsymbol{\sigma}-\mathcal{R}_h\boldsymbol{\sigma})\| \le C\,h\,\|\nabla\cdot\boldsymbol{\sigma}\|_1$$

uniformly in λ (cf. [11, Proposition III.3.9]). The proof is concluded using the regularity estimate (2.5) and the quasioptimality result in Corollary 4.2. □

Due to (3.14), the norm $|||(\,\cdot\,,\,\cdot\,)|||_h$ in Theorem 4.2 degenerates for the trace part as $\lambda\to\infty$. With Lemma 3.2 we get the following stronger result.

COROLLARY 4.4. *Under the same assumptions as in Theorem* 4.3 *we have the error estimate*

$$\left(\sum_{K\in\mathcal{T}_h} \|\mathcal{C}^{1/2}\boldsymbol{\epsilon}(\mathbf{u}-\mathbf{u}_h)\|_{0,K}^2+\|\boldsymbol{\sigma}-\boldsymbol{\sigma}_h\|_{H(\text{div};\Omega)}^2\right)^{1/2} \le C\,h\,\|\mathbf{f}\|. \tag{4.10}$$

Remark. The approximation results (4.9) and (4.10) are also valid for the case $k=2$. For the quadratic nonconforming elements $\mathbf{V}_h$, the existence of an interpolation operator $\mathcal{I}_h : H_D^1(\Omega)^2 \to \mathbf{V}_h$ with the desired properties follows along the same lines as in [10, section 9.4]. The crucial ingredient in the proof there is the property

$$\text{div}\,\mathbf{u}=0 \implies \text{div}\,(\mathcal{I}_h\mathbf{u})|_T=0\ \forall T\in\mathcal{T}_h,$$

which is shown in [17, pp. 513 and 514]. The interpolation result for the quadratic Raviart–Thomas elements also follows from [11, Proposition III.3.9].

Remark. The definition of $|||(\,\cdot\,,\,\cdot\,)|||_h$ involves the term

$$\sum_{K\in\mathcal{T}_h} \|C^{1/2}\boldsymbol{\epsilon}(\mathbf{v})\|_{0,K}^2.$$

For our approximation results (4.9) and (4.10) to be meaningful, we need to show that this defines a norm on $H_D^1(\Omega)+\mathbf{V}_h$. If $\Gamma_N\ne\emptyset$, this is not true for linear Crouzeix–Raviart elements, in general (cf. [11, section VI.3]). For nonconforming finite element spaces of higher degree, however, a discrete Korn's inequality can be shown (see [16]), giving us the desired result.

5. A Helmholtz decomposition. We establish a Helmholtz decomposition for any $\boldsymbol{\tau} \in \mathbf{X}$. To this end, define $\mathbf{q} \in H_D^1(\Omega)^d$ satisfying

$$\left\{ \begin{array}{rll} \nabla \cdot \big(\mathcal{C}\nabla \mathbf{q}\big) = \nabla \cdot \boldsymbol{\tau} & \text{in} & \Omega, \\ \mathbf{q} = \mathbf{0} & \text{on} & \Gamma_D, \\ \mathbf{n} \cdot \big(\mathcal{C}\nabla \mathbf{q}\big) = \mathbf{0} & \text{on} & \Gamma_N. \end{array} \right. \tag{5.1}$$

Its weak form is to find $\mathbf{q} \in H_D^1(\Omega)^d$ such that

$$\lambda(\nabla \cdot \mathbf{q},\, \nabla \cdot \boldsymbol{\xi}) + (\nabla \mathbf{q},\, \nabla \boldsymbol{\xi}) = (\nabla \cdot \boldsymbol{\tau},\, \boldsymbol{\xi}) \quad \forall\, \boldsymbol{\xi} \in H_D^1(\Omega)^d. \tag{5.2}$$

Let $L_D^2(\Omega)$ denote $L_0^2(\Omega) = \{v \in L^2(\Omega) \,:\, \int_\Omega v\, dx = 0\}$ if $\Gamma_N = \emptyset$, or $L^2(\Omega)$ otherwise. We will make use of the following lemma (see, e.g., [15]).

LEMMA 5.1. *For any $p \in L_D^2(\Omega)$, one has*

$$\|p\| \le C \sup_{\mathbf{v} \in H_D^1(\Omega)^d} \frac{(p, \nabla \cdot \mathbf{v})}{\|\mathbf{v}\|_1}. \tag{5.3}$$

LEMMA 5.2. *The solution of* (5.2) *satisfies the following regularity estimate:*

$$\lambda \|\nabla \cdot \mathbf{q}\| + \|\mathbf{q}\|_1 \le C\, \|\nabla \cdot \boldsymbol{\tau}\|_{-1,D}. \tag{5.4}$$

Proof. Taking $\boldsymbol{\xi} = \mathbf{q}$ in (5.2) and using the Poincaré inequality, one has

$$\lambda \|\nabla \cdot \mathbf{q}\|^2 + \|\mathbf{q}\|_1^2 \le C\, \|\nabla \cdot \boldsymbol{\tau}\|_{-1,D}^2. \tag{5.5}$$

It follows from Lemma 5.1 that

$$\lambda \|\nabla \cdot \mathbf{q}\| \le C \sup_{\mathbf{v} \in H_D^1(\Omega)^d} \frac{(\lambda \nabla \cdot \mathbf{q},\, \nabla \cdot \mathbf{v})}{\|\mathbf{v}\|_1} = C \sup_{\mathbf{v} \in H_D^1(\Omega)^d} \frac{(\nabla \cdot \boldsymbol{\tau},\, \mathbf{v}) - (\nabla \mathbf{q},\, \nabla \mathbf{v})}{\|\mathbf{v}\|_1},$$

which, together with the Cauchy–Schwarz inequality and (5.5), implies (5.4). □

First, let us consider the case in which $d = 2$. We use standard curl notation for two dimensions by identifying $\Re^2$ with the (x, y)-plane in $\Re^3$. Thus, the curl of $\mathbf{v} = (v_1,\, v_2)^t$ means the scalar function

$$\nabla \times \mathbf{v} = \partial_1 v_2 - \partial_2 v_1,$$

and $\nabla^\perp$ denotes its formal adjoint:

$$\nabla^\perp v = \begin{pmatrix} \partial_2 v \\ -\partial_1 v \end{pmatrix}.$$

Since $\boldsymbol{\tau} - \mathcal{C}\nabla \mathbf{q}$ is divergence-free, there exists $\boldsymbol{\phi} \in H_N^1(\Omega)^2$ such that

$$\boldsymbol{\tau} = \mathcal{C}\nabla \mathbf{q} + \nabla^\perp \boldsymbol{\phi},$$

where $\boldsymbol{\phi}$ satisfies that

$$\left\{ \begin{array}{rll} \nabla \times \big(\mathcal{C}^{-1}\nabla^\perp \boldsymbol{\phi}\big) = \nabla \times \big(\mathcal{C}^{-1}\boldsymbol{\tau}\big) & \text{in} & \Omega, \\ \mathbf{n} \times \big(\mathcal{C}^{-1}\nabla^\perp \boldsymbol{\phi}\big) = \mathbf{n} \times \big(\mathcal{C}^{-1}\boldsymbol{\tau}\big) & \text{on} & \Gamma_D, \\ \boldsymbol{\phi} = \mathbf{0} & \text{on} & \Gamma_N. \end{array} \right. \tag{5.6}$$

It is easy to see that

$$(\mathcal{C}^{-1}\nabla^\perp \boldsymbol{\phi},\, \nabla^\perp \boldsymbol{\phi}) = (\mathcal{C}^{-1}\boldsymbol{\tau},\, \nabla^\perp \boldsymbol{\phi}) \le \|\mathcal{C}^{-\frac{1}{2}}\boldsymbol{\tau}\|\, \|\mathcal{C}^{-\frac{1}{2}}\nabla^\perp \boldsymbol{\phi}\|,$$

which implies that

$$\frac{1}{2\mu}\left(\|\nabla^\perp \boldsymbol{\phi}\|^2 - \frac{\lambda}{2(\lambda+\mu)}\|\nabla\times\boldsymbol{\phi}\|^2\right) = (\mathcal{C}^{-1}\nabla^\perp \boldsymbol{\phi},\, \nabla^\perp \boldsymbol{\phi}) \le \|\mathcal{C}^{-\frac{1}{2}}\boldsymbol{\tau}\|^2. \tag{5.7}$$

LEMMA 5.3. *For any $\boldsymbol{\tau}\in\mathbf{X}$ and $d=2$, we have the following decomposition:*

$$\boldsymbol{\tau} = \mathcal{C}\nabla\mathbf{q} + \nabla^\perp\boldsymbol{\phi}, \tag{5.8}$$

where $\mathbf{q}\in H^1_D(\Omega)^2$ and $\boldsymbol{\phi}\in H^1_N(\Omega)^2$ satisfy (5.1) *and* (5.6)*, respectively. Moreover, we have that*

$$\|\mathrm{tr}\,\boldsymbol{\tau}\| \le C\left(\|\mathcal{C}^{-\frac{1}{2}}\boldsymbol{\tau}\| + \|\nabla\cdot\boldsymbol{\tau}\|_{-1,D}\right). \tag{5.9}$$

Proof. Since

$$\mathbf{b}^t\nabla\mathbf{q} = \nabla\cdot\mathbf{q} \quad\text{and}\quad \mathbf{b}^t\nabla^\perp\boldsymbol{\phi} = -\nabla\times\boldsymbol{\phi},$$

applying the trace operator to (5.8) gives that

$$\mathrm{tr}\,\boldsymbol{\tau} = 2(\lambda+\mu)\nabla\cdot\mathbf{q} - \nabla\times\boldsymbol{\phi}.$$

By Lemma 5.2, (5.7), and the fact that $\frac{\lambda}{\lambda+\mu} < 1$, to show the validity of (5.9), it then suffices to prove that

$$\|\nabla\times\boldsymbol{\phi}\| \le C\left(\|\nabla^\perp\boldsymbol{\phi}\|^2 - \frac{1}{2}\|\nabla\times\boldsymbol{\phi}\|^2\right)^{\frac{1}{2}}. \tag{5.10}$$

If $\Gamma_N = \emptyset$, then $\nabla\times\boldsymbol{\phi}\in L^2_0(\Omega)$ since

$$\int_\Omega \nabla\times\boldsymbol{\phi}\,dx = 2(\lambda+\mu)\int_\Omega \nabla\cdot\mathbf{q}\,dx - \int_\Omega \mathrm{tr}\,\boldsymbol{\tau}\,dx = 0,$$

where we have used the divergence theorem and $\mathbf{q}=\mathbf{0}$ on $\partial\Omega$ for the first integral, $\boldsymbol{\tau}\in\mathbf{X}$ for the second. Since $(\nabla^\perp\boldsymbol{\phi},\,\nabla\mathbf{v}) = 0$ for all $\mathbf{v}\in H^1_D(\Omega)^2$, it follows from the Cauchy–Schwarz inequality that for any $\mathbf{v}\in H^1_D(\Omega)^2$

$$\begin{aligned}(\nabla\times\boldsymbol{\phi},\,\nabla\cdot\mathbf{v}) &= \big((\nabla\times\boldsymbol{\phi})\mathbf{b},\,\nabla\mathbf{v}\big) = \big((\nabla\times\boldsymbol{\phi})\mathbf{b} + 2\nabla^\perp\boldsymbol{\phi},\,\nabla\mathbf{v}\big)\\ &\le \|(\nabla\times\boldsymbol{\phi})\mathbf{b} + 2\nabla^\perp\boldsymbol{\phi}\|\,\|\nabla\mathbf{v}\| = 2\left(\|\nabla^\perp\boldsymbol{\phi}\|^2 - \frac{1}{2}\|\nabla\times\boldsymbol{\phi}\|^2\right)^{\frac{1}{2}}\|\nabla\mathbf{v}\|.\end{aligned}$$

Hence, by Lemma 5.1, we have

$$\|\nabla\times\boldsymbol{\phi}\| \le C\sup_{\mathbf{v}\in H^1_D(\Omega)^d}\frac{(\nabla\times\boldsymbol{\phi},\nabla\cdot\mathbf{v})}{\|\mathbf{v}\|_1} \le C\left(\|\nabla^\perp\boldsymbol{\phi}\|^2 - \frac{1}{2}\|\nabla\times\boldsymbol{\phi}\|^2\right)^{\frac{1}{2}}.$$

This completes the proof of (5.10) and, hence, the lemma. □

In the case that $d = 3$, since $\boldsymbol{\tau} - \mathcal{C}\nabla\mathbf{q}$ is divergence-free, there exists $\boldsymbol{\Phi} = (\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \boldsymbol{\phi}_3) \in H(\mathbf{curl};\,\Omega)^3$ such that

$$\boldsymbol{\tau} = \mathcal{C}\nabla\mathbf{q} + \nabla\times\boldsymbol{\Phi},$$

where $\boldsymbol{\Phi}$ satisfies that

$$(5.11)\qquad \left\{\begin{array}{cl} \nabla\times\left(\mathcal{C}^{-1}\nabla\times\boldsymbol{\Phi}\right) = \nabla\times\left(\mathcal{C}^{-1}\boldsymbol{\tau}\right) & \text{in } \Omega, \\ \nabla\cdot\boldsymbol{\Phi} = \mathbf{0} & \text{in } \Omega, \\ \mathbf{n}\times\left(\mathcal{C}^{-1}\nabla^{\perp}\boldsymbol{\Phi}\right) = \mathbf{n}\times\left(\mathcal{C}^{-1}\boldsymbol{\tau}\right) & \text{on } \Gamma_D, \\ \mathbf{n}\times\boldsymbol{\Phi} = \mathbf{0} & \text{on } \Gamma_N. \end{array}\right.$$

An argument similar to that for $d = 2$ gives that

$$(5.12)\qquad \frac{1}{2\mu}\left(\|\nabla\times\boldsymbol{\Phi}\|^2 - \frac{\lambda}{3\lambda+2\mu}\|\mathbf{b}^t\nabla\times\boldsymbol{\Phi}\|^2\right) = (\mathcal{C}^{-1}\nabla\times\boldsymbol{\Phi},\,\nabla\times\boldsymbol{\Phi}) \le \|\mathcal{C}^{-\frac{1}{2}}\boldsymbol{\tau}\|^2\,.$$

LEMMA 5.4. *For any $\boldsymbol{\tau} \in \mathbf{X}$ and $d = 3$, we have the following decomposition:*

$$(5.13)\qquad \boldsymbol{\tau} = \mathcal{C}\nabla\mathbf{q} + \nabla\times\boldsymbol{\Phi},$$

where $\mathbf{q} \in H^1_D(\Omega)^2$ and $\boldsymbol{\Phi} \in H(\mathbf{curl};\,\Omega)^3$ satisfy (5.1) *and* (5.11)*, respectively. Moreover, the estimate in* (5.9) *is valid.*

Proof. Again, it suffices to show that

$$(5.14)\qquad \|\mathbf{b}^t\nabla\times\boldsymbol{\Phi}\| \le C\left(\|\nabla\times\boldsymbol{\Phi}\|^2 - \frac{1}{3}\|\mathbf{b}^t\nabla\times\boldsymbol{\Phi}\|^2\right)^{\frac{1}{2}}.$$

An argument similar to that in the proof of Lemma 5.3 implies that

$$\mathbf{b}^t\nabla\times\boldsymbol{\Phi} \in L^2_D(\Omega) \quad\text{and}\quad (\nabla\times\boldsymbol{\Phi},\,\nabla\mathbf{v}) = 0 \quad \forall\ \mathbf{v}\in H^1_D(\Omega)^3.$$

Since

$$\|(\mathbf{b}^t\nabla\times\boldsymbol{\Phi})\mathbf{b} - 3\nabla\times\boldsymbol{\Phi}\| = 3\left(\|\nabla\times\boldsymbol{\Phi}\|^2 - \frac{1}{3}\|\mathbf{b}^t\nabla\times\boldsymbol{\Phi}\|^2\right)^{\frac{1}{2}},$$

it then follows from Lemma 5.3 that

$$\begin{aligned}\|\mathbf{b}^t\nabla\times\boldsymbol{\Phi}\| &\le C\sup_{\mathbf{v}\in H^1_D(\Omega)^d}\frac{(\mathbf{b}^t\nabla\times\boldsymbol{\Phi},\nabla\cdot\mathbf{v})}{\|\mathbf{v}\|_1} \le C\,\|(\mathbf{b}^t\nabla\times\boldsymbol{\Phi})\mathbf{b} - 3\nabla\times\boldsymbol{\Phi}\| \\ &\le C\left(\|\nabla\times\boldsymbol{\Phi}\|^2 - \frac{1}{3}\|\mathbf{b}^t\nabla\times\boldsymbol{\Phi}\|^2\right)^{\frac{1}{2}}.\end{aligned}$$

This completes the proof of (5.10) and, hence, the lemma. □

6. A numerical example. We conclude this paper with a simple numerical example. On the unit square $\Omega = (-1,1)\times(-1,1)$, we consider the system (3.2), (3.3) with

$$\Gamma_D = [-1,1]\times\{-1\}, \qquad \Gamma_N = ([-1,1]\times\{1\})\cup\{-1,1\}\times[-1,1]$$

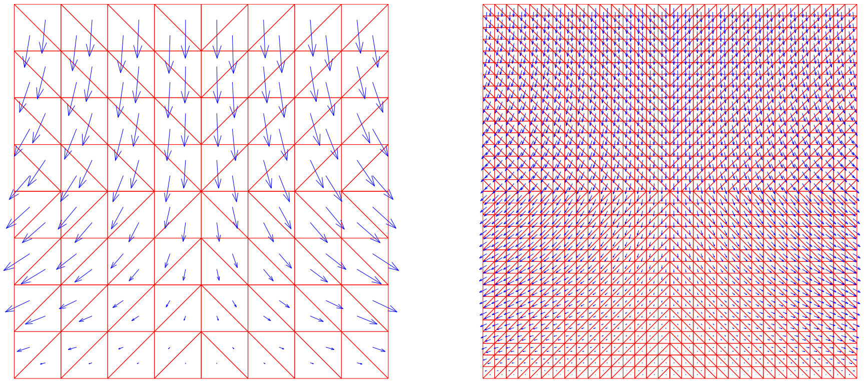


Fig. 6.1. *Displacement field on a uniform triangulation.*

Table 6.1
$G_h(\mathbf{u}_h, \boldsymbol{\sigma}_h; \mathbf{f})$ *for different values of* λ.

h	# triangles	# d.o.f.	$\lambda = 10$	$\lambda = 1000$	$\lambda = 100000$
1	8	76	$2.785 \cdot 10^{-1}$	$3.366 \cdot 10^{-1}$	$3.374 \cdot 10^{-1}$
1/2	32	296	$1.205 \cdot 10^{-1}$	$1.508 \cdot 10^{-1}$	$1.512 \cdot 10^{-1}$
1/4	128	1168	$4.817 \cdot 10^{-2}$	$6.130 \cdot 10^{-2}$	$6.147 \cdot 10^{-2}$
1/8	512	4640	$1.917 \cdot 10^{-2}$	$2.456 \cdot 10^{-2}$	$2.463 \cdot 10^{-2}$
1/16	2048	18496	$7.736 \cdot 10^{-2}$	$1.003 \cdot 10^{-2}$	$1.005 \cdot 10^{-2}$
1/32	8192	73856	$3.160 \cdot 10^{-3}$	$4.174 \cdot 10^{-3}$	$4.187 \cdot 10^{-3}$
1/64	32768	295168	$1.303 \cdot 10^{-3}$	$1.766 \cdot 10^{-3}$	$1.772 \cdot 10^{-3}$

and with $\mathbf{f} = (0, -1)$, i.e., a unit volume force pointing downward. The Lamé parameter μ is always 1 in this example. We compute the least-squares finite element approximation for a sequence of triangulations resulting from uniform refinement. The displacement field for $\lambda = 1000$ is shown in Figure 6.1 (for $h = 1/4$ on the left and for $h = 1/16$ on the right).

Table 6.1 shows the least-squares functional for different mesh sizes h and different values of the Lamé parameters λ. Obviously, the convergence is uniform as $\lambda \to \infty$, as indicated by the theory. Also shown is the number of triangles and the total number of degrees of freedom (for displacement and stress) in the system.

More numerical results including more sophisticated test examples will be presented in a companion paper [12], which focusses on adaptive refinement strategies.

Acknowledgment. We thank Travis Austin for helpful discussions.

REFERENCES

[1] M. Amara and J. M. Thomas, *Equilibrium finite elements for the linear elasticity problem*, Numer. Math., 33 (1979), pp. 367–383.

[2] D. N. Arnold, F. Brezzi, and J. Douglas, *PEERS: A new mixed finite element for plane elasticity*, Japan J. Appl. Math., 1 (1984), pp. 347–367.

[3] D. N. Arnold, J. Douglas, and C. P. Gupta, *A family of higher order mixed finite element methods for plane elasticity*, Numer. Math., 45 (1984), pp. 1–22.

[4] D. N. Arnold and R. S. Falk, *A new mixed formulation for elasticity*, Numer. Math., 53 (1988), pp. 13–30.

[5] D. N. Arnold and R. Winther, *Mixed finite elements for elasticity*, Numer. Math., 92 (2002), pp. 401–419.

[6] K. E. Atkinson and W. Han, *Theoretical Numerical Analysis*, Springer, New York, 2001.

[7] P. B. Bochev and M. D. Gunzburger, *Finite element methods of least-squares type*, SIAM Rev., 40 (1998), pp. 789–837.

[8] D. Braess, *Finite Elements*, Cambridge University Press, Cambridge, UK, 1997.

[9] J. H. Bramble, R. D. Lazarov, and J. E. Pasciak, *A least-squares approach based on a discrete minus one inner product for first order systems*, Math. Comp., 66 (1997), pp. 935–955.

[10] S. C. Brenner and L. R. Scott, *The Mathematical Theory of Finite Element Methods*, Springer, New York, 1994.

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

[12] Z. Cai, J. Korsawe, and G. Starke, *Adaptive least squares mixed finite element computations for the stress-displacement formulation of linear elasticity*, SIAM J. Sci. Comput., (2002), submitted.

[13] Z. Cai, T. A. Manteuffel, S. F. McCormick, and S. V. Parter, *First-order system least squares (FOSLS) for planar linear elasticity: Pure traction problem*, SIAM J. Numer. Anal., 35 (1998), pp. 320–335.

[14] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North–Holland, Amsterdam, 1978.

[15] G. Duvaut and J. L. Lions, *Inequalities in Mechanics and Physics*, Springer, New York, 1976.

[16] R. S. Falk, *Nonconforming finite element methods for the equations of linear elasticity*, Math. Comp., 57 (1991), pp. 529–550.

[17] M. Fortin and M. Soulie, *A non-conforming piecewise quadratic finite element on triangles*, Internat. J. Numer. Methods Engrg., 19 (1983), pp. 505–520.

[18] B. Jiang, *The Least-Squares Finite Element Method*, Springer, Berlin, 1998.

[19] B. Jiang and J. Wu, *The least-squares finite element method in elasticity—Part* I: *Plane stress or strain with drilling degrees of freedom*, Internat. J. Numer. Methods Engrg., 53 (2002), pp. 621–636.

[20] R. Stenberg, *A family of mixed finite elements for the elasticity problem*, Numer. Math., 53 (1988), pp. 513–538.

SIAM J. Numer. Anal.
Vol. 41, No. 2, pp. 731–750

CONDITIONING OF HIERARCHIC p-VERSION NÉDÉLEC ELEMENTS ON MESHES OF CURVILINEAR QUADRILATERALS AND HEXAHEDRA[*]

MARK AINSWORTH[†] AND JOE COYLE[†]

Abstract. The conditioning of a set of hierarchic basis functions for p-version edge element approximation of the space $H(\mathbf{curl})$ is studied. Theoretical bounds are obtained on the location of the eigenvalues and on the growth of the condition numbers for the mass, curl-curl, and stiffness matrices that naturally arise from Galerkin approximation of Maxwell's equations. The theory is applicable to meshes of curvilinear quadrilaterals or hexahedra in two and three dimensions, respectively, including the case in which the local order of approximation is nonuniform. Throughout, the theory is illustrated with numerical examples that show that the theoretical asymptotic bounds are sharp and are attained within the range of practical computation.

Key words. eigenvalue bounds, hierarchic basis, edge finite elements, Maxwell equations

AMS subject classifications. 78-08, 65N30

PII. S003614290239590X

1. Introduction.

The p-version of the finite element method is an established tool for the numerical approximation of problems arising in mechanics. The use of higher order *edge* finite elements for problems in electromagnetic applications such as approximation of Maxwell's equations, though less well established, has witnessed a steadily increasing interest since the early works of Nédélec [16] and Bossavit [3, 4, 5]. Let $H(\mathbf{curl};\Omega)$ denote the space

$$H(\mathbf{curl};\Omega) = \left\{ \boldsymbol{u} \in \boldsymbol{L}^2(\Omega) : \mathbf{curl}\, \boldsymbol{u} \in \boldsymbol{L}^2(\Omega) \right\},$$

where Ω is a bounded, curvilinear polyhedral domain in $\mathbb{R}^d$, $d = 2, 3$. The space $H(\mathbf{curl};\Omega)$ arises naturally in many physical models, such as Maxwell's equations. While the traces of the tangential components of functions belonging to $H(\mathbf{curl};\Omega)$ are continuous across any interface, the normal components may be discontinuous. Consequently, the space $H(\mathbf{curl};\Omega)$ is a proper subspace of $\boldsymbol{H}^1(\Omega)$. It is vital that any Galerkin approximation should be based on a finite dimensional subspace that matches these continuity properties. For instance, a standard conforming finite element approximation of the space $\boldsymbol{H}^1(\Omega)$ is known to lead to spurious solutions in quite commonly occurring situations [7]. A finite dimensional subspace of $H(\mathbf{curl};\Omega)$ suitable for the Galerkin approximation of Maxwell's equations may be constructed using the p-version of the finite element method based on Nédélec, or edge, finite elements [16].

Higher order edge elements are employed in the engineering literature [19, 20], where the degree of element is typically fixed in the range $p = 2, 3, 4$ and convergence is sought through mesh refinement. However, the p-version means, at least in principle, that the mesh is fixed and that the degree of approximation p tends to infinity. The mathematical analysis of the p-version edge finite elements was considered by

[*] Received by the editors March 6, 2002; accepted for publication (in revised form) September 27, 2002; published electronically May 6, 2003. This research was supported by the Engineering and Physical Sciences Research Council of Great Britain, grant GR/M59426.

http://www.siam.org/journals/sinum/41-2/39590.html

[†] Department of Mathematics, Strathclyde University, 26 Richmond St., Glasgow G1 1XH, Scotland (M.Ainsworth@strath.ac.uk, ra.jcoy@maths.strath.ac.uk).

Monk [14, 15], while significant contributions to the practical implementation of p- and hp-version edge finite elements were made by Demkowicz and coworkers [8, 18].

The improved accuracy and rate of convergence obtained using p-version procedures comes at a price. An efficient practical implementation of higher order methods requires the use of hierarchic basis functions. Generally, it is found that the conditioning of the matrices that arise from discretization using higher order elements degenerates quite rapidly in comparison with those of h-version procedures. The conditioning of the stiffness and mass matrices on a tensor product reference element using a particular hierarchic basis for H^1 conforming p-version approximation was studied by Maitre and Pourquier [13]. Subsequently, Olsen and Douglas [17] studied the conditioning of hierarchic bases on tensor product elements in general and conjectured that, regardless of the choice of basis, the condition numbers grow as $\mathcal{O}(p^{4d})$ or faster in d space dimensions. Hu, Guo, and Katz [11] disproved this conjecture by exhibiting a hierarchical basis where the condition number of the stiffness matrix grows as $\mathcal{O}(p^{4(d-1)})$.

Comparatively little is known concerning the conditioning of the matrices arising from p-version edge element approximation of Maxwell's equations. The hierarchic basis presented by Rachowicz and Demkowicz was studied numerically in [1], where it was observed that the condition number degrades exponentially fast with increasing polynomial order, even in two spatial dimensions. An alternative hierarchic basis presented in [1] was observed numerically to have superior conditioning properties. Nevertheless, there has been no theoretical analysis of the conditioning of hierarchic p-version edge elements.

The aim of the present work is to address this problem directly and to study the conditioning theoretically and establish bounds on the growth explicitly in terms of the polynomial degree. The situation for Maxwell's equations is rather different from the cases considered in the works mentioned above. For instance, the underlying space is $H(\mathbf{curl})$ rather than H^1, meaning that the basis functions are vectorial in nature and possess only continuity of traces in the tangential components. One implication of this is that the basis functions on the physical elements are constructed using a covariant transformation from the reference element, rather than a standard pull-back construction employed in the H^1-conforming situation. Furthermore, the operators involved in the Maxwell equations are different and naturally lead [12] to the *mass*

$$M(\boldsymbol{E},\boldsymbol{F}) = \int_\Omega \varepsilon \boldsymbol{E}\cdot\boldsymbol{F}\,\mathrm{d}V \tag{1}$$

and *curl-curl*

$$S(\boldsymbol{E},\boldsymbol{F}) = \int_\Omega \mu^{-1}\,\mathbf{curl}\,\boldsymbol{E}\cdot\mathbf{curl}\,\boldsymbol{F}\,\mathrm{d}V \tag{2}$$

bilinear forms. Here, the permeability μ and permittivity ε are real, scalar-valued functions that are assumed to be bounded above and below in Ω; i.e., there exist positive constants c_1 and C_1 such that

$$c_1 \le \mu(\boldsymbol{x}), \varepsilon(\boldsymbol{x}) \le C_1 \qquad \text{for all } \boldsymbol{x}\in\Omega. \tag{3}$$

Moreover, for transient simulations, a difference approximation of the time derivatives would typically lead to the need to invert a *stiffness* matrix given by

$$\boldsymbol{A} = \boldsymbol{S} + \omega^2\boldsymbol{M}, \tag{4}$$

where ω would be inversely proportional to the time step Δt. The exponential rate of convergence of the p-version spatial discretization means that the order p will generally be modest in comparison with the choice of the time step, i.e., $\Delta t^{-1} \gg p$.

We derive bounds on the asymptotic behavior of the eigenvalues and on the growth of the condition numbers for each of the above matrices for the family of hierarchic basis functions presented in [1]. The theory is applicable to meshes of curvilinear quadrilaterals or hexahedra in two and three dimensions, respectively, and allows for nonuniform local order of approximation. By analogy with the results obtained by Hu, Guo, and Katz [11], it is shown that the condition number of the mass and curl-curl matrices grows as $\mathcal{O}(p^{4(d-1)})$ and $\mathcal{O}(p^{2(d-1)})$, respectively, in d spatial dimensions. The diagonally scaled stiffness matrix has a condition number that grows as

$$C \max\left(1, \frac{p}{\omega}\right)^2 p^{2(d-1)}$$

so that, in the typical case where $\omega \propto \Delta t^{-1} \gg p$, the condition number grows as $\mathcal{O}(p^{2(d-1)})$. Throughout, the theory is illustrated with numerical examples that show that the theoretical asymptotic bounds are sharp and are attained within the range of practical computation.

2. Statement of the results. Let $\mathcal{M}$ be a partitioning of Ω into curvilinear quadrilaterals or hexahedra [6] such that the nonempty intersection of distinct elements is either a single common face, edge, or vertex of both elements. Each element $K \in \mathcal{M}$ is the image of a reference element $\widehat{K} = (-1,1)^d$ under a differentiable bijection $\boldsymbol{F}_K : \widehat{K} \to K$. It is assumed that positive constants exist such that the Jacobian matrix $\boldsymbol{J}_K$ of the mapping satisfies

$$c_{2,K} \leq \det \boldsymbol{J}_K(\boldsymbol{\xi}) \leq C_{2,K} \tag{5}$$

and

$$\sigma\left(\boldsymbol{J}_K^{-1}\boldsymbol{J}_K^{-\top}\right) \subset [c_{3,K}, C_{3,K}] \tag{6}$$

for all $\boldsymbol{\xi} \in \widehat{K}$, where $\sigma(\boldsymbol{A})$ denotes the spectrum [10] of the matrix $\boldsymbol{A}$.

A finite element in the sense of Ciarlet [6] is represented by a triple $(\mathcal{P}, K, \Sigma)$. The space $\widehat{\mathcal{P}}$ associated with the Nédélec element of order p on the reference element is given by

$$\widehat{\mathcal{P}} = \begin{cases} \mathbb{Q}_{p,p+1} \times \mathbb{Q}_{p+1,p}, & d = 2, \\ \mathbb{Q}_{p,p+1,p+1} \times \mathbb{Q}_{p+1,p,p+1} \times \mathbb{Q}_{p+1,p+1,p}, & d = 3, \end{cases}$$

where

$$\mathbb{Q}_{p,q} = \left\{x^i y^j : 0 \leq i \leq p, 0 \leq j \leq q\right\}$$

and

$$\mathbb{Q}_{p,q,r} = \left\{x^i y^j z^k : 0 \leq i \leq p, 0 \leq j \leq q, 0 \leq k \leq r\right\}.$$

The set of degrees of freedom $\widehat{\Sigma}$ is specified implicitly by the choice of basis. Let $\{L_i\}_{i=0}^{p}$ denote normalized Legendre polynomials, so that $\|L_i\|_{(-1,1)} = 1$, and define the set $\{\ell_i\}_{i=0}^{p+1}$ as follows:

$$\ell_0(s) = \frac{1}{2}(1-s), \quad \ell_1(s) = \frac{1}{2}(1+s)$$

and

$$\ell_i(s) = \int_{-1}^{s} L_{i-1}(t)\,\mathrm{d}t, \quad i = 2, \dots, p+1.$$

The basis functions on the quadrilateral reference element $\widehat{K} = (-1,1)^2$ are chosen to be

$$\left.\begin{aligned} &L_i(\xi_1)\ell_j(\xi_2)\boldsymbol{e}_1 \\ &\ell_j(\xi_1)L_i(\xi_2)\boldsymbol{e}_2 \end{aligned}\right\} \quad i = 0, \dots, p,\ j = 0, \dots, p+1, \tag{7}$$

while for the hexahedral reference element $\widehat{K} = (-1,1)^3$ the basis functions are given by

$$\left.\begin{aligned} &L_i(\xi_1)\ell_j(\xi_2)\ell_k(\xi_3)\boldsymbol{e}_1 \\ &\ell_j(\xi_1)L_i(\xi_2)\ell_k(\xi_3)\boldsymbol{e}_2 \\ &\ell_j(\xi_1)\ell_k(\xi_2)L_i(\xi_3)\boldsymbol{e}_3 \end{aligned}\right\} \quad i = 0, \dots, p,\ j, k = 0, \dots, p+1, \tag{8}$$

where $\boldsymbol{e}_1, \dots, \boldsymbol{e}_3$ denote the unit Cartesian vectors. The dimensions of $\widehat{\mathcal{P}}$ are given by $d(p+1)(p+2)^{d-1}$ for $d = 2, 3$.

The Nédélec element $(\mathcal{P}, K, \Sigma)$ on a physical domain K is constructed from the reference element as follows. First, observe that the electric field $\widehat{\boldsymbol{E}}$ on a reference element is related to the field $\boldsymbol{E}$ on the physical element by the *covariant* transformation [12],

$$\boldsymbol{E}(\boldsymbol{x})|_K = \boldsymbol{J}_K^{-\top}\widehat{\boldsymbol{E}}(\boldsymbol{\xi}), \quad \boldsymbol{x} = \boldsymbol{F}_K(\boldsymbol{\xi}). \tag{9}$$

Consequently, the global basis function $\boldsymbol{\phi}$ corresponding to the local basis function $\widehat{\boldsymbol{\phi}}$ on the reference element is defined by

$$\boldsymbol{\phi}(\boldsymbol{x})|_K = \boldsymbol{J}_K^{-\top}\widehat{\boldsymbol{\phi}}(\boldsymbol{\xi}). \tag{10}$$

The degrees of freedom Σ on the global element are implicit in the choice of basis. The degrees of freedom shared by more than one element may be shown to correspond to tangential moments of the field on the edges and faces of the element similar to the degrees of freedom employed by Nédélec [16] in the case of uniform order approximation. However, some care must be exercised if the use of a nonuniform polynomial order of approximation p_K on each element K in the partition $\mathcal{M}$ is permitted. The continuity properties of the space $H(\mathbf{curl};\Omega)$ mean that the order of approximation must be appropriately restrained at common interfaces and edges between neighboring elements. This is accomplished by applying the *minimum rule*, whereby the order of approximation on the element of higher local order is reduced to match that of the neighboring elements. We refer to [2] for a discussion of the minimum rule in the context of $\boldsymbol{H}^1(\Omega)$ conforming approximation, and to [8] for the $H(\mathbf{curl};\Omega)$ situation. The polynomial orders are collected into a degree vector $\boldsymbol{p} = \{p_K : K \in \mathcal{P}\}$, with the maximum order of approximation denoted by $p_{\max}$.

One feature of the basis presented above that is important for efficient practical implementation is that it is *hierarchical.* An alternative hierarchical basis will be found in [18]. However, numerical evidence presented in [1] indicates that the latter choice leads to extremely poorly conditioned matrices. In fact, numerical evidence suggests that the condition number degenerates exponentially fast with the polynomial order p. The conditioning of the basis described above is the subject of Theorem 1.

The spectral condition number $\kappa(\boldsymbol{A})$ of a square matrix $\boldsymbol{A}$ is defined by

$$\kappa(\boldsymbol{A}) = \frac{\lambda_{\max}(\boldsymbol{A})}{\lambda_{\min}(\boldsymbol{A})}. \tag{11}$$

Many numerical algorithms work, either explicitly or implicitly, with the diagonally scaled matrix $\widetilde{\boldsymbol{A}}$ defined by

$$\widetilde{\boldsymbol{A}} = \boldsymbol{D}^{-1/2}\boldsymbol{A}\boldsymbol{D}^{-1/2},$$

where $\boldsymbol{D}$ denotes the diagonal of $\boldsymbol{A}$. The next result presents bounds on the growth of the spectral condition number of various matrices that arise in the Galerkin approximation of Maxwell's equations. The mass matrix $\boldsymbol{M}$ is real, symmetric, and positive definite and therefore has positive eigenvalues. However, the curl-curl matrix $\boldsymbol{S}$ is semidefinite, and in this case, we bound the ratio $\kappa'(\boldsymbol{S})$ of the largest eigenvalue to the smallest nonzero eigenvalue.

THEOREM 1. *Let $\mathcal{M}$ be a partition with variable order of approximation given by $\boldsymbol{p}$, and with maximum order $p_{\max}$. Then, there exist positive constants $C^*_{\mathcal{M},\boldsymbol{M}}, C^*_{\mathcal{M},\boldsymbol{S}}$, and $C^*_{\mathcal{M},\boldsymbol{A}}$, depending on the constants defined in* (5)–(6) *but independent of the polynomial order, such that*

(1) *for the global mass matrix,*

$$\left.\begin{aligned} \kappa\big(\boldsymbol{M}^{(d)}_{\mathcal{M},\boldsymbol{p}}\big) &\le C^*_{\mathcal{M},\boldsymbol{M}} p_{\max}^{4(d-1)} \\ \kappa\big(\widetilde{\boldsymbol{M}}^{(d)}_{\mathcal{M},\boldsymbol{p}}\big) &\le C^*_{\mathcal{M},\boldsymbol{M}} p_{\max}^{2(d-1)} \end{aligned}\right\} \quad d = 2, 3; \tag{12}$$

(2) *for the global curl-curl matrix,*

$$\left.\begin{aligned} \kappa'\big(\boldsymbol{S}^{(d)}_{\mathcal{M},\boldsymbol{p}}\big) &\le C^*_{\mathcal{M},\boldsymbol{S}} p_{\max}^{4(d-2)} \\ \kappa'\big(\widetilde{\boldsymbol{S}}^{(d)}_{\mathcal{M},\boldsymbol{p}}\big) &\le C^*_{\mathcal{M},\boldsymbol{S}} p_{\max}^{2(d-2)} \end{aligned}\right\} \quad d = 2, 3; \tag{13}$$

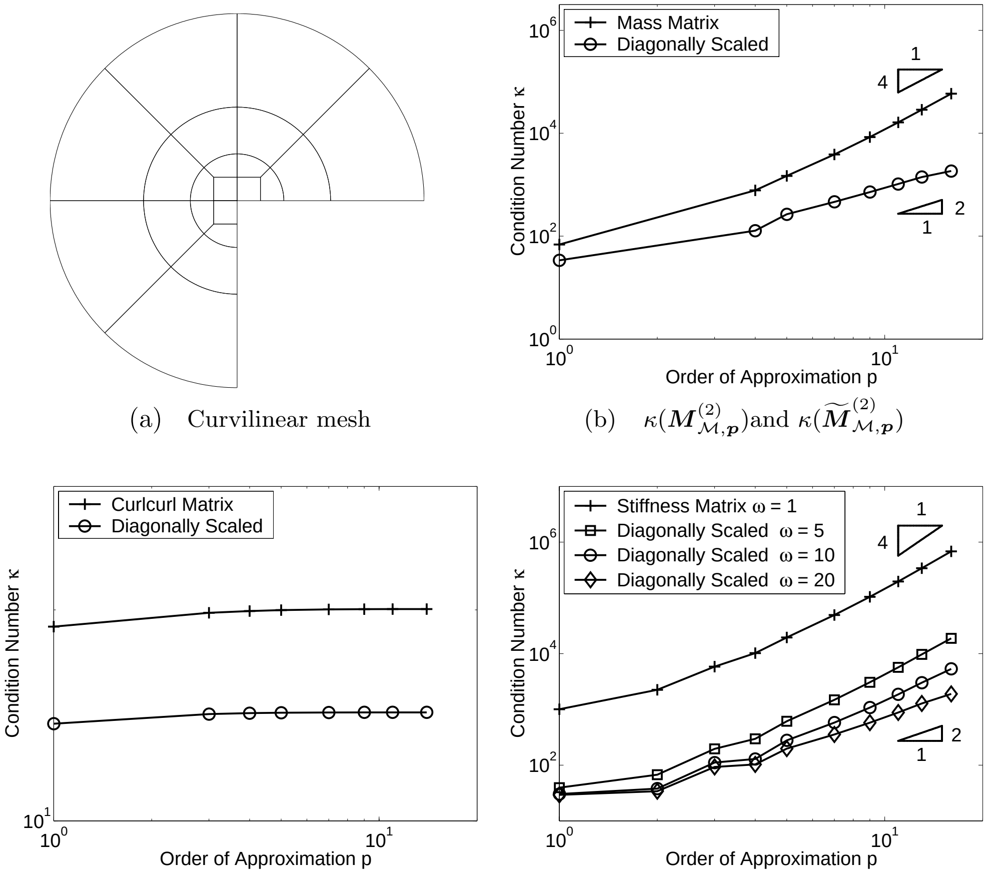
(3) *and for the global stiffness matrix,*

$$\left.\begin{aligned} \kappa\big(\boldsymbol{A}^{(d)}_{\mathcal{M},\boldsymbol{p}}\big) &\le C^*_{\mathcal{M},\boldsymbol{A}} p_{\max}^{4(d-1)} \\ \kappa\big(\widetilde{\boldsymbol{A}}^{(d)}_{\mathcal{M},\boldsymbol{p}}\big) &\le C^*_{\mathcal{M},\boldsymbol{A}} \max\left(1, \frac{p_{\max}}{\omega}\right)^2 p_{\max}^{2(d-1)} \end{aligned}\right\} \quad d = 2, 3. \tag{14}$$

The proof of this result is deferred until the next section. First, we present a simple example to illustrate the results in the case of uniform polynomial degree approximation, i.e., $p_K = p$ for all $K \in \mathcal{M}$, using the mesh shown in Figure 1(a). This mesh is typical of the type of geometrically graded meshes needed to achieve exponential rates of convergence. The permittivity and permeability are chosen to be unity throughout the domain. The numerical results agree with theoretical predictions. In particular, the theorem predicts a transition from growth of order $\mathcal{O}(p^4)$ to $\mathcal{O}(p^2)$ in the condition number of the diagonally scaled stiffness matrix $\widetilde{\boldsymbol{A}}^{(d)}_{\mathcal{M},\boldsymbol{p}}$ as the value of ω is increased. This behavior is also observed in practice, as seen in Figure 1(d).

3. Proofs of the results.

This section is organized as follows. First, bounds are established for the eigenvalues and condition numbers of the mass, curl-curl, and stiffness matrices on a single reference element. Theorem 1 is then deduced from these results.



(a) Curvilinear mesh

(b) $\kappa(\boldsymbol{M}^{(2)}_{\mathcal{M},\boldsymbol{p}})$and $\kappa(\widetilde{\boldsymbol{M}}^{(2)}_{\mathcal{M},\boldsymbol{p}})$

(c) $\kappa(\boldsymbol{S}^{(2)}_{\mathcal{M},\boldsymbol{p}})$ and $\kappa(\widetilde{\boldsymbol{S}}^{(2)}_{\mathcal{M},\boldsymbol{p}})$

(d) $\kappa(\boldsymbol{A}^{(2)}_{\mathcal{M},\boldsymbol{p}})$ and $\kappa(\widetilde{\boldsymbol{A}}^{(2)}_{\mathcal{M},\boldsymbol{p}})$

FIG. 1. *Mesh used to illustrate the results in* (12)–(14) *of Theorem* 1 *in two dimensions. The variation of the condition numbers versus the order of approximation using the mesh in* (a) *for uniform polynomial order $p = 0, \dots, 15$ is given for* (b) *the global mass and diagonally scaled global mass matrices,* (c) *the global curl-curl and diagonally scaled global curl-curl matrices, and* (d) *the global stiffness and diagonally scaled global stiffness matrices. The theoretical rates predicted in Theorem* 1 *are also indicated.*

3.1. Analysis of the mass matrix. The first result concerns the mass matrix.

LEMMA 2. *Let $\boldsymbol{M}^{(d)}_{\widehat{K},p}$ and $\widetilde{\boldsymbol{M}}^{(d)}_{\widehat{K},p}$ denote the mass and diagonally scaled mass matrices in d spatial dimensions with order p approximation. Then there exist positive constants $c_{\widehat{K},\boldsymbol{M}}$ and $C_{\widehat{K},\boldsymbol{M}}$ independent of p such that*

$$\lambda_{\min}\big(\boldsymbol{M}^{(d)}_{\widehat{K},p}\big) \geq c_{\widehat{K},\boldsymbol{M}}p^{-4(d-1)}, \quad \lambda_{\max}\big(\boldsymbol{M}^{(d)}_{\widehat{K},p}\big) \leq C_{\widehat{K},\boldsymbol{M}} \tag{15}$$

and

$$\lambda_{\min}\big(\widetilde{\boldsymbol{M}}^{(d)}_{\widehat{K},p}\big) \geq c_{\widehat{K},\boldsymbol{M}}p^{-2(d-1)}, \quad \lambda_{\max}\big(\widetilde{\boldsymbol{M}}^{(d)}_{\widehat{K},p}\big) \leq C_{\widehat{K},\boldsymbol{M}}. \tag{16}$$

Therefore, there exists a positive constant $C^*_{\widehat{K},\boldsymbol{M}}$ independent of p such that

$$\kappa\big(\boldsymbol{M}^{(d)}_{\widehat{K},p}\big) \leq C^*_{\widehat{K},\boldsymbol{M}}p^{4(d-1)} \tag{17}$$

and

$$\kappa\big(\widetilde{\boldsymbol{M}}^{(d)}_{\widehat{K},p}\big) \le C^{*}_{\widehat{K},\boldsymbol{M}} p^{2(d-1)}. \tag{18}$$

Proof. The bounds (17) and (18) on the condition numbers are immediate consequences of (15) and (16), thanks to (11) and setting $C^{*}_{\widehat{K},\boldsymbol{M}} = C_{\widehat{K},\boldsymbol{M}}/c_{\widehat{K},\boldsymbol{M}}$. It remains to prove (15) and (16). The basis functions in (7) and (8) are multiples of the Cartesian vectors. As a result, the mass matrix can be written in block diagonal form:

$$\boldsymbol{M}^{(2)}_{\widehat{K},p} = \operatorname{blockdiag}\big(\boldsymbol{M}^{(2)}_{1,p}, \boldsymbol{M}^{(2)}_{2,p}\big) \tag{19}$$

and

$$\boldsymbol{M}^{(3)}_{\widehat{K},p} = \operatorname{blockdiag}\big(\boldsymbol{M}^{(3)}_{1,p}, \boldsymbol{M}^{(3)}_{2,p}, \boldsymbol{M}^{(3)}_{3,p}\big), \tag{20}$$

where $\boldsymbol{M}^{(d)}_{i,p}$ represents the coupling between the basis functions that are multiples of the ith Cartesian vectors. In the case of $d = 2$, order the basis functions as follows:

$$\left.\begin{aligned} \boldsymbol{\psi}^1_{j+1+i(p+2)} &= L_i(\xi_1)\ell_j(\xi_2)\boldsymbol{e}_1 \\ \boldsymbol{\psi}^2_{i+1+j(p+1)} &= \ell_j(\xi_1)L_i(\xi_2)\boldsymbol{e}_2 \end{aligned}\right\} \quad i = 0,\dots,p,\ j = 0,\dots,p+1. \tag{21}$$

That is to say, we order the functions by looping over the index corresponding to the ξ_2 variable first. The block matrices in (19) can then be rewritten by observing that

$$\big(\boldsymbol{M}^{(2)}_{1,p}\big)_{j+1+i(p+2),n+1+m(p+2)} = \delta_{im} \int_{-1}^{1} \ell_j(\xi_2)\ell_n(\xi_2)\,\mathrm{d}\xi_2$$

and

$$\big(\boldsymbol{M}^{(2)}_{2,p}\big)_{i+1+j(p+1),m+1+n(p+1)} = \delta_{im} \int_{-1}^{1} \ell_j(\xi_1)\ell_n(\xi_1)\,\mathrm{d}\xi_1$$

for $i, m = 0,\dots,p$ and $j, n = 0,\dots,p+1$, where δ_{im} is the Kronecker symbol. Hence,

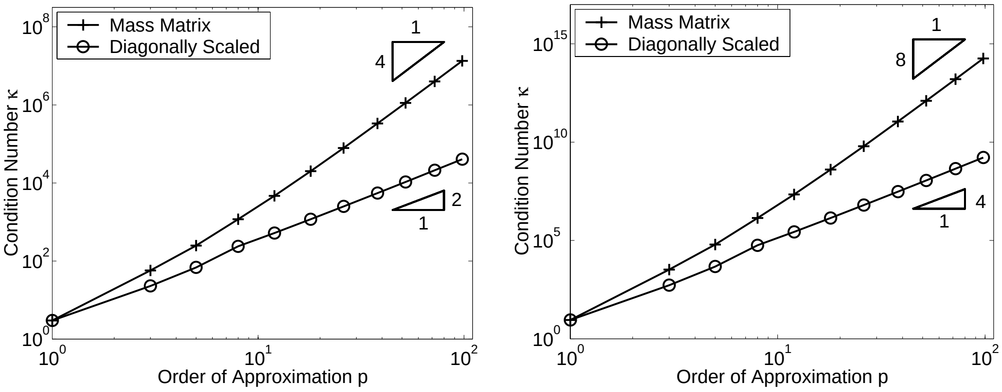
$$\boldsymbol{M}^{(2)}_{1,p} = \boldsymbol{I}_{p+1} \otimes \boldsymbol{\ell}_{p+1} \quad \text{and} \quad \boldsymbol{M}^{(2)}_{2,p} = \boldsymbol{\ell}_{p+1} \otimes \boldsymbol{I}_{p+1},$$

where $\otimes$ denotes the Kronecker product [10], $\boldsymbol{I}_{p+1}$ is the $p+1$ by $p+1$ identity matrix, and $\boldsymbol{\ell}_{p+1}$ is the mass matrix in one dimension with entries given by

$$(\boldsymbol{\ell}_{p+1})_{ij} = \int_{-1}^{1} \ell_i(s)\ell_j(s)\,\mathrm{d}s, \quad i, j = 0,\dots,p+1.$$

Likewise, for $d = 3$, we order the basis functions by first looping over the indices corresponding to the ξ_3 variable and then over the indices corresponding to the ξ_2 variable. The block matrices in (20) take the form

$$\begin{aligned} \boldsymbol{M}^{(3)}_{1,p} &= \boldsymbol{I}_{p+1} \otimes \boldsymbol{\ell}_{p+1} \otimes \boldsymbol{\ell}_{p+1}, \\ \boldsymbol{M}^{(3)}_{2,p} &= \boldsymbol{\ell}_{p+1} \otimes \boldsymbol{I}_{p+1} \otimes \boldsymbol{\ell}_{p+1}, \\ \boldsymbol{M}^{(3)}_{3,p} &= \boldsymbol{\ell}_{p+1} \otimes \boldsymbol{\ell}_{p+1} \otimes \boldsymbol{I}_{p+1}. \end{aligned}$$



(a) $\kappa(\boldsymbol{M}^{(2)}_{\widehat{K},p})$ and $\kappa(\widetilde{\boldsymbol{M}}^{(2)}_{\widehat{K},p})$ (b) $\kappa(\boldsymbol{M}^{(3)}_{\widehat{K},p})$ and $\kappa(\widetilde{\boldsymbol{M}}^{(3)}_{\widehat{K},p})$

FIG. 2. *Variation of the condition numbers versus the order of approximation of the mass and diagonally scaled mass matrices for* (a) *two and* (b) *three dimensions on the reference element. These results agree with those predicted by* (17) *and* (18) *for* $d = 2$ *and* $d = 3$*, respectively.*

By Theorem 4.2.12 of [10], the spectrum of each block matrix is given by

$$\sigma\big(\boldsymbol{M}^{(d)}_{i,p}\big) = \sigma\left(\boldsymbol{I}_{p+1}\right) \otimes \overbrace{\sigma(\boldsymbol{\ell}_{p+1}) \otimes \cdots \otimes \sigma(\boldsymbol{\ell}_{p+1})}^{d-1},$$

and since $\sigma(\boldsymbol{I}_{p+1}) = \{1\}$, we obtain

$$\lambda_{\min}\big(\boldsymbol{M}^{(d)}_{\widehat{K},p}\big) = \lambda_{\min}\big(\boldsymbol{\ell}_{p+1}\big)^{d-1} \quad \text{and} \quad \lambda_{\max}\big(\boldsymbol{M}^{(d)}_{\widehat{K},p}\big) = \lambda_{\max}\left(\boldsymbol{\ell}_{p+1}\right)^{d-1}. \tag{22}$$

The bounds (15) then follow by recalling (see [13]) that there exist positive constants c and C independent of p such that

$$\lambda_{\min}\left(\boldsymbol{\ell}_{p+1}\right) \geq cp^{-4} \quad \text{and} \quad \lambda_{\max}\left(\boldsymbol{\ell}_{p+1}\right) \leq C. \tag{23}$$

Similar arguments show that for the diagonally scaled mass matrix,

$$\widetilde{\boldsymbol{M}}^{(d)}_{\widehat{K},p} = \operatorname{diag}\big(\widetilde{\boldsymbol{M}}^{(d)}_{1,p}, \ldots, \widetilde{\boldsymbol{M}}^{(d)}_{d,p}\big), \quad d = 2, 3,$$

and hence,

$$\lambda_{\min}\big(\widetilde{\boldsymbol{M}}^{(d)}_{\widehat{K},p}\big) = \lambda_{\min}\big(\widetilde{\boldsymbol{\ell}}_{p+1}\big)^{d-1} \quad \text{and} \quad \lambda_{\max}\big(\widetilde{\boldsymbol{M}}^{(d)}_{\widehat{K},p}\big) = \lambda_{\max}\big(\widetilde{\boldsymbol{\ell}}_{p+1}\big)^{d-1}.$$

The bounds (16) follow by recalling (see [13]) that

$$\lambda_{\min}\big(\widetilde{\boldsymbol{\ell}}_{p+1}\big) \geq cp^{-2} \quad \text{and} \quad \lambda_{\max}\big(\widetilde{\boldsymbol{\ell}}_{p+1}\big) \leq C \tag{24}$$

for the diagonally scaled one dimensional mass matrix. The result follows as claimed by choosing $c_{\widehat{K},\boldsymbol{M}} = \min\left(c, c^2\right)$ and $C_{\widehat{K},\boldsymbol{M}} = \max\left(C, C^2\right)$. □

In Figure 2 the variation of the condition numbers of $\boldsymbol{M}^{(d)}_{\widehat{K},p}$ and $\widetilde{\boldsymbol{M}}^{(d)}_{\widehat{K},p}$ versus order of approximation for $p = 0, \ldots, 100$ is shown for $d = 2$ in (a) and $d = 3$ in (b). It will be observed that the estimates in Lemma 2 are sharp.

3.2. Analysis of the curl-curl matrix. The next result gives bounds for the nontrivial eigenvalues of the curl-curl matrix on a single reference element.

LEMMA 3. *Let $\boldsymbol{S}^{(d)}_{\widehat{K},p}$ and $\widetilde{\boldsymbol{S}}^{(d)}_{\widehat{K},p}$ denote the curl-curl and diagonally scaled curl-curl matrices in d spatial dimensions with order p approximation. Then there exist positive constants $c_{\widehat{K},\boldsymbol{S}}$ and $C_{\widehat{K},\boldsymbol{S}}$ independent of p such that*

$$\lambda'_{\min}\big(\boldsymbol{S}^{(d)}_{\widehat{K},p}\big) \geq c_{\widehat{K},\boldsymbol{S}}\, p^{-4(d-2)}, \quad \lambda_{\max}\big(\boldsymbol{S}^{(d)}_{\widehat{K},p}\big) \leq C_{\widehat{K},\boldsymbol{S}} \tag{25}$$

and

$$\lambda'_{\min}\big(\widetilde{\boldsymbol{S}}^{(d)}_{\widehat{K},p}\big) \geq c_{\widehat{K},\boldsymbol{S}}\, p^{-2(d-2)}, \quad \lambda_{\max}\big(\widetilde{\boldsymbol{S}}^{(d)}_{\widehat{K},p}\big) \leq C_{\widehat{K},\boldsymbol{S}}, \tag{26}$$

where $\lambda'_{\min}$ denotes the smallest nonzero eigenvalue. Therefore, there exists a positive constant $C^{*}_{\widehat{K},\boldsymbol{S}}$ independent of p such that

$$\kappa'\big(\boldsymbol{S}^{(d)}_{\widehat{K},p}\big) \leq C^{*}_{\widehat{K},\boldsymbol{S}}\, p^{4(d-2)} \tag{27}$$

and

$$\kappa'\big(\widetilde{\boldsymbol{S}}^{(d)}_{\widehat{K},p}\big) \leq C^{*}_{\widehat{K},\boldsymbol{S}}\, p^{2(d-2)}, \tag{28}$$

where κ' denotes the ratio of $\lambda_{\max}$ to $\lambda'_{\min}$.

Proof. As before, it suffices to prove (25) and (26). We begin by considering the two dimensional case. The curl-curl matrix is less straightforward to analyze than the mass matrix. Nevertheless, the basis functions given in (7) may be partitioned into four sets that are mutually orthogonal with respect to the $H(\mathbf{curl})$ semi-inner product, as follows:

$$\begin{aligned}
S_1 &= \operatorname{span}\{L_0(\xi_1)\ell_j(\xi_2)\boldsymbol{e}_1 : j \in \{0,1\}\} \\
&\quad \oplus \operatorname{span}\{\ell_j(\xi_1)L_0(\xi_2)\boldsymbol{e}_2 : j \in \{0,1\}\}, \\
S_2 &= \operatorname{span}\{L_i(\xi_1)\ell_j(\xi_2)\boldsymbol{e}_1 : i \in \{1,\dots,p\}; j \in \{0,1\}\} \\
&\quad \oplus \operatorname{span}\{\ell_j(\xi_1)L_0(\xi_2)\boldsymbol{e}_2 : j \in \{2,\dots,p+1\}\}, \\
S_3 &= \operatorname{span}\{L_0(\xi_1)\ell_j(\xi_2)\boldsymbol{e}_1 : j \in \{2,\dots,p+1\}\} \\
&\quad \oplus \operatorname{span}\{\ell_j(\xi_1)L_i(\xi_2)\boldsymbol{e}_2 : i \in \{1,\dots,p\}; j \in \{0,1\}\}, \\
S_4 &= \operatorname{span}\{L_i(\xi_1)\ell_j(\xi_2)\boldsymbol{e}_1 : i \in \{1,\dots,p\}; j \in \{2,\dots,p+1\}\} \\
&\quad \oplus \operatorname{span}\{\ell_j(\xi_1)L_i(\xi_2)\boldsymbol{e}_2 : i \in \{1,\dots,p\}; j \in \{2,\dots,p+1\}\}.
\end{aligned}$$

Hence

$$\boldsymbol{S}^{(2)}_{\widehat{K},p} = \operatorname{blockdiag}(\boldsymbol{S}_1, \boldsymbol{S}_2, \boldsymbol{S}_3, \boldsymbol{S}_4),$$

where $\boldsymbol{S}_i$ is the curl-curl matrix corresponding to the nonzero coupling between basis functions in S_i. Simple computation reveals that

$$\begin{aligned}
\boldsymbol{S}_1 &= \frac{1}{2}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \otimes \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}, \\
\boldsymbol{S}_2 &= \frac{1}{2}\begin{bmatrix} 1 & -1 & \sqrt{2} \\ -1 & 1 & -\sqrt{2} \\ \sqrt{2} & -\sqrt{2} & 2 \end{bmatrix} \otimes \boldsymbol{I}_p, \\
\boldsymbol{S}_3 &= \boldsymbol{I}_p \otimes \frac{1}{2}\begin{bmatrix} 1 & -1 & \sqrt{2} \\ -1 & 1 & -\sqrt{2} \\ \sqrt{2} & -\sqrt{2} & 2 \end{bmatrix}, \\
\boldsymbol{S}_4 &= \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \otimes (\boldsymbol{I}_p \otimes \boldsymbol{I}_p).
\end{aligned}$$

The spectra of these matrices are found, again using Theorem 4.2.12 of [10], to be

$$\begin{aligned}
&\sigma\left(\boldsymbol{S}_1\right) = \{2, 0, 0, 0\}, \\
&\sigma\left(\boldsymbol{S}_2\right) = \{2, 0\} \qquad \text{(multiplicity of } p \text{ and } 2p\text{, respectively)}, \\
&\sigma\left(\boldsymbol{S}_3\right) = \sigma(\boldsymbol{S}_2), \\
&\sigma\left(\boldsymbol{S}_4\right) = \{2, 0\} \qquad \text{(multiplicity of } p^2 \text{ for each)}.
\end{aligned}$$

Hence, the nonzero eigenvalues satisfy

$$\lambda'_{\min}\big(\boldsymbol{S}^{(2)}_{\widehat{K},p}\big) = 2 \quad \text{and} \quad \lambda_{\max}\big(\boldsymbol{S}^{(2)}_{\widehat{K},p}\big) = 2. \tag{29}$$

The same arguments apply equally well to the case when diagonal scaling is applied. In particular, the matrices $\widetilde{\boldsymbol{S}}_2, \dots, \widetilde{\boldsymbol{S}}_4$ agree with $\boldsymbol{S}_2, \dots, \boldsymbol{S}_4$, while $\widetilde{\boldsymbol{S}}_1 = 2\boldsymbol{S}_1$. The eigenvalues of $\widetilde{\boldsymbol{S}}_1$ are therefore given by $\{4, 0, 0, 0\}$, so that

$$\lambda'_{\min}\big(\widetilde{\boldsymbol{S}}^{(2)}_{\widehat{K},p}\big) = 2 \quad \text{and} \quad \lambda_{\max}\big(\widetilde{\boldsymbol{S}}^{(2)}_{\widehat{K},p}\big) = 4. \tag{30}$$

Together, these results give (25) and (26) in the case $d = 2$. Before proceeding with the case $d = 3$, it is useful to note that maintaining the *same* ordering of the degrees of freedom as described in (21) yields the alternative form

$$\boldsymbol{S}^{(2)}_{\widehat{K},p} = \begin{bmatrix} \boldsymbol{I}_{p+1} \otimes \boldsymbol{\ell}'_{p+1} & -\widetilde{\boldsymbol{L}}_p \otimes \widetilde{\boldsymbol{L}}_p^{\top} \\ -\widetilde{\boldsymbol{L}}_p^{\top} \otimes \widetilde{\boldsymbol{L}}_p & \boldsymbol{\ell}'_{p+1} \otimes \boldsymbol{I}_{p+1} \end{bmatrix}, \tag{31}$$

where $\widetilde{\boldsymbol{L}_p}$ is the matrix with entries given by

$$(\widetilde{\boldsymbol{L}_p})_{ij} = \int_{-1}^{1} L_i(s)\ell_j(s)\,\mathrm{d}s, \quad i = 0, \dots, p,\ j = 0, \dots, p+1,$$

and $\boldsymbol{\ell}'_{p+1}$ is the one dimensional stiffness matrix

$$(\boldsymbol{\ell}'_{p+1})_{ij} = \int_{-1}^{1} \ell'_i(s)\ell'_j(s)\,\mathrm{d}s, \quad i, j = 0, \dots, p+1.$$

Adopting the ordering of the basis functions used in the proof of Lemma 2 for the three dimensional mass matrix, the three dimensional curl-curl matrix takes the form

$$\boldsymbol{S}^{(3)}_{\widehat{K},p} = \boldsymbol{S}_{\mathbf{1}} + \boldsymbol{S}_2 + \boldsymbol{S}_3, \tag{32}$$

where

$$\boldsymbol{S}_1 = \begin{bmatrix} \boldsymbol{I}_{p+1} \otimes \boldsymbol{\ell}'_{p+1} \otimes \boldsymbol{\ell}_{p+1} & -\widetilde{\boldsymbol{L}}_p \otimes \widetilde{\boldsymbol{L}}_p^{\top} \otimes \boldsymbol{\ell}_{p+1} & \mathbf{0} \\ -\widetilde{\boldsymbol{L}}_p^{\top} \otimes \widetilde{\boldsymbol{L}}_p \otimes \boldsymbol{\ell}_{p+1} & \boldsymbol{\ell}'_{p+1} \otimes \boldsymbol{I}_{p+1} \otimes \boldsymbol{\ell}_{p+1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix},$$

$$\boldsymbol{S}_2 = \begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\ell}_{p+1} \otimes \boldsymbol{I}_{p+1} \otimes \boldsymbol{\ell}'_{p+1} & -\boldsymbol{\ell}_{p+1} \otimes \widetilde{\boldsymbol{L}}_p \otimes \widetilde{\boldsymbol{L}}_p^{\top} \\ \mathbf{0} & -\boldsymbol{\ell}_{p+1} \otimes \widetilde{\boldsymbol{L}}_p^{\top} \otimes \widetilde{\boldsymbol{L}}_p & \boldsymbol{\ell}_{p+1} \otimes \boldsymbol{I}_{p+1} \otimes \boldsymbol{\ell}'_{p+1} \end{bmatrix},$$

and

$$\boldsymbol{S}_3 = \begin{bmatrix} \boldsymbol{I}_{p+1} \otimes \boldsymbol{\ell}_{p+1} \otimes \boldsymbol{\ell}'_{p+1} & \mathbf{0} & -\widetilde{\boldsymbol{L}}_p \otimes \boldsymbol{\ell}_{p+1} \otimes \widetilde{\boldsymbol{L}}_p^{\top} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \\ -\widetilde{\boldsymbol{L}}_p^{\top} \otimes \boldsymbol{l}_{p+1} \otimes \widetilde{\boldsymbol{L}}_p & \mathbf{0} & \boldsymbol{\ell}'_{p+1} \otimes \boldsymbol{\ell}_{p+1} \otimes \boldsymbol{I}_{p+1} \end{bmatrix}.$$

Observe, in particular, that $\boldsymbol{S}_1$ is related to the two dimensional curl-curl matrix $\boldsymbol{S}^{(2)}_{\widehat{K},p}$ defined in (31) by the rule

$$\boldsymbol{S}_1 = \boldsymbol{S}^{(2)}_{\widehat{K},p} \otimes \boldsymbol{\ell}_{p+1}. \tag{33}$$

By symmetry, the same expression holds for $\boldsymbol{S}_2$ and $\boldsymbol{S}_3$, and hence, by Theorem 4.2.12 of [10],

$$\sigma\left(\boldsymbol{S}_i\right) = \sigma\bigl(\boldsymbol{S}^{(2)}_{\widehat{K},p}\bigr) \otimes \sigma\left(\boldsymbol{\ell}_{p+1}\right), \quad i = 1, \dots, 3. \tag{34}$$

Let $\vec{\alpha} = (\vec{\alpha}_1, \vec{\alpha}_2, \vec{\alpha}_3) \in \mathbb{R}^{3(p+1)(p+2)^2}$ denote the values of the degrees of freedom in the approximation on the reference element. Decompose

$$\vec{\alpha} = \frac{1}{2}\bigl(\vec{\beta}_1 + \vec{\beta}_2 + \vec{\beta}_3\bigr), \tag{35}$$

where

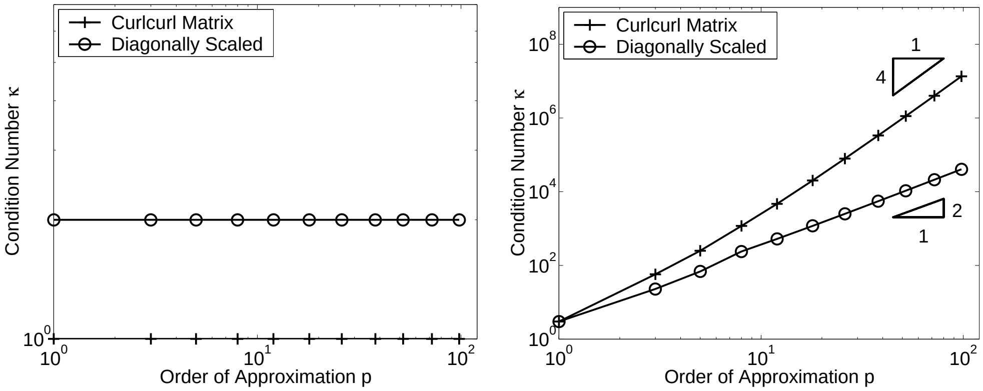
$$\vec{\beta}_1 = \begin{pmatrix} \vec{\alpha}_1 \\ \vec{\alpha}_2 \\ 0 \end{pmatrix}, \quad \vec{\beta}_2 = \begin{pmatrix} 0 \\ \vec{\alpha}_2 \\ \vec{\alpha}_3 \end{pmatrix}, \text{ and } \vec{\beta}_3 = \begin{pmatrix} \vec{\alpha}_1 \\ 0 \\ \vec{\alpha}_3 \end{pmatrix}.$$

Observing that

$$\vec{\alpha}^{\top} S_i \vec{\alpha} = \vec{\beta}_i^{\top} S_i \vec{\beta}_i, \quad i = 1, \dots, 3, \tag{36}$$

we deduce from (34) that

$$\vec{\alpha}^{\top} \boldsymbol{S}_i \vec{\alpha} \geq \lambda_{\min}\left(\boldsymbol{\ell}_{p+1}\right) \lambda_{\min}\bigl(\boldsymbol{S}^{(2)}_{\widehat{K},p}\bigr) |\vec{\beta}_i|^2 = 2\lambda_{\min}\left(\boldsymbol{\ell}_{p+1}\right) |\vec{\beta}_i|^2$$



(a) $\kappa'(\boldsymbol{S}^{(2)}_{\widehat{K},p})$ and $\kappa'(\widetilde{\boldsymbol{S}}^{(2)}_{\widehat{K},p})$ (b) $\kappa'(\boldsymbol{S}^{(3)}_{\widehat{K},p})$ and $\kappa'(\widetilde{\boldsymbol{S}}^{(3)}_{\widehat{K},p})$

FIG. 3. *Variation of the condition numbers versus the order of approximation of the curl-curl and diagonally scaled curl-curl matrices for* (a) *two and* (b) *three dimensions on the reference element. These results agree with those predicted by* (27) *and* (28) *for* $d = 2$ *and* $d = 3$*, respectively.*

and

$$\vec{\alpha}^\top \boldsymbol{S}_i \vec{\alpha} \leq \lambda_{\max}\left(\boldsymbol{\ell}_{p+1}\right) \lambda_{\max}\big(\boldsymbol{S}^{(2)}_{\widehat{K},p}\big) |\vec{\beta}_i|^2 = 2\lambda_{\max}\left(\boldsymbol{\ell}_{p+1}\right) |\vec{\beta}_i|^2.$$

Thus, using (32), (36) and the fact that $\sum_{i=1}^3 |\vec{\beta}_i|^2 = 2\sum_{i=1}^3 |\vec{\alpha}_i|^2$, we obtain

$$\vec{\alpha}^\top \boldsymbol{S}^{(3)}_{\widehat{K},p} \vec{\alpha} \geq 2\lambda_{\min}\left(\boldsymbol{\ell}_{p+1}\right) \sum_{i=1}^{3} |\vec{\beta}_i|^2 = 4\lambda_{\min}(\boldsymbol{\ell}_{p+1})|\vec{\alpha}|^2 \tag{37}$$

and

$$\vec{\alpha}^\top \boldsymbol{S}^{(3)}_{\widehat{K},p} \vec{\alpha} \leq 2\lambda_{\max}\left(\boldsymbol{\ell}_{p+1}\right) \sum_{i=1}^{3} |\vec{\beta}_i|^2 = 4\lambda_{\max}\left(\boldsymbol{l}_{p+1}\right) |\vec{\alpha}|^2. \tag{38}$$

Analogous arguments may be used in the case of diagonal scaling to obtain

$$\vec{\alpha}^\top \widetilde{\boldsymbol{S}}^{(3)}_{\widehat{K},p} \vec{\alpha} \geq 4\lambda_{\min}\big(\widetilde{\boldsymbol{\ell}}_{p+1}\big)|\vec{\alpha}|^2 \quad \text{and} \quad \vec{\alpha}^\top \widetilde{\boldsymbol{S}}^{(3)}_{\widehat{K},p} \vec{\alpha} \leq 8\lambda_{\max}\big(\widetilde{\boldsymbol{\ell}}_{p+1}\big)|\vec{\alpha}|^2. \tag{39}$$

Thus, for the nonzero eigenvalues, using (23) and (24) in conjunction with (37)–(39) yields

$$\lambda'_{\min}\big(\boldsymbol{S}^{(3)}_{\widehat{K},p}\big) \geq 4cp^{-4}, \qquad \lambda_{\max}\big(\boldsymbol{S}^{(3)}_{\widehat{K},p}\big) \leq 4C \tag{40}$$

and

$$\lambda'_{\min}\big(\widetilde{\boldsymbol{S}}^{(3)}_{\widehat{K},p}\big) \geq 4cp^{-2}, \qquad \lambda_{\max}\big(\widetilde{\boldsymbol{S}}^{(3)}_{\widehat{K},p}\big) \leq 8C. \tag{41}$$

Setting $c_{\widehat{K},\boldsymbol{S}} = 4c$ and $C_{\widetilde{K},\boldsymbol{S}} = 8C$ establishes (25) and (26) for $d = 2, 3$, and the lemma is proved. □

In Figure 3 the condition numbers of $\boldsymbol{S}^{(d)}_{\widehat{K},p}$ and $\widetilde{\boldsymbol{S}}^{(d)}_{\widehat{K},p}$ versus order of approximation for $p = 0, \ldots, 100$ are shown for $d = 2$ in (a) and $d = 3$ in (b). Once again, it will be observed that the results in Lemma 3 are sharp.

3.3. Analysis of the stiffness matrix. Finally, we present bounds for the eigenvalues of the stiffness matrix on a single reference element.

LEMMA 4. *Let $\boldsymbol{A}^{(d)}_{\widehat{K},p}$ and $\widetilde{\boldsymbol{A}}^{(d)}_{\widehat{K},p}$ denote the stiffness and diagonally scaled stiffness matrices in d spatial dimensions with order p approximation. Then there exist positive constants $c_{\widehat{K},\boldsymbol{A}}$ and $C_{\widehat{K},\boldsymbol{A}}$ independent of p such that*

$$\lambda_{\min}\big(\boldsymbol{A}^{(d)}_{\widehat{K},p}\big) \geq c_{\widehat{K},\boldsymbol{A}}p^{-4(d-1)}, \quad \lambda_{\max}\big(\boldsymbol{A}^{(d)}_{\widehat{K},p}\big) \leq C_{\widehat{K},\boldsymbol{A}} \tag{42}$$

and

$$\lambda_{\min}\big(\widetilde{\boldsymbol{A}}^{(d)}_{\widehat{K},p}\big) \geq c_{\widehat{K},\boldsymbol{A}} \min\left(1, \frac{\omega}{p}\right)^2 p^{-2(d-1)}, \quad \lambda_{\max}\big(\widetilde{\boldsymbol{A}}^{(d)}_{\widehat{K},p}\big) \leq C_{\widehat{K},\boldsymbol{A}}. \tag{43}$$

Therefore, there exists a positive constant $C^*_{K,\boldsymbol{A}}$ independent of p such that

$$\kappa\big(\boldsymbol{A}^{(d)}_{\widehat{K},p}\big) \leq C^*_{\widehat{K},\boldsymbol{A}} \frac{p^{4(d-1)}}{\omega^2} \tag{44}$$

and

$$\kappa\big(\widetilde{\boldsymbol{A}}^{(d)}_{\widehat{K},p}\big) \leq C^*_{\widehat{K},\boldsymbol{A}} \max\left(1, \frac{p}{\omega}\right)^2 p^{2(d-1)}. \tag{45}$$

Proof. Once again, it suffices to prove (42) and (43). We obtain upper bounds for the maximum eigenvalues by applying the results in Lemmas 2 and 3 for the maximum eigenvalues of the mass and curl-curl matrices to deduce

$$\vec{\alpha}^\top \boldsymbol{A}^{(d)}_{\widehat{K},p}\vec{\alpha} = \vec{\alpha}^\top \boldsymbol{S}^{(d)}_{\widehat{K},p}\vec{\alpha} + \omega^2\vec{\alpha}^\top \boldsymbol{M}^{(d)}_{\widehat{K},p}\vec{\alpha} \leq 2\max\big(C_{\widehat{K},\boldsymbol{S}}, \omega^2 C_{\widehat{K},\boldsymbol{M}}\big)|\vec{\alpha}|^2 \tag{46}$$

and hence

$$\lambda_{\max}\big(\boldsymbol{A}^{(d)}_{\widehat{K},p}\big) \leq 2\max\big(C_{\widehat{K},\boldsymbol{S}}, \omega^2 C_{\widehat{K},\boldsymbol{M}}\big).$$

Equally well, using the bounds in Lemmas 2 and 3 for the maximum eigenvalues of the diagonally scaled matrices gives

$$\begin{aligned}
\vec{\alpha}^\top \boldsymbol{A}^{(d)}_{\widehat{K},p}\vec{\alpha} &= \vec{\alpha}^\top \left[\boldsymbol{S}^{(d)}_{\widehat{K},p} + \omega^2 \boldsymbol{M}^{(d)}_{\widehat{K},p}\right]\vec{\alpha} \\
&\leq \vec{\alpha}^\top \left[C_{\widehat{K},\boldsymbol{S}}\mathrm{diag}\big(\boldsymbol{S}^{(d)}_{\widehat{K},p}\big) + \omega^2 C_{\widehat{K},\boldsymbol{M}}\mathrm{diag}\big(\boldsymbol{M}^{(d)}_{\widehat{K},p}\big)\right]\vec{\alpha} \\
&\leq \max\big(C_{\widehat{K},\boldsymbol{S}}, C_{\widehat{K},\boldsymbol{M}}\big)\vec{\alpha}^\top \mathrm{diag}\big(\boldsymbol{A}^{(d)}_{\widehat{K},p}\big)\vec{\alpha} \qquad (47)
\end{aligned}$$

and hence

$$\lambda_{\max}\big(\widetilde{\boldsymbol{A}}^{(d)}_{\widehat{K},p}\big) \leq \max\big(C_{\widehat{K},\boldsymbol{S}}, C_{\widehat{K},\boldsymbol{M}}\big).$$

The minimum eigenvalue of the stiffness matrix is bounded using the result in (15) for the minimum eigenvalue as follows,

$$\vec{\alpha}^\top \boldsymbol{A}^{(d)}_{\widehat{K},p}\vec{\alpha} \geq \omega^2\vec{\alpha}^\top \boldsymbol{M}^{(d)}_{\widehat{K},p}\vec{\alpha} \geq \omega^2 c_{\widehat{K},\boldsymbol{M}}p^{-4(d-1)}|\vec{\alpha}|^2,$$

and hence

$$\lambda_{\min}\big(\boldsymbol{A}^{(d)}_{\widehat{K},p}\big) \geq \omega^2 c_{\widehat{K},\boldsymbol{M}} p^{-4(d-1)}.$$

It remains to derive the lower bound on the minimum eigenvalue of the diagonally scaled stiffness matrix. The bound in Lemma 2 for the minimum eigenvalue of the diagonally scaled mass matrix gives

$$\vec{\alpha}^\top \boldsymbol{M}^{(d)}_{\widehat{K},p}\vec{\alpha} \geq cp^{-2(d-1)}\vec{\alpha}^\top \mathrm{diag}\big(\boldsymbol{M}^{(d)}_{\widehat{K},p}\big)\vec{\alpha}. \tag{48}$$

A direct computation using standard properties of Legendre polynomials reveals that if $\boldsymbol{\phi}$ is any of the basis functions defined in (7) or (8), then

$$\|\mathbf{curl}\,\boldsymbol{\phi}\|^2_{L_2(\widehat{K})} \leq Cp^2 \|\boldsymbol{\phi}\|^2_{L_2(\widehat{K})},$$

or, expressed in terms of matrices,

$$\vec{\alpha}^\top \mathrm{diag}\big(\boldsymbol{S}^{(d)}_{\widehat{K},p}\big)\vec{\alpha} \leq Cp^2\vec{\alpha}^\top \mathrm{diag}\big(\boldsymbol{M}^{(d)}_{\widehat{K},p}\big)\vec{\alpha}.$$

With the aid of (48) we deduce that

$$\vec{\alpha}^\top \boldsymbol{M}^{(d)}_{\widehat{K},p}\vec{\alpha} \geq cp^{-2d}\vec{\alpha}^\top \mathrm{diag}\left(\boldsymbol{S}^{(d)}_{\widehat{K},p}\right)\vec{\alpha}. \tag{49}$$

The bounds (48) and (49) are used to obtain the lower bound on the eigenvalue as follows,

$$\begin{aligned}
\vec{\alpha}^\top \boldsymbol{A}^{(d)}_{\widehat{K},p}\vec{\alpha} &\geq \omega^2\vec{\alpha}^\top \boldsymbol{M}^{(d)}_{\widehat{K},p}\vec{\alpha} \\
&\geq cp^{-2(d-1)}\vec{\alpha}^\top\left[\frac{\omega^2}{p^2}\,\mathrm{diag}\big(\boldsymbol{S}^{(d)}_{\widehat{K},p}\big) + \omega^2\mathrm{diag}\big(\boldsymbol{M}^{(d)}_{\widehat{K},p}\big)\right]\vec{\alpha} \\
&\geq cp^{-2(d-1)}\min\left(1,\frac{\omega}{p}\right)^2\vec{\alpha}^\top\mathrm{diag}\big(\boldsymbol{S}^{(d)}_{\widehat{K},p} + \omega^2\boldsymbol{M}^{(d)}_{\widehat{K},p}\big)\vec{\alpha} \\
&= cp^{-2(d-1)}\min\left(1,\frac{\omega}{p}\right)^2\vec{\alpha}^\top\mathrm{diag}\big(\boldsymbol{A}^{(d)}_{\widehat{K},p}\big)\vec{\alpha}, \qquad (50)
\end{aligned}$$

and the result follows as claimed. □

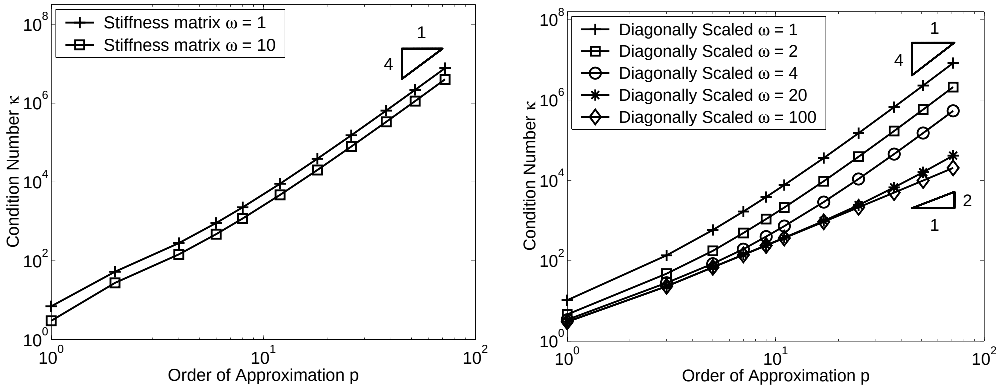
Figures 4 and 5 show the computed condition numbers of $\boldsymbol{A}^{(d)}_{\widehat{K},p}$ and $\widetilde{\boldsymbol{A}}^{(d)}_{\widehat{K},p}$ versus order of approximation on the reference element for $p = 0,\dots,80$, for $d = 2$ and $d = 3$, respectively. As before, the bounds are seen to be sharp. Observe the change in asymptotic behavior with increasing values of the coefficient ω from $\mathcal{O}(p^4)$ to $\mathcal{O}(p^2)$ in Figure 4, and from $\mathcal{O}(p^6)$ to $\mathcal{O}(p^4)$ in Figure 5, as predicted in (45).

3.4. Proof of Theorem 1. (1) Let $\boldsymbol{M}^{(d)}_{K,p}$ denote the mass matrix over a physical element $K \in \mathcal{M}$ corresponding to order of approximation p. Then, for a discrete electric field $\boldsymbol{E}$ on the physical element, relation (9) implies that

$$\boldsymbol{M}_K(\boldsymbol{E},\boldsymbol{E}) = \int_K \varepsilon|\boldsymbol{E}|^2\,\mathrm{d}\boldsymbol{x} = \int_{\widehat{K}} \varepsilon|\boldsymbol{J}_K^{-\top}\widehat{\boldsymbol{E}}|^2|\det(\boldsymbol{J}_K)|\,\mathrm{d}\boldsymbol{\xi}.$$

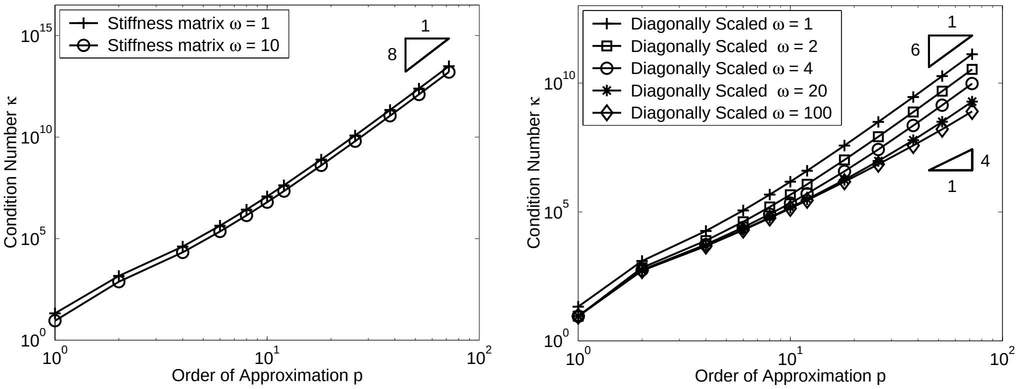
Applying the bounds in (3) and (5)–(6) leads to the conclusion

$$c_1c_{2,K}c_{3,K}\lambda_{\min}\big(\boldsymbol{M}^{(d)}_{\widehat{K},p}\big)\boldsymbol{I} \leq \boldsymbol{M}^{(d)}_{K,p} \leq C_1C_{2,K}C_{3,K}\lambda_{\max}\big(\boldsymbol{M}^{(d)}_{\widehat{K},p}\big)\boldsymbol{I}. \tag{51}$$



(a) $\kappa(\boldsymbol{A}^{(2)}_{\widehat{K},p})$ (b) $\kappa(\widetilde{\boldsymbol{A}}^{(2)}_{\widehat{K},p})$

FIG. 4. *Variation of the condition numbers versus the order of approximation of the stiffness matrix* (a) *and diagonally scaled stiffness matrix* (b) *in two dimensions on the reference element. Note that in* (a) *only the results for the extreme values of ω have been shown, for clarity, since the curves for intermediate values are found to lie in between. These results agree with those predicted in* (44) *and* (45) *for $d = 2$.*



(a) $\kappa(\boldsymbol{A}^{(3)}_{\widehat{K},p})$ (b) $\kappa(\widetilde{\boldsymbol{A}}^{(3)}_{\widehat{K},p})$

FIG. 5. *Variation of the condition numbers versus the order of approximation of the stiffness matrix* (a) *and diagonally scaled stiffness matrix* (b) *in three dimensions on the reference element. As before, in* (a), *only the results for the extreme values of ω have been shown, for clarity, since the curves for intermediate values are found to lie in between. These results agree with those predicted by* (44) *and* (45) *for $d = 3$.*

Lemma 2 implies that

$$\lambda_{\min}\big(\boldsymbol{M}^{(d)}_{K,p}\big) \geq c_{K,\boldsymbol{M}} p^{-4(d-1)} \quad \text{and} \quad \lambda_{\max}\big(\boldsymbol{M}^{(d)}_{K,p}\big) \leq C_{K,\boldsymbol{M}}, \tag{52}$$

where

$$c_{K,\boldsymbol{M}} = c_1 c_{2,K} c_{3,K} c_{\widehat{K},\boldsymbol{M}} \quad \text{and} \quad C_{K,\boldsymbol{M}} = C_1 C_{2,K} C_{3,K} C_{\widehat{K},\boldsymbol{M}}.$$

As a consequence of (52),

$$\kappa\big(\boldsymbol{M}^{(d)}_{K,p}\big) \leq C^*_{K,M} p^{4(d-1)}, \tag{53}$$

where $C^*_{K,M} = C_{K,\boldsymbol{M}}/c_{K,\boldsymbol{M}}$. The same argument may be applied in the case of diagonal scaling to deduce that

$$\lambda_{\min}\big(\widetilde{\boldsymbol{M}}^{(d)}_{K,p}\big) \geq c_{K,\boldsymbol{M}}p^{-2(d-1)}, \qquad \lambda_{\max}\big(\widetilde{\boldsymbol{M}}^{(d)}_{K,p}\big) \leq C_{K,\boldsymbol{M}} \tag{54}$$

and

$$\kappa\big(\widetilde{\boldsymbol{M}}^{(d)}_{K,p}\big) \leq C^*_{K,M}p^{2(d-1)}. \tag{55}$$

Let $\vec{\alpha}_{\boldsymbol{p}} \in \mathbb{R}^N$ denote the values of the global degrees of freedom for the nonuniform order of approximation $\boldsymbol{p}$ over $\mathcal{M}$ with maximum order $p = p_{\max}$, and let $\vec{\alpha}_p \in \mathbb{R}^{N'}$ denote the value of the global degrees of freedom for the uniform order of approximation p over $\mathcal{M}$. The global mass matrix $\boldsymbol{M}^{(d)}_{\mathcal{M},p}$, corresponding to the uniform order approximation p over the entire mesh, satisfies

$$\vec{\alpha}_p^\top \boldsymbol{M}^{(d)}_{\mathcal{M},p}\vec{\alpha}_p = \vec{\alpha}_p^\top \left(\sum_{K\in\mathcal{M}} \Lambda_K \boldsymbol{M}^{(d)}_{K,p}\Lambda_K^\top\right)\vec{\alpha}_p = \sum_{K\in\mathcal{M}} \vec{\beta}^\top_{K,p}\boldsymbol{M}^{(d)}_{K,p}\vec{\beta}_{K,p},$$

where $\vec{\beta}_{K,p} = \Lambda_K^\top\vec{\alpha}_p$ and the Λ_K is the usual connectivity matrix representing the mapping from the local to the global degrees of freedom for each element $K \in \mathcal{M}$. In particular, this implies that there are positive constants c_6 and C_6 such that

$$c_6|\vec{\alpha}_p|^2 \leq \sum_{K\in\mathcal{M}} |\vec{\beta}_{K,p}|^2 \leq C_6|\vec{\alpha}_p|^2. \tag{56}$$

Using (52), it follows that

$$\vec{\alpha}_p^\top \boldsymbol{M}^{(d)}_{\mathcal{M},p}\vec{\alpha}_p \leq \sum_{K\in\mathcal{M}} \lambda_{\max}\left(\boldsymbol{M}^{(d)}_{K,p}\right)|\vec{\beta}_{K,p}|^2 \leq \left(\max_{K\in\mathcal{M}} C_{K,\boldsymbol{M}}\right)\sum_{K\in\mathcal{M}}|\vec{\beta}_{K,p}|^2$$

and

$$\vec{\alpha}_p^\top \boldsymbol{M}^{(d)}_{\mathcal{M},p}\vec{\alpha}_p \geq \sum_{K\in\mathcal{M}} \lambda_{\min}\left(\boldsymbol{M}^{(d)}_{K,p}\right)|\vec{\beta}_{K,p}|^2 \geq \left(\min_{K\in\mathcal{M}} c_{K,\boldsymbol{M}}\right)p^{-4(d-1)}\sum_{K\in\mathcal{M}}|\vec{\beta}_{K,p}|^2.$$

We use (56) to deduce

$$c_6p^{-4(d-1)}\left(\min_{K\in\mathcal{M}} c_{K,\boldsymbol{M}}\right)|\vec{\alpha}_p|^2 \leq \vec{\alpha}_p^\top\boldsymbol{M}_{\mathcal{M},p}\vec{\alpha}_p \leq C_6\left(\max_{K\in\mathcal{M}} C_{K,\boldsymbol{M}}\right)|\vec{\alpha}_p|^2,$$

which implies

$$\lambda_{\min}\big(\boldsymbol{M}^{(d)}_{\mathcal{M},p}\big) \geq c'_{\mathcal{M},\boldsymbol{M}}p^{-4(d-1)} \quad \text{and} \quad \lambda_{\max}\big(\boldsymbol{M}^{(d)}_{\mathcal{M},p}\big) \leq C'_{\mathcal{M},\boldsymbol{M}},$$

where $c'_{\mathcal{M},\boldsymbol{M}} = c_6\min_{K\in\mathcal{M}} c_{K,\boldsymbol{M}}$ and $C'_{\mathcal{M},\boldsymbol{M}} = C_6\max_{K\in\mathcal{M}} C_{K,\boldsymbol{M}}$. The result for diagonal scaling

$$\lambda_{\min}\big(\widetilde{\boldsymbol{M}}^{(d)}_{\mathcal{M},p}\big) \geq c'_{\mathcal{M},\boldsymbol{M}}p^{-2(d-1)} \quad \text{and} \quad \lambda_{\max}\big(\widetilde{\boldsymbol{M}}^{(d)}_{\mathcal{M},p}\big) \leq C'_{\mathcal{M},\boldsymbol{M}}$$

is established in the same way.

The eigenvalues of the global mass matrix corresponding to the nonuniform order approximation $\boldsymbol{p}$ over $\mathcal{M}$ satisfy

$$\lambda_{\max}\left(\boldsymbol{M}^{d}_{\mathcal{M},\boldsymbol{p}}\right) = \max_{\vec{\alpha}_{\boldsymbol{p}}\in\mathbb{R}^{N}} \frac{\vec{\alpha}_{\boldsymbol{p}}^{\top}\boldsymbol{M}^{(d)}_{\mathcal{M},\boldsymbol{p}}\vec{\alpha}_{\boldsymbol{p}}}{|\vec{\alpha}_{\boldsymbol{p}}|^{2}} \le \max_{\vec{\alpha}_{p}\in\mathbb{R}^{N'}} \frac{\vec{\alpha}_{p}^{\top}\boldsymbol{M}^{(d)}_{\mathcal{M},p}\vec{\alpha}_{p}}{|\vec{\alpha}_{p}|^{2}} \le C'_{\mathcal{M},\boldsymbol{M}} \tag{57}$$

and

$$\begin{aligned}\lambda_{\min}\left(\boldsymbol{M}^{d}_{\mathcal{M},\boldsymbol{p}}\right) = \min_{\vec{\alpha}_{\boldsymbol{p}}\in\mathbb{R}^{N}} \frac{\vec{\alpha}_{\boldsymbol{p}}^{\top}\boldsymbol{M}^{(d)}_{\mathcal{M},\boldsymbol{p}}\vec{\alpha}_{\boldsymbol{p}}}{|\vec{\alpha}_{\boldsymbol{p}}|^{2}} &\ge \min_{\vec{\alpha}_{p}\in\mathbb{R}^{N'}} \frac{\vec{\alpha}_{p}^{\top}\boldsymbol{M}^{(d)}_{\mathcal{M},p}\vec{\alpha}_{p}}{|\vec{\alpha}_{p}|^{2}}\\ &\ge p^{-4(d-1)}c'_{\mathcal{M},\boldsymbol{M}},\end{aligned} \tag{58}$$

and hence the first part of (12) holds with $C^{*}_{\mathcal{M},\boldsymbol{M}} = C'_{\mathcal{M},\boldsymbol{M}}/c'_{\mathcal{M},\boldsymbol{M}}$. The second part of (12) is proved using (54) in the same fashion.

(2) Define the following skew-symmetric form (see [16]), using the discrete electric field $\widehat{\boldsymbol{E}}$ on the reference element:

$$\boldsymbol{C}(\widehat{\boldsymbol{E}}) = \begin{bmatrix} 0 & \widehat{E}_{1,2} & \widehat{E}_{1,3} \\ -\widehat{E}_{1,2} & 0 & \widehat{E}_{2,3} \\ -\widehat{E}_{1,3} & -\widehat{E}_{2,3} & 0 \end{bmatrix}, \tag{59}$$

where

$$\widehat{\boldsymbol{E}} = (\widehat{E}_1, \widehat{E}_2, \widehat{E}_3)^{\top} \quad \text{and} \quad \widehat{E}_{i,j} = \frac{\partial \widehat{E}_j}{\partial \xi_i} - \frac{\partial \widehat{E}_i}{\partial \xi_j}.$$

Straightforward calculations reveal that

$$|\mathbf{curl}\, \widehat{\boldsymbol{E}}|^2 = \frac{1}{2}\|\boldsymbol{C}(\widehat{\boldsymbol{E}})\|_F^2,$$

where $\|\cdot\|_F$ denotes the Frobenius norm (see [9]). Furthermore, using (9), it follows that

$$\boldsymbol{C}(\boldsymbol{E}) = \boldsymbol{J}_K^{-\top}\boldsymbol{C}(\widehat{\boldsymbol{E}})\boldsymbol{J}_K^{-1}$$

for the discrete electric field $\boldsymbol{E}$ on the physical element $K \in \mathcal{M}$, and hence,

$$|\mathbf{curl}\, \boldsymbol{E}|^2 = \frac{1}{2}\|\boldsymbol{J}_K^{-\top}\boldsymbol{C}(\widehat{\boldsymbol{E}})\boldsymbol{J}_K^{-1}\|_F^2. \tag{60}$$

Using the inequality (see, e.g., problem 23, section 5.6 of [9])

$$\frac{1}{d^2}\|\boldsymbol{J}_K^{T}\|_2^{-2}\|\boldsymbol{C}(\widehat{\boldsymbol{E}})\|_F^2\|\boldsymbol{J}_K\|_2^{-2} \le \|\boldsymbol{J}_K^{-\top}\boldsymbol{C}(\widehat{\boldsymbol{E}})\boldsymbol{J}_K^{-1}\|_F^2 \le d^2\|\boldsymbol{J}_K^{-\top}\|_2^2\|\boldsymbol{C}(\widehat{\boldsymbol{E}})\|_F^2\|\boldsymbol{J}_K^{-1}\|_2^2,$$

where $\|\cdot\|_2$ denotes the spectral norm (see [9]), establishes that

$$\frac{c_3^2}{d^2}|\mathbf{curl}\, \widehat{\boldsymbol{E}}|^2 \le \frac{1}{2}\|\boldsymbol{J}_K^{-\top}\boldsymbol{C}(\widehat{\boldsymbol{E}})\boldsymbol{J}_K^{-1}\|_F^2 \le d^2C_3^2|\mathbf{curl}\, \widehat{\boldsymbol{E}}|^2 \tag{61}$$

using (5)–(6). By integrating over the physical element K, we deduce

$$S_K(\boldsymbol{E},\boldsymbol{E}) = \int_K \mu^{-1}|\mathbf{curl}\, \boldsymbol{E}|^2\,\mathrm{d}\boldsymbol{x} = \frac{1}{2}\int_{\widehat{K}} \mu^{-1}\|\boldsymbol{J}_K^{-\top}\boldsymbol{C}(\widehat{\boldsymbol{E}})\boldsymbol{J}_K^{-1}\|_F^2\ |\det(\boldsymbol{J}_K)|\,\mathrm{d}\boldsymbol{\xi},$$

and applying (3) and (61) yields

$$\lambda_{\max}\big(\boldsymbol{S}^{(d)}_{K,p}\big) \leq d^2 C_1 C_{2,K} C_{3,K}^2 \lambda_{\max}\big(\boldsymbol{S}^{(d)}_{\widehat{K},p}\big) \tag{62}$$

and

$$\lambda_{\min}\big(\boldsymbol{S}^{(d)}_{K,p}\big) \geq \frac{c_1 c_{2,K} c_{3,K}^2}{d^2} \lambda_{\min}\big(\boldsymbol{S}^{(d)}_{\widehat{K},p}\big).$$

As a result of (25),

$$\lambda_{\min}\big(\boldsymbol{S}^{(d)}_{K,p}\big) \geq c_{K,\boldsymbol{S}} p^{-4(d-2)} \quad \text{and} \quad \lambda_{\max}\big(\boldsymbol{S}^{(d)}_{K,p}\big) \leq C_{K,\boldsymbol{S}}, \tag{63}$$

where

$$c_{K,\boldsymbol{S}} = \frac{1}{9} c_1 c_{2,K} c_{3,K}^2 c_{\widehat{K},\boldsymbol{S}} \quad \text{and} \quad C_{K,\boldsymbol{S}} = 9 C_1 C_{2,K} C_{3,K}^2 C_{\widehat{K},\boldsymbol{S}}.$$

Hence

$$\kappa\big(\boldsymbol{S}^{(d)}_{K,p}\big) \leq C^*_{K,\boldsymbol{S}} p^{4(d-2)}, \tag{64}$$

where $C^*_{K,\boldsymbol{M}} = C_{K,\boldsymbol{S}}/c_{K,\boldsymbol{S}}$. The same argument may be applied in the case of diagonal scaling:

$$\lambda_{\min}\big(\widetilde{\boldsymbol{S}}^{(d)}_{K,p}\big) \geq c_{K,\boldsymbol{S}} p^{-2(d-2)}, \qquad \lambda_{\max}\big(\widetilde{\boldsymbol{S}}^{(d)}_{K,p}\big) \leq C_{K,\boldsymbol{S}} \tag{65}$$

and

$$\kappa\big(\widetilde{\boldsymbol{S}}^{(d)}_{K,p}\big) \leq C^*_{K,\boldsymbol{S}} p^{2(d-2)}. \tag{66}$$

The global curl-curl matrix $\boldsymbol{S}^{(d)}_{\mathcal{M},p}$ corresponding to the uniform order approximation p over $\mathcal{M}$ satisfies

$$\left(\min_{K\in\mathcal{M}} c_{K,\boldsymbol{S}}\right) \sum_{K\in\mathcal{M}} |\vec{\beta}_{K,p}|^2 \leq \vec{\alpha}_p^\top \boldsymbol{S}^{(d)}_{\mathcal{M},p} \vec{\alpha}_p \leq \left(\max_{K\in\mathcal{M}} C_{K,\boldsymbol{S}}\right) \sum_{K\in\mathcal{M}} |\vec{\beta}_{K,p}|^2.$$

Thus, using (56) results in

$$c_6 p^{-4(d-2)} \left(\min_{K\in\mathcal{M}} c_{K,\boldsymbol{S}}\right) |\vec{\alpha}_p|^2 \leq \vec{\alpha}_p^\top \boldsymbol{S}^{(d)}_{\mathcal{M},p} \vec{\alpha}_p \leq C_6 \left(\max_{K\in\mathcal{M}} C_{K,\boldsymbol{S}}\right) |\vec{\alpha}_p|^2,$$

which leads to

$$\lambda_{\min}\big(\boldsymbol{S}^{(d)}_{\mathcal{M},p}\big) \geq c'_{\mathcal{M},\boldsymbol{S}} p^{-4(d-2)} \quad \text{and} \quad \lambda_{\max}\big(\boldsymbol{S}^{(d)}_{\mathcal{M},p}\big) \leq C'_{\mathcal{M},\boldsymbol{S}},$$

where $c'_{\mathcal{M},\boldsymbol{S}} = c_4 \min_{K\in\mathcal{M}} c_{K,\boldsymbol{S}}$ and $C'_{\mathcal{M},\boldsymbol{S}} = C_4 \max_{K\in\mathcal{M}} C_{K,\boldsymbol{S}}$.

Applying arguments similar to those used in (57) and (58), the global curl-curl matrix corresponding to the nonuniform order approximation $\boldsymbol{p}$ over $\mathcal{M}$ satisfies

$$\lambda_{\min}\big(\boldsymbol{S}^{(d)}_{\mathcal{M},\boldsymbol{p}}\big) \geq c'_{\mathcal{M},\boldsymbol{S}} p^{-4(d-2)} \quad \text{and} \quad \lambda_{\max}\big(\boldsymbol{S}^{(d)}_{\mathcal{M},\boldsymbol{p}}\big) \leq C'_{\mathcal{M},\boldsymbol{S}},$$

and the first part of (13) holds with $C^*_{\mathcal{M},\boldsymbol{S}} = C'_{\mathcal{M},\boldsymbol{S}}/c'_{\mathcal{M},\boldsymbol{S}}$. The second part is proved using (65) in the same manner.

(3) Let $\vec{\alpha}_{K,p}$ denote the values of the degrees of freedom in the approximation over $K \in \mathcal{M}$. Retracing the steps in deriving (46) and using the estimates (52) and (63) yields

$$\vec{\alpha}_{K,p}^{\top}\boldsymbol{A}_{K,p}^{(d)}\vec{\alpha}_{K,p} \leq 2\max(C_{K,\boldsymbol{S}}, \omega^2 C_{K,\boldsymbol{M}})|\vec{\alpha}_{K,p}|^2,$$

while following the steps leading to (47) using the estimates (54) and (65) yields

$$\vec{\alpha}_{K,p}^{\top}\boldsymbol{A}_{K,p}^{(d)}\vec{\alpha}_{K,p} \leq \max(C_{K,\boldsymbol{S}}, C_{K,\boldsymbol{M}})\vec{\alpha}_{K,p}^{\top}\mathrm{diag}\big(\boldsymbol{A}_{K,p}^{(d)}\big)\vec{\alpha}_{K,p},$$

which implies that

$$\lambda_{\max}\big(\boldsymbol{A}_{K,p}^{(d)}\big) \leq C_{K,\boldsymbol{A}} \quad \text{and} \quad \lambda_{\max}\big(\widetilde{\boldsymbol{A}}_{K,p}^{(d)}\big) \leq C_{K,\boldsymbol{A}}.$$

Similarly, using (52) gives a bound for the minimum eigenvalue of the stiffness matrix

$$\vec{\alpha}_{K,p}^{\top}\boldsymbol{A}_{K,p}^{(d)}\vec{\alpha}_{K,p} \geq \omega^2\vec{\alpha}_{K,p}^{\top}\boldsymbol{M}_{K,p}^{(d)}\vec{\alpha}_{K,p} \geq \omega^2 c_{K,\boldsymbol{M}}p^{-4(d-1)}|\vec{\alpha}_{K,p}|^2.$$

The arguments used to obtain (48) and (49) may be modified to deduce that

$$\vec{\alpha}_{K,p}^{\top}\boldsymbol{M}_{K,p}^{(d)}\vec{\alpha}_{K,p} \geq cp^{-2(d-1)}\vec{\alpha}_{K,p}^{\top}\mathrm{diag}\big(\boldsymbol{M}_{K,p}^{(d)}\big)\vec{\alpha}_{K,p}$$

and

$$\vec{\alpha}_{K,p}^{\top}\boldsymbol{M}_{K,p}^{(d)}\vec{\alpha}_{K,p} \geq cp^{-2d}\vec{\alpha}_{K,p}^{\top}\mathrm{diag}\big(\boldsymbol{S}_{K,p}^{(d)}\big)\vec{\alpha}_{K,p}.$$

Then, following the steps in (50), we obtain

$$\vec{\alpha}_{K,p}^{\top}\boldsymbol{A}_{K,p}^{(d)}\vec{\alpha}_{K,p} \geq cp^{-2(d-1)}\min\left(1, \frac{\omega}{p}\right)^2\vec{\alpha}_{K,p}^{\top}\mathrm{diag}\left(\boldsymbol{A}_{K,p}^{(d)}\right)\vec{\alpha}_{K,p},$$

which leads to a bound on the smallest eigenvalue of the diagonally scaled stiffness matrix.

The above bounds for the eigenvalues of matrices on a single element are used to obtain bounds for the global matrices corresponding to uniform order of approximation p by summing element contributions as before. Then, arguments analogous to (57) and (58) are used to deduce bounds for nonuniform order of approximation $\boldsymbol{p}$.

This completes the proof of Theorem 1.

REFERENCES

[1] M. Ainsworth and J. Coyle, *Hierarchic hp-edge element families for Maxwell's equations on hybrid quadrilateral/triangular meshes*, Comput. Methods Appl. Mech. Engrg., 190 (2001), pp. 6709–6733.

[2] M. Ainsworth and B. Senior, *Aspects of an adaptive hp-finite element method: Adaptive strategy, conforming approximation and efficient solvers*, Comput. Methods Appl. Mech. Engrg., 150 (1997), pp. 65–87.

[3] A. Bossavit, *A rationale for edge-elements in* 3*-d fields computations*, IEEE Trans. Magnetics, 24 (1988), pp. 74–79.

[4] A. Bossavit, *Whitney forms—A class of finite elements for* 3*-dimensional computations in electromagnetism*, IEEE Proceedings A, 135 (1988), pp. 493–500.

[5] A. Bossavit and I. Mayergoyz, *Edge elements in scattering problems*, IEEE Trans. Magnetics, 25 (1989), pp. 2816–2821.

[6] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, Elsevier–North Holland, Amsterdam, 1978.

[7] M. Costabel, *A coercive bilinear form for Maxwell's equations*, J. Math. Anal. Appl., 157 (1991), pp. 527–541.

[8] L. Demkowicz and L. Vardapetyan, *Modeling of electromagnetic absorption/scattering problems using hp-adaptive finite elements*, Comput. Methods Appl. Mech. Engrg., 152 (1998), pp. 103–124.

[9] R. Horn and C. Johnson, *Matrix Analysis*, Cambridge University Press, London, 1985.

[10] R. Horn and C. Johnson, *Topics in Matrix Analysis*, Cambridge University Press, London, 1989.

[11] N. Hu, X.-Z. Guo, and I. Katz, *Bounds for the eigenvalues and condition number in the p-version of the finite element method*, Math. Comp., 67 (1998), pp. 1423–1450.

[12] J. Jin, *The Finite Element Method in Electromagnetics*, John Wiley & Sons, New York, 1993.

[13] J.-F. Maitre and O. Pourquier, *Conditionnements et préconditionnements diagonaux pour la p-version des méthodes d'éléments finis pour des problèmes elliptiques du second order*, C. R. Acad. Sci. Paris Sér. I Math., 318 (1994), pp. 583–586.

[14] P. Monk, *An analysis of Nédélec's method for the spatial discretization of Maxwell equations*, J. Comput. Appl. Math., 47 (1993), pp. 101–121.

[15] P. Monk, *On the p-extension and hp-extension of Nédélec curl-conforming elements*, J. Comput. Appl. Math., 53 (1994), pp. 117–137.

[16] J. Nédélec, *Mixed finite elements in* $\mathbb{R}^3$, Numer. Math., 93 (1980), pp. 315–341.

[17] E. T. Olsen and J. Douglas, Jr., *Bounds on spectral condition numbers of matrices in the p-version of the finite element method*, Numer. Math., 69 (1995), pp. 333–352.

[18] W. Rachowicz and L. Demkowicz, *A two dimensional hp-adaptive finite element package for electromagnetics*, Comput. Methods Appl. Mech. Engrg., 93 (1999), pp. 315–341.

[19] J. Wang and N. Ida, *Curvilinear and higher order "edge" elements in electromagnetic field computations*, IEEE Trans. Magnetics, 29 (1993), pp. 1491–1494.

[20] J. P. Webb, *Hierarchal vector basis functions of arbitrary order for triangular and tetrahedral finite elements*, IEEE Trans. Antennas and Propagation, 47 (1999), pp. 1244–1253.

SIAM J. NUMER. ANAL.
Vol. 41, No. 2, pp. 751–766

POLYNOMIAL INTERPOLATION ON THE UNIT SPHERE*

YUAN XU†

Abstract. The problem of interpolation at $(n+1)^2$ points on the unit sphere S^2 by spherical polynomials of degree at most n is studied. Many sets of points that admit unique interpolation are given explicitly. The proof is based on a method of factorization of polynomials. A related problem of interpolation by trigonometric polynomials is also solved.

Key words. interpolation, spherical polynomials, unit sphere

AMS subject classifications. 41A05, 41A63, 65D05

PII. S003614290139946X

1. Introduction.

The purpose of this paper is to study polynomial interpolation on the unit sphere $S^2 = \{x : \|x\| = 1\}$ of $\mathbb{R}^3$, where $\|x\|$ is the Euclidean norm of $\mathbb{R}^3$. Let Π_n^2 denote the space of polynomials of degree at most n in 2 variables, and let $\mathcal{P}_n^3$ denote the space of homogeneous polynomials of degree n in 3 variables. The notation $\Pi_n(S^2)$ denotes the space of spherical polynomials of 3 variables, that is, the restriction of polynomials of 3 variables in Π_n^3 on S^2. It is known that

$$\dim \Pi_n(S^2) = (n+1)^2, \qquad n \geq 0.$$

We study the following polynomial interpolation problem on S^2:

Problem 1. Let $X = \{\mathbf{a}_i : 1 \leq i \leq (n+1)^2\}$ be a set of distinct points on S^2. Find conditions on X such that there is a unique polynomial $T \in \Pi_n(S^2)$ satisfying

$$T(\mathbf{a}_i) = f_i, \qquad \mathbf{a}_i \in X, \quad 1 \leq i \leq (n+1)^2,$$

for any given data $\{f_i\}$.

If there is a *unique* solution to the interpolation problem, we say that the problem is *poised* and that X solves Problem 1. In terms of a basis of $\Pi_n(S^2)$, the interpolation conditions give linear equations for the coefficients of T. It follows that the problem has a unique solution if and only if $T(\mathbf{a}_i) = 0$ for all $\mathbf{a}_i \in X$ implies $T = 0$, which holds if and only if the determinant of the linear system of equations is nonzero. The determinant is a polynomial of the interpolation points; hence it is nonzero for almost all choices of interpolation points. In other words, Problem 1 is poised for almost all choices of X. On the other hand, given a set of points, it is often difficult to determine if it leads to unique interpolation. The question is related to several other problems; see the discussion below. In many applications, one would like to have poised sets of points explicitly given. The main result of this paper provides families of such points that admit unique interpolation by $\Pi_n(S^2)$ for all n.

We need some basic facts about spherical harmonics (see [3, 10], for example). The harmonic polynomials are homogeneous polynomials Y in $\mathcal{P}_n^3$ that satisfy $\Delta Y = 0$, where Δ is the usual Laplace operator, and $\Delta = \partial_1^2 + \partial_2^2 + \partial_3^2$ with $\partial_i = \partial^2/\partial x_i^2$. The

[*] Received by the editors December 11, 2001; accepted for publication (in revised form) November 2, 2002; published electronically May 6, 2003. This work was partially supported by the National Science Foundation under grant DMS-0201669.

http://www.siam.org/journals/sinum/41-2/39946.html

[†] Department of Mathematics, University of Oregon, Eugene, OR 97403-1222 (yuan@math.uoregon.edu).

spherical harmonics are the restriction of the harmonic polynomials on S^2. Let $\mathcal{H}_n^3$ denote the space of spherical harmonics of degree n in three variables. It is known that

$$\dim \mathcal{H}_n^3 = 2n+1, \qquad n \ge 0.$$

Let $\{S_{k,n} : 1 \le k \le 2n+1\}$ be an orthonormal basis of $\mathcal{H}_n^3$. The reproducing kernel $Y_n(\mathbf{x},\mathbf{y})$ of $\mathcal{H}_n^3$ is defined by

$$Y_n(\mathbf{x},\mathbf{y}) = \sum_{k=1}^{2n+1} S_{k,n}(\mathbf{x}) S_{k,n}(\mathbf{y}), \qquad \mathbf{x},\mathbf{y} \in S^2.$$

In particular, for each fixed $\mathbf{y} \in S^2$, $Y_n(\mathbf{x},\mathbf{y})$ is a spherical harmonic, the so-called zonal harmonic of degree n. Let $C_n^\lambda(t)$ be the Gegenbauer polynomial of degree n normalized so that $C_n^\lambda(1) = \binom{n+2\lambda-1}{n}$. Then Y_n satisfies

$$Y_n(\mathbf{x},\mathbf{y}) = (2n+1) C_n^{1/2}(\langle \mathbf{x},\mathbf{y}\rangle), \qquad \mathbf{x},\mathbf{y} \in S^2.$$

In $L^2(S^2)$, the space $\Pi_n(S^2)$ admits a unique orthogonal decomposition $\Pi_n(S^2) = \bigoplus_{k=0}^n \mathcal{H}_k^3$ in terms of the spaces of spherical harmonics. The reproducing kernel of $\Pi_n(S^2)$ in $L^2(S^2)$ is given by

$$K_n(\mathbf{x},\mathbf{y}) = \sum_{k=0}^n Y_k(\mathbf{x},\mathbf{y}) = \sum_{k=0}^n (2k+1) C_k^{1/2}(\langle \mathbf{x},\mathbf{y}\rangle).$$

For each fixed $\mathbf{y} \in S^2$, $K_n(\mathbf{x},\mathbf{y})$ is a polynomial in $\Pi_n(S^2)$ that depends only on $\langle \mathbf{x},\mathbf{y}\rangle$.

Apart from its own interest, the interpolation problem is related to other problems concerning spherical polynomials. As an example, let $X = \{\mathbf{a}_i : 1 \le i \le (n+1)^2\}$ be a set of distinct points on S^2; one can ask the question of when $\{K_n(\mathbf{x},\mathbf{a}_i), 1 \le i \le (n+1)^2\}$ is a basis of $\Pi_n(S^2)$. It turns out that this problem is equivalent to Problem 1. The following proposition is folklore and can be proved easily using the formulae for K_n and Y_n given above.

Proposition 1.1. *Let $X = \{\mathbf{a}_i : 1 \le i \le (n+1)^2\}$ be a set of distinct points. Then X is a solution of Problem* 1 *if and only if $\{K_n(\mathbf{x},\mathbf{a}_i), 1 \le i \le (n+1)^2\}$ forms a basis of $\Pi_n(S^2)$.*

The problem of $\{K_n(\mathbf{x},\mathbf{a}_i)\}$ being a basis has been considered recently in [4] for possible application in wavelets, since such a basis is better localized than the usual orthonormal basis. The interpolation problem is also closely related to constructing cubature formulae on the sphere (see section 3), and there have been considerable efforts computing good interpolation points for that purpose; see [12] and the references therein. Recently, the polynomial interpolation has also been studied in [6] in connection with radial basis interpolation on the sphere and in [11] in connection with scattered data interpolation on the sphere. In [15], we showed that there is a close relation between interpolation on S^d and on the unit ball B^d of $\mathbb{R}^d$. In particular, it shows that many sets of points that allow unique polynomial interpolation on $B^2 = \{(x,y) : x^2+y^2 \le 1\}$ can be used to generate symmetric points on S^2 that solve Problem 1, where symmetry means that the points are symmetric with respect to a coordinate plane. The interpolation on B^2 has been studied in [1, 2] recently. The result in [1] leads to a set of symmetric points on S^2 in which points are equidistant points on parallel circles, and each circle has the same number of points. However, the more general result in [2] cannot be used to give points on S^2, since the correspondence between points on B^2 and on S^2 requires that there are $n+1$ points on the boundary circle of B^2.

In the following section, we present a theorem that gives many sets of points that solve Problem 1. The $(n+1)^2$ points all lie on $n+1$ distinct latitudes (parallel circles on S^2), and on each latitude there are an odd number of equidistant points. The number of points need not be the same on each latitude, and there is no restriction on the distribution of latitudes. The proof relies on a method of factorization, which is closely related to the method used for polynomial interpolation on the unit ball B^2 in [2, 7]. The use of equidistant points allows us to reduce the problem of interpolation on the sphere to an interpolation problem by a family of special trigonometric polynomials, which turns out to be equivalent to a Hermite–Birkhoff interpolation by trigonometric polynomials.

As far as we know, apart from a result that is a simple consequence of Bezout's theorem (see Proposition 2.1 below), these families of points are first examples of poised interpolation points that are given explicitly for all n.

The paper is organized as follows. We prove the factorization theorem in the following section. The main result and various examples are given in section 3, which also includes a result on cubature formulae. The related trigonometric Hermite–Birkhoff interpolation is discussed in section 4.

2. Preliminary and results on factorization of polynomials. On S^2 it is more convenient to work with spherical coordinates:

$$x = \sin\theta\sin\phi, \quad y = \sin\theta\cos\phi, \quad z = \cos\theta, \qquad 0 \le \phi \le 2\pi, \quad 0 \le \theta \le \pi.$$

For a polynomial $T \in \Pi(S^2)$ we introduce the notation $\widetilde{T}$ defined by

$$\widetilde{T}_n(\theta,\phi) = T_n(\sin\theta\sin\phi, \sin\theta\cos\phi, \cos\theta), \quad 0 \le \phi \le 2\pi, \quad 0 \le \theta \le \pi.$$

If $X = \{(x_i, y_i, z_i) : 1 \le i \le M\}$ is a set of points on S^2, we also use the notation $\widetilde{X} = \{(\theta_i, \phi_i) : 1 \le i \le M\}$ for the corresponding set in spherical coordinates.

The notation $S^2(a) := \{(x,y,z) : (x,y,z) \in S^2,\ z = a\}$, $-1 < a < 1$, denotes the circle on S^2 resulting from the intersection of S^2 with the plane $z = a$ (called the latitude at $z = a$).

An orthogonal basis of $\mathcal{H}_n^3$ can be given in terms of the Gegenbauer polynomials. In spherical coordinates, define

$$\begin{aligned} Y_{k,n}^{(1)}(x,y,z) &= C_{n-k}^{k+1/2}(\cos\theta)(\sin\theta)^k\cos k\phi, \qquad 0 \le k \le n, \\ Y_{k,n}^{(2)}(x,y,z) &= C_{n-k}^{k+1/2}(\cos\theta)(\sin\theta)^k\sin k\phi, \qquad 1 \le k \le n. \end{aligned}$$

Then $\mathcal{H}_n^3 = \operatorname{span}\{Y_{k,n}^{(1)},\ 0 \le k \le n, \text{ and } Y_{k,n}^{(2)},\ 1 \le k \le n\}$. Since $\Pi_n(S^2) = \bigoplus_{k=0}^n \mathcal{H}_k^3$, every polynomial $T_n \in \Pi_n(S^2)$ can be written as

$$T_n(x,y,z) = \sum_{j=0}^{n}\sum_{k=0}^{j} C_{j-k}^{k+1/2}(\cos\theta)(\sin\theta)^k(a_{k,j}\cos k\phi + b_{k,j}\sin k\phi).$$

Changing the order of sums in the above expression, we see that

$$\widetilde{T}_n(\theta,\phi) = a_0(\cos\theta) + \sum_{k=1}^{n}[a_k(\cos\theta)(\sin\theta)^k\cos k\phi + b_k(\cos\theta)(\sin\theta)^k\sin k\phi], \tag{2.1}$$

where $a_k(t)$ and $b_k(t)$ are polynomials of degree $n-k$ in one variable,

$$a_k(\cos\theta) = \sum_{j=0}^{n-k} a_{k,k+j} C_j^{k+1/2}(\cos\theta), \quad \text{and} \quad b_k(\cos\theta) = \sum_{j=0}^{n-k} b_{k,k+j} C_j^{k+1/2}(\cos\theta).$$

Since C_j^λ is a polynomial of degree exactly j, a_k and b_k are generic polynomials of degree $n-k$. Notice that for a fixed θ, the polynomial $\widetilde{T}(\theta,\phi)$ is a trigonometric polynomial of degree n in ϕ. We will use the fact that the interpolation on $2n+1$ distinct points inside $[0,2\pi)$ by trigonometric polynomials of degree n is unique.

The proof of our main result is based on a factorization of the polynomial that vanishes on the interpolation points. In the simplest case, an analogue of such a factorization is akin to the Bezout theorem for algebraic curves. That Bezout's theorem can be used to establish the uniqueness of the interpolation is folklore (see, for example, [5]). For interpolation on the sphere, we use it to give a proof of the following proposition (see [6]), which gives a prelude of the factorization method that leads to our main result.

Proposition 2.1. *Let $z_0, z_1, \ldots, z_n$ be $n+1$ distinct elements in $(-1,1)$. If X consists of $2k+1$ distinct points on the latitude $S^2(z_k)$ for $0 \le k \le n$, then X solves Problem 1 in $\Pi_n(S^2)$.*

Proof. It is sufficient to prove that if $T_n \in \Pi_n(S^2)$ vanishes on X, then T_n is identically zero. Using the expression of $\widetilde{T}_n$ in (2.1) and the fact that T_n vanishes on $2n+1$ points of $S^2(z_n)$, it follows from the uniqueness of the trigonometric interpolation that $a_k(z_n) = 0$ and $b_k(z_n) = 0$, $0 \le k \le n$. Consequently, $a_n(z) = b_n(z) = 0$ and $a_k(z) = (z - z_n)a_k^*(z)$ and $b_k(z) = (z - z_n)b_k^*(z)$ for $0 \le k \le n-1$ so that T_n satisfies a factorization $T_n(x,y,z) = (z - z_n)T_{n-1}(x,y,z)$, where $T_{n-1} \in \Pi_{n-1}(S^2)$. Evidently, we can continue this process for $z_{n-1}, \ldots, z_0$ and conclude that $T_n(x,y,z) = (z - z_n)\ldots(z - z_0)T^*(x,y,z)$. However T_n is a polynomial of degree n so that $T_n \equiv 0$. □

The analogue of this proposition holds for higher dimensional spheres, as shown in [6]. The points in the above proposition have little symmetry on S^2, since no two latitudes have the same number of points. It is worthwhile to emphasize that the position of the points on each latitude is completely arbitrary. In our result below, the points on each latitude are equidistant points. In essence, our main result is based on a more general factorization that holds, however, only for equidistant points.

Let $\Theta_{\alpha,m}$ denote a set of $2m+1$ equidistant points,

$$\Theta_{\alpha,m} = \{\theta_j^\alpha : \theta_j^\alpha = (2j+\alpha)\pi/(2m+1), \quad j = 0,1,\ldots,2m, \quad 0 \le \alpha < 2\}. \tag{2.2}$$

Under the mapping $\phi \mapsto e^{i\phi}$, these points can be considered as points on the unit circle. The presence of the number α means that the equidistant points are defined up to a rotation. The following simple fact plays an important role in the development below.

Lemma 2.2. *Let $n = 2m$ or $n = 2m-1$. For $\phi \in \Theta_{\alpha,m}$,*

$$\widetilde{T}_n(\theta,\phi) = a_0(\cos\theta) + \sum_{k=1}^{m} \left[\left(a_k(\cos\theta)(\sin\theta)^k + u_{2m-k+1}(\cos\theta)(\sin\theta)^{2m-k+1}\right)\cos k\phi \right.$$
$$\left. + \left(b_k(\cos\theta)(\sin\theta)^k + v_{2m-k+1}(\cos\theta)(\sin\theta)^{2m-k+1}\right)\sin k\phi\right],$$

where

$$\begin{aligned} u_{2m-k+1}(\cos\theta) \ &= a_{2m-k+1}(\cos\theta)\cos\alpha\pi + b_{2m-k+1}(\cos\theta)\sin\alpha\pi, \\ v_{2m-k+1}(\cos\theta) \ &= a_{2m-k+1}(\cos\theta)\sin\alpha\pi - b_{2m-k+1}(\cos\theta)\cos\alpha\pi, \end{aligned}$$

and we assume that $a_{2m}(t) = b_{2m}(t) = 0$ *if* $n = 2m - 1$.

Proof. The proof amounts to using the fact that

$$\begin{aligned} \cos(2m-k+1)\phi \ &= \cos(\alpha\pi - k\phi) = \cos\alpha\pi\cos k\phi + \sin\alpha\pi\sin k\phi, \\ \sin(2m-k+1)\phi \ &= \sin(\alpha\pi - k\phi) = \sin\alpha\pi\cos k\phi - \cos\alpha\pi\sin k\phi \end{aligned}$$

for $\phi \in \Theta_{\alpha,m}$, and we rewrite $\widetilde{T}(\theta,\phi)$ accordingly. $\square$

In particular, using the uniqueness of trigonometric interpolation, this shows that if $\widetilde{T}_n(\theta_j,\phi_i) = 0$ for $\phi_i \in \Theta_{\alpha,m}$, $0 \le i \le 2m$, then $a_0(\theta_j) = 0$ and

$$\begin{aligned} &a_k(\cos\theta_j)(\sin\theta_j)^k + u_{2m-k+1}(\cos\theta_j)(\sin\theta_j)^{2m-k+1} = 0, \\ &b_k(\cos\theta_j)(\sin\theta_j)^k + v_{2m-k+1}(\cos\theta_j)(\sin\theta_j)^{2m-k+1} = 0. \end{aligned} \tag{2.3}$$

Hence, we need to consider the poisedness of interpolation by trigonometric polynomials of the form

$$p(\cos\theta)(\sin\theta)^k + q(\cos\theta)(\sin\theta)^{2m-k+1},$$

where $p(t)$ and $q(t)$ are polynomials of degree $n-k$ and $n-2m+k-1$, respectively. For this purpose, we need a lemma which is elementary but somewhat unexpected. It contains the following elementary formula as its simplest case:

$$\frac{d^5}{dt^5}\Big[(1-t^2)^{3/2}(at+b)\Big] = 45\frac{a+bt}{(1-t^2)^{7/2}}.$$

It is this simple formula, stumbled upon using a computer algebra system, that leads to the lemma below. We will use the Pochhammer symbol $(a)_n = a(a+1)\dots(a+n-1)$.

LEMMA 2.3. *Let* m *and* k *be positive integers, and let* $1 \le k \le 2m$. *Let* $q_{k-1}(t)$ *be an algebraic polynomial of degree* $k-1$. *Then*

$$(1-t^2)^{m+1/2}\Big(\frac{d}{dt}\Big)^{2m-k+1}\Big((1-t^2)^{m-k+1/2}q_{k-1}(t)\Big) = q^*_{k-1}(t),$$

where q^*_{k-1} *is a polynomial of degree* $k-1$ *such that* $q_{k-1}(t) = 0$ *if and only if* $q^*_{k-1}(t) = 0$. *In fact, if* $q_{k-1}(t) = \sum_{l=0}^{k-1} a_l(1-t)^l$, *then* q^*_{k-1} *is given by*

$$q^*_{k-1}(t) = (-1)^{k-1}2^{2m-k+1}\sum_{l=0}^{k-1}(-m+l+1/2)_{2m-k+1}\, a_l\, s_l(t),$$

where s_l *are polynomials of degree* $k-1$ *that are defined by*

$$s_l(t) = (1-t)^l\sum_{j=0}^{k-l-1}\frac{(-k+l+1)_j(k-2m-1)_j}{(-m+l+1/2)_j j!}\Big(\frac{1-t}{2}\Big)^j.$$

Proof. We need to recall the definition of Jacobi polynomials. Let α and β be two arbitrary real numbers. Then the Jacobi polynomial $P_n^{(\alpha,\beta)}$ is given by Rodrigue's formula [14, p. 67, (4.3.1)],

$$P_n^{(\alpha,\beta)}(t) = \frac{(-1)^n}{n!2^n}(1-t)^{-\alpha}(1+t)^{-\beta}\Big(\frac{d}{dt}\Big)^n\left[(1-t)^{n+\alpha}(1+t)^{n+\beta}\right].$$

Let $q_{k-1}(t) = \sum_{l=0}^{k-1} a_l(1-t)^l$. It follows that

$$\begin{aligned}\phi(t) :=&(1-t^2)^{m+1/2}\Big(\frac{d}{dt}\Big)^{2m-k+1}\Big((1-t^2)^{m-k+1/2}q_{k-1}(t)\Big)\\ =&\sum_{l=0}^{k-1} a_l(1-t^2)^{m+1/2}\Big(\frac{d}{dt}\Big)^{2m-k+1}\Big((1-t)^{m-k+l+1/2}(1+t)^{m-k+1/2}\Big).\end{aligned}$$

Using Rodrigue's formula with $n = 2m-k+1$, $\alpha = -m+l-1/2$, and $\beta = -m-1/2$, we get

$$\phi(t) = \sum_{l=0}^{k-1} a_l(-1)^{k-1}2^{2m-k+1}(2m-k+1)!(1-t)^l P_{2m-k+1}^{(-m-1/2+l,-m-1/2)}(t).$$

It is known that the Jacobi polynomials with negative parameters satisfy the following relation [14, p. 64, (4.22.3)]:

$$\binom{n}{j-1}P_n^{(\alpha,\beta)}(t) = \binom{n+\alpha}{n-j+1}P_{j-1}^{(\alpha,\beta)}(t) \qquad \text{if} \quad n+\alpha+\beta+j=0,$$

which shows that $P_n^{(\alpha,\beta)}(t)$ is in fact a polynomial of degree $j-1$ with $j=-n-\alpha-\beta$. The condition $n+\alpha+\beta+j=0$ is satisfied in our case with n,α,β as defined above and $j=k-l$. Consequently, we conclude that

$$\begin{aligned}\phi(t) = \sum_{l=0}^{k-1} a_l(-1)^{k-1}2^{2m-k+1}(k-l-1)!&\frac{(-m+l+1/2)_{2m-k+1}}{(-m+l+1/2)_{k-l-1}}\\ &\times(1-t)^l P_{k-l-1}^{(-m-1/2+l,-m-1/2)}(t),\end{aligned}$$

after some simplification of the constants. This shows that ϕ is indeed a polynomial of degree $k-1$, which we called q_{k-1}^* in the statement. The explicit formula of q_{k-1}^* is derived from the explicit formula of $P_n^{(\alpha,\beta)}(t)$ [14, p. 62, (4.21.2)],

$$P_n^{(\alpha,\beta)}(t) = \binom{n+\alpha}{n}{}_2F_1\left(-n, n+\alpha+\beta+1;\alpha+1;\frac{1-t}{2}\right),$$

where ${}_2F_1$ is the hypergeometric function defined by

$${}_2F_1(a,b;c;z) = \sum_{k=0}^{\infty}\frac{(a)_k(b)_k}{(c)_k k!}z^k, \qquad |z|<1.$$

Finally, let $s_l(t)$ be the polynomials in the statement. Writing $s_0(t), s_1(t), \ldots, s_{k-1}(t)$ in terms of $1, (1-t), \ldots, (1-t)^{k-1}$, the transition matrix is triangular with nonzero diagonal elements. Hence, the polynomials $s_0, s_1, \ldots, s_{k-1}$ are linearly independent. Consequently, $q_{k-1}^*(t) = 0$ if and only if $q_{k-1}(t) = 0$. $\square$

Lemma 2.4. *Let k and m be positive integers, and let $k \leq 2m$. Let p_{2m-k} be a polynomial of degree $2m-k$, and let q_{k-1} be a polynomial of degree $k-1$. If $\phi(t) = p_{2m-k}(t) + (1-t^2)^{m-k+1/2} q_{k-1}(t)$ vanishes on $2m+1$ distinct points in $[-1,1]$, then $\phi(t) \equiv 0$.*

Proof. Since ϕ has $2m+1$ zeros, Rolle's theorem implies that $\phi^{(2m-k+1)}(t)$ has at least k zeros inside $(-1,1)$. Using the previous lemma,

$$\phi^{(2m-k+1)}(t) = \left(\frac{d}{dt}\right)^{2m-k+1} \left[(1-t^2)^{m-k+1/2} q_{k-1}(t)\right] = (1-t^2)^{-m-1/2} q^*_{k-1}(t)$$

so that $q^*_{k-1}(t)$ has k zeros inside $(-1,1)$. Since q^*_{k-1} is a polynomial of degree $k-1$, we have $q^*_{k-1}(t) \equiv 0$. Consequently, $q_{k-1}(t) \equiv 0$ by the previous lemma. Therefore, $\phi(t) = p_{2m-k}(t)$, which must be zero since it is a polynomial of degree $2m-k$ and it vanishes on $2m+1$ points. □

Corollary 2.5. *Let k and m be positive integers, and let $k \leq 2m$. Let p_{2m-k} be a polynomial of degree $2m-k$, and let q_{k-1} be a polynomial of degree $k-1$. If*

$$\psi(\theta) = p_{2m-k}(\cos\theta)(\sin\theta)^k + (\sin\theta)^{2m-k+1} q_{k-1}(\cos\theta)$$

vanishes on $2m+1$ distinct points in $(0,\pi)$, then $\psi(\theta) \equiv 0$.

Proof. Using the fact that $\sin\theta$ is positive on $(0,\pi)$, the stated result follows from the previous proposition with $\phi(t) = \psi(\theta)/(\sin\theta)^k$ and $t = \cos\theta$. □

There is another way to state the result in Corollary 2.5. A system of functions $\{g_1, \dots, g_n\}$ is called a Chebyshev system on the interval $(a,b) \subset \mathbb{R}$ if for any set of nonzero real numbers $c_1, \dots, c_n$ the function $c_1 g_1 + \cdots + c_n g_n$ has at most $n-1$ zeros in (a,b).

Corollary 2.6. *Let k and m be integers, and let $k \leq 2m$. The system of functions*

$$(\sin\theta)^k\{1, \cos\theta, \dots, \cos(2m-k)\theta\} \cup (\sin\theta)^{2m-k+1}\{1, \cos\theta, \dots, \cos(k-1)\theta\}$$

is a Chebyshev system on $(0,\pi)$. Equivalently, the system of functions

$$\{1, t, \dots, t^{2m-k}\} \cup (1-t^2)^{2m-2k+1}\{1, t, \dots, t^{k-1}\}$$

is a Chebyshev system on $(0,1)$.

Even in the case of $k=1$, which states that $\{1, \cos\theta, \dots, \cos(2m-1)\theta, (\sin\theta)^{2m-1}\}$ is a Chebyshev system on $(0,\pi)$, this corollary is not obvious.

The following factorization theorem holds the key to our main result.

Theorem 2.7. *Let m be an integer, let $m \leq s \leq 2m+1$, and let $n = 2m$ or $n = 2m-1$. Let $\theta_0, \theta_1, \dots, \theta_{2\lambda}$ be distinct numbers in $(0,\pi)$, where $\lambda = s-m$. If $T \in \Pi_s(S^2)$ satisfies*

$$\widetilde{T}(\theta_j, \phi_i) = 0, \qquad 0 \leq j \leq 2\lambda, \quad 0 \leq i \leq 2m, \quad \phi_i \in \Theta_{\alpha,m},$$

where α is a number in $[0,2)$, then there is a spherical polynomial $T^* \in \Pi_{n-2\lambda-1}(S^2)$ such that

$$T(x,y,z) = \prod_{j=0}^{2\lambda} (z - \cos\theta_j) T^*(x,y,z).$$

In particular, $T^*(x,y,z) = 0$ if $s = 2m$ or $s = 2m-1$.

Proof. Using the formula (2.1), we write

$$\widetilde{T}(\theta,\phi) = a_0(\cos\theta) + \sum_{k=1}^{s} \left[a_k(\cos\theta)(\sin\theta)^k \cos k\phi + b_k(\cos\theta)(\sin\theta)^k \sin k\phi\right],$$

where a_k and b_k are polynomials of degree $s-k$. By Lemma 2.2, for $\phi_i \in \Theta_{\alpha,m}$,

$$\begin{aligned}\widetilde{T}(\theta_j,\phi_i) = a_0(\cos\theta_j) &+ \sum_{k=1}^{m} (\sin\theta_j)^k \left(a_k(\cos\theta_j)\cos k\phi_i + b_k(\cos\theta_j)\sin k\phi_i\right) \\ &+ \sum_{k=2m+1-s}^{m} \big(u_{2m-k+1}(\cos\theta_j)(\sin\theta_j)^{2m-k+1}\cos k\phi_i \\ &\qquad + v_{2m-k+1}(\cos\theta_j)(\sin\theta_j)^{2m-k+1}\sin k\phi_i\big)\end{aligned}$$

for $0 \le i \le 2m$ and $0 \le j \le 2\lambda$. For each fixed j, $\widetilde{T}(\theta_j,\phi_i) = 0$ shows that the trigonometric polynomial $\widetilde{T}(\theta_j,\cdot)$ of degree m vanishes on $2m+1$ points; the uniqueness of the trigonometric interpolation implies the following two cases.

Case 1: $0 \le k \le 2m-s$.

$$a_k(\cos\theta_i)(\sin\theta_j)^k = 0, \quad b_k(\cos\theta_j)(\sin\theta_j)^k = 0, \qquad 0 \le j \le 2\lambda,$$

where we assume $b_0(\cos\theta) = 0$. Since $\theta_j \in (0,\pi)$, this shows that $a_k(\cos\theta_j) = 0$ and $b_k(\sin\theta_j) = 0$. Consequently, there exist polynomials $a_k^*(t)$ and $b_k^*(t)$, both of degree $n-k-(2\lambda+1)$, such that $a_k(t) = \prod_{j=0}^{2\lambda}(t-\cos\theta_j)a_k^*(t)$ and $b_k(t) = \prod_{j=0}^{2\lambda}(t-\cos\theta_j)b_k^*(t)$.

Case 2: $2m-s+1 \le k \le m$.

$$\begin{aligned} a_k(\cos\theta_j)(\sin\theta_j)^k + u_{2m-k+1}(\cos\theta_j)(\sin\theta_j)^{2m-k+1} &= 0, \\ b_k(\cos\theta_j)(\sin\theta_j)^k + u_{2m-k+1}(\cos\theta_j)(\sin\theta_j)^{2m-k+1} &= 0, \end{aligned} \qquad 0 \le j \le 2\lambda.$$

By Corollary 2.5, this shows that $a_k(t) = b_k(t) = 0$ and $u_{2m-k+1}(t) = v_{2m-k+1}(t) = 0$. The definition of u_{2m-k+1} and v_{2m-k+1} then shows

$$\begin{aligned} a_{2m-k+1}(\cos\theta)\cos\alpha\pi + b_{2m-k+1}(\cos\theta)\sin\alpha\pi &= 0, \\ a_{2m-k+1}(\cos\theta)\sin\alpha\pi - b_{2m-k+1}(\cos\theta)\cos\alpha\pi &= 0, \end{aligned}$$

which implies that $a_{2m-k+1}(t) = b_{2m-k+1}(t) = 0$.

Together these two cases show that we have the following factorization:

$$\begin{aligned}\widetilde{T}(\theta,\phi) &= \prod_{j=0}^{2\lambda}(\cos\theta - \cos\theta_j) \\ &\quad \times \left(a_0^*(\cos\theta) + \sum_{j=1}^{2m-s} \left[a_k^*(\cos\theta)(\sin\theta)^k\cos k\phi + b_k^*(\cos\theta)(\sin\theta)^k \sin k\phi\right]\right),\end{aligned}$$

which completes the proof. □

The fact that the interpolation points are equidistant on the circle is essential in this theorem. One important property of the factorization is that it allows us to repeat the argument to get a complete factorization of the polynomial.

THEOREM 2.8. *Let n and σ be positive integers. Let $\lambda_1, \dots, \lambda_\sigma$ be nonnegative integers. Define $n_k = n_{k-1} - (2\lambda_k + 1)$ for $1 \le k \le \sigma$ with $n_0 = n$. Assume that $n_k \ge 0$ for $1 \le k \le \sigma - 1$. If $T \in \Pi_n(S^2)$ satisfies*

$$\widetilde{T}(\theta_{j,k}, \phi_{i,k}) = 0, \quad 0 \le j \le 2\lambda_k,\ 0 \le i \le 2(n_{k-1} - \lambda_k),\ 1 \le k \le \sigma,$$

where $\theta_{j,k}$, $0 \le j \le 2\lambda_k$ and $1 \le k \le \sigma$, are distinct numbers in $(0, \pi)$ and $\phi_{i,k} \in \Theta_{\alpha_k, n_{k-1}-\lambda_k}$ with $\alpha_k \in [0, 2)$, then there exists a polynomial $T^* \in \Pi_{n_\sigma}(S^2)$ such that

$$T(x, y, z) = \prod_{k=1}^{\sigma} \prod_{j=0}^{2\lambda_k} (z - \cos\theta_{j,k}) T^*(x, y, z).$$

In particular, $T(x, y, z) \equiv 0$ if $n_\sigma < 0$.

Proof. We apply the factorization in the theorem repeatedly with $s = n_{k-1}$, $m = n_{k-1} - \lambda_k$, and $\lambda = \lambda_k$ for $k = 1, 2, \dots, \sigma$. □

We note that the interpolation points in the corollary are located on σ groups of latitudes $\{S^2(z_{j,k}) : 0 \le j \le 2\lambda_k\}$, $1 \le k \le \sigma$ and $z_{j,k} = \cos\theta_{j,k}$, and latitudes in different groups contain different number of nodes. More precisely, each of the latitudes in the kth group, $S^2(z_{0,k}), S^2(z_{1,k}), \dots, S^2(z_{2\lambda_k,k})$, contains $2(n_{k-1} - \lambda_k) + 1$ equidistant points.

3. Interpolation on the unit sphere. Our main result on interpolation follows from the result on factorization. The following formula is used to show that the interpolation condition matches the dimension of the polynomial space:

$$\dim \Pi_s(S^2) = \dim \Pi_{s-2\lambda-1}(S^2) + (2\lambda + 1)(2s - 2\lambda + 1). \tag{3.1}$$

THEOREM 3.1. *Let n and σ be positive integers such that $n + 1 - \sigma$ is an even integer and $\sigma \le n + 1$. Let $\lambda_1, \dots, \lambda_\sigma$ be nonnegative integers such that*

$$\lambda_1 + \cdots + \lambda_\sigma = \frac{n + 1 - \sigma}{2}. \tag{3.2}$$

Define $n_k = n_{k-1} - (2\lambda_k + 1)$ for $1 \le k \le \sigma - 1$ with $n_0 = n$. Let

$$\widetilde{X} = \{(\theta_{j,k}, \phi_{i,k}) : 0 \le j \le 2\lambda_k,\ 0 \le i \le 2(n_{k-1} - \lambda_k),\ 1 \le k \le \sigma\},$$

where $\theta_{j,k}$, $0 \le j \le 2\lambda_k$ and $1 \le k \le \sigma$, are distinct numbers in $(0, \pi)$ and $\phi_{i,k} \in \Theta_{\alpha_k, n_{k-1}-\lambda_k}$ with $\alpha_k \in [0, 2)$. Then the set X solves Problem 1 in $\Pi_n(S^2)$.

Proof. Again it is sufficient to prove that the dimension of $\Pi_n(S^2)$ matches the interpolation conditions, and if $T \in \Pi_n(S^2)$ vanishes on X, then $T(x, y, z) \equiv 0$. Under the condition (3.2), it follows that

$$n_\sigma := n_{\sigma-1} - (2\lambda_\sigma + 1) = n - (2\lambda_1 + 1) - \cdots - (2\lambda_\sigma + 1) = -1 < 0.$$

Hence, the factorization theorem in Theorem 2.8 shows that $T(x, y, z) \equiv 0$. Moreover, it follows from (3.1) that

$$\dim \Pi_n(S^2) = (n + 1)^2 = \sum_{k=1}^{\sigma} (2\lambda_k + 1)(2n_{k-1} - 2\lambda_k + 1)$$

so that the interpolation condition matches the dimension of $\Pi_n(S^2)$. □

For a fixed n this theorem contains a number of different interpolation processes. In fact, for each positive integer n, the number of sets X contained in Theorem 3.1 depends on the number of nonnegative integer solutions of (3.2), where σ is also a variable which satisfies the condition that $n+1-\sigma$ is a nonnegative even integer. Every such solution of (3.2) leads to a set of interpolation points that solves Problem 1. The number of solutions of such an equation grows exponentially as n goes to infinity. Moreover, the order of $\lambda_1, \ldots, \lambda_\sigma$ matters; that is, different permutations of a solution $\lambda_1, \ldots, \lambda_\sigma$ of (3.2) give different sets of interpolation points.

Among the solutions of (3.2), one extreme case is $\sigma = n+1$, for which the equation has only one solution, $\lambda_1 = \cdots = \lambda_{n+1} = 0$. In this case, $n_k = n-k$, and the interpolation points are located on $n+1$ latitudes $S^2(z_0)$, $S^2(z_1), \ldots, S^2(z_n)$ and the latitude $S^2(z_k)$ contains $2k+1$ points. Hence, this is just a special case of Proposition 2.1.

The other extreme case is $\sigma = 1$. In this case, n needs to be even to keep $n+1-\sigma$ even. Assume $n = 2m$. There is only one solution of the equation, which is $\lambda_1 = (n-\sigma+1)/2 = m$. We state this case as a corollary.

Corollary 3.2. *Let m be a positive integer. Let $\theta_0, \theta_1, \ldots, \theta_{2m}$ be $2m+1$ distinct numbers in $(0,\pi)$. Let X be defined by*

$$\widetilde{X} = \{(\theta_i, \phi_i) : 0 \le j \le 2m,\ 0 \le i \le 2m\},$$

where $\phi_i \in \Theta_{\alpha,m}$ with $\alpha \in [0, 2\pi)$. Then X solves Problem 1 in $\Pi_{2m}(S^2)$.

In this case, the interpolation points are located on $2m+1$ latitudes $S^2(z_i)$ with $z_i = \cos\theta_i$; each latitude has $2m+1$ equidistant points. In the special case that $\{\theta_0, \ldots, \theta_{2m}\}$ are symmetric (that is, $\theta_{2m-i} = \pi - \theta_i$), the corollary has been established in [15] using the fact that the uniqueness of the interpolation on S^2 of the symmetric set can be related to the interpolation on the unit disc B^2 of $\mathbb{R}^2$, which allows us to use the result in [1] for interpolation on B^2. The first nontrivial case of this corollary is $m = 1$, for which there are 3 latitudes, each with 3 points.

To illustrate the main result, we list all sets of interpolation points contained in Theorem 3.1 for small n. For $n = 1$, there is only one set, for which points are located on 2 latitudes with 3 points and 1 point, respectively, which is a special case of Proposition 2.1. The case $n = 2$ contains two cases, $\sigma = 1$ and $\sigma = 3$; the case $\sigma = 1$ is a special case of Corollary 3.2 with 3 latitudes and each with 3 points; the case $\sigma = 3$ is a special case of Proposition 2.1 of 3 latitudes with 5, 3, and 1 points, respectively. In general, for each σ there can be multiple solutions of (3.2) and the order matters. The cases of $n = 3$ and $n = 4$ are given below.

Example 3.3. $n = 3$. *That $n+1-\sigma$ is even implies that σ can be either 2 or 4. The set X contains 16 points.*

1. $\sigma = 2$: $\lambda_1 + \lambda_2 = 1$ *has two solutions (order matters).*
   (a) $\lambda_1 = 1, \lambda_2 = 0$: 3 *latitudes each with* 5 *points and* 1 *latitude with* 1 *point;*
   (b) $\lambda_1 = 0, \lambda_2 = 1$: 1 *latitude with* 7 *points and* 3 *latitudes each with* 3 *points.*
2. $\sigma = 4$: $\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 = 0$ *has only one solution, $\lambda_i = 0$. This is the special case of Proposition* 2.1 *of* 4 *latitudes with* 7, 5, 3, *and* 1 *points, respectively.*

Example 3.4. $n = 4$. *That $n+1-\sigma$ is even implies that σ can be either* 1, 3, *or* 5. *Here X contains* 25 *points.*

1. $\sigma = 1$: $\lambda_1 = 5$ *is a special case of Corollary* 3.2; 5 *latitudes each with* 5 *points.*
2. $\sigma = 3$: $\lambda_1 + \lambda_2 + \lambda_3 = 1$ *has three solutions.*

(a) $\lambda_1 = 1, \lambda_2 = 0, \lambda_3 = 0$: 3 *latitudes each with* 7 *points,* 1 *latitude with* 3 *points, and* 1 *latitude with* 1 *point;*
(b) $\lambda_1 = 0, \lambda_2 = 1, \lambda_3 = 0$: 1 *latitude with* 9 *points,* 3 *latitudes each with* 5 *points, and* 1 *latitude with* 1 *point;*
(c) $\lambda_1 = 0, \lambda_2 = 0, \lambda_3 = 1$: 1 *latitude with* 9 *points,* 1 *latitude with* 7 *points, and* 3 *latitudes each with* 3 *points;*

3. $\sigma = 5$: $\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 + \lambda_5 = 0$ *has only one solution,* $\lambda_i = 0$. *This is the special case of Proposition* 2.1 *of* 5 *latitudes with* 9, 7, 5, 3, *and* 1 *points, respectively.*

Thus, for $n = 3$, Theorem 3.1 contains 3 sets of 16 points that solve Problem 1, and for $n = 4$ it contains 5 sets of 25 points. In the next two cases, $n = 5$ gives 8 sets of 36 points and $n = 6$ gives 13 sets of 49 points. In general, the number of different sets for each degree is a Fibonacci sequence. (I thank a referee for pointing out this fact.)

PROPOSITION 3.5. *Let* γ_n *denote the number of different sets in Theorem* 3.1 *that solves Problem* 1 *in* $\Pi_n(S^2)$. *Then* $\{\gamma_n\}$ *is a Fibonacci sequence:* $\lambda_n = \lambda_{n-1} + \lambda_{n-2}$, $\lambda_1 = 1$, *and* $\lambda_2 = 2$.

Proof. Denote by Ω_n the set of solutions of the equation $\lambda_1 + \cdots + \lambda_\sigma = (n + 1 - \sigma)/2$, where $\lambda_i \in \mathbb{N}_0$, $\sigma \in \mathbb{N}_0$ and $n + 1 - \sigma$ is even. Let $(\lambda_1, \ldots, \lambda_\sigma) \in \Omega_n$. If $\lambda_\sigma = 0$, then $\lambda_1 + \cdots + \lambda_{\sigma-1} = (n - 1 + 1 - (\sigma - 1))/2$ so that $(\lambda_1, \ldots, \lambda_{\sigma-1}) \in \Omega_{n-1}$, and this defines a one-to-one mapping from Ω_n to Ω_{n-1}. If $\lambda_\sigma > 0$, then $\lambda_1 + \cdots + \lambda_{\sigma-1} + (\lambda_\sigma - 1) = (n - 2 + 1 - \sigma)/2$ so that $(\lambda_1, \ldots, \lambda_{\sigma-1}, \lambda_\sigma - 1) \in \Omega_{n-2}$. Again this defines a one-to-one mapping from Ω_n to Ω_{n-2}. Since $\gamma_n = \#\Omega_n$, this shows that $\lambda_n = \lambda_{n-1} + \lambda_{n-2}$. □

Let us consider the interpolation process in Corollary 3.2 again. In this case, the interpolation points are on $2m + 1$ latitudes and each latitude has $2m + 1$ points. This seems to indicate that the interpolation polynomial should be a product type, that is, a product of two interpolation polynomials of one variable. However, since polynomials in $\Pi_n^2(S^2)$ have to have the form of (2.1), this is not the case. To find the formula for the interpolation polynomials, we need to find a formula for interpolation using the Chebyshev system in Corollary 2.6.

If we integrate an interpolation polynomial in Theorem 3.1 over the sphere, we get a cubature formula that is exact for polynomials of degree n; that is, the cubature formula is of degree n. In the case where points are those in Corollary 3.2, the formula takes a particularly simple form and can be explicitly given.

PROPOSITION 3.6. *Let* m *be a positive integer. Let* $\theta_0, \ldots, \theta_{2m}$ *be distinct numbers in* $(0, \pi)$, *and let* α *be a number in* $[0, 2)$. *Then for all* $f \in \Pi_{2m}(S^2)$,

$$\int_{S^2} f(x, y, z) d\omega = \sum_{j=0}^{2m} \frac{\lambda_j}{2m + 1} \sum_{i=0}^{2m} \widetilde{f}(\theta_j, \phi_i), \qquad \phi_i = \frac{(2i + \alpha)\pi}{2m + 1},$$

where λ_j *are given by*

$$\lambda_j = \int_{-1}^{1} \prod_{i=0, i \neq j}^{2m} \frac{t - \cos\theta_i}{\cos\theta_j - \cos\theta_i} dt.$$

Proof. Let the interpolation polynomial T_{2m} be of the form (2.1). We use the quadrature formula

$$\frac{1}{2\pi}\int_0^{2\pi}\tau(t)dt=\frac{1}{2m+1}\sum_{j=0}^{2m}\tau(\theta_j^\alpha),\qquad \theta_j^\alpha=\frac{(2j+\alpha)\pi}{2m+1}, \tag{3.3}$$

which is known to hold for every trigonometric polynomial of degree $2m$. For $\alpha=0$ this is the classical result in [16, Vol. 2, p. 8], and for $\alpha\neq 0$ it follows from

$$\int_0^{2\pi}\tau(\theta+t)dt=\int_0^{2\pi}\tau(t)dt,$$

which holds for every θ and for every trigonometric polynomial τ. Using the formula (3.3) and the interpolation property of T_{2m}, it follows that

$$a_0(\cos\theta_j)=\frac{1}{2\pi}\int_0^{2\pi}\widetilde{T}_{2m}(\theta_j,\phi)d\phi=\frac{1}{2m+1}\sum_{i=0}^{2m}\widetilde{f}(\theta_j,\phi_i),\qquad \phi_i=\frac{(2i+\alpha)\pi}{2m+1},$$

for every θ_j, $0\le j\le 2m$. Consequently, a_0 is uniquely determined by these interpolation conditions. It follows that

$$a_0(t)=\sum_{j=0}^{2m}\left(\frac{1}{2m+1}\sum_{i=0}^{2m}\widetilde{f}(\theta_j,\phi_i)\right)\ell_j(t),\qquad \ell_j(t)=\prod_{i=0,i\neq j}^{2m}\frac{t-\cos\theta_i}{\cos\theta_j-\cos\theta_i}.$$

Using the change of variable formula

$$\int_{S^2}f(x,y,z)d\omega=\int_0^\pi\int_0^{2\pi}\widetilde{f}(\theta,\phi)\sin\theta d\theta d\phi,$$

the integral of T_{2m} over the surface of the sphere is equal to

$$\int_{S^2}T_{2m}(x,y,z)d\omega=2\pi\int_0^\pi a_0(\cos\theta)\sin\theta d\theta=2\pi\int_{-1}^1 a_0(t)dt.$$

The stated formula follows from the formula of $a_0(t)$ derived above. □

In particular, if $\cos\theta_j$ are chosen so that λ_j are nonnegative, then the formula is a nonnegative cubature formula. For example, this holds if $\cos\theta_j$ are the zeros of the Legendre polynomial P_{2m+1} of degree $2m+1$ or the zeros of quasi-Legendre orthogonal polynomial $P_{2m+1}+\alpha P_{2m}$ with mild conditions imposed on $\alpha\in\mathbb{R}$. The nodes of such a formula are located on circles, just as the usual product-type formula. However, these cubature formulae are different from the usual product formulae. As it is well-known [13] that the usual way of deriving the product formula is to treat it as the product of Gaussian quadrature for trigonometric polynomials and the Gaussian quadrature formula for the unit weight on $[-1,1]$, the result works for all linear combinations of $\cos j\theta(a_k\cos k\phi+b_k\sin k\phi)$ of degree $2m+1$, which contains the spherical polynomials of degree $2m+1$ as a subset. In our case, the cubature formula is derived by interpolation, and it holds for spherical polynomials of degree up to $2m$, which are trigonometric polynomials that are of the special form (2.1). Our formula uses twice as many nodes as the usual product formula, but we still have the freedom of choosing the latitudes (that is, θ_j) on which the interpolation points lie and, in the case of several groups of latitudes, the rotations (that is, α_k) of points on the latitudes for different groups. A proper choice of θ_j and α_k may lead to some cubature formula of higher degree. Furthermore, every set of interpolation points in Theorem 3.1 leads to a cubature formula of degree n, and one can ask the question of how to find a cubature formula of the highest degree by choosing proper θ_j and α_k.

4. Trigonometric interpolation. In order to construct a formula for the interpolation polynomial in Theorem 3.1, we need to understand the interpolation by trigonometric polynomials in Corollary 2.6. It turns out that this interpolation is closely related to a nontrivial Hermite–Birkhoff interpolation problem by trigonometric polynomials, which we discuss in this section.

Let $\mathcal{T}_n$ denote the space of trigonometric polynomials of degree n,

$$\mathcal{T}_n = \operatorname{span}\left\{ a_0 + \sum_{k=1}^{n} (a_k \cos\theta + b_k \sin\theta) : a_k, b_k \in \mathbb{R} \right\}.$$

A Birkhoff interpolation problem is usually described using the notion of *incidence matrices*, which are matrices whose entries are 0 and 1. Let $E = (e_{l,j})$ be such a matrix with s rows $l = 1, 2, \ldots, s$ and n columns with $j = 0, 1, \ldots, n-1$. Then

$$T^{(j)}(t_l) = f_{l,j} \qquad \text{if} \quad e_{l,j} = 1, \quad T \in \mathcal{T}_n,$$

describes an interpolation process on the points $t_1, \ldots, t_s$. An incidence matrix E of $s \times n$ is said to satisfy the Pólya condition if

$$\sum_{k=1}^{j} \sum_{l=1}^{s} e_{l,k} \geq j+1, \qquad j = 0, 1, \ldots, n.$$

A sequence in an incidence matrix is a sequence of consecutive 1's in a row of E, say the lth row with $e_{l,k} = 1$ for $k = i+1, \ldots, i+j$, $e_{l,i} = 0$, and $e_{l,i+j+1} = 0$, and it is an odd sequence if j is odd. A supported sequence is a sequence such that there are nonzero elements of E in both its upper and lower left sides; that is, there are $e_{i_1,j_1} = 1$ and $e_{i_2,j_2} = 1$ with $i_1 < l$, $i_2 > l$, $j_1 \leq i$, and $j_2 \leq i$. For example, the matrix

$$E = \begin{bmatrix} 1 & 0 & 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 & 0 \end{bmatrix}$$

has, in its second row, two supported odd sequences of length 3 and 1, respectively. Since the trigonometric polynomials are periodic, the interpolation matrix should be considered periodic: Its last row should precede the first row; that is, the points in Θ are considered to be circular in the sense that the last point θ_s is in front of the first point θ_1. For example, if the matrix E in the above example is an incidence matrix for trigonometric interpolation, then its first row contains an odd supported sequence. Thus, for the interpolation by trigonometric polynomials, the supported sequence is a sequence that does not start from the first row, assuming that the matrix satisfies the Pólya condition. Although taking derivatives does not reduce the degree of a trigonometric polynomial, the analogue of the Atkinson–Sharma theorem ([8] or [9, p. 23]) still holds, giving the above understanding of the support sequence. The theorem states that a Birkhoff interpolation is poised if all its odd supported sequences begin in column 0 and E satisfies the Pólya condition.

We consider the following Hermite–Birkhoff interpolation problem.

Problem 2. Let m be a positive integer. Let $\Theta = \{\theta_i : 1 \leq i \leq 2m+1\}$ be a set of distinct points in $(0, \pi)$. Let k be an integer such that $1 \leq k \leq 2m$. Find conditions on Θ such that the interpolation problem

$$\begin{aligned} &T(\theta_i) = f_i, \qquad i = 1, 2, \dots 2m+1, \\ &T^{(j)}(0) = g_{0,j}, \quad T^{(j)}(\pi) = g_{\pi,j}, \quad j = 0, 1, \dots, k-2, k-1 \quad \text{and} \\ &\qquad\qquad\qquad\qquad\qquad\qquad j = k+1, k+3, \dots, 2m-k-1, \end{aligned}$$

has a unique solution in $\mathcal{T}_{2m}$ for any data f_i, $g_{0,j}$, and $g_{\pi,j}$.

Let Θ be the set as above. Assume that $0 < \theta_1 < \theta_2 < \cdots < \theta_{2m+1} < \pi$, and define $\theta_0 = 0$ and $\theta_{2m+2} = \pi$. Then the above problem is described by the incidence matrix E with $2m+3$ rows, whose elements $e_{l,j}$ are defined by

$$e_{0,j} = e_{2m+2,j} = 1, \qquad 0 \le j \le k-1 \ \text{ and } \ j = k+1, k+3, \dots, 2m-k-1,$$

and $e_{l,0} = 1$ for $l = 1, 2, \dots, 2m+1$; all other $e_{l,j}$ are 0. By the definition, there are exactly $2(2m)+1$ interpolation conditions, which is the same as the dimension of $\mathcal{T}_{2m}$. Clearly, the matrix E for this interpolation satisfies the Pólya condition. This matrix, however, has many odd supported sequences that do not begin in column 0; the analogue of the Atkinson–Sharma theorem does not apply. It turns out that Corollary 2.5 can be used to show that it is poised if Θ is a subset of $(0, \pi)$. The key ingredient is the following lemma.

LEMMA 4.1. *Let m be a positive integer. The trigonometric polynomial $T \in \mathcal{T}_{2m}$ satisfies the conditions*

$$\begin{aligned} T^{(j)}(0) = T^{(j)}(\pi) = 0, \quad & j = 0, 1, \dots, k-2, k-1 \quad \textit{and} \\ & j = k+1, k+3, \dots, 2m-k-1, \end{aligned}$$

for a fixed integer k, $0 \le k \le 2m$, if and only if T takes the form

$$T(\theta) = (\sin\theta)^k p(\cos\theta) + (\sin\theta)^{2m-k+1} q(\cos\theta), \tag{4.1}$$

where p is a polynomial of degree $2m-k$ and q is a polynomial of $k-1$.

Proof. First assume that T is of the special form (4.1). Let $T_1(\theta) = (\sin\theta)^k p(\cos\theta)$ and $T_2(\theta) = (\sin\theta)^{2m-k+1} q(\cos\theta)$. Since T_1 is an odd function if k is odd and an even function if k is even, and an odd trigonometric polynomial vanishes at 0 and π, it follows from taking derivatives that

$$T_1^{(j)}(0) = T_1^{(j)}(\pi) = 0, \quad 0 \le j \le k-1 \ \text{ and } \ j = k+1, k+3, k+5, \dots .$$

Similarly, since T_2 is odd if k is odd and even if k is even, it follows that

$$T_2^{(j)}(0) = T_2^{(j)}(\pi) = 0, \quad 0 \le j \le 2m-k \ \text{ and } \ j = 2m-k+2, 2m-k+4, \dots .$$

Since $T = T_1 + T_2$ and $0 \le k \le m$, the stated result follows from the above two displayed equations.

On the other hand, the conditions $T^{(j)}(0) = T^{(j)}(\pi) = 0$ for $0 \le j \le k-1$ imply that

$$T(\theta) = (\sin\theta)^k S(\theta), \qquad S \in T_{2m-k}.$$

This can be considered as a simple consequence of the fact that the Hermite interpolation problem is unique for the trigonometric interpolation (consider the Hermite interpolation conditions on 0 and π, together with Lagrange interpolation on $2(2m-k)+1$

distinct points). Applying the Leibniz rule to T implies that

$$T^{(k+2l-1)}(\theta) = \sum_{j=0}^{k+2l-1} \binom{k+2l-1}{j} S^{(j)}(\theta) \frac{d^{k+2l-1-j}}{d\theta^{k+2l-1-j}} (\sin\theta)^k.$$

Since $(d^{k+2l-1-j}/d\theta^{k+2l-1-j})(\sin\theta)^k$ is even if j is odd and odd if j is even, it follows that these terms are nonzero at $\theta = 0$ or $\theta = \pi$ only if $j = 1, 3, 5, \ldots, 2l-1$, where $l = 1, 2, 3, \ldots, m-k$. Consequently, using induction on l shows that the conditions $T^{(j)}(0) = T^{(j)}(\pi) = 0$ for $j = k+1, k+3, \ldots, 2m-k-1$ are equivalent to the following conditions on S:

$$S^{(2l-1)}(0) = S^{(2l-1)}(\pi) = 0, \qquad l = 1, 2, 3, \ldots, m-k. \tag{4.2}$$

Since $S \in T_{2m-k}$, we can write $S(\theta) = a_0 + \sum_{j=1}^{2m-k}(a_j \cos j\theta + b_j \sin j\theta)$. Because the conditions (4.2) involve only odd derivatives, it applies only to the odd part of S. Hence, we need only to show that if $S_o(\theta) := \sum_{j=1}^{2m-k} b_j \sin j\theta$ satisfies (4.2), then $S_o(\theta) = (\sin\theta)^{2m-2k+1} q(\cos\theta)$. Since S_o automatically satisfies $S_o^{(2j)}(0) = S_o^{(2j)}(\pi) = 0$ for all j, it follows that it satisfies the Hermite interpolation conditions $S_o^{(j)}(0) = S_o^{(j)}(\pi) = 0$ for $0 \le j \le 2m-2k$. Consequently, it follows that $S_o(\theta) = (\sin\theta)^{2m-2k} R(\theta)$, $R \in \mathcal{T}_{k+2m}$. However, S_o is odd, so it must be R. Consequently, we can write $R(\theta) = \sin\theta q(\cos\theta)$ so that $S(\theta) = p(\cos\theta) + (\sin\theta)^{m-2k+1} q(\theta)$, which completes the proof. □

THEOREM 4.2. *Let m be a positive integer. Let $\theta_1 < \theta_2 < \cdots < \theta_{2m+1}$ be a set of distinct points in $(0, \pi)$. Then Problem 2 has a unique solution in $\mathcal{T}_{2m}$.*

Proof. Let T be a polynomial that satisfies $T^{(j)}(x_i) = 0$ for $e_{i,j} = 1$. The above lemma shows that T is of the form (4.2), and the problem reduces to showing that if

$$T(\theta) = (\sin\theta)^k p(\cos\theta) + (\sin\theta)^{2m-k+1} q(\cos\theta)$$

vanishes on $\theta_1, \ldots, \theta_{2m+1}$, then $T(\theta) \equiv 0$. This, however, is a consequence of Corollary 2.6. □

Acknowledgment. The author thanks the referees for their careful review and valuable comments.

REFERENCES

[1] B. BOJANOV AND Y. XU, *On a Hermite interpolation by polynomials of two variables*, SIAM J. Numer. Anal., 39 (2002), pp. 1780–1793.

[2] B. BOJANOV AND Y. XU, *Polynomial interpolation of two variables based on a factorization method*, J. Approx. Theory, to appear.

[3] C. F. DUNKL AND Y. XU, *Orthogonal Polynomials of Several Variables*, Cambridge University Press, Cambridge, UK, 2001.

[4] N. L. FERNÁNDEZ, *Polynomial Bases on the Sphere*, Internat. Ser. Numer. Math. 142, Birkhäuser, Basel, 2002, pp. 39–52.

[5] M. GASCA AND T. SAUER, *Polynomial interpolation in several variables*, Adv. Comput. Math., 12 (2000), pp. 377–410.

[6] M. VON GOLITSCHEK AND W. A. LIGHT, *Interpolation by polynomials and radial basis functions on spheres*, Constr. Approx., 17 (2001), pp. 1–18.

[7] H. HAKOPIAN AND S. ISMAEIL, *On a bivariate interpolation problem*, J. Approx. Theory, 116 (2002), pp. 76–99.

[8] D. J. JOHNSON, *The trigonometric Hermite-Birkhoff interpolation problem*, Trans. Amer. Math. Soc., 212 (1975), pp. 365–374.

[9] G. G. Lorentz, K. Jetter, and S. D. Riemenschneider, *Birkhoff Interpolation*, Addison-Wesley, Reading, MA, 1983.

[10] C. Müller, *Analysis of Spherical Symmetries in Euclidean Spaces*, Springer, New York, 1997.

[11] F. J. Narcowich and J. D. Ward, *Scattered data interpolation on spheres: Error estimates and locally supported basis functions*, SIAM J. Math. Anal, 33 (2002), pp. 1393–1410.

[12] I. H. Sloan and R. S. Womersley, *How good can polynomial interpolation on the sphere be?*, Adv. Comput. Math., 14 (2001), pp. 195–226.

[13] A. Stroud, *Approximate Calculation of Multiple Integrals*, Prentice–Hall, Englewood Cliffs, NJ, 1971.

[14] G. Szego, *Orthogonal Polynomials*, 4th ed., Amer. Math. Soc. Colloq. Publ. 23, AMS, Providence, RI, 1975.

[15] Y. Xu, *Polynomial interpolation on the unit ball and on the unit sphere*, Adv. Comput. Math., to appear.

[16] A. Zygmund, *Trigonometric Series*, Cambridge University Press, Cambridge, UK, 1959.

SIAM J. NUMER. ANAL.
Vol. 41, No. 2, pp. 767–795

SEMICIRCULANT PRECONDITIONING OF ELLIPTIC OPERATORS*

SANG DONG KIM† AND SEYMOUR V. PARTER‡

Abstract. In this work we consider the semicirculant preconditioning of elliptic differential operators of the form

$$Lu := -\epsilon \Delta u + a u_x + b u_y + c u$$

in two cases: $0 < \epsilon \ll 1$ and $\epsilon \equiv 1$. The paper [*Numer. Math.*, 81 (1998), pp. 211–249] provided extremely interesting and useful results in the first case. On the other hand, those appear to contradict basic results on preconditioning given in [*SIAM J. Numer. Anal.*, 27 (1990), pp. 656–694]. We reobtain the results of [*Numer. Math.*, 81 (1998), pp. 211–249] by a new approach which we believe to be more transparent. We also clarify the situation regarding the apparent contradiction with [*SIAM J. Numer. Anal.*, 27 (1990), pp. 656–694]. Finally, we describe the distribution of the preconditioned eigenvalues in the uniformly elliptic case, $\epsilon \equiv 1$.

Key words. preconditioning, difference equations, limiting operator, convection-diffusion equation

AMS subject classifications. 65N, 65F

PII. S0036142902403000

1. Introduction.

This work was motivated by the paper [LH], which discussed the semicirculant preconditioning of two-dimensional convection diffusion equations. The results on the distribution of the eigenvalues of the preconditioned system are extremely interesting. On the other hand, these results seemed to contradict a fundamental principle enunciated in [MP]. The basic results of [MP] imply that if one preconditions a discrete elliptic operator L_h with Dirichlet boundary conditions by another discrete elliptic operator B_h and one has results such as those of [LH], then B_h must also impose Dirichlet boundary conditions. However, it would appear that the semicirculant preconditioner has some Dirichlet boundary conditions and some periodic boundary conditions. Finally, since part of the attraction of the semicirculant preconditioner is the fact that this preconditioner is easily inverted, one is led to a study of semicirculant preconditioners for the uniformly elliptic case.

In this paper we clarify the apparent contradiction between the results of [LH] and [MP]. Simply put, the limiting operator of the difference schemes is *not* elliptic. It is the limiting hyperbolic operator when ϵ, the singular perturbation parameter, goes to zero. Since the solution of the elliptic convection-diffusion equation converges to the solution of this hyperbolic equation, the method and the preconditioning approach are reasonable. Of course, there is a boundary layer in both the solution of the differential equation and the solution of the difference equation. This method does not capture the correct boundary layer. However, that is a small point.

*Received by the editors February 22, 2002; accepted for publication (in revised form) November 20, 2002; published electronically May 12, 2003.
http://www.siam.org/journals/sinum/41-2/40300.html

†Department of Mathematics, Kyungpook National University, Taegu, 702-701, Korea (skim@knu.ac.kr). The research of this author was supported by Korea Research Foundation grant KRF-2002-070-C00014.

‡Department of Computer Sciences, Department of Mathematics, University of Wisconsin-Madison, Madison, WI 53706 (parter@cs.wisc.edu).

While this clarification is both scientifically important and satisfying, it is not the main thrust of our work. Our major effort is directed at the development and exposition of a method for studying these problems which we believe is both more transparent and elementary. Using this approach we both reobtain the results of [LH] and describe the distribution of the eigenvalues of the semicirculant preconditioned system in the uniformly elliptic case. As we might expect from the results of [CC] for the full circulant preconditioner, roughly $(\frac{1}{h})$ of these eigenvalues grow like $O(\frac{1}{h})$. Still, as we shall see, for many of those eigenvalues the coefficients of the growth are quite small. Thus, given the ease of inverting the preconditioner semicirculant, preconditioning may be a useful approach for some problems. While this work focuses on the distributions of the eigenvalues, we are well aware that in these nonsymmetric problems eigenvalues do not tell the whole story as compared with the effectiveness of the preconditioning strategy; see [G].

Let $f \in C(\overline{\Omega})$, and consider the equations

$$Lu = -\epsilon\Delta u + au_x + bu_y + du = f \quad \text{in} \quad \Omega, \tag{1.1}$$

$$u = 0 \quad \text{on} \quad \partial\Omega. \tag{1.2}$$

Here Ω is the unit square $[0,1] \times [0,1]$ and the coefficients a, b, d, and ϵ are constant. Moreover,

$$d \geq 0, \quad \epsilon > 0. \tag{1.3}$$

Let m_1, m_2 be positive integers, and set

$$h_1 = \frac{1}{m_1+1}, \quad h_2 = \frac{1}{m_2+1}, \quad \varphi = \frac{h_2}{h_1}. \tag{1.4}$$

The usual centered second order finite-difference scheme which approximates (1.1) is given by

$$\begin{aligned} -\frac{\epsilon}{h_2^2}\{\, & Cu_{k-1,j} + Au_{k,j} + Bu_{k+1,j} + \gamma u_{k,j-1} \\ & +\alpha u_{k,j} + \beta u_{k,j+1}\,\} = f_{k,j}, \quad 1 \leq k \leq m_1,\ 1 \leq j \leq m_2, \end{aligned} \tag{1.5}$$

where

$$A = -2\varphi^2, \tag{1.6}$$

$$B = \varphi^2\left(1 - \frac{ah_1}{2\epsilon}\right), \quad C = \varphi^2\left(1 + \frac{ah_1}{2\epsilon}\right), \tag{1.7}$$

$$\alpha = -2 - \frac{dh_2^2}{\epsilon}, \tag{1.8}$$

$$\beta = 1 - \frac{bh_2}{2\epsilon}, \quad \gamma = 1 + \frac{bh_2}{2\epsilon}. \tag{1.9}$$

Here $u_{k,j}$ is the value of the approximant at the point (x_k, y_j) and

$$f_{k,j} = f(x_k, y_j). \tag{1.10}$$

Throughout this paper we assume the mesh Peclet number condition

$$\left|\frac{bh_2}{2\epsilon}\right| < 1, \quad \left|\frac{ah_1}{2\epsilon}\right| < 1. \tag{1.11}$$

The difference operator L_h on the left-hand side of (1.5) is easily described as

$$L_h := -\frac{\epsilon}{h_2^2}\{T_2 \otimes I_{m_1} + I_{m_2} \otimes T_1\}, \tag{1.12}$$

where I_{m_k} are the appropriate identity matrices, and T_1 and T_2 are tridiagonal matrices of order m_1 and m_2, respectively. In particular,

$$T_1 = \begin{bmatrix} A & B & & & \\ C & A & B & & \\ & \ddots & \ddots & \ddots & \\ & & & & B \\ & & & C & A \end{bmatrix} \tag{1.13}$$

and

$$T_2 = \begin{bmatrix} \alpha & \beta & & & \\ \gamma & \alpha & \beta & & \\ & \ddots & \ddots & \ddots & \\ & & & & \beta \\ & & & \gamma & \alpha \end{bmatrix}. \tag{1.14}$$

The semicircular preconditioner is given by

$$S := -\frac{\epsilon}{h_2^2}\{\overline{C} \otimes I_{m_1} + I_{m_2} \otimes T_1\}, \tag{1.15}$$

where $\overline{C}$ is the circulant

$$\overline{C} = \begin{bmatrix} \alpha & \beta & & & \gamma \\ \gamma & \alpha & \beta & & \\ & \ddots & \ddots & \ddots & \\ & & & & \beta \\ \beta & & & \gamma & \alpha \end{bmatrix} = T_2 + \begin{bmatrix} 0 & \cdots & 0 & \gamma \\ 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & \cdots & 0 \\ \beta & 0 & \cdots & 0 \end{bmatrix}. \tag{1.16}$$

Our basic problem is the study of the eigenvalues λ which satisfy

$$\lambda S U = (S - Q \otimes I_{m_1})U, \tag{1.17}$$

where

$$Q := -\frac{\epsilon}{h_2^2}\begin{bmatrix} 0 & \cdots & 0 & \gamma \\ 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & \cdots & 0 \\ \beta & 0 & \cdots & 0 \end{bmatrix}. \tag{1.18}$$

The reader familiar with [LH] will realize that we interchanged the roles of x and y. That small change leads to this form of the error term Q, which appears to be simpler than the corresponding term in [LH].

Of course, we can ignore the $(-\epsilon/h_2^2)$ term and deal with the eigenvalue problem

$$\lambda S_0 U = (S_0 - Q_0 \otimes I_{m_1})U, \tag{1.19}$$

where

$$S_0 = \{\overline{C} \otimes I_{m_1} + I_{m_2} \otimes T_1\} \tag{1.20}$$

and

$$Q_0 = \begin{bmatrix} 0 & \cdots & 0 & \gamma \\ 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & \cdots & 0 \\ \beta & 0 & \cdots & 0 \end{bmatrix}. \tag{1.21}$$

Since Q_0 is of rank 2, it is clear that there will be $m_1m_2 - 2m_1$ eigenvalues λ which are exactly equal to one. We then focus on the remaining $2m_1$ eigenvalues.

In our approach we reduce the problem to m_1 problems of the form

$$\lambda C_0 u = T_0 u, \tag{1.22}$$

where T_0 is a particular $m_2 \times m_2$ tridiagonal matrix and C_0 is a related $m_2 \times m_2$ circulant. This is done as follows. We rewrite (1.17) as

$$\lambda\{\overline{C} \otimes I_{m_1} + I_{m_2} \otimes T_1\}U = \{T_2 \otimes I_{m_1} + I_{m_2} \otimes T_1\}U. \tag{1.23}$$

Assume that (τ, F) is an eigenpair for T_1. That is,

$$T_1 F = \tau F. \tag{1.24}$$

We seek an eigenvector of (1.23) in the form $u \otimes F$. We obtain

$$\lambda[(\overline{C} + \tau I_{m_2})u \otimes F] = (T_2 + \tau I_{m_2})u \otimes F. \tag{1.25}$$

Thus, we set

$$C_0 = \overline{C} + \tau I_{m_2},$$
$$T_0 = T_2 + \tau I_{m_2}.$$

Since T_1 has m_1 distinct eigenvalues, and hence m_1 linearly independent eigenvectors, we have indeed found m_1 such simpler (one-dimensional) problems.

In dealing with these problems we consider two distinct cases.

Case 1. $0 < E \ll 1$. This is the case considered in [LH]. In this work, as in [LH], we assume

$$\frac{h_s}{\epsilon} = G_s, \quad \text{a constant}, \quad s = 1, 2. \tag{1.26}$$

In this case the eigenvalues of (1.23) fall into three groups. There are exactly $m_1 m_2 - 2m_1$ eigenvalues equal to one. And, as $h_1, h_2 \to 0$, m_1 of the remaining eigenvalues cluster about an interval (a_1, b_1) with

$$\frac{1}{2} \leq a_1 < b_1 < 1$$

while the other m_1 eigenvalues cluster about a finite interval (c_1, d_1) with

$$1 < c_1 < d_1 < \infty.$$

These are exactly the results of [LH].

In this case the difference approximation is a poor approximation to the elliptic convection-diffusion equation [F], [WH]. On the other hand, if we imagine a sequence (or family) of computations in which (1.26) holds and $h_1 \to 0$, $h_2 \to 0$, the solutions of the difference equations converge—in every subdomain away from the edges which have the boundary layers—to an appropriate solution of the reduced hyperbolic equation

$$au_x + bu_z + du = f. \tag{1.27}$$

We will prove this in the appendix. However, it is important to point out that while that proof can be extended from the constant coefficient case to some problems with variable coefficients there are many important cases for which it cannot be extended. An example of such a case is the case of an interior "stagnation point," i.e., a point $(\overline{x}, \overline{y})$ at which

$$a(\overline{x}, \overline{y}) = b(\overline{x}, \overline{y}) = 0.$$

Case 2. $\epsilon = 1$. In this case we are dealing with a uniformly elliptic problem and simply let $h_1 \to 0$, $h_2 \to 0$.

In this case there are exactly $m_2 m_1 - 2m_1$ eigenvalues which are equal to one. There are m_1 eigenvalues in the interval $(.38, 1)$. We believe the correct interval is $(.5, 1)$, but we cannot prove this sharper result. There are m_1 eigenvalues greater than one. The larger of these grows like cm_2. However, many of the coefficients of growth are small. Indeed, $[m_1/2]$ of these eigenvalues are in the interval $[1, 1 + \frac{1+\sqrt{2}}{2\varphi^2}]$. These results are important for most regular problems. In addition, they are relevant for convection-diffusion equations where (1.26) does not hold. For example, the paper [LW] draws its inspiration and motivation from [LH]. However, we believe our results are equally relevant to those computations.

Analytically, the distinction between the two cases is that in Case 1 the limiting operator is *not* elliptic while in Case 2 the limiting operator is elliptic. Algebraically, the distinction concerns the eigenvalues τ. As we shall see in section 5, the eigenvalues in Case 1 satisfy

$$0 < 2\varphi^2(1 - \delta) \leq |\tau_j| \leq 2\varphi^2(1 + \delta),$$

where

$$\delta = \sqrt{1 - \left(\frac{aG_1}{2}\right)^2}.$$

On the other hand, in Case 2 the eigenvalues τ_j range from $O(h^2)$ to $O(1)$. It is the small eigenvalues τ_j which lead to the large eigenvalues λ.

In section 2 we develop the basic theory for finding the eigenvalues of $C_0^{-1}T_0$. In section 3 we use the elementary theory of one-dimensional difference equations to further extend the theory and obtain the required asymptotic estimates needed to deal with the reaction-diffusion equations. In section 4 we turn to an analysis of the problem in the case $\epsilon = 1$, i.e., the uniformly elliptic case. In section 5 we apply the results of sections 2, 3, and 4 to resolve the two-dimensional problems in both cases. In section 6 we discuss some computational results.

2. The basic theory.

In this section we turn to the study of (1.22) and develop the theory for finding the eigenvalues λ of the matrix $C_0^{-1}T_0$. With this in mind we replace α by

$$\hat{\alpha} = -2(1 + D_0). \tag{2.1}$$

We are interested in a wide range of values of D_0, not necessarily small. The matrix T_0 is the $m_2 \times m_2$ tridiagonal matrix given by

$$T_0 = \begin{bmatrix} \hat{\alpha} & \beta & & & \\ \gamma & \hat{\alpha} & \beta & & \\ & \ddots & \ddots & \ddots & \\ & & & & \beta \\ & & & \gamma & \hat{\alpha} \end{bmatrix} \tag{2.2}$$

and C_0 is the circulant

$$C_0 = T_0 + Q_0. \tag{2.3}$$

Thus, we are concerned with the eigenvalues λ of

$$\lambda C_0 U = TU = (C_0 - Q_0)U \tag{2.4}$$

or

$$\lambda U = (I - C_0^{-1} Q_0)U. \tag{2.5}$$

Therefore (m_2-2) eigenvalues are exactly one and there are two nontrivial eigenvalues which are of the form

$$\lambda = 1 - \rho, \tag{2.6}$$

where ρ is a nonzero eigenvalue of the problem

$$\rho C_0 U = Q_0 U. \tag{2.7}$$

We begin our discussion with the following problem: find two m_2 vectors v and w such that

$$C_0 v = \begin{bmatrix} \sigma \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \end{bmatrix}, \quad C_0 w = \begin{bmatrix} 0 \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ \sigma \end{bmatrix}, \tag{2.8}$$

where σ will be determined later.

LEMMA 2.1. *Consider the (m_2+1) vector $\hat{v}$ of the form*

$$\hat{v} = \begin{bmatrix} 1 \\ v_1 \\ v_2 \\ \cdot \\ \cdot \\ \cdot \\ v_{m_2-1} \\ 1 \end{bmatrix} = \begin{bmatrix} v_0 \\ v_1 \\ v_2 \\ \cdot \\ \cdot \\ \cdot \\ v_{m_2-1} \\ v_{m_2} \end{bmatrix} \tag{2.9}$$

which satisfies the (m_2-1) equations

$$\gamma v_{k-1} + \hat{\alpha} v_k + \beta v_{k+1} = 0, \quad k = 1, 2, \ldots, (m_2-1). \tag{2.10}$$

Let

$$v = \begin{bmatrix} 1 \\ v_1 \\ v_2 \\ \cdot \\ \cdot \\ \cdot \\ v_{m_2-1} \end{bmatrix}, \quad w = \begin{bmatrix} v_1 \\ v_2 \\ \cdot \\ \cdot \\ \cdot \\ v_{m_2-1} \\ 1 \end{bmatrix}. \tag{2.11}$$

Then

$$C_0 v = \begin{bmatrix} \sigma \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \end{bmatrix}, \quad C_0 w = \begin{bmatrix} 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \\ \sigma \end{bmatrix}, \tag{2.12}$$

where

$$\sigma = \hat{\alpha} + \beta v_1 + \gamma v_{m_2-1}. \tag{2.13}$$

Moreover,

$$0 < v_j \leq 1. \tag{2.14}$$

Proof. Direct verification yields (2.12) and (2.13). Since the Peclet condition (1.11) holds and β, γ are positive, the bound (2.14) follows from a standard maximum principle (convexity) argument. □

We seek an eigenvector U of (2.7) of the form

$$U = xv + yw. \tag{2.15}$$

Then

$$C_0(xv + yw) = \begin{bmatrix} x\sigma \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \\ y\sigma \end{bmatrix} \tag{2.16}$$

and

$$Q_0(xv+yw) = \begin{bmatrix} (\gamma v_{m_2-1})x & + & \gamma y \\ & 0 & \\ & \cdot & \\ & \cdot & \\ & \cdot & \\ & 0 & \\ \beta x & + & (\beta v_1)y \end{bmatrix}. \tag{2.17}$$

LEMMA 2.2. *Let M be the matrix*

$$M = \begin{bmatrix} \gamma v_{m_2-1} & 0 & \cdots & 0 & \gamma \\ 0 & \cdots & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & \cdots & \cdots & 0 \\ \beta & 0 & \cdots & 0 & \beta v_1 \end{bmatrix}. \tag{2.18}$$

Let $[x, 0, \ldots, 0, y]^T$ be an eigenvector of M with associated nonzero eigenvalues μ. Then

$$U = (xv + yw) \tag{2.19}$$

satisfies

$$\left(\frac{\mu}{\sigma}\right) C_0(xv+yw) = \frac{\mu}{\sigma} C_0 U = Q_0 U. \tag{2.20}$$

Thus, U is an eigenvector and $\mu/\sigma = \rho$ is an eigenvalue of the eigenvalue problem (2.7).

Proof. From (2.16) and (2.17) we have

$$\frac{\mu}{\sigma} C_0(xv+yw) = \mu \begin{bmatrix} x \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \\ y \end{bmatrix} \tag{2.21}$$

and

$$Q_0(xv+yw) = M \begin{bmatrix} x \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \\ y \end{bmatrix} = \mu \begin{bmatrix} x \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \\ y \end{bmatrix}. \tag{2.22}$$

Therefore the lemma is proven. ☐

THEOREM 2.3. *Let*

$$R = (\gamma v_{m_2-1} + \beta v_1). \tag{2.23}$$

Then the nonzero eigenvalues of M are given by

$$\mu = \frac{R \pm \sqrt{R^2 + 4\gamma\beta(1 - v_1 v_{m_2-1})}}{2} \tag{2.24}$$

and the eigenvalues ρ of (2.7) *are given by*

$$\rho = \frac{R \pm \sqrt{R^2 + 4\gamma\beta(1 - v_1 v_{m_2-1})}}{2(-2 + R - 2D_0)} \tag{2.25}$$

and the value "0" $(m_2 - 2)$ times.

Proof. The theorem is proven by a direct computation. □

LEMMA 2.4. *Let D_0 be a number of order* 1. *Consider ρ_+ given by*

$$\rho_+ = \frac{R + \sqrt{R^2 + 4\gamma\beta(1 - v_1 v_{m_2-1})}}{2(-2 + R - 2D_0)}. \tag{2.26}$$

Then $\rho_+ < 0$ and

$$|\rho_+| \leq \frac{1 + \sqrt{2}}{2D_0}. \tag{2.27}$$

Proof. The lemma is proven by observing the following:

$$0 \leq R \leq 2, \quad 0 \leq 4\gamma\beta(1 - v_1 v_{m_2-1}) \leq 4,$$

and

$$| - 2 + R - 2D_0| \geq 2D_0. \qquad \square$$

3. The difference equations. In this section we study the eigenvalues ρ_+ and ρ_- and their dependence on the parameter D_0. This involves a rather technical discussion of the properties of the vector $\hat{v}$ described by (2.9) and (2.10). This discussion uses differentiation with respect to $(2D_0)$.

As before, we assume (1.11). Since h_2 and ϵ are always positive, we have

$$\frac{|b|h_2}{2\epsilon} < 1. \tag{3.1}$$

As is well known [H] the $\{v_j; j = 1, 2, \ldots, (m_2 - 1)\}$ are linear combinations of the roots S_1, S_2 of the quadratic equation

$$\beta S^2 + \hat{\alpha} S + \gamma = 0. \tag{3.2}$$

LEMMA 3.1. *Under the assumption* (1.11) *the roots of* (3.2)

$$S_1 = \frac{-\hat{\alpha} + \sqrt{\hat{\alpha}^2 - 4\gamma\beta}}{2\beta}, \quad S_2 = \frac{-\hat{\alpha} - \sqrt{\hat{\alpha}^2 - 4\gamma\beta}}{2\beta} \tag{3.3}$$

are both positive. If $D_0 > 0$, then

$$0 < S_2 < 1 < S_1. \tag{3.4}$$

Finally, the solution of (2.10) *subject to the boundary condition* $v_0 = v_{m_2} = 1$ *is given by*

$$v_k = \frac{(S_1^{m_2} - 1)S_2^k + (1 - S_2^{m_2})S_1^k}{S_1^{m_2} - S_2^{m_2}}. \tag{3.5}$$

Proof. The formulae (3.3) are elementary. Since

$$\hat{\alpha}^2 - 4\gamma\beta > 4 - 4\left(1 - \frac{(bh_2)^2}{4\epsilon^2}\right) = \left(\frac{bh_2}{\epsilon}\right)^2 \geq 0,$$

the roots S_1 and S_2 are distinct and positive. Furthermore,

$$S_1 \geq \frac{2(1 + D_0) + |\frac{bh_2}{\epsilon}|}{2 + |\frac{bh_2}{\epsilon}|} \geq 1 \tag{3.6}$$

and strictly greater than one if $D_0 > 0$.

If $b \leq 0$, then

$$S_1 S_2 = \frac{\gamma}{\beta} = \frac{1 - |\frac{bh_2}{2\epsilon}|}{1 + |\frac{bh_2}{2\epsilon}|} \leq 1. \tag{3.7}$$

Since $S_1 > 1$, we see that $S_2 < 1$.

On the other hand, if $b > 0$, we observe that

$$\theta_1 = \frac{1}{S_2}, \quad \theta_2 = \frac{1}{S_1} \tag{3.8}$$

are the roots of the equation $\gamma\theta^2 + \hat{\alpha}\theta + \beta = 0$. The previous argument shows that

$$\theta_1 = \frac{1}{S_2} > 1, \quad \theta_2 = \frac{1}{S_1} < 1.$$

Thus, we have established (3.4). Finally, the formula (3.5) is verified by evaluation of v_0 and v_{m_2}. □

Let "·" represent differentiation with respect to $(2D_0)$. That is, for any quantity m,

$$\frac{\partial}{\partial(2D_0)} m := \dot{m}. \tag{3.9}$$

LEMMA 3.2. *For* $D_0 > 0$ *we have*

$$\dot{S}_1 = \left[1 + \frac{-\hat{\alpha}}{\sqrt{\hat{\alpha}^2 - 4\gamma\beta}}\right] / 2\beta = \left[\frac{S_1}{\sqrt{\hat{\alpha}^2 - 4\gamma\beta}}\right] > 0, \tag{3.10}$$

$$\dot{S}_2 = \left[1 - \frac{-\hat{\alpha}}{\sqrt{\hat{\alpha}^2 - 4\gamma\beta}}\right] / 2\beta = \left[-\frac{S_2}{\sqrt{\hat{\alpha}^2 - 4\gamma\beta}}\right] < 0. \tag{3.11}$$

Moreover,

$$\dot{v}_1 < 0, \quad \dot{v}_{m_2 - 1} < 0. \tag{3.12}$$

Proof. The formulae (3.10), (3.11) are established by a simple computation.

Let us consider $v_1 = v_1(D_0)$. We have

$$v_1 = \frac{(S_1^{m_2} - 1)S_2 + (1 - S_2^{m_2})S_1}{S_1^{m_2} - S_2^{m_2}}. \tag{3.13}$$

A computation yields

$$\dot{v}_1 = Z_1 + Z_2, \tag{3.14}$$

where

$$Z_1 = \frac{\dot{S}_1(1 - S_2^{m_2})\{(S_1^{m_2} - S_2^{m_2}) - m_2(S_1 - S_2)S_1^{m_2-1}\}}{(S_1^{m_2} - S_2^{m_2})^2}, \tag{3.15}$$

$$Z_2 = \frac{\dot{S}_2(S_1^{m_2} - 1)\{(S_1^{m_2} - S_2^{m_2}) - m_2(S_1 - S_2)S_2^{m_2-1}\}}{(S_1^{m_2} - S_2^{m_2})^2}. \tag{3.16}$$

We have

$$S_1^{m_2} - S_2^{m_2} = S_1^{m_2}\left(1 - \frac{S_2}{S_1}\right)\sum_{k=0}^{m_2-1}\left(\frac{S_2}{S_1}\right)^k.$$

Thus,

$$(S_1^{m_2} - S_2^{m_2}) < m_2(S_1 - S_2)S_1^{m_2-1}.$$

Since $\dot{S}_1 > 0$ and $0 < S_2 < 1$, we see that

$$Z_1 < 0. \tag{3.17}$$

As for Z_2, we have

$$S_1^{m_2} - S_2^{m_2} = S_2^{m_2-1}(S_1 - S_2)\sum_{k=0}^{m_2-1}\left(\frac{S_1}{S_2}\right)^k.$$

Thus,

$$(S_1^{m_2} - S_2^{m_2}) > m_2 S_2^{m_2-1}(S_1 - S_2).$$

Since $\dot{S}_2 < 0$ and $S_1 > 1$, we see that

$$Z_2 < 0. \tag{3.18}$$

Therefore we have

$$\dot{v}_1 < 0. \tag{3.19}$$

The proof that $\dot{v}_{m_2-1} < 0$ follows from the observation that

$$v_{m_2-1} = \frac{(\theta_1^{m_2} - 1)\theta_2 + (1 - \theta_2^{m_2})\theta_1}{\theta_1^{m_2} - \theta_2^{m_2}},$$

$$\theta_1 = \frac{-\hat{\alpha} + \sqrt{\hat{\alpha}^2 - 4\gamma\beta}}{2\gamma} = \frac{1}{S_2},$$

$$\theta_2 = \frac{-\hat{\alpha} - \sqrt{\hat{\alpha}^2 - 4\gamma\beta}}{2\gamma} = \frac{1}{S_1},$$

and the previous argument applies. □

Lemma 3.3. *Let R and ρ_+ be defined as in the previous section. Then*

$$\dot{R} < 0, \quad -\dot{\rho}_+ < 0. \tag{3.20}$$

Proof. We have $\dot{R} < 0$ because of Lemma 3.2. And after some algebra we see that

$$-\dot{\rho}_+ \leq \frac{2[2R\dot{R} - 2\gamma\beta(\dot{v}_1 v_{m_2-1} + v_1\dot{v}_{m_2-1})][2+2D_0-R]}{[2(2+2D_0-R)]^2\sqrt{R^2+4\gamma\beta(1-v_1v_{m_2-1})}} \tag{3.21}$$
$$- 2\frac{(1-\dot{R})[R+\sqrt{R^2+4\gamma\beta(1-v_1v_{m_2-1})}]}{[2(2+2D_0-R)]^2}.$$

The second term of (3.21) is negative. Hence we need prove only that

$$2R\dot{R} - 2\gamma\beta(\dot{v}_1 v_{m_2-1} + v_1\dot{v}_{m_2-1}) < 0.$$

However, after some algebra we see that this reduces to

$$2(\gamma^2 v_{m_2-1}\dot{v}_{m_2-1} + \beta^2 v_1\dot{v}_1) < 0. \qquad \square \tag{3.22}$$

Remark 3.4. Since $\rho_+ < 0$, this implies that

$$|\rho_+| \text{ decreases as } D_0 \text{ increases.} \tag{3.23}$$

We believe that

$$|\rho_-| \text{ decreases as } D_0 \text{ increases,}$$

but we cannot prove it.

Remark 3.5. This lemma, and (3.23) in particular, enables us to study the largest eigenvalues in Case 2 by studying the smallest eigenvalues τ_j which are $O(h^2)$. In those cases we can estimate v_1 and v_{m_2-1} by studying a particular boundary value problem for an ordinary differential equation. We shall see this in section 4.

While the next theorem is valid in both cases, it is used primarily in Case 1, where it yields the asymptotic distribution of the eigenvalues.

Theorem 3.6. *Assume that*

$$D_0 = D_0(h_2) \geq D_1 > 0, \tag{3.24}$$

where D_1 is a constant. Further assume that

$$D_0(h_2) \to \overline{D}_0 \quad as \quad h_2 \to 0. \tag{3.25}$$

Observe that as $h_2 \to 0$ the quantities γ and β have limits γ^∞, β^∞. In Case 1*, where* (1.26) *holds, we have*

$$\beta \equiv 1 - \frac{b}{2}G_2 = \beta^\infty, \quad \gamma \equiv 1 + \frac{h}{2}G_2 = \gamma^\infty. \tag{3.26}$$

In Case 2*, where $\epsilon \equiv 1$, we have*

$$\beta \to 1 = \beta^\infty, \quad \gamma \to 1 = \gamma^\infty \quad as \;\; h_1, h_2 \to 0. \tag{3.27}$$

Furthermore,

$$S_1 \to S_1^\infty = \frac{(1+\overline{D}_0) + \sqrt{(1+\overline{D}_0)^2 - \gamma^\infty \beta^\infty}}{\beta^\infty}, \tag{3.28}$$

$$S_2 \to S_2^\infty = \frac{(1+\overline{D}_0) - \sqrt{(1+\overline{D}_0)^2 - \gamma^\infty \beta^\infty}}{\beta^\infty}. \tag{3.29}$$

Then as $m_2 \to \infty$ *(i.e.,* $h_2 \to 0$*) we have*

$$-\rho_+ = |\rho_+| \to -\rho_+^\infty = \frac{(1+\overline{D}_0) + \sqrt{\gamma^\infty \beta^\infty} - Z}{2Z}, \tag{3.30}$$

$$\rho_- = |\rho_-| \to \rho_-^\infty = \frac{Z + \sqrt{\gamma^\infty \beta^\infty} - (1+\overline{D}_0)}{2Z}, \tag{3.31}$$

where

$$Z = \sqrt{(1+\overline{D}_0)^2 - \gamma^\infty \beta^\infty}. \tag{3.32}$$

Proof. Under the hypotheses (3.24) and (3.25) the remarks before (3.30) and (3.31) are obvious. We proceed to prove (3.30), (3.31).

Since

$$v_1 = \frac{(S_1^{m_2} - 1)S_2 + (1 - S_2^{m_2})S_1}{S_1^{m_2} - S_2^{m_2}} \tag{3.33}$$

and

$$v_{m_2-1} = \frac{(1 - S_1^{-m_2})S_2^{m_2-1} + (1 - S_2^{m_2})S_1^{-1}}{1 - \left(\frac{S_2}{S_1}\right)^{m_2}}, \tag{3.34}$$

we see that as $m_2 \to \infty$ ($h_2 \to 0$) we have

$$v_1 \to v_1^\infty = S_2^\infty, \tag{3.35}$$

$$v_{m_2-1} \to v_{m_2-1}^\infty = \frac{1}{S_1^\infty}. \tag{3.36}$$

Hence

$$R \to R^\infty = \frac{\gamma^\infty}{S_1^\infty} + \beta^\infty S_2^\infty. \tag{3.37}$$

However,

$$\frac{\gamma^\infty}{S_1^\infty} = \beta^\infty S_2^\infty. \tag{3.38}$$

Hence

$$-\rho_+^\infty = \frac{R^\infty + 2\sqrt{\gamma^\infty \beta^\infty}}{2[2(1 - \beta^\infty S_2^\infty) + 2\overline{D}_0]}, \tag{3.39}$$

$$\rho_-^\infty = \frac{2\sqrt{\gamma^\infty \beta^\infty} - R^\infty}{2[2(1-\rho^\infty S_2^\infty) + 2\overline{D}_0]}. \tag{3.40}$$

The theorem now follows from algebraic manipulation using the fact that

$$\beta^\infty S_2^\infty = (1+\overline{D}_0) - \sqrt{(1+\overline{D}_0)^2 - \gamma^\infty \beta^\infty}. \qquad \square$$

Corollary 3.7.

$$-\dot{\rho}_+^\infty < 0, \quad \dot{\rho}_-^\infty < 0, \tag{3.41}$$

and

$$\rho_-^\infty \leq 1/2. \tag{3.42}$$

4. The case $\epsilon = 1$ in one dimension. In this section we consider the case $\epsilon \equiv 1$ and $h_2 = \frac{1}{m_2+1} \to 0$. At first we consider the case where

$$2D_0 = \hat{d} h_2^2$$

and $\hat{d}$ is a fixed constant of modest size. In this case the solutions of the equations

$$\frac{1}{h_2^2}[\gamma v_{k-1} + \hat{\alpha} v_k + \beta v_{k+1}] = 0, \quad k = 1, 2, \ldots, (m_2 - 1), \tag{4.1}$$

$$v_0 = v_{m_2} = 1 \tag{4.2}$$

approximate the function $u_h(x)$ which satisfies

$$-u_h'' + b u_h' + \hat{d} u_h = 0, \quad 0 \leq x \leq 1 - h_2, \tag{4.3}$$

$$u_h(0) = u_h(1-h_2) = 1. \tag{4.4}$$

We use this fact to obtain reasonable estimates for ρ_+, ρ_-, and

$$\lambda_+ = 1 - \rho_-, \quad \lambda_- = 1 - \rho_+. \tag{4.5}$$

The solution of (4.3) and (4.4) is

$$u_h(x) = \frac{(e^{M_1(1-h_2)} - 1)e^{M_2 x} + (1 - e^{M_2(1-h_2)})e^{M_1 x}}{e^{M_1(1-h_2)} - e^{M_2(1-h_2)}}, \tag{4.6}$$

where

$$M_1 = \frac{b + \sqrt{b^2 + 4\hat{d}}}{2}, \quad M_2 = \frac{b - \sqrt{b^2 + 4\hat{d}}}{2}. \tag{4.7}$$

Therefore

$$u_h'(0) = \frac{M_2(e^{M_1(1-h_2)} - 1) + M_1(1 - e^{M_2(1-h_2)})}{e^{M_1(1-h_2)} - e^{M_2(1-h_2)}} \tag{4.8}$$

and

$$u_h'(1-h_2) = \frac{M_2(e^{M_1(1-h_2)}-1)e^{M_2(1-h_2)} + M_1(1-e^{M_2(1-h_2)})e^{M_1(1-h_2)}}{e^{M_1(1-h_2)} - e^{M_2(1-h_2)}}. \tag{4.9}$$

We observe that

$$u_h'(0) < 0, \quad u_h'(1-h_2) > 0.$$

Moreover, there is a d_0 and an $H > 0$ such that for $\hat{d} \geq d_0$ and $0 < h_2 < H$ we have

$$|u'(0)| > \frac{|b|}{2}, \quad |u'(1)| > \frac{|b|}{2}.$$

Hence, for $\hat{d} \geq d_0$ we have

$$\beta v_1 \approx \left(1 - \frac{bh_2}{2}\right)(1 + u_h'(0)h_2) < 1 \tag{4.10}$$

and

$$\gamma v_{m_2-1} \approx \left(1 + \frac{bh_2}{2}\right)(1 - u_h'(1-h_2)h_2) < 1. \tag{4.11}$$

Since v_1 and v_{m_2-1} decrease with increasing D_0, once (4.10) and (4.11) hold for some $D_0 > 0$, they hold for all larger D_0. These inequalities will be important in the latter part of this section, where we discuss the case of large D_0 and we cannot use (4.6) to approximate v_1 and v_{m_2-1}.

We recall that

$$\sigma = \hat{\alpha} + \beta v_1 + \gamma v_{m_2-1}.$$

Since

$$v_1 = u_h(h_2) + O(h_2^3), \quad v_{m_2-1} = u_h(1-2h_2) + O(h_2^3), \tag{4.12}$$

we see that

$$\begin{aligned} \sigma = &\left[\beta u_h(h_2) - \left[1 + \frac{\hat{d}h_2^2}{2}\right] u_h(0)\right] \\ &+ \left[\gamma u_h(1-2h_2) - \left[1 + \frac{\hat{d}h_2^2}{2}\right] u_h(1-h_2)\right] + O(h_2^3). \end{aligned} \tag{4.13}$$

A careful computation using the differential equation shows that

$$\sigma = s_h h_2 + O(h_2^3), \tag{4.14}$$

where

$$s_h = u_h'(0) - u_h'(1-h_2). \tag{4.15}$$

Therefore, after a lengthy computation we see that

$$s_h = s_0\,[1 + K_0 h_2 + K_1 h_2], \tag{4.16}$$

where

$$s_0 = -\frac{e^{M_1} + e^{M_2} - 1 - e^b}{e^{M_1} - e^{M_2}} \sqrt{b^2 + 4\hat{d}} = u_0'(0) - u_0'(1), \tag{4.17}$$

$$K_0 = \frac{be^b - (M_1 e^{M_1} + M_2 e^{M_2})}{(e^{M_1} - 1)(1 - e^{M_2})}, \tag{4.18}$$

and

$$K_1 = \frac{M_1 e^{M_1} - M_2 e^{M_2}}{e^{M_1} - e^{M_2}}. \tag{4.19}$$

We also require the quantity

$$W = [u_0'(0) + u_0'(1)]^2 - 2b[u_0'(0) + u_0'(1)]. \tag{4.20}$$

We are now prepared to state and prove the main theorem for this case.

Theorem 4.1. *For $2D_0 = \hat{d}h_2^2$ and h_2 small we have*

$$\rho_+ = \frac{1}{2} + \frac{2}{\sigma} + O(h_2) \tag{4.21}$$

and

$$\rho_- = \frac{1}{2} + O(h_2). \tag{4.22}$$

Therefore

$$\lambda_+ = 1 - \rho_+ = \frac{1}{2} + \frac{2}{|s_h| h_2} + O(h_2), \tag{4.23}$$

or

$$\lambda_+ = \frac{1}{2} + \frac{2(K_0 + K_1)}{s_0} + \frac{2}{|s_0| h_2} + O(h_2), \tag{4.24}$$

and

$$\lambda_- = 1 - \rho_- = \frac{1}{2} + \frac{h_2^2}{|s_h|}\Big(\frac{\hat{d}}{2} - \frac{W}{8}\Big) + O(h_2) = \frac{1}{2} + O(h_2). \tag{4.25}$$

Proof. Since (see (2.26))

$$|\rho_+| = \frac{R + \sqrt{(\beta v_1 + \gamma v_{m_2-1})^2 + 4\gamma\beta(1 - v_1 v_{m_2-1})}}{2|\sigma|},$$

therefore

$$|\rho_+| = \frac{R + \sqrt{(\beta v_1 - \gamma v_{m_2-1})^2 + 4 - (bh_2)^2}}{2|\sigma|}. \tag{4.26}$$

Straightforward computations yield

$$R = 2 + \sigma + \hat{d}h_2^2 \tag{4.27}$$

and

$$(\beta v_1 - \gamma v_{m_2-1})^2 = h_2^2\{W + b^2\} + O(h_2^2). \tag{4.28}$$

Hence

$$\rho_+ \approx \frac{2 + \sigma + \hat{d}h_2^2 + \sqrt{4 + Wh_2^2 + O(h_2^2)}}{2\sigma}.$$

Thus,

$$\rho_+ \approx \frac{2}{\sigma} + \frac{1}{2} + O\Big(\frac{h_2^2}{\sigma}\Big). \tag{4.29}$$

Hence we have proven (4.21). A similar calculation yields (4.22). Finally, (4.24) follows from (4.29), (4.14), and (4.16). □

We now turn to those cases where $2D_0$ is not necessarily small. We recall that for $d \geq d_0$ we have (4.10) and (4.11).

THEOREM 4.2. *Let*

$$\gamma v_{m_2-1} < 1, \quad \beta v_1 < 1. \tag{4.30}$$

Consider ρ_- *given by*

$$\rho_- = \frac{R - \sqrt{R^2 + 4\gamma\beta(1 - v_1 v_{m_2-1})}}{2(-2 + R - 2D_0)}. \tag{4.31}$$

Then

$$|\rho_-| \leq \frac{\sqrt{5} - 1}{2} \approx .618. \tag{4.32}$$

Proof. Let

$$1 - \gamma v_{m_2-1} = p, \quad 1 - \beta v_1 = r. \tag{4.33}$$

Observe that

$$R^2 + 4\gamma\beta(1 - v_1 v_{m_2-1}) = (\gamma v_{m_2-1} - \beta v_1)^2 + 4\gamma\beta \tag{4.34}$$

and

$$\gamma v_{m_2-1} - \beta v_1 = r - p. \tag{4.35}$$

Let

$$\theta = \frac{\sqrt{5} - 1}{2}. \tag{4.36}$$

The assertion (4.32) is equivalent to the assertion

$$\frac{\sqrt{(r-p)^2 + 4\gamma\beta} - R}{2[(2 - R) + 2D_0]} \leq \theta.$$

Now $2 - R = (p + r)$, so this assertion is equivalent to the statement

$$\sqrt{(r-p)^2 + 4\gamma\beta} \leq 2\theta[2 - R + 2D_0] + R$$

or

$$\sqrt{(r-p)^2+4\gamma\beta} \le (2\theta-1)(p+q)+2(1+2\theta D_0).$$

Since both sides of this inequality are positive, this inequality is equivalent to

$$(r-p)^2+4\gamma\beta = (2\theta-1)^2(p+r)^2+4(2\theta-1)(1+2\theta D_0)(p+r) \\ +4(1+2\theta D_0)^2.$$

This inequality is equivalent to

$$4\gamma\beta \le [(2\theta-1)^2-1](p^2+r^2)+2[(2\theta-1)^2+1]pr \\ +4(2\theta-1)(1+2\theta D_0)(p+r)+4(1+2\theta D_0)^2.$$

Since $|p|$ and $|r|$ are each less than one and $4\gamma\beta \le 4$, it is sufficient to prove that

$$4 \le [(2\theta-1)^2+4(2\theta-1)-1][p^2+r^2]+4.$$

However, $2\theta-1=(\sqrt{5}-2)$ is a root of

$$m^2+4m-1.$$

Hence the inequality is proven. □

Remark 4.3. Notice that while this estimate is sufficient for our purposes, we have made no effort to make a sharp estimate.

5. Two dimensions. We now return to the basic eigenvalue problem (1.23). From the discussion following (1.22) we must determine the eigenvalues τ_j of T_1. However, these are known. We have the following lemma.

LEMMA 5.1. *Consider the matrix T_1 given by* (1.13) *with A,B,C given by* (1.6) *and* (1.7). *The eigenvalues of T_1 are*

(5.1) $$\tau_j = A+2\sqrt{BC}\cos \pi j h_1.$$

Proof. This is a well-known result; see [P]. □

Our first basic theorem is for the case of the convection-diffusion equation studied in [LH]. However, we must remark that due to differences in notation we have not attempted to show the exact equivalence of these formulae and those of [LH].

Unfortunately, the description of these eigenvalues requires quite a bit of notation. Our basic result is the following theorem.

THEOREM 5.2. *Let $|a|>0$. Let $0<\epsilon \ll 1$, and let*

(5.2) $$\delta = \sqrt{1-\left(\frac{aG_1}{2}\right)^2} = \varphi^2\sqrt{BC},$$

(5.3) $$D_m = \varphi^2(1-\delta),$$

(5.4) $$D_M = \varphi^2(1+\delta),$$

(5.5) $$X_m = 1+D_m,$$

$$X_M = 1 + D_M, \tag{5.6}$$

$$\delta_1 = \sqrt{1 - \left(\frac{bG_2}{2}\right)^2}, \tag{5.7}$$

$$Y_m = \sqrt{X_m^2 - \delta_1^2}, \tag{5.8}$$

$$Y_M = \sqrt{X_M^2 - \delta_1^2}, \tag{5.9}$$

$$c_1 = 1 + \frac{X_M + \delta_1 - Y_M}{2Y_M}, \tag{5.10}$$

$$d_1 = 1 + \frac{X_m + \delta_1 - Y_m}{2Y_m}, \tag{5.11}$$

$$a_1 = 1 - \frac{Y_M + \delta_1 - X_M}{2Y_M}, \tag{5.12}$$

$$b_1 = 1 - \frac{Y_m + \delta_1 - X_m}{2Y_m}. \tag{5.13}$$

Then the eigenvalues λ_j of (1.23) *subject to the conditions* (1.26) *fall into three groups:* (i) *There are $m_1m_2 - 2m_1$ eigenvalues which are exactly one.* (ii) *There are m_1 eigenvalues which lie in an interval $(\overline{a}_1(h), \overline{b}_1(h))$ and*

$$\overline{a}_1(h) \to a_1 \geq \frac{1}{2} \quad as \ \ h_1, h_2 \to 0 \tag{5.14}$$

and

$$\overline{b}_1(h) \to b_1 \leq 1 \quad as \ \ h_1, h_2 \to 0. \tag{5.15}$$

(iii) *Finally, there are m_1 eigenvalues which lie in a finite interval $(\overline{c}_1(h), \overline{d}_1(h))$ and*

$$\overline{c}_1(h) \to c_1 \geq 1 \quad as \ \ h_1, h_2 \to 0 \tag{5.16}$$

and

$$\overline{d}_1(h) \to d_1 < \infty \quad \text{as} \ \ h_1, h_2 \to 0. \tag{5.17}$$

Proof. In this case the eigenvalues τ_j are given by

$$\tau_j = 2\varphi^2(1 - \delta \cos \pi j h_1). \tag{5.18}$$

Therefore the argument in the introduction shows that we are led to study the problems of sections 2 and 3 with

$$D_0 = D_0(j) = \varphi^2(1 - \delta \cos \pi j h_1) + \frac{dh_2^2}{2\epsilon}.$$

And for all $j = 1, 2, \ldots, m_1$ we have

$$D_m \leq D_0(j) \leq D_M. \tag{5.19}$$

We now apply Theorem 3.6 with $D_0 = D_m$ and $D_0 = D_m$, together with (3.41) and (3.42), to see that the limiting eigenvalues all fall within the indicated intervals. □

We now turn to the uniformly elliptic case with $\epsilon \equiv 1$.

THEOREM 5.3. *Let $\epsilon \equiv 1$. Let j be a fixed integer, and set*

$$\hat{d}(j) = d - \frac{\tau_j}{h_2^2}. \tag{5.20}$$

Let

$$M_1 = M_1(j) = \frac{b + \sqrt{b^2 + 4\hat{d}(j)}}{2}, \quad M_2 = M_2(j) = \frac{b - \sqrt{b^2 + 4\hat{d}(j)}}{2} \tag{5.21}$$

and

$$K_0(j) = \frac{be^b - [M_1(j)e^{M_1(j)} + M_2(j)e^{M_2(j)}]}{(e^{M_1(j)} - 1)(1 - e^{M_2(j)})}, \tag{5.22}$$

$$K_1(j) = \frac{M_1(j)e^{M_1(j)} - M_2(j)e^{M_2(j)}}{e^{M_1(j)} - e^{M_2(j)}}. \tag{5.23}$$

Let

$$\overline{s} = \overline{s}(j) = -\frac{e^{M_1} + e^{M_2} - 1 - e^b}{e^{M_1} - e^{M_2}}\sqrt{b + 4\hat{d}(j)}. \tag{5.24}$$

Let

$$\lambda_1 > \lambda_2 > \cdots > \lambda_{m_1}$$

be the m_1 largest eigenvalues of (1.23). *Then as $h_1, h_2 \to 0$ we have*

$$\lambda_j \approx \frac{1}{2} + \frac{2[K_0(j) + K_1(j)]}{\overline{s}(j)} + \frac{1}{h_2}\frac{2}{|s(j)|}. \tag{5.25}$$

Furthermore, for $1 \leq j < s \leq m_1$ we have

$$1 \leq \lambda_s \leq \lambda_j. \tag{5.26}$$

Finally, for $m_1/2 \leq j \leq m_1$ we have

$$1 \leq \lambda_j \leq 1 + \frac{1 + \sqrt{2}}{2\varphi^2}. \tag{5.27}$$

Proof. For each j we have the situation described in Theorem 4.1 with $\hat{d}$ given by $\hat{d}(j)$ in (5.20). Hence (5.25) follows from Theorem 4.1. The estimate (5.26) follows from Lemma 3.3. Finally, the estimate (5.27) follows from the fact that

$$|\tau_j| \geq 2\varphi^2, \quad \frac{m_1}{2} < j \leq m_1,$$

and from (2.27). □

Remark 5.4. We note that

$$\frac{\tau_j}{h_2^2} \to -\left(\frac{a^2}{4} + (\pi j)^2\right) = -\mu_j. \tag{5.28}$$

And the quantity μ_j is precisely the jth eigenvalue of the operator

$$L_a := -\left(\frac{d}{du}\right)^2 + a\frac{d}{du}.$$

Furthermore, for very large j

$$\overline{s}(j) \to \sqrt{b^2 + 4\mu_j} \approx 2(j\pi).$$

Hence

$$\lambda_j \approx K_2 + \frac{1}{h_2}\frac{1}{(j\pi)},$$

where K_2 is a constant. Thus, for j modestly large the coefficient of $\frac{1}{h_2}$ is quite small. And (5.26) and (5.27) assure us that it is even smaller for larger values of j.

THEOREM 5.5. *There are m_1 eigenvalues $\lambda(h)$ of* (1.23) *which satisfy*

$$.38 < \frac{3-\sqrt{5}}{2} \leq \lambda \leq 1 \tag{5.29}$$

as $h_1, h_2 \to 0$.

Proof. For any fixed j we have the results of Theorem 4.2 which yields

$$\lambda_- \approx \frac{1}{2} - h_2\overline{S},$$

where $\overline{S}$ is a constant depending on b and j.

On the other hand, if j is fixed and so large that (4.30) holds, we have (4.32) which implies that (5.29) holds for all j. □

6. Numerical tests. In this section we provide a series of numerical examples to illustrate the theories developed in the previous sections. For this, we take the following examples by recalling (1.4) and (1.11).

Example 1. Consider

$$-\epsilon\Delta u + 2u_x + u_y.$$

We fix $\frac{h}{\epsilon} = \frac{3}{4}$, where $h = h_1 = h_2$ so that $\epsilon = \frac{4}{3}h$. The non-one eigenvalues are in either $[0.7136, .84762]$ or $[1.2186, 1.67316]$, which can be predicted by Theorem 5.2. The two groups of eigenvalues are listed in Tables 6.1 and 6.2.

Example 2. Consider

$$-\epsilon\Delta u + 3u_x + 3u_y + u.$$

We fix

$$\frac{h}{\epsilon} = \frac{1}{6}, \quad \text{where} \quad h = h_1 = h_2$$

TABLE 6.1
Two groups of eigenvalues for $-\epsilon\Delta u + 2u_x + u_y$.

Mesh size	Eigenvalues	Eigenvalues
$h = \frac{1}{12}, \quad \epsilon = \frac{1}{9}$	0.71835511876813	1.22163948354662
	0.73076629515934	1.22913287556744
	0.74816845617320	1.24219275206102
	⋮	⋮
	0.83684051876625	1.50835544156721
	0.84287352767439	1.58621137822330
	0.84643374383331	1.65101906814222
$h = \frac{1}{20} \quad \epsilon = \frac{1}{15}$	0.71482281489471	1.22007306632850
	0.71976536398126	1.22272233383939
	0.72742605877215	1.22721010848919
	⋮	⋮
	0.84377959972260	1.59928967271870
	0.84591431356486	1.63764397287832
	0.84718698934437	1.66385884311555

TABLE 6.2
Two groups of eigenvalues for $-\epsilon\Delta u + 2u_x + u_y$.

Mesh size	Eigenvalues	Eigenvalues
$h = \frac{1}{36}, \quad \epsilon = \frac{1}{27}$	0.71363243881612	1.21946709926894
	0.71520966610960	1.22027908908089
	0.71777733892708	1.22163902770585
	⋮	⋮
	0.84643342008918	1.64796148998121
	0.84708778274694	1.66165790078332
	0.84747937636160	1.67023428901938

so that

$$\epsilon = 6h.$$

The non-one eigenvalues are in either $[0.58861, .8563394]$ or $[1.2014, 3.23054]$, which can be predicted by Theorem 5.2. The two groups of eigenvalues are listed in Tables 6.3 and 6.4.

Example 3. Consider

$$-\Delta u + u.$$

The two groups of extreme eigenvalues are listed in Table 6.5. The range of these non-one eigenvalues can be predicted by Theorems 5.3 and 5.5.

Example 4. Consider

$$-\Delta u + 2u_x + u_y.$$

Table 6.3

Two groups of eigenvalues for $-\epsilon\Delta u + 3u_x + 3u_y + u$.

Mesh size	Eigenvalues	Eigenvalues
$h = \frac{1}{12}, \quad \epsilon = \frac{1}{2}$	0.62466691796258	1.20374405122117
	0.65546439052744	1.21266616467924
	0.69692692594231	1.22864225975214
	⋮	⋮
	0.84310400457523	1.77404662541921
	0.85079543914318	2.15993704327438
	0.85524282838158	2.98996959219607
$h = \frac{1}{20}, \quad \epsilon = \frac{3}{10}$	0.60537755820952	1.20214214350094
	0.61998545186220	1.20527177800520
	0.64351646258473	1.21062284753986
	⋮	⋮
	0.85180405592886	2.24806343511870
	0.85447327937503	2.63972089973388
	0.85605326117245	3.17985080512051

Table 6.4

Two groups of eigenvalues for $-\epsilon\Delta u + 3u_x + 3u_y + u$.

Mesh size	Eigenvalues	Eigenvalues
$h = \frac{1}{36}, \quad \epsilon = \frac{1}{6}$	0.59440675231608	1.20158915529747
	0.60043847405159	1.20254607834540
	0.61030104435121	1.20415353499653
	⋮	⋮
	0.85503620670062	2.77188289202506
	0.85584854532074	3.01565361766728
	0.85633390013561	3.22519300391089

Table 6.5

Intervals of non-one eigenvalues for $-\Delta u + u$.

h_1	h_2	Interval	of eigenvalues	Interval	of eigenvalues
$\frac{1}{10}$	$\frac{1}{10}$	[0.59004667366385,	0.85158439347714]	[1.21106725151139,	3.92860278756037]
	$\frac{1}{20}$	[0.54471305741227,	0.72162412759633]	[1.62803608757833,	7.17580980406249]
	$\frac{1}{40}$	[0.52225190888377,	0.62000575794387]	[2.58323341635383,	13.72238351656002]
	$\frac{1}{80}$	[0.51109595032917,	0.56134263742518]	[4.57546877591445,	26.84016379297525]
	$\frac{1}{160}$	[0.50553997118222,	0.53084591888113]	[8.60479998696234,	53.08770653678894]
$\frac{1}{20}$	$\frac{1}{20}$	[0.54480311242263,	0.85306283371772]	[1.20808925812872,	7.15212321759623]
	$\frac{1}{40}$	[0.52229798679507,	0.72311100784202]	[1.62051844692941,	13.67584277325651]
	$\frac{1}{80}$	[0.51111919462630,	0.62095166870805]	[2.56694130531961,	26.74789578807177]
$\frac{1}{40}$	$\frac{1}{40}$	[0.52230953204310,	0.85343085930757]	[1.20735192871894,	13.66424738691479]
	$\frac{1}{80}$	[0.51112501865209,	0.72348284346403]	[1.61865410393456,	26.72490726700744]

TABLE 6.6
Two groups of eigenvalues for $-\Delta u + 2u_x + u_y$.

Mesh size	Eigenvalues	Eigenvalues	Predicted eigenvalues
$h_1 = \frac{1}{10}$, $h_2 = \frac{1}{20}$	0.54542861516161	1.62987429948952	
	⋮	⋮	
	0.68750237545680	2.26976641446903	2.1979
	0.70340882342242	2.74846688963437	2.6923
	0.71461453038915	3.74970109457185	3.7099
	0.72126356537295	7.16806444851755	7.1365
$h_1 = \frac{1}{10}$, $h_2 = \frac{1}{40}$	0.52260538394637	2.58744503148715	
	⋮	⋮	
	0.59911298302300	3.93229722429224	3.8958
	0.60866013977106	4.91275919905165	4.8844
	0.61557719684619	6.93558127427726	6.9157
	0.61976363553775	13.70450516030257	13.6890
$h_1 = \frac{1}{10}$, $h_2 = \frac{1}{80}$	0.51127161201154	4.58419759593318	
	⋮	⋮	
	0.55030179229328	7.30991697075110	7.2916
	0.55531702253128	9.28283232359111	9.2686
	0.55898081759268	13.33712479068527	13.3272
	0.56121153499154	26.80162392976290	26.7939
$h_1 = \frac{1}{10}$, $h_2 = \frac{1}{160}$	0.50562753543833	8.62241607624211	
	⋮	⋮	
	0.52524671556309	14.09235001577338	14.0832
	0.52778614033191	18.04414211391962	18.0370
	0.52964529138036	26.15516009926207	26.1502
	0.53077901958173	53.00764397304166	53.0038

TABLE 6.7
Growth rates of the first five largest eigenvalues for $-\Delta u + 2u_x + u_y$.

jth largest eigenvalue	1st	2nd	3rd	4th	5th
Computed rates	0.32682	0.15929	0.10821	0.08313	0.06857
	0.32743	0.16004	0.10925	0.08444	0.07012
	0.32756	0.16023	0.10952	0.08478	0.07053
Predicted rate	0.3276	0.1603	0.1096	0.0849	0.0707

We list a sequence of eigenvalues in Table 6.6. We also list several of the predicted growth rates of the first five largest eigenvalues with their computed growth rates in Table 6.7 when $h_1 = \frac{1}{10}$ and h_2 takes $\frac{1}{20}$, $\frac{1}{40}$, $\frac{1}{80}$, and $\frac{1}{160}$. These predictions are based on (4.24), and the computed rates are calculated using Table 6.6. For example, for $h_1 = \frac{1}{10}$ fixed we assume these larger eigenvalues can be written as

$$X_j + Y_j \frac{1}{h_2} = \lambda(j, h_2).$$

Then, a first approximation to Y_j is given by

$$Y_j = \frac{\lambda(j, \frac{1}{40}) - \lambda(j, \frac{1}{20})}{20}.$$

Appendix. Consider the case where $0 < \epsilon \ll 1$ and

$$h_1/\epsilon = G_1, \quad h_2/\epsilon = G_2. \tag{A.1}$$

For definitiveness we assume that

$$a < 0, \quad b < 0. \tag{A.2}$$

Thus the boundary layers in (1.1) occur on $\partial\Omega_b$ given by

$$\partial\Omega_b := \{(x,0); 0 < x < 1\} \cup \{(0,y); 0 < y < 1\}. \tag{A.3}$$

Let $\partial\Omega_H = \partial\Omega \backslash \partial\Omega_b$ be the hyperbolic boundary, i.e.,

$$\partial\Omega_H = \{(x,1); 0 < x < 1\} \cup \{(1,y); 0 < y < 1\}. \tag{A.4}$$

Let $u = \{u_{k,j}\}$ be the solution of (1.5) with zero Dirichlet boundary values. Let $\Phi(x,y)$ be the solution of

$$a\Phi_x + b\Phi_y = f \tag{A.5}$$

with boundary condition

$$\Phi = 0 \ \text{ on } \ \partial\Omega_H. \tag{A.6}$$

The basic result of this appendix is the following theorem.

THEOREM A.1. *Let* $0 < \delta < 1$. *Let*

$$\Omega_\delta := \{(x,y) : \delta < x, y < 1\}. \tag{A.7}$$

Then, for every such δ, *if* $(x_k, y_j) \in \Omega_\delta$, *then*

$$\text{Max } |u_{k,j} - \Phi(x_k, y_j)| \to 0 \ \text{ as } \ h_1, h_2 \to 0. \tag{A.8}$$

The proof follows from the argument below.

Let

$$u = u^I + u^{II}, \tag{A.9}$$

where

$$L_h u^I = f \ \text{ in } \ \Omega, \tag{A.10}$$

$$u^I = \Phi \ \text{ on } \ \partial\Omega, \tag{A.11}$$

and

$$L_h u^{II} = 0 \ \text{ in } \ \Omega, \tag{A.12}$$

$$u^{II} = -\Phi \quad \text{on} \quad \partial\Omega. \tag{A.13}$$

We observe that (A.13) means

$$u^{II} = 0 \quad \text{on} \quad \partial\Omega_H, \quad u^{II} = -\Phi \quad \text{on} \quad \partial\Omega_b. \tag{A.14}$$

LEMMA A.2. *There is a constant $K_3 > 0$, depending only on a, such that*

$$\|u\|_\infty \le K_3 \|f\|_\infty. \tag{A.15}$$

Proof. Let

$$\theta = 1 + \frac{h_1}{a} \approx e^{-\frac{h_1}{|a|}}, \tag{A.16}$$

and set

$$u_{k,j} = \theta^k v_{k,j}. \tag{A.17}$$

Then $v = \{v_{k,j}\}$ satisfies the equation

$$\begin{aligned} -\frac{\epsilon}{h_2^2}\Big\{\frac{C}{\theta} v_{k-1,j} - 2\varphi^2 v_{k,j} + B\theta v_{k+1,j} + \gamma v_{k,j-1} \\ +\ \alpha v_{k,j} + \beta v_{k,j+1}\Big\} = \theta^{-k} f_{k,j}. \end{aligned} \tag{A.18}$$

We approximate $\frac{1}{\theta}$ by $1 - \frac{h_1}{a}$. Let

$$C_1 = C/\theta, \quad B_1 = B\theta, \tag{A.19}$$

$$g = \left(1 - \frac{h_1^2}{2\epsilon}\right) = 1 - \frac{1}{2} G_1 h_1, \tag{A.20}$$

$$P = \left(\frac{a h_1}{2\epsilon} - \frac{h_1}{a}\right)/g = \left(\frac{a G_1}{2} - \frac{h_1}{a}\right)/g. \tag{A.21}$$

Then an algebraic computation shows that

$$C_1 = \varphi^2 g(1 + P), \quad B_1 = \varphi^2 g(1 - P), \tag{A.22}$$

$$-2\varphi^2 = -2\varphi^2 g - \frac{h_1^2}{\epsilon}\varphi^2. \tag{A.23}$$

Since

$$\frac{h_1^2}{\epsilon}\varphi^2 = \frac{h_2^2}{\epsilon},$$

we see that v satisfies

$$\begin{aligned} -\frac{\epsilon}{h_2^2}\{\varphi^2 g[(1+P)v_{k-1,j} - 2v_{k,j} + (1-P)v_{k+1,j}] \\ +\ [\gamma v_{k,j-1} + \alpha v_{k,j} + \beta v_{k,j+1}]\} + v_{k,j} = \theta^{-k} f_{k,j}. \end{aligned} \tag{A.24}$$

An easy application of maximum principle arguments shows that

$$\|v\|_\infty \le \theta^{-m_1}\|f\|_\infty \approx e^{\frac{1}{|a|}}\|f\|_\infty. \tag{A.25}$$

From (A.17) we see that

$$\|u\|_\infty \le \|v\|_\infty \le K_3\|f\|_\infty, \tag{A.26}$$

and the lemma is proven. □

LEMMA A.3. *Let $W > 0$ be a constant. Let $w = \{w_{k,j}\}$ satisfy*

$$L_h w = 0, \tag{A.27}$$

$$w = 0 \ \ on \ \ \partial\Omega_H, \tag{A.28}$$

$$|w| \le W \ \ on \ \ \partial\Omega_b. \tag{A.29}$$

Let

$$\theta_1 = \sqrt{C/B}, \quad \theta_2 = \sqrt{\gamma/\beta}. \tag{A.30}$$

Note. Since $a < 0, b < 0$, we have

$$\theta_1 < 1, \quad \theta_2 < 1. \tag{A.31}$$

Then

$$|w_{k,j}| \le \theta_1^k\theta_2^j|W|. \tag{A.32}$$

Proof. Let

$$w_{k,j} = \theta_1^k\theta_2^j v_{k,j}. \tag{A.33}$$

Then v satisfies

$$-\frac{\epsilon}{h_2^2}\{\overline{C}v_{k-1,j} - 2\varphi^2 v_{k,j} + \overline{C}v_{k+1,j} \ \ \overline{\gamma}v_{k,j-1} + \alpha v_{k,j} + \overline{\gamma}v_{k,j+1}\} = 0. \tag{A.34}$$

And, since $w = 0$ on $\partial\Omega_H$, we have

$$v = w \ \ \text{on} \ \ \partial\Omega. \tag{A.35}$$

Here

$$\overline{C} = \sqrt{CB} = \varphi^2\sqrt{1 - \left(\frac{aG_1}{2}\right)^2} \le \varphi^2, \tag{A.36}$$

$$\overline{\gamma} = \sqrt{\gamma\beta} = \sqrt{1 - \left(\frac{bG_2}{2}\right)^2} \le 1. \tag{A.37}$$

An application of the maximum principle now shows that v takes on its maximum on $\partial\Omega$. Hence

$$\|v\|_\infty \le W$$

and

$$|w_{k,j}| \leq \theta_1^k \theta_2^j W. \quad \square \tag{A.38}$$

Proof of Theorem A.1. To complete the proof of Theorem A.1 we must show that

$$\|u^I - \Phi\|_\infty \to \quad \text{as} \quad h_1, h_2 \to 0. \tag{A.39}$$

For the sake of completeness we consider two cases. First we deal with the easy case where

$$\Phi \in C^4(\overline{\Omega}).$$

Then we deal with the general case.

THEOREM A.4. *Let* $\Phi \in C^4(\overline{\Omega})$. *Then there is a constant* K *such that*

$$\|u^I - \Phi\|_\infty \leq K[h_1^2 + h_2^2]. \tag{A.40}$$

Proof. In this case we see that

$$\|L_h u^I - L_h \Phi\|_\infty \leq K(h_1^2 + h_2^2).$$

Hence the theorem follows from Lemma A.2. $\square$

THEOREM A.5. *Let* $f \in (C\overline{\Omega})$. *Then*

$$\|\Phi - u^I\|_\infty \to 0 \quad \textit{as} \quad h_2 = \epsilon / G_2 \to 0. \tag{A.41}$$

Proof. Since $f \in C(\overline{\Omega})$, we know that $\Phi \in C^1(\overline{\Omega})$. Standard "mollifier" arguments (see [N]) show that for each ϵ, h_1, h_2 with $\epsilon / h_s = G_s$, a fixed number, there exists a function $\Phi_\epsilon(x, y)$ and a number $\eta_\epsilon > 0$ such that

$$\|\Delta \Phi_\epsilon\| \leq \epsilon^{-1/2}, \tag{A.42}$$

$$\|\Phi - \Phi_\epsilon\|_\infty + \left\| \frac{\partial}{\partial x}[\Phi - \Phi_\epsilon] \right\|_\infty + \left\| \frac{\partial}{\partial y}[\Phi - \Phi_\epsilon] \right\|_\infty \leq \eta_\epsilon, \tag{A.43}$$

$$\eta_\epsilon \to 0 \quad \text{as} \quad \epsilon \to 0. \tag{A.44}$$

Therefore

$$L_h \Phi_\epsilon = f_\epsilon, \tag{A.45}$$

where

$$\|f - f_\epsilon\|_\infty \leq [|a| + |b| + |c|]\eta_\epsilon + \epsilon^{1/2}. \tag{A.46}$$

Hence, since

$$\|L_h(u^I - \Phi_\epsilon)\|_\infty \leq \|f - f_\epsilon\|,$$

Lemma A.2 yields

$$\|u^I - \Phi_\epsilon\|_\infty \leq K[(|a| + |b| + |c|)\eta_\epsilon + \epsilon^{1/2}]. \tag{A.47}$$

Finally,

$$\|u^I - \Phi\|_\infty \leq \|u^I - \Phi_\epsilon\|_\infty + \|\Phi_\epsilon - \Phi\|_\infty .$$

Therefore

$$\|u^I - \Phi\|_\infty \leq K[|a| + |b| + |c|]\eta_\epsilon + \eta_\epsilon + \epsilon^{1/2},$$

and the theorem is proven. □

REFERENCES

[CC] R. Chan and T. Chan, *Circulant preconditioners for elliptic problems*, Numer. Linear Algebra Appl., 1 (1992), pp. 77–101.

[F] D. Funaro, *A note on second-order finite-difference schemes on uniform meshes for advection-diffusion equations*, Numer. Methods Partial Differential Equations, 15 (1999), pp. 581–588.

[G] A. Greenbaum, V. Pták, and Z. Strakos, *Any nonincreasing convergence curve is possible for GMRES*, SIAM J. Matrix Anal. Appl., 17 (1996) pp. 465–469.

[WH] W. Heinrichs, *Defect correction for convection-dominated flow*, SIAM J. Sci. Comput., 17 (1996), pp. 1082–1091.

[LH] L. Hemmingsson, *A semi-circulant preconditioner for the convection-diffusion equations*, Numer. Math., 81 (1998), pp. 211–249.

[LW] L. Hemmingsson-Fränden and A. Wathen, *A nearly optimal preconditioner for the Navier-Stokes equations*, Numer. Linear Algebra Appl., 8 (2001), pp. 229–243.

[H] F. B. Hildebrand, *Finite-Difference Equations and Simulations*, Prentice-Hall, Englewood Cliffs, NJ, 1968.

[MP] T. A. Manteuffel and S. V. Parter, *Preconditioning and boundary conditions*, SIAM J. Numer. Anal., 27 (1990), pp. 656–694.

[N] J. Nitsche and J. C. C. Nitsche, *Error estimates for the numerical solution of elliptic differential equations*, Arch. Rational Mech. Anal., 5 (1960), pp. 293–306.

[P] S. V. Parter, *Stability, convergence, and pseudo-stability of finite-difference equations for an overdetermined problem*, Numer. Math., 4 (1962), pp. 277–292.

SIAM J. Numer. Anal.
Vol. 41, No. 2, pp. 796–798

CORRIGENDUM: FOURIER SPECTRAL APPROXIMATION TO A DISSIPATIVE SYSTEM MODELING THE FLOW OF LIQUID CRYSTALS*

QIANG DU†, BENYU GUO‡, AND JIE SHEN§

PII. S003614290241653X

The purpose of this note is to correct an error in the proof of Proposition 2.4 in [1]. The inequality $||g(d_M)||_1 \leq c||d_M||^2_{L^4}|d_M|_1$ on line 18 of page 741 in [1] is not correct. We now revise the proof and the result of Proposition 2.4 as follows. Indeed,

$$||g(d_M)||_1^2 \leq c\int_\Omega |d_M|^4(\nabla d_M)^2dx.$$

By integration by parts, the Cauchy inequality, and (2.10) in [1], we obtain

$$||g(d_M)||_1^2 \leq c||d_M||^5_{L^{10}}|d_M|_2 \leq c||d_M||^5_{\frac{2n}{5}}|d_M|_2 \leq cM^{2n-5}||d_M||_1^5|d_M|_2.$$

Thus, by (2.18) of [1], we have

$$||(P_M-I)g(d_M)|| \leq cM^{\frac{2n-7}{2}}||d_M||_1^{\frac{5}{2}}|d_M|_2^{\frac{1}{2}}.$$

Next, by virtue of the imbedding inequality and (2.10) of [1],

$$\begin{aligned}2\lambda|G| &\leq 2\lambda||u_M||_{L^3}||\nabla d_M||_{L^6}||(P_M-I)g(d_M)|| \leq c\lambda M^{\frac{2n-7}{2}}||u_M||_{\frac{n}{6}}||d_M||_1^{\frac{5}{2}}||d_M||_2^{\frac{3}{2}}\\ &\leq c\lambda M^{\frac{2n-7}{2}}||u_M||^{\frac{3}{4}}_{\frac{n}{6}}||u_M||_1^{\frac{1}{4}}||d_M||_1^{\frac{5}{2}}||d_M||_2^{\frac{3}{2}}\\ &\leq c\lambda M^{\frac{9n-28}{8}}||u_M||^{\frac{3}{4}}||u_M||_1^{\frac{1}{4}}||d_M||_1^{\frac{5}{2}}||d_M||_2^{\frac{3}{2}}\\ &\leq c\lambda M^{\frac{9n-28}{40}}||u_M||_1^{\frac{1}{4}}\cdot M^{\frac{3(9n-28)}{80}}||d_M||_2^{\frac{3}{2}}\cdot M^{\frac{3(9n-28)}{160}}||u_M||^{\frac{3}{4}}\cdot M^{\frac{9n-28}{16}}||d_M||_1^{\frac{5}{2}}\\ &\leq c\lambda(M^{\frac{9n-28}{5}}||u_M||_1^2+M^{\frac{9n-28}{20}}||d_M||_2^2+M^{\frac{3(9n-28)}{10}}||u_M||^{12}\cdot M^{9n-28}||d_M||_1^{40}).\end{aligned}$$

On the other hand, we have

$$\begin{aligned}2\lambda\int_\Omega F(d_M)dx &\geq \frac{\lambda}{2\varepsilon^2}(||d_M||^4_{L^4}-2||d_M||^2+(2\pi)^n)\\ &\geq \frac{\lambda}{2\varepsilon^2}\left(\frac{1}{(2\pi)^n}||d_M||^4-2||d_M||^2+(2\pi)^n\right)\\ &\geq \frac{\lambda}{2\varepsilon^2(2\pi)^n}(||d_M||^2-(2\pi)^n(1+\varepsilon^2))^2+\lambda||d_M||^2-\frac{\lambda(2\pi)^n}{2}(2+\varepsilon^2)\\ &\geq \lambda||d_M||^2-\tfrac{\lambda}{2}(2\pi)^n(2+\varepsilon^2).\end{aligned}$$

*Received by the editors October 20, 2002; accepted for publication November 26, 2002; published electronically May 12, 2003.

http://www.siam.org/journals/sinum/41-2/41653.html

†Department of Mathematics, Penn State University, 307 McAllister Bldg., State College, PA 16802 (qdu@math.psu.edu).

‡Department of Mathematics, Shanghai Normal University, 100 Guilin Road, Shanghai 200234, People's Republic of China, (byguo@guomai.sh.cn).

§Department of Mathematics, Purdue University, West Lafayette, IN 47907 (shen@math.purdue.edu).

Moreover, by (2.23) of [1],

$$\lambda\gamma||\Delta d_M - P_M f(d_M)||^2 = \lambda\gamma\left(|d_M|_2^2 + ||P_M f(d_M)||^2 - \frac{2}{\varepsilon^2}|d_M|_1^2\right) \geq \lambda\gamma|d_M|_2^2 - \frac{2\lambda\gamma}{\varepsilon^2}|d_M|_1^2.$$

Substituting the above three estimates into (2.17) of [1] and integrating the resulting inequality with respect to t, we find that for $n \leq 3$ and M sufficiently large

(1)
$$\begin{aligned}\widetilde{E}(t) &\equiv E(t) + \int_0^t \left(\frac{\nu}{4}|u_M(s)|_1^2 + \lambda\gamma||\Delta d_M(s) - P_M f(d_M(s))||^2\right) ds \\ &\leq \sigma_0 + \int_0^t \left(\frac{2\lambda\gamma}{\varepsilon^2}||d_M(s)||_1^2 + c_4 M^{\frac{3}{10}(9n-28)}||u_M(s)||^{12} + c_4 M^{9n-28}||d_M(s)||_1^{40}\right) ds,\end{aligned}$$

where

$$\begin{aligned}&E(t) = ||u_M(t)||^2 + \lambda||d_M(t)||_1^2, \\ &\sigma_0 = ||u_0||^2 + \lambda|d_0|_1^2 + 3\lambda\int_\Omega F(d_0)dx + \lambda(2\pi)^n(\varepsilon^2 + 1 + \varepsilon\sqrt{\varepsilon^2+1}).\end{aligned} \tag{2}$$

Finally, we apply Lemma 2.3 of [1] to the above inequality to obtain for $n \leq 3$

$$\widetilde{E}(t) \leq \sigma_0 e^{\left(\frac{2\lambda\gamma}{\epsilon^2} + c_4 M^{-\frac{3}{10}}\right)t}. \tag{3}$$

In fact, we can derive improved results for Proposition 2.4 in the two-dimensional case (i.e., $n = 2$). Indeed, using the imbedding theory and (2.9) and (2.10) in [1], we obtain for any $\delta > 0$

$$||g(d_M)||_1 \leq c||d_M||_{L^\infty}^2|d_M|_1 \leq c||d_M||_{1+\frac{\delta}{2}}^2|d_M|_1 \leq cM^{\frac{\delta}{2}}||d_M||_1^3.$$

Thus, by (2.18) of [1],

$$||(P_M - I)g(d_M)|| \leq cM^{\frac{\delta-2}{2}}||d_M||_1^3.$$

By virtue of imbedding theory and the Cauchy inequality,

$$\begin{aligned}2\lambda|G| &\leq ||u_M||_{L^\infty}||d_M||_1||(P_M - I)g(d_M)|| \leq cM^{\delta-1}||u_M||_1||d_M||_1^4 \\ &\leq \frac{\nu}{2}|u_M|_1^2 + c_4M^{\frac{1}{2}(\delta-1)}||u_M||^2 + c_4M^{\frac{3}{2}(\delta-1)}||d_M||_1^8.\end{aligned}$$

Using the above estimate instead of (2.19) in [1] and repeating the same procedure as in the proof of Proposition 2.4, we obtain the following revised result.

Proposition 2.4 (revised). *Let $\widetilde{E}(t)$, $E(t)$, and σ_0 be defined in (1)–(2). Then, for $n = 3$, we have*

$$\widetilde{E}(t) \leq \sigma_0 e^{\left(\frac{2\lambda\gamma}{\epsilon^2} + c_4 M^{-\frac{3}{10}}\right)t};$$

for $n = 2$, we have for any small $\delta > 0$,

$$\begin{aligned}&E(t) + \int_0^t \left(\frac{\nu}{2}|u_M(s)|_1^2 + 2\lambda\gamma||\Delta d_M(s) - P_M f(d_M(s))||^2\right) ds \leq \sigma_0 e^{c_4 M^{\frac{1}{2}(\delta-1)}t}, \\ &E(t) + \int_0^t \left(\frac{\nu}{2}|u(s)|_1^2 + 2\lambda\gamma|d_M(s)|_2^2\right) ds \leq \left(1 + \frac{4\gamma}{\varepsilon^2}\right)\sigma_0 e^{c_4 M^{\frac{1}{2}(\delta-1)}t}.\end{aligned}$$

Remark 1. The revised result improves the result of Proposition 2.4 in [1] when $n = 2$. We can use the revised Proposition 2.4 to prove directly the existence of a global solution for (2.3) when $n = 2$, and of a local solution for (2.3) when $n = 3$. We can also use the same techniques as in [2] to prove the existence of a global solution for (2.3) when $n = 3$.

Remark 2. There is a similar error in the proof of Theorem 3.1: the estimate (3.20) is not correct. However, we can revise the proof for Theorem 3.1 as above and show that the result of Theorem 3.1 still holds.

REFERENCES

[1] Q. Du, B. Guo, and J. Shen, *Fourier spectral approximation to a dissipative system modeling the flow of liquid crystals*, SIAM J. Numer. Anal., 39 (2001), pp. 735–762.

[2] F.-H. Lin and C. Liu, *Nonparabolic dissipative systems modeling the flow of liquid crystals*, Comm. Pure Appl. Math., 48 (1995), pp. 501–537.

SIAM J. NUMER. ANAL.
Vol. 41, No. 3, pp. 799–826

AN ADAPTIVE FINITE ELEMENT METHOD WITH PERFECTLY MATCHED ABSORBING LAYERS FOR THE WAVE SCATTERING BY PERIODIC STRUCTURES*

ZHIMING CHEN[†] AND HAIJUN WU[†]

Abstract. We develop a finite element adaptive strategy with error control for the wave scattering by periodic structures. The unbounded computational domain is truncated to a bounded one by an extension of the perfectly matched layer (PML) technique, which attenuates both the outgoing and evanescent waves in the PML region. PML parameters such as the thickness of the layer and the medium property are determined through sharp a posteriori error estimates. Numerical experiments are included to illustrate the competitive behavior of the proposed adaptive method.

Key words. adaptivity, perfectly matched layer, a posteriori error analysis, grating optics

AMS subject classifications. 65N30, 78A45, 35Q60

PII. S0036142902400901

1. Introduction.

We consider the prediction of the scattered modes that arise when an electromagnetic wave is incident on some periodic structure. The media are assumed to be nonmagnetic, and the magnetic permeability μ is constant everywhere. Then the electromagnetic fields in the whole space are governed by the following time harmonic (time dependence $e^{-\mathbf{i}\omega t}$) Maxwell equations:

$$\nabla \times \mathbf{E} - \mathbf{i}\omega\mu\mathbf{H} = 0, \tag{1.1}$$

$$\nabla \times \mathbf{H} + \mathbf{i}\omega\varepsilon\mathbf{E} = 0. \tag{1.2}$$

Here $\mathbf{E}$ and $\mathbf{H}$ are the electric and the magnetic field vectors, respectively. The physical structure is described by the dielectric coefficient $\varepsilon(x)$, $x = (x_1, x_2, x_3)$. In this paper we restrict ourselves to the two-dimensional setting (a one-dimensional (1D) grating problem); the medium and the grating surface are assumed to be constant in the x_2 direction. The more complicated biperiodic diffraction (three-dimensional) problems will be considered in a separate work. We assume that the dielectric coefficient $\varepsilon(x) = \varepsilon(x_1, x_3)$ is periodic in the x_1 direction with period $L > 0$:

$$\varepsilon(x_1 + nL, x_3) = \varepsilon(x_1, x_3) \qquad \forall\, x_1, x_3 \in \mathbf{R},\ \ n \text{ integer}.$$

The dielectric coefficient $\varepsilon(x)$ may be complex. We assume $\operatorname{Im}\varepsilon(x) \geq 0$ and $\operatorname{Re}\varepsilon(x) > 0$ whenever $\operatorname{Im}\varepsilon(x) = 0$. It is natural to assume that ε is constant away from a region $\{(x_1, x_3) : b_2 < x_3 < b_1\}$ that includes the structure; that is, there exist constants ε_1 and ε_2 such that

$$\begin{aligned} \varepsilon(x_1, x_3) &= \varepsilon_1 \quad \text{in } \ \Omega_1 = \{(x_1, x_3) : x_3 \geq b_1\}, \\ \varepsilon(x_1, x_3) &= \varepsilon_2 \quad \text{in } \ \Omega_2 = \{(x_1, x_3) : x_3 \leq b_2\}. \end{aligned}$$

In practical applications, we have $\varepsilon_1 > 0$, but ε_2 may be complex according to the substrate material used in Ω_2. Depending on the direction and polarization of the

*Received by the editors January 14, 2002; accepted for publication November 18, 2002; published electronically May 22, 2003.

http://www.siam.org/journals/sinum/41-3/40090.html

[†]LSEC, Institute of Computational Mathematics, Chinese Academy of Sciences, Beijing 100080, PR China (zmchen@lsec.cc.ac.cn, whj@lsec.cc.ac.cn). The first author was supported in part by the China NSF under grant 10025102 and by the China MOST under grant G1999032802.

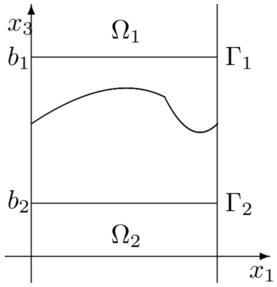


FIG. 1.1. *Geometry of the grating problem.*

incident plane wave, the Maxwell equations can be simplified by considering the two fundamental polarizations: the transverse electric (TE) polarization and the transverse magnetic (TM) polarization. In the TE case, the electric field $\mathbf{E}$ is parallel to the x_2 axis: $\mathbf{E} = (0, u, 0)^T \in \mathbf{R}^3$, where $u = u(x_1, x_3)$ satisfies the Helmholtz equation

$$\Delta u + k^2(x)u = 0 \quad \text{in} \quad \mathbf{R}^2. \tag{1.3}$$

Here $k^2(x) = \omega^2\varepsilon(x)\mu$ is the magnitude of the wave vector. Similarly, in the TM case, the magnetic field $\mathbf{H}$ is parallel to the x_2 axis: $\mathbf{H} = (0, u, 0)^T \in \mathbf{R}^3$, where $u = u(x_1, x_3)$ satisfies the equation

$$\operatorname{div}\left(\frac{1}{k^2(x)}\nabla u\right) + u = 0 \quad \text{in} \quad \mathbf{R}^2. \tag{1.4}$$

Scattering theory in periodic structures, which has many important applications in microoptics, has recently received considerable attention in the applied mathematical community. We refer to Dobson and Friedman [14], Abboud [1], Dobson [13], Bao, Dobson, and Cox [4], Bao [5], and Bao, Cao, and Yang [7] for the existence, uniqueness, and numerical approximations of solutions to grating problems. A good introduction to the problem of electromagnetic diffraction through periodic structures can be found in Petit [19]. More recent review on the diffractive optics technology and its mathematical modeling can be found in Bao, Cowsar, and Masters [6].

The purpose of this paper is to develop efficient numerical methods for solving the 1D grating problem for both the TE (1.3) and the TM (1.4) polarizations. In doing so, the first difficulty is to truncate the domain into a bounded computational domain. The finite element method studied in [4] and [7] is based on variational formulation on the bounded domain Ω, with periodic condition in the x_1 direction and the transparent boundary condition on the top and bottom boundaries Γ_1 and Γ_2 (see Figure 1.1). The transparent boundary condition is obtained by insisting that the solution u of (1.3) or (1.4) be composed of bounded outgoing plane waves in Ω_1 and Ω_2 plus the incident wave u_{I} in Ω_1. The derived transparent boundary condition is represented as a quasi-differential operator and is nonlocal. In practical computations, the infinite series in the definition of the quasi-differential operator must be truncated. The second difficulty is the singularity of the solutions. Usually, the grating surface is piecewise smooth, and across the surface the dielectric coefficient $\varepsilon(x)$ is discontinuous. Thus the solution of (1.4) will have singularities which slow down the finite element convergence when using uniform mesh refinements. Even in the TE case (1.3), when there are lossy materials beneath the grating surface, the transmitted waves decay exponentially. This makes uniform mesh refinements uneconomical.

The purpose of this paper is twofold: First we explore the possibility of applying the recently introduced perfectly matched layer (PML) technique to deal with the first difficulty in truncating the unbounded domain. Second we explore the possibility of using an a posteriori error analysis to design an efficient adaptive method with error control which adaptively determines the finite element meshes and the PML parameters such as the thickness of the PML region and the medium property inside the region. We hope the ideas developed in this paper will be useful for solving other scattering problems on unbounded domains.

A posteriori error estimates are quantities computable in terms of the discrete solution and data, and they measure the actual discrete errors without knowledge of the limit solutions. They are essential in designing algorithms for mesh modifications, which equidistribute the computational effort and optimize the computation. Ever since the pioneering work of Babuška and Rheinboldt [3], the adaptive finite element methods based on a posteriori error estimates have become crucial in scientific and engineering computations. The ability to control error and the asymptotically optimal approximation property (see, e.g., Morin, Nochetto, and Siebert [18], Chen and Dai [10]) make the adaptive finite element method attractive for complicated physical and industrial processes (cf., e.g., Chen and Dai [9], Chen, Nochetto, and Schmidt [11]). For efforts to solve scattering problems using adaptive methods based on a posteriori error estimates, we refer to the recent work of Monk [16] and Monk and Süli [17].

Under the assumption that the exterior solution is composed of outgoing waves only, the basic idea of the PML technique is to surround the computational domain with a finite thickness layer of a specially designed model medium, which would either slow down or attenuate all the waves that propagate from inside the computational domain. Since the work of Berenger [8], which proposed a PML for use with the time dependent Maxwell equations, various constructions of PML absorbing layers have been proposed and studied in the literature. We refer to Turkel and Yefet [22] for a review on various proposed models, and Lassas and Somersalo [15] for the study of mathematical properties of the PML equations. In practical applications involving the PML method, there is a judicial compromise between a thin layer, which requires a rapid variation of the artificial material property, and a thick layer, which requires more grid points and hence more computer time and more storage (see, e.g., Collino and Monk [12]). In this paper, we propose to use an a posteriori error estimate to determine the PML parameters. Moreover, the derived a posteriori error estimate has the nice feature of exponentially decaying in terms of the distance to the computational domain. This property leads to coarse mesh size away from the computational domain and thus makes the total computational cost insensitive to the thickness of the PML absorbing layer.

The layout of the paper is as follows. In section 2 we first recall the transparent boundary condition for 1D grating problems and introduce our PML formulation, which extends the standard PML condition for Helmholtz equations, in that our PML condition attenuates both the outgoing plane waves and the evanescent waves. This extension allows us to reduce the computational domain and thus reduce the computational cost. Existence, uniqueness, and convergence of the PML formulation are considered. In section 3 we introduce the finite element discretization. In section 4 we derive a sharp a posteriori error estimate, which lays down the basis of the adaptive method. In section 5 we present parallel results for the case in which $\operatorname{Im}\varepsilon_2 > 0$. Finally, in section 6 we discuss the implementation of the adaptive method and present several numerical examples to illustrate the competitive behavior of the method.

2. The PML formulation. In this section we shall introduce variational formulations for the 1D grating problem (1.3) and (1.4) using the PML technique, which are suitable for further finite element approximations. As the discussions for the TE polarization and TM polarization are parallel, we shall concentrate on the TE polarization first, and state the corresponding results on the TM polarization without proof. In sections 2–4, we consider the case in which $\operatorname{Im}\varepsilon_2 = 0$, and thus $k_j = \omega\sqrt{\varepsilon_j\mu}$ is real, for $j = 1, 2$. The parallel results for the case in which $\operatorname{Im}\varepsilon_2 > 0$ will be presented in section 5.

2.1. TE polarization. We start by recalling the variational formulation with transparent boundary condition in [4], which is the basis of the analysis in this paper. Let $u_{\mathrm{I}} = e^{\mathbf{i}\alpha x_1 - \mathbf{i}\beta x_3}$ be the incoming plane wave that is incident upon the grating surface from the top, where $\alpha = k_1 \sin\theta, \beta = k_1\cos\theta$, and $-\pi/2 < \theta < \pi/2$ is the incident angle. We are interested in quasi-periodic solutions u, that is, solutions u of (1.3) such that $u_\alpha = ue^{-\mathbf{i}\alpha x_1}$ are periodic in x_1 with period $L > 0$.

Define $\Gamma_j = \{(x_1, x_3) : 0 < x_1 < L, x_3 = b_j\}, j = 1, 2$. We wish to reduce the problem to the bounded domain

$$\Omega = \{(x_1, x_3) : 0 < x_1 < L \text{ and } b_2 < x_3 < b_1\}.$$

The radiation condition for the diffraction problem insists that u is composed of bounded outgoing plane waves in Ω_1 and Ω_2, plus the incident wave u_{I} in Ω_1.

For each integer n, let $\alpha_n = 2\pi n/L$; since u_α is periodic in the x_1 direction with period $L > 0$, it has a Fourier series expansion

$$u_\alpha(x_1, x_3) = \sum_{n\in Z} u_\alpha^{(n)}(x_3)e^{\mathbf{i}\alpha_n x_1}, \qquad u_\alpha^{(n)}(x_3) = \frac{1}{L}\int_0^L u_\alpha e^{-\mathbf{i}\alpha_n x_1}dx_1.$$

Thus we have the expansion

$$u(x_1, x_3) = u_\alpha e^{\mathbf{i}\alpha x_1} = \sum_{n\in Z} u_\alpha^{(n)}(x_3)e^{\mathbf{i}(\alpha_n+\alpha)x_1}. \tag{2.1}$$

Since u satisfies the Helmholtz equation $\Delta u + k_1^2 u = 0$ in Ω_1, we deduce that

$$\sum_{n\in Z}\left[\left(k_1^2 - (\alpha_n+\alpha)^2\right)u_\alpha^{(n)}(x_3) + \frac{d^2}{dx_3^2}u_\alpha^{(n)}(x_3)\right]e^{\mathbf{i}(\alpha_n+\alpha)x_1} = 0,$$

which yields

$$\left(k_1^2 - (\alpha_n+\alpha)^2\right)u_\alpha^{(n)}(x_3) + \frac{d^2}{dx_3^2}u_\alpha^{(n)}(x_3) = 0 \quad \text{for} \quad x_3 \geq b_1. \tag{2.2}$$

For any integer $n \in Z$ and $j = 1, 2$, we define

$$\beta_j^n = \beta_j^n(\alpha) = \begin{cases} \left(k_j^2 - (\alpha_n+\alpha)^2\right)^{1/2} & \text{if } k_j^2 \geq (\alpha_n+\alpha)^2, \\ \mathbf{i}\left((\alpha_n+\alpha)^2 - k_j^2\right)^{1/2} & \text{if } k_j^2 < (\alpha_n+\alpha)^2. \end{cases} \tag{2.3}$$

Note that $\beta_1^0 = \beta$ by definition. We assume that $k_j^2 \neq (\alpha_n+\alpha)^2$ for all $n \in Z, j = 1, 2$. This condition excludes "resonance." All the solutions of the ODE (2.2) can then be written as

$$u_\alpha^{(n)}(x_3) = A_1^n e^{\mathbf{i}\beta_1^n x_3} + B_1^n e^{-\mathbf{i}\beta_1^n x_3},$$

with complex constants A_1^n and B_1^n. The assumption that only bounded outgoing plane waves except $u_{\rm I}$ exist in Ω_1 implies $B_1^n = 0$ for $n \neq 0$. Thus we deduce from (2.1) the following Rayleigh expansion in Ω_1:

$$u = u_{\rm I} + \sum_{n\in Z} A_1^n e^{\mathbf{i}(\alpha_n+\alpha)x_1+\mathbf{i}\beta_1^n x_3}, \qquad x \in \Omega_1. \tag{2.4}$$

Similarly, we have the following Rayleigh expansion in Ω_2:

$$u = \sum_{n\in Z} A_2^n e^{\mathbf{i}(\alpha_n+\alpha)x_1-\mathbf{i}\beta_2^n x_3}, \qquad x \in \Omega_2. \tag{2.5}$$

For any quasi-periodic function f which has the expansion $f = \sum_{n\in Z} f^{(n)}\ e^{\mathbf{i}(\alpha_n+\alpha)x_1}$, the following Dirichlet to Neumann operator T_j is introduced in [4]:

$$(T_j f)(x_1) = \sum_{n\in Z} \mathbf{i}\beta_j^n f^{(n)} e^{\mathbf{i}(\alpha_n+\alpha)x_1}, \quad 0 < x_1 < L\ ,\ j = 1, 2. \tag{2.6}$$

With this notation in mind, simple calculation shows that the Rayleigh expansion u in $\Omega_j, j = 1, 2$, defined in (2.4) and (2.5) satisfies, respectively, the following relations:

$$\frac{\partial(u-u_{\rm I})}{\partial\nu} - T_1(u-u_{\rm I}) = 0 \quad \text{on} \quad \Gamma_1, \qquad \frac{\partial u}{\partial \nu} - T_2 u = 0 \quad \text{on} \quad \Gamma_2, \tag{2.7}$$

where ν stands for the unit outer normal to $\partial\Omega$. These are the transparent boundary conditions used in [4]. To define a variational formulation for the 1D grating problem (1.3) using the boundary conditions (2.7), we first introduce the following subspace of $H^1(\Omega)$, which includes all the quasi-periodic functions:

$$X(\Omega) = \{w \in H^1(\Omega) : w(0, x_3) = e^{-\mathbf{i}\alpha L} w(L, x_3) \ \text{ for } b_2 < x_3 < b_1\}.$$

Define the sesquilinear form $b : X(\Omega) \times X(\Omega) \to \mathbf{C}$ as follows:

$$b(\varphi, \psi) = \int_\Omega \left(\nabla\varphi\nabla\bar\psi - k^2(x)\varphi\bar\psi\right) dx - \sum_{j=1}^{2} \int_{\Gamma_j} (T_j\varphi)\bar\psi dx_1. \tag{2.8}$$

Note that $\frac{\partial u_{\rm I}}{\partial\nu} - T_1 u_{\rm I} = -2\mathbf{i}\beta u_{\rm I}$ on Γ_1; the weak formulation of the 1D grating problem in the TE polarization then reads as follows: Given incoming plane wave $u_{\rm I} = e^{\mathbf{i}\alpha x_1 - \mathbf{i}\beta x_3}$, seek $u \in X(\Omega)$ such that

$$b(u, \psi) = -\int_{\Gamma_1} 2\mathbf{i}\beta u_{\rm I}\bar\psi dx_1 \quad \forall\ \psi \in X(\Omega). \tag{2.9}$$

Recall that $k^2(x) = \omega^2\varepsilon(x)\mu$. The existence of a unique solution u to (2.9) is proved for all but a sequence of countable frequencies ω_j with $|\omega_j| \to +\infty$. Further uniqueness results can be obtained for any frequency ω if the dielectric coefficient $\varepsilon(x)$ has nonzero imaginary part in some subdomains in Ω. In this paper, we shall not elaborate on this issue, and we assume in the following that the variational problem (2.9) has a unique solution. Then the general theory in Babuška and Aziz [2, Chapter 5] implies that there exists a constant $\gamma > 0$ such that the following inf-sup condition holds:

$$\sup_{0\neq\psi\in H^1(\Omega)} \frac{|b(\varphi,\psi)|}{\|\,\psi\,\|_{H^1(\Omega)}} \geq \gamma\ \|\,\varphi\,\|_{H^1(\Omega)} \quad \forall\ \varphi \in X(\Omega). \tag{2.10}$$

Now we turn to the introduction of absorbing PML layers. We surround our computational domain Ω with two PML layers of thickness δ_1 and δ_2 in Ω_1 and Ω_2, respectively. The specially designed model medium in the PML layers should basically be chosen so that either the wave never reaches its external boundary or the amplitude of the reflected wave is so small that it does not essentially contaminate the solution in Ω. Let $s(x_3) = s_1(x_3) + \mathbf{i}s_2(x_3)$ be the model medium property which satisfies

$$s_1, s_2 \in C(\mathbf{R}), \quad s_1 \geq 1, s_2 \geq 0, \quad \text{and } s(x_3) = 1 \text{ for } b_2 \leq x_3 \leq b_1. \tag{2.11}$$

Here we remark that, in contrast to the original PML condition which takes $s_1 \equiv 1$ in the PML region, we allow a variable s_1 in order to attenuate both the outgoing and evanescent waves there. The advantage of this extension makes our method insensitive to the distance of the PML region from the structure. Following the general idea in designing PML absorbing layers, we introduce the PML regions

$$\begin{aligned}\Omega_1^{\mathrm{PML}} &= \{(x_1, x_3) : 0 < x_1 < L \text{ and } b_1 < x_3 < b_1 + \delta_1\},\\ \Omega_2^{\mathrm{PML}} &= \{(x_1, x_3) : 0 < x_1 < L \text{ and } b_2 - \delta_2 < x_3 < b_2\},\end{aligned}$$

and the PML differential operator

$$\mathcal{L} := \frac{\partial}{\partial x_1}\left(s(x_3)\frac{\partial}{\partial x_1}\right) + \frac{\partial}{\partial x_3}\left(\frac{1}{s(x_3)}\frac{\partial}{\partial x_3}\right) + k^2(x)s(x_3).$$

The PML equations in the PML region are

$$\mathcal{L}(\hat{u} - u_{\mathrm{I}}) = 0 \quad \text{in } \Omega_1^{\mathrm{PML}}, \tag{2.12}$$

$$\mathcal{L}\hat{u} = 0 \quad \text{in } \Omega_2^{\mathrm{PML}}. \tag{2.13}$$

The equation satisfied by the PML solution $\hat{u}$ in the domain Ω is the original Helmholtz equation $\Delta\hat{u} + k^2(x)\hat{u} = 0$. Let $D = \{(x_1, x_3) : 0 < x_1 < L, b_2 - \delta_2 < x_3 < b_1 + \delta_1\}$. Due to the assumption (2.11), we can now formulate the PML model which we are going to solve in this paper:

$$\mathcal{L}\hat{u} = -g \quad \text{in} \quad D, \tag{2.14}$$

with the quasi-periodic boundary condition $\hat{u}(0, x_3) = e^{-\mathbf{i}\alpha L}\hat{u}(L, x_3)$ for $b_2 - \delta_2 < x_3 < b_1 + \delta_1$, and the Dirichlet condition $\hat{u} = u_{\mathrm{I}}$ on $\Gamma_1^{\mathrm{PML}} = \{(x_1, x_3) : 0 < x_1 < L, x_3 = b_1 + \delta_1\}$, $\hat{u} = 0$ on $\Gamma_2^{\mathrm{PML}} = \{(x_1, x_3) : 0 < x_1 < L, x_3 = b_2 - \delta_2\}$. Here the source function is

$$g = \begin{cases} -\mathcal{L}u_{\mathrm{I}} & \text{in } \Omega_1^{\mathrm{PML}},\\ 0 & \text{elsewhere.}\end{cases}$$

For any $G \subset D$, define

$$X(G) = \{w \in H^1(G) : w_\alpha = we^{-\mathbf{i}\alpha x_1} \text{ is periodic in } x_1 \text{ with period } L\},$$

and introduce the sesquilinear form $a_G : X(G) \times X(G) \to \mathbf{C}$ as

$$a_G(\varphi, \psi) = \int_G \left(s(x_3)\frac{\partial\varphi}{\partial x_1}\frac{\partial\bar{\psi}}{\partial x_1} + \frac{1}{s(x_3)}\frac{\partial\varphi}{\partial x_3}\frac{\partial\bar{\psi}}{\partial x_3} - k^2(x)s(x_3)\varphi\bar{\psi}\right)dx.$$

Define $\overset{\circ}{X}(D) = \{w \in X(D), w = 0 \text{ on } \Gamma_1^{\mathrm{PML}} \cup \Gamma_2^{\mathrm{PML}}\}$. Then the weak formulation of the PML model reads as follows: Find $\hat{u} \in X(D)$ such that $\hat{u} = u_{\mathrm{I}}$ on $\Gamma_1^{\mathrm{PML}}, \hat{u} = 0$ on Γ_2^{PML}, and

$$a_{D}(\hat{u}, \psi) = \int_D g\bar{\psi}dx \qquad \forall\, \psi \in \overset{\circ}{X}(D). \tag{2.15}$$

Our next objective is to prove the existence and uniqueness of the above problem and derive an error estimate between $\hat{u}$ and u, the solution of the original 1D grating problem (2.9). To achieve the goal, we first find an equivalent formulation of (2.15) in the domain Ω. Similar to the argument leading to the Rayleigh expansion (2.4), we deduce from (2.12) that

$$\hat{u} = u_{\mathrm{I}} + \sum_{n \in Z} \left(A_1^n e^{\mathbf{i}\beta_1^n \int_{b_1}^{x_3} s(\tau)d\tau} + B_1^n e^{-\mathbf{i}\beta_1^n \int_{b_1}^{x_3} s(\tau)d\tau} \right) e^{\mathbf{i}(\alpha_n+\alpha)x_1} \quad \text{in } \Omega_1^{\mathrm{PML}}. \tag{2.16}$$

If we write $\hat{u}(x_1, b_1) = u_{\mathrm{I}}(x_1, b_1) + \sum_{n\in Z} \hat{u}_{\alpha}^{(n)}(b_1)e^{\mathbf{i}(\alpha_n+\alpha)x_1}$ on Γ_1, then the constants A_1^n, B_1^n can be uniquely determined by the additional boundary condition $\hat{u} = u_{\mathrm{I}}$ on Γ_1^{PML} through following equations:

$$\begin{aligned} &A_1^n + B_1^n = \hat{u}_{\alpha}^n(b_1), \\ &A_1^n e^{\mathbf{i}\beta_1^n \int_{b_1}^{b_1+\delta_1} s(\tau)d\tau} + B_1^n e^{-\mathbf{i}\beta_1^n \int_{b_1}^{b_1+\delta_1} s(\tau)d\tau} = 0. \end{aligned}$$

Thus we conclude from (2.16) that

$$\hat{u} = u_{\mathrm{I}} + \sum_{n\in Z} \frac{\zeta_1^n(x_3)}{\zeta_1^n(b_1)} \hat{u}_{\alpha}^{(n)}(b_1) e^{\mathbf{i}(\alpha_n+\alpha)x_1} \quad \text{in} \quad \Omega_1^{\mathrm{PML}}, \tag{2.17}$$

where $\zeta_1^n(x_3) = e^{-\mathbf{i}\beta_1^n \int_{x_3}^{b_1+\delta_1} s(\tau)d\tau} - e^{\mathbf{i}\beta_1^n \int_{x_3}^{b_1+\delta_1} s(\tau)d\tau}$. Similarly, we deduce from (2.13) that

$$\hat{u} = \sum_{n\in Z} \frac{\zeta_2^n(x_3)}{\zeta_2^n(b_2)} \hat{u}_{\alpha}^{(n)}(b_2) e^{\mathbf{i}(\alpha_n+\alpha)x_1} \quad \text{in} \quad \Omega_2^{\mathrm{PML}}, \tag{2.18}$$

where $\zeta_2^n(x_3) = e^{-\mathbf{i}\beta_2^n \int_{b_2-\delta_2}^{x_3} s(\tau)d\tau} - e^{\mathbf{i}\beta_2^n \int_{b_2-\delta_2}^{x_3} s(\tau)d\tau}$. Similar to (2.6), for any quasi-periodic function f which has the expansion $f = \sum_{n\in Z} f^{(n)}e^{\mathbf{i}(\alpha_n+\alpha)x_1}$, we define the following Dirichlet to Neumann operator T_j^{PML}:

$$\left(T_j^{\mathrm{PML}} f\right)(x_1) = \sum_{n\in Z} \mathbf{i}\beta_j^n \coth(-\mathbf{i}\beta_j^n \sigma_j) f^{(n)} e^{\mathbf{i}(\alpha_n+\alpha)x_1}, \tag{2.19}$$

where $\coth(\tau) = \frac{e^{\tau}+e^{-\tau}}{e^{\tau}-e^{-\tau}}$ and

$$\sigma_1 = \int_{b_1}^{b_1+\delta_1} s(\tau)d\tau, \qquad \sigma_2 = \int_{b_2-\delta_2}^{b_2} s(\tau)d\tau. \tag{2.20}$$

Then we know easily from (2.17), (2.18) that

$$\frac{\partial(\hat{u} - u_{\mathrm{I}})}{\partial\nu} - T_1^{\mathrm{PML}}(\hat{u} - u_{\mathrm{I}}) = 0 \quad \text{on} \quad \Gamma_1, \qquad \frac{\partial \hat{u}}{\partial \nu} - T_2^{\mathrm{PML}}\hat{u} = 0 \quad \text{on} \quad \Gamma_2. \tag{2.21}$$

This motivates us to introduce the sesquilinear form $b^{\text{PML}} : X(\Omega) \times X(\Omega) \to \mathbf{C}$,

$$b^{\text{PML}}(\varphi, \psi) = \int_\Omega (\nabla\varphi\nabla\bar{\psi} - k^2(x)\varphi\bar{\psi})dx - \sum_{j=1}^{2}\int_{\Gamma_j}(T_j^{\text{PML}}\varphi)\bar{\psi}dx_1, \tag{2.22}$$

and introduce the following variational problem: Find $\vartheta \in X(\Omega)$ such that

$$b^{\text{PML}}(\vartheta, \psi) = -\int_{\Gamma_1} \mathbf{i}\beta(1 + \coth(-\mathbf{i}\beta\sigma_1))u_{\text{I}}\bar{\psi}dx_1 \qquad \forall\, \psi \in X(\Omega), \tag{2.23}$$

where we have used the fact that $\frac{\partial u_{\text{I}}}{\partial \nu} - T_1^{\text{PML}}u_{\text{I}} = -\mathbf{i}\beta(1 + \coth(-\mathbf{i}\beta\sigma_1))u_{\text{I}}$ on Γ_1. The following lemma establishes the relation of this variational problem to the PML model problem (2.15).

Lemma 2.1. *Any solution $\hat{u}$ of the problem* (2.15) *restricted to Ω is a solution of* (2.23)*. Conversely, any solution ϑ of the problem* (2.23) *can be uniquely extended to the whole domain D to be a solution of* (2.15)*.*

Proof. This proof is standard based on the constructions given in (2.17) and (2.18). We omit the details. □

Let $\Delta_j^n = |k_j^2 - (\alpha_n + \alpha)^2|^{1/2}$ and $U_j = \{n : k_j^2 > (\alpha_n + \alpha)^2\}$, $j = 1, 2$. Then we have $\beta_j^n = \Delta_j^n$ for $n \in U_j$, and $\beta_j^n = \mathbf{i}\Delta_j^n$ for $n \notin U_j$. Let

$$\Delta_j^- = \min\{\Delta_j^n : n \in U_j\}, \qquad \Delta_j^+ = \min\{\Delta_j^n : n \notin U_j\}. \tag{2.24}$$

The following lemma plays a key role in the subsequent analysis.

Lemma 2.2. *For any $\varphi, \psi \in X(\Omega)$, we have*

$$\left|\int_{\Gamma_j}(T_j\varphi - T_j^{\text{PML}}\varphi)\bar{\psi}dx_1\right| \le M_j\|\,\varphi\,\|_{L^2(\Gamma_j)}\|\,\psi\,\|_{L^2(\Gamma_j)},$$

where

$$M_j = \max\left(\frac{2\Delta_j^-}{e^{2\sigma_j^I\Delta_j^-} - 1}, \frac{2\Delta_j^+}{e^{2\sigma_j^R\Delta_j^+} - 1}\right)$$

and σ_j^R, σ_j^I are the real and imaginary parts of σ_j defined in (2.20)*, that is, $\sigma_j = \sigma_j^R + \mathbf{i}\sigma_j^I$.*

Proof. For any $\varphi, \psi \in X(\Omega)$, their traces on Γ_j have the following expansions:

$$\varphi(x_1, b_j) = \sum_{n\in Z}\varphi_\alpha^{(n)}(b_j)e^{\mathbf{i}(\alpha_n+\alpha)x_1}, \qquad \psi(x_1, b_j) = \sum_{n\in Z}\psi_\alpha^{(n)}(b_j)e^{\mathbf{i}(\alpha_n+\alpha)x_1}.$$

The $\varphi_\alpha^{(n)}$ and $\psi_\alpha^{(n)}$ are the Fourier coefficients of periodic functions $\varphi_\alpha(x_1, b_j) = \varphi(x_1, b_j)e^{-\mathbf{i}\alpha x_1}$ and $\psi_\alpha(x_1, b_j) = \psi(x_1, b_j)e^{-\mathbf{i}\alpha x_1}$. The orthogonality property of Fourier series yields

$$\|\varphi\|^2_{L^2(\Gamma_j)} = \|\varphi_\alpha\|^2_{L^2(\Gamma_j)} = L\sum_{n\in Z}|\varphi_\alpha^{(n)}(b_j)|^2,$$

$$\|\psi\|^2_{L^2(\Gamma_j)} = \|\psi_\alpha\|^2_{L^2(\Gamma_j)} = L\sum_{n\in Z}|\psi_\alpha^{(n)}(b_j)|^2.$$

By the orthogonality of Fourier series, we also have

$$(2.25)\qquad \int_{\Gamma_j}(T_j\varphi - T_j^{\mathrm{PML}}\varphi)\bar{\psi}dx = L\sum_{n\in Z}\mathbf{i}\beta_j^n(1-\coth(-\mathbf{i}\beta_j^n\sigma_j))\varphi_\alpha^{(n)}(b_j)\bar{\psi}_\alpha^{(n)}(b_j).$$

For $n\in U_j$, we have $\beta_j^n = \Delta_j^n > 0$; thus

$$\begin{aligned}|\mathbf{i}\beta_j^n(1-\coth(-\mathbf{i}\beta_j^n\sigma_j))| &= \left|\frac{2\beta_j^n}{1-e^{-2\mathbf{i}\beta_j^n\sigma_j}}\right| = \frac{2\Delta_j^n}{|e^{2\mathbf{i}\Delta_j^n\sigma_j^R} - e^{2\Delta_j^n\sigma_j^I}|}\\ &\le \frac{2\Delta_j^n}{e^{2\Delta_j^n\sigma_j^I}-1} \le \frac{2\Delta_j^-}{e^{2\Delta_j^-\sigma_j^I}-1} \le M_j,\end{aligned}$$

where we have used the facts that $\Delta_j^n \ge \Delta_j^-$ for $n\in U_j$ and that the function $\xi(\tau) = 2\tau/(e^{2\tau}-1)$ is decreasing for $\tau > 0$. A similar argument shows that, for $n\notin U_j$,

$$\left|\mathbf{i}\beta_j^n(1-\coth(-\mathbf{i}\beta_j^n\sigma_j))\right| \le \frac{2\Delta_j^+}{e^{2\Delta_j^+\sigma_j^R}-1} \le M_j.$$

This completes the proof upon using the Cauchy inequality in (2.25). □

Now let us take a closer look at the structure of constant M_j, which controls the modeling error of the PML equation towards the original 1D grating problem (see Theorems 2.4 and 2.5 below). Once the incoming plane wave $u_{\mathrm{I}} = e^{\mathbf{i}\alpha x_1 - \mathbf{i}\beta x_3}$ is fixed, the numbers Δ_j^-, Δ_j^+ are fixed according to (2.24). Thus the constant M_j approaches zero exponentially as the PML parameters σ_j^R, σ_j^I tend to infinity. From the definition (2.20) we know that σ_j^R, σ_j^I can be calculated by the medium property $s(x_3)$, which is usually taken as a power function:

$$s(x_3) = \begin{cases} 1+\sigma_1^m\Big(\dfrac{x_3-b_1}{\delta_1}\Big)^m & \text{if } x_3\ge b_1,\\ 1+\sigma_2^m\Big(\dfrac{b_2-x_3}{\delta_2}\Big)^m & \text{if } x_3\le b_2,\end{cases}\qquad m\ge 1.$$

Thus we have

$$(2.26)\qquad \sigma_j^R = \left(1+\frac{\operatorname{Re}\sigma_j^m}{m+1}\right)\delta_j,\qquad \sigma_j^I = \frac{\operatorname{Im}\sigma_j^m}{m+1}\,\delta_j.$$

It is obvious that either enlarging the thickness δ_j of the PML layers or enlarging the medium parameters $\operatorname{Re}\sigma_j^m$ and $\operatorname{Im}\sigma_j^m$ will reduce the PML approximation error.

LEMMA 2.3. *For any $\psi\in X(\Omega)$, we have*

$$\|\,\psi\,\|_{L^2(\Gamma_j)} \le \|\,\psi\,\|_{H^{1/2}(\Gamma_j)} \le \hat{C}\|\,\psi\,\|_{H^1(\Omega)},$$

with $\hat{C} = \sqrt{1+(b_2-b_1)^{-1}}$. Here if $\psi(x_1,b_j) = \sum_{n\in Z}\psi_\alpha^{(n)}(b_j)e^{\mathbf{i}(\alpha_n+\alpha)x_1}$ on Γ_j,

$$\|\,\psi\,\|_{H^{1/2}(\Gamma_j)} = \left(L\sum_{n\in Z}(1+|\alpha_n+\alpha|^2)^{1/2}|\psi_\alpha^{(n)}(b_j)|^2\right)^{1/2}.$$

Proof. Since $\psi\in X(\Omega)$, we have the expansion

$$\psi(x_1,x_3) = \sum_{n\in Z}\psi_\alpha^{(n)}(x_3)e^{\mathbf{i}(\alpha_n+\alpha)x_1}\quad\text{in}\quad\Omega.$$

Thus

$$\|\psi\|^2_{H^1(\Omega)} = L\sum_{n\in Z}\int_{b_2}^{b_1}\left((1+|\alpha_n+\alpha|^2)|\psi_\alpha^{(n)}(x_3)|^2 + \left|\frac{d}{dx_3}\psi_\alpha^{(n)}(x_3)\right|^2\right)dx_3.$$

Now the proof follows from the identity

$$|\psi_\alpha^{(n)}(b_j)|^2 = |\psi_\alpha^{(n)}(x_3)|^2 + \int_{b_j}^{x_3}\frac{d}{dx_3}|\psi_\alpha^{(n)}(\tau)|^2 d\tau, \quad b_2 \le x_3 \le b_1,$$

and the Cauchy inequality. This completes the proof. □

THEOREM 2.4. *Let $\gamma > 0$ be the constant in the* inf-sup *condition* (2.10) *and* $(M_1 + M_2)\hat{C}^2 < \gamma$. *Then the PML variational problem has a unique solution* $\hat{u}$. *Moreover, we have the following error estimate:*

$$\begin{aligned} |\!|\!|u-\hat{u}|\!|\!|_\Omega &:= \sup_{0\ne\psi\in H^1(\Omega)}\frac{|b(u-\hat{u},\psi)|}{\|\,\psi\,\|_{H^1(\Omega)}} \\ &\le \hat{C}M_1\|\,\hat{u}-u_{\rm I}\,\|_{L^2(\Gamma_1)} + \hat{C}M_2\|\,\hat{u}\,\|_{L^2(\Gamma_2)}. \end{aligned} \tag{2.27}$$

Proof. By Lemma 2.1 we need to show only that the variational problem (2.23) has a unique solution. We resort to the general existence and uniqueness result for sesquilinear forms in [2, Chapter 5]. The key point is to show the inf-sup condition for the sesquilinear form $b^{\rm PML} : X(\Omega)\times X(\Omega)\to \mathbf{C}$ defined in (2.22). Thanks to Lemmas 2.2 and 2.3 and the assumption $(M_1+M_2)\hat{C}^2 < \gamma$, this is now obvious:

$$\begin{aligned} |b^{\rm PML}(\varphi,\psi)| &\ge |b(\varphi,\psi)| - \sum_{j=1}^2\left|\int_{\Gamma_j}(T_j\varphi - T_j^{\rm PML}\varphi)\bar\psi dx_1\right| \\ &\ge |b(\varphi,\psi)| - (M_1+M_2)\hat{C}^2\|\,\varphi\,\|_{H^1(\Omega)}\|\,\psi\,\|_{H^1(\Omega)} \qquad \forall\varphi,\psi\in X(\Omega). \end{aligned}$$

It remains to prove the estimate (2.27). By (2.8), (2.9), (2.22), (2.23), and Lemma 2.1 we conclude that

$$\begin{aligned} b(u-\hat{u},\psi) &= -\int_{\Gamma_1}2\mathbf{i}\beta u_{\rm I}\bar\psi dx_1 + \int_{\Gamma_1}\mathbf{i}\beta(1+\coth(-\mathbf{i}\beta\sigma_1))u_{\rm I}\bar\psi dx_1 \\ &\qquad + b^{\rm PML}(\hat{u},\psi) - b(\hat{u},\psi) \\ &= \int_{\Gamma_1}(T_1 - T_1^{\rm PML})(\hat{u}-u_{\rm I})\bar\psi dx_1 \\ &\qquad + \int_{\Gamma_2}(T_2 - T_2^{\rm PML})\hat{u}\bar\psi dx_1 \qquad \forall\psi\in X(\Omega). \end{aligned} \tag{2.28}$$

This completes the proof of the theorem upon using Lemmas 2.2 and 2.3. □

To conclude, we remark that the error estimate (2.27) is a posteriori in nature as it depends on the PML solution $\hat{u}$. This makes a posteriori error control possible (see section 3 for details).

2.2. TM polarization. In this subsection we state the parallel results for the grating problem (1.4). First we introduce the sesquilinear form $b_{\rm TM} : X(\Omega)\times X(\Omega)\to \mathbf{C}$ as follows:

$$b_{\rm TM}(\varphi,\psi) = \int_\Omega\left(\frac{1}{k^2(x)}\nabla\varphi\nabla\bar\psi - \varphi\bar\psi\right)dx - \sum_{j=1}^2\int_{\Gamma_j}\frac{1}{k_j^2}(T_j\varphi)\bar\psi dx_1. \tag{2.29}$$

The variational form for the 1D grating problem in the TM polarization then reads as follows: Given incoming plane wave $u_{\rm I} = e^{\mathbf{i}\alpha x_1 - \mathbf{i}\beta x_3}$, seek $u^{\rm TM} \in X(\Omega)$ such that

$$b_{\rm TM}(u^{\rm TM}, \psi) = -\int_{\Gamma_1} \frac{2\mathbf{i}\beta}{k_1^2} u_{\rm I}\bar{\psi}dx_1 \qquad \forall\, \psi \in X(\Omega). \tag{2.30}$$

We again assume the above variational problem has a unique solution, and thus there exists a constant $\gamma_{\rm TM} > 0$ such that

$$\sup_{0\neq\psi\in H^1(\Omega)} \frac{|b_{\rm TM}(\varphi,\psi)|}{\|\,\psi\,\|_{H^1(\Omega)}} \geq \gamma_{\rm TM}\|\,\varphi\,\|_{H^1(\Omega)} \qquad \forall\, \varphi \in X(\Omega). \tag{2.31}$$

The sesquilinear form $a_{\rm TM} : X(D) \times X(D) \to \mathbf{C}$ associated with the PML problem is

$$a_{\rm TM}(\varphi,\psi) = \int_\Omega \left(\frac{1}{k^2(x)} s(x_3) \frac{\partial\varphi}{\partial x_1}\frac{\partial\bar{\psi}}{\partial x_1} + \frac{1}{k^2(x)}\frac{1}{s(x_3)}\frac{\partial\varphi}{\partial x_3}\frac{\partial\bar{\psi}}{\partial x_3} - s(x_3)\varphi\bar{\psi} \right) dx,$$

and the weak formulation of the PML problem reads as the following: Find $\hat{u}^{\rm TM} \in X(D)$ such that $\hat{u}^{\rm TM} = u_{\rm I}$ on $\Gamma_1^{\rm PML}$, $\hat{u}^{\rm TM} = 0$ on $\Gamma_2^{\rm PML}$, and

$$a_{\rm TM}(\hat{u}^{\rm TM}, \psi) = \int_D g_{\rm TM}\bar{\psi}dx \quad \forall\, \psi \in \overset{\circ}{X}(D), \tag{2.32}$$

where $g_{\rm TM} = g/k_1^2$. The following theorem is an analogue of Theorem 2.4.

THEOREM 2.5. *Assume that* $\sum_{j=1}^2 M_j\hat{C}^2/k_j^2 < \gamma_{\rm TM}$. *Then the PML variational problem* (2.32) *has a unique solution* $\hat{u}^{\rm TM}$. *Moreover, we have the following error estimate:*

$$\begin{aligned} |||u^{\rm TM} - \hat{u}^{\rm TM}|||_\Omega^{\rm TM} &= \sup_{0\neq\psi\in H^1(\Omega)} \frac{|b_{\rm TM}(u^{\rm TM} - \hat{u}^{\rm TM}, \psi)|}{\|\,\psi\,\|_{H^1(\Omega)}} \\ &\leq \left(\frac{M_1\hat{C}}{k_1^2}\right)\|\,\hat{u}^{\rm TM} - u_{\rm I}\,\|_{L^2(\Gamma_1)} + \left(\frac{M_2\hat{C}}{k_2^2}\right)\|\,\hat{u}^{\rm TM}\,\|_{L^2(\Gamma_2)}. \end{aligned}$$

3. The discrete problem.

In this section we introduce the finite element approximations of the PML problems (2.15) and (2.32). Let $\mathcal{M}_h$ be a regular triangulation of the domain D. Remember that any triangle $T \in \mathcal{M}_h$ is considered as closed. We assume that any element T must be completely included in $\overline{\Omega_1^{\rm PML}}$, $\overline{\Omega_2^{\rm PML}}$, or $\overline{\Omega}$. To define a finite element space whose functions are quasi-periodic in the x_1 direction, we also require that if $(0, z)$ is a node on the left boundary, then (L, z) also be a node on the right boundary, and vice versa. Let $V_h(D) \subset X(D)$ be the conforming linear finite element space and $\overset{\circ}{V}_h(D) = V_h(D) \bigcap \overset{\circ}{X}(D)$. Denote by $I_h : C(\bar{D}) \to V_h(D)$ the standard finite element interpolation operator.

The finite element approximation to the PML problem (2.15) reads as follows: Find $\hat{u}_h \in V_h(D)$ such that $\hat{u}_h = I_h u_{\rm I}$ on $\Gamma_1^{\rm PML}$, $\hat{u}_h = 0$ on $\Gamma_2^{\rm PML}$, and

$$a_D(\hat{u}_h, \psi_h) = \int_D g\bar{\psi}_h dx \qquad \forall\, \psi_h \in \overset{\circ}{V}_h(D). \tag{3.1}$$

Following the general theory in [2, Chapter 5], the existence of a unique solution of the discrete problem (3.1) and the finite element convergence analysis depend on the following discrete inf-sup condition:

$$\sup_{0\neq\psi_h\in\overset{\circ}{V}_h(D)} \frac{|a_D(\varphi_h,\psi_h)|}{\|\,\psi_h\,\|_{H^1(D)}} \geq \gamma_D\ \|\,\varphi_h\,\|_{H^1(D)} \quad \forall\, \varphi_h \in \overset{\circ}{V}_h(D), \tag{3.2}$$

where the constant $\gamma_{_D} > 0$ is independent of the finite element mesh size. Since the continuous problem (2.15) has a unique solution by Theorem 2.4, the sesquilinear form $a_{_D} : X(D) \times X(D) \to \mathbf{C}$ satisfies the continuous inf-sup condition. Then a general argument of Schatz [20] implies that (3.2) is valid for sufficiently small mesh size $h < h^*$. Based on (3.2), an appropriate a priori error estimate can also be derived that depends on the regularity of the PML solution $\hat{u}$. In this paper, we are interested in a posteriori error estimates and the associated adaptive algorithm. Thus in the following we simply assume that the discrete problem (3.1) has a unique solution $\hat{u}_h \in V_h(D)$.

Let

$$A(x) = \begin{pmatrix} A_{11} & 0 \\ 0 & A_{22} \end{pmatrix} = \begin{pmatrix} s(x_3) & 0 \\ 0 & 1/s(x_3) \end{pmatrix}, \qquad B(x) = k^2(x)s(x_3).$$

Then the definitions of $\mathcal{L}$ and $a_{_D}$ can be rewritten as

$$\mathcal{L} = \text{div}\ (A(x)\nabla) + B(x),$$
$$a_{_D}(\varphi, \psi) = \int_D \left(A(x)\nabla\varphi\nabla\bar{\psi} - B(x)\varphi\bar{\psi}\right) dx.$$

For any $T \in \mathcal{M}_h$, we denote by h_T its diameter. Let $\mathcal{B}_h$ denote the set of all sides that do not lie on Γ_j^{PML}, $j = 1, 2$. For any $e \in \mathcal{B}_h$, h_e stands for its length. For any $T \in \mathcal{M}_h$, we introduce the residual

$$R_T := \mathcal{L}\hat{u}_h|_T + g|_T = \begin{cases} \mathcal{L}(\hat{u}_h|_T - u_{\text{I}}|_T) & \text{if } T \subset \overline{\Omega_1^{\text{PML}}}, \\ \mathcal{L}\hat{u}_h|_T & \text{otherwise.} \end{cases} \tag{3.3}$$

For any interior side $e \in \mathcal{B}_h$ which is the common side of T_1 and $T_2 \in \mathcal{M}_h$, we define the jump residual across e as

$$J_e = (A\nabla\hat{u}_h|_{T_1} - A\nabla\hat{u}_h|_{T_2}) \cdot \nu_e, \tag{3.4}$$

using the convention that the unit normal vector ν_e to e points from T_2 to T_1. Define $\Gamma_{\text{left}} = \{(x_1, x_3) : x_1 = 0, b_2 - \delta_2 < x_3 < b_1 + \delta_1\}$ and $\Gamma_{\text{right}} = \{(x_1, x_3) : x_1 = L, b_2 - \delta_2 < x_3 < b_1 + \delta_1\}$. If $e = \Gamma_{\text{left}} \cap \partial T$ for some element $T \in \mathcal{M}_h$ and e' the corresponding side on Γ_{right} which is also a side of some element T', then we define the jump residual as

$$\begin{aligned} J_e &= A_{11}\left[\frac{\partial}{\partial x_1}(\hat{u}_h|_T) - e^{-\mathbf{i}\alpha L} \cdot \frac{\partial}{\partial x_1}(\hat{u}_h|_{T'})\right], \\ J_{e'} &= A_{11}\left[e^{\mathbf{i}\alpha L} \cdot \frac{\partial}{\partial x_1}(\hat{u}_h|_T) - \frac{\partial}{\partial x_1}(\hat{u}_h|_{T'})\right]. \end{aligned} \tag{3.5}$$

For any $T \in \mathcal{M}_h$, denote by $\eta_{_T}$ the local error estimator, which is defined as follows:

$$\eta_{_T} = \max_{x \in \tilde{T}} \rho(x_3) \cdot \left[h_T\|R_T\|_{L^2(T)} + \left(\frac{1}{2}\sum_{e \subset T} h_e\|\, J_e \,\|_{L^2(e)}^2\right)^{1/2}\right], \tag{3.6}$$

where $\tilde{T}$ is the union of all elements having nonempty intersection with T and

$$\rho(x_3) = \begin{cases} |s(x_3)|e^{-R_j(x_3)} & \text{if } x \in \overline{\Omega_j^{\text{PML}}}, \\ 1 & \text{if } x \in \Omega, \end{cases}$$

with $R_j(x_3)$, $j=1,2$, being defined in (4.5)–(4.6) below.

The following theorem is the main result in this paper.

THEOREM 3.1. *There exists a constant $C>0$, depending only on the minimum angle of the mesh $\mathcal{M}_h$, such that the following a posteriori error estimate is valid:*

$$|||u-\hat{u}_h|||_\Omega \le \hat{C}M_1\|\hat{u}_h-u_{\rm I}\|_{L^2(\Gamma_1)}+\hat{C}M_2\|\hat{u}_h\|_{L^2(\Gamma_2)} + \hat{C}M_3\|\, I_h u_{\rm I}-u_{\rm I}\,\|_{L^2(\Gamma_1^{\rm PML})}+C(1+C_1+C_2)\left(\sum_{T\in\mathcal{M}_h}\eta_T^2\right)^{1/2},$$

where the constants $M_j (j=1,2)$, $\hat{C}$, C_j, and M_3 are defined in Lemmas 2.2, 2.3, 4.3, *and* 4.4, *respectively.*

The proof of this theorem will be given in section 4. We notice that when the PML parameters σ_j^R and σ_j^I tend to infinity, the constants M_j decay exponentially and the constants C_j remain bounded. The important exponential decay factors $e^{-R_j(x_3)}$ in the PML region $\Omega_j^{\rm PML}$ allow us to take thicker PML layers without introducing unnecessary fine meshes away from the computational domain. Recall that thicker PML layers allow a smaller PML medium property, which enhances numerical stability.

To conclude this section, we state the parallel results for the TM polarization. The finite element approximation to the TM polarization problem (2.32) reads as follows: Find $\hat{u}_h^{\rm TM}\in V_h(D)$ such that $\hat{u}_h^{\rm TM}=I_h u_{\rm I}$ on $\Gamma_1^{\rm PML}$, $\hat{u}_h^{\rm TM}=0$ on $\Gamma_2^{\rm PML}$, and

$$a_{\rm TM}(\hat{u}_h^{\rm TM},\psi_h)=\int_D g_{\rm TM}\bar{\psi}_h dx \quad \forall\, \psi_h\in \overset{\circ}{V}_h(D). \tag{3.7}$$

Let $A_{\rm TM}(x)=A(x)/k^2(x)$, $B_{\rm TM}(x)=B(x)/k^2(x)$, and $\mathcal{L}_{\rm TM}={\rm div}\,(A_{\rm TM}(x)\nabla)+B_{\rm TM}(x)$. Then we have the following theorem, parallel to Theorem 3.1, whose proof is omitted.

THEOREM 3.2. *There exists a constant $C>0$, depending only on the minimum angle of the mesh $\mathcal{M}_h$, such that the following a posteriori error estimate is valid:*

$$|||u^{\rm TM}-\hat{u}_h^{\rm TM}|||_\Omega^{\rm TM} \le \left(\frac{\hat{C}M_1}{k_1^2}\right)\|\hat{u}_h^{\rm TM}-u_{\rm I}\|_{L^2(\Gamma_1)}+\left(\frac{\hat{C}M_2}{k_2^2}\right)\|\hat{u}_h^{\rm TM}\|_{L^2(\Gamma_2)} + \left(\frac{\hat{C}M_3}{k_1^2}\right)\|\, I_h u_{\rm I}-u_{\rm I}\,\|_{L^2(\Gamma_1^{\rm PML})}+C(1+C_1+C_2)\left(\sum_{T\in\mathcal{M}_h}\eta_T^2\right)^{1/2},$$

where the constants $M_j(j=1,2)$, $\hat{C}$, C_j, and M_3 are defined in Lemmas 2.2, 2.3, 4.3, *and* 4.4, *respectively. Here η_T is defined as in* (3.6), *with $A,\mathcal{L},g$, and $\hat{u}_h$ being replaced by $A_{\rm TM},\mathcal{L}_{\rm TM},g_{\rm TM}$, and $\hat{u}_h^{\rm TM}$, respectively.*

4. A posteriori error analysis.

In this section we prove the a posteriori error estimates in Theorem 3.1.

4.1. Error representation formula.

For any $\psi\in X(\Omega)$, we extend it to be a function in $X(D)$ denoted by $\tilde{\psi}$ as follows:

$$\tilde{\psi}(x_1,x_3)=\sum_{n\in Z}\frac{\bar{\zeta}_j^n(x_3)}{\bar{\zeta}_j^n(b_j)}\psi_\alpha^{(n)}(b_j)e^{\mathbf{i}(\alpha_n+\alpha)x_1} \qquad \text{in } \Omega_j^{\rm PML},\ j=1,2, \tag{4.1}$$

where $\zeta_j^n(x_3)$ are defined in (2.17) and (2.18), and $\psi_\alpha^{(n)}(b_j)$ are the Fourier coefficients of the function $\psi_\alpha = \psi e^{-\mathbf{i}\alpha x_1}$ on Γ_j; that is,

$$\psi(x_1, b_j) = \sum_{n\in Z} \psi_\alpha^{(n)}(b_j) e^{\mathbf{i}(\alpha_n+\alpha)x_1}. \tag{4.2}$$

It is easy to see that $\tilde{\psi} = \psi$ on Γ_j and $\mathcal{L}\bar{\tilde{\psi}} = 0$ in $\Omega_j^{\rm PML}$.

Lemma 4.1. *Let ν_j be the unit outer normal to $\Omega_j^{\rm PML}$. Then for any $\varphi, \psi \in X(\Omega)$ we have*

$$\int_{\Gamma_j} T_j^{\rm PML}\varphi\bar{\psi}dx_1 = -\int_{\Gamma_j}\varphi\frac{\partial\bar{\tilde{\psi}}}{\partial\nu_j}dx_1. \tag{4.3}$$

Proof. Define

$$\varphi(x_1, b_j) = \sum_{n\in Z}\varphi_\alpha^{(n)}(b_j)e^{\mathbf{i}(\alpha_n+\alpha)x_1}.$$

Then, by the definition of $T_j^{\rm PML}$ in (2.19) and the orthogonality property of Fourier series, we have

$$\int_{\Gamma_j} T_j^{\rm PML}\varphi\bar{\psi}dx_j = L\sum_{n\in Z}\mathbf{i}\beta_j^n\coth(-\mathbf{i}\beta_j^n\sigma_j)\varphi_\alpha^{(n)}(b_j)\bar{\psi}_\alpha^{(n)}(b_j).$$

On the other hand, by direct calculation from (4.1), we have

$$\begin{aligned}
-\int_{\Gamma_j}\varphi\frac{\partial\bar{\tilde{\psi}}}{\partial\nu_j}dx_1 &= (-1)^{j-1}\int_{\Gamma_j}\varphi\frac{\partial\bar{\tilde{\psi}}}{\partial x_3}dx_1\\
&= (-1)^{j-1}L\sum_{n\in Z}\frac{d}{dx_3}\left(\frac{\zeta_j^n(x_3)}{\zeta_j^n(b_j)}\right)\Bigg|_{x_3=b_j}\varphi_\alpha^{(n)}(b_j)\bar{\psi}_\alpha^{(n)}(b_j)\\
&= L\sum_{n\in Z}\mathbf{i}\beta_j^n\coth(-\mathbf{i}\beta_j^n\sigma_j)\varphi_\alpha^{(n)}(b_j)\bar{\psi}_\alpha^{(n)}(b_j).
\end{aligned}$$

This completes the proof. □

Whenever no confusion of the notation is incurred, we shall write $\tilde{\psi}$ as ψ in $\Omega_j^{\rm PML}$ in what follows.

Lemma 4.2 (error representational formula). *For any $\psi \in X(\Omega)$, which is extended to be a function in $X(D)$ according to* (4.1), *and $\psi_h \in \mathring{V}_h(D)$, we have*

$$\begin{aligned}
b(u-\hat{u}_h, \psi) &= \int_D g(\overline{\psi-\psi_h})dx - a_D(\hat{u}_h, \psi-\psi_h)\\
&\quad + \int_{\Gamma_1}(T_1 - T_1^{\rm PML})(\hat{u}_h - u_{\rm I})\bar{\psi}dx_1 + \int_{\Gamma_2}(T_2 - T_2^{\rm PML})\hat{u}_h\bar{\psi}dx_1\\
&\quad + \int_{\Gamma_1^{\rm PML}}\frac{1}{s(x_3)}\frac{\partial\bar{\psi}}{\partial x_3}(I_h u_{\rm I} - u_{\rm I})dx_1.
\end{aligned} \tag{4.4}$$

Proof. First by (2.28), (2.22), and (2.8) we have

$$\begin{aligned}
b(u-\hat{u}_h,\psi) &= b(u-\hat{u},\psi)+b(\hat{u}-\hat{u}_h,\psi)\\
&= \int_{\Gamma_1}(T_1-T_1^{\rm PML})(\hat{u}-u_{\rm I})\bar{\psi}dx_1+\int_{\Gamma_2}(T_2-T_2^{\rm PML})\hat{u}\bar{\psi}dx_1\\
&\quad + b^{\rm PML}(\hat{u}-\hat{u}_h,\psi)-\sum_{j=1}^{2}\int_{\Gamma_j}(T_j-T_j^{\rm PML})(\hat{u}-\hat{u}_h)\bar{\psi}dx_1\\
&= \int_{\Gamma_1}(T_1-T_1^{\rm PML})(\hat{u}_h-u_{\rm I})\bar{\psi}dx_1+\int_{\Gamma_2}(T_2-T_2^{\rm PML})\hat{u}_h\bar{\psi}dx_1\\
&\quad + b^{\rm PML}(\hat{u}-\hat{u}_h,\psi).
\end{aligned}$$

Next, by (2.22) and Lemma 4.1, we obtain

$$\begin{aligned}
b^{\rm PML}(\hat{u}-\hat{u}_h,\psi) &= a_{\Omega}(\hat{u}-\hat{u}_h,\psi)-\sum_{j=1}^{2}\int_{\Gamma_j}T_j^{\rm PML}(\hat{u}-\hat{u}_h)\bar{\psi}dx_1\\
&= a_{\Omega}(\hat{u}-\hat{u}_h,\psi)+\sum_{j=1}^{2}\int_{\Gamma_j}(\hat{u}-\hat{u}_h)\frac{\partial\bar{\psi}}{\partial\nu_j}dx_1.
\end{aligned}$$

Recall that ν_j is the unit outer normal to $\partial\Omega_j^{\rm PML}$. Since $\mathcal{L}\bar{\psi}=0$ in $\Omega_j^{\rm PML}$, we deduce by the Green formula that

$$a_{\Omega_j^{\rm PML}}(\hat{u}-\hat{u}_h,\psi)=\int_{\Gamma_j}(\hat{u}-\hat{u}_h)\frac{\partial\bar{\psi}}{\partial\nu_j}dx_1+\int_{\Gamma_j^{\rm PML}}\frac{1}{s(x_3)}\frac{\partial\bar{\psi}}{\partial\nu_j}(\hat{u}-\hat{u}_h)dx_1.$$

Thus, by using (2.15) and (3.1), we conclude that

$$\begin{aligned}
b^{\rm PML}(\hat{u}-\hat{u}_h,\psi) &= a_D(\hat{u}-\hat{u}_h,\psi)-\int_{\Gamma_j^{\rm PML}}\frac{1}{s(x_3)}\frac{\partial\bar{\psi}}{\partial\nu_j}(\hat{u}-\hat{u}_h)dx_1\\
&= \int_D g(\overline{\psi-\psi_h})dx-a_D(\hat{u}_h,\psi-\psi_h)-\int_{\Gamma_1^{\rm PML}}\frac{1}{s(x_3)}\frac{\partial\bar{\psi}}{\partial x_3}(\hat{u}-\hat{u}_h)dx_1,
\end{aligned}$$

where we have used the fact that $\hat{u}=\hat{u}_h=0$ on $\Gamma_2^{\rm PML}$. This completes the proof. □

We remark that evaluating the various terms in the error representation formula in suitable Sobolev norms would yield the desired a posteriori error estimate in Theorem 3.1. To achieve the goal, we need to prove stability estimates for the extension (4.1) of the function ψ in $\Omega_j^{\rm PML}$.

4.2. Estimates for the extension. We begin by introducing the notation

$$R_1(x_3)=\min\left(\Delta_1^-\int_{b_1}^{x_3}s_2(\tau)d\tau,\Delta_1^+\int_{b_1}^{x_3}s_1(\tau)d\tau\right),\quad x_3\ge b_1, \tag{4.5}$$

$$R_2(x_3)=\min\left(\Delta_2^-\int_{x_3}^{b_2}s_2(\tau)d\tau,\Delta_2^+\int_{x_3}^{b_2}s_1(\tau)d\tau\right),\quad x_3\le b_2. \tag{4.6}$$

The objective of this section is to prove the following two lemmas.

LEMMA 4.3. *For any $\psi \in X(\Omega)$, let ψ be extended to the whole domain D according to* (4.1). *Then we have the following estimates, for $j = 1, 2$:*

$$\| s^{-1} e^{R_j} \nabla \psi \|_{L^2(\Omega_j^{\mathrm{PML}})} \le C_j \| \psi \|_{H^1(\Omega)},$$

where

$$C_j = \hat{C} \max\left(\frac{2k_j \delta_j^{1/2}}{1 - e^{-2\Delta_j^- \sigma_j^I}}, \quad \frac{2(1 + 2\delta_j(\Delta_j^+ + k_j))^{1/2}}{1 - e^{-2\Delta_j^+ \sigma_j^R}} \right).$$

Proof. We define

$$r_1(x_3) = \int_{x_3}^{b_1+\delta_1} s(\tau) d\tau, \qquad r_2(x_3) = \int_{b_2-\delta_2}^{x_3} s(\tau) d\tau.$$

Then we have $\zeta_j^n(x_3) = e^{-\mathbf{i}\beta_j^n r_j(x_3)} - e^{\mathbf{i}\beta_j^n r_j(x_3)}$ and consequently

$$\frac{d\bar{\zeta}_j^n}{dx_3} = \mathbf{i}\bar{\beta}_j^n (-1)^j \bar{s}(x_3) \left[e^{\mathbf{i}\bar{\beta}_j^n \bar{r}_j(x_3)} + e^{-\mathbf{i}\bar{\beta}_j^n \bar{r}_j(x_3)} \right].$$

By direct calculation, we deduce from (4.1) that

$$\begin{aligned} \int_0^L |\nabla \psi|^2 dx_1 = L \sum_{n \in Z} |\alpha_n + \alpha|^2 |e^{-\mathbf{i}\beta_j^n r_j(x_3)} - e^{\mathbf{i}\beta_j^n r_j(x_3)}|^2 |\zeta_j^n(b_j)|^{-2} |\psi_\alpha^{(n)}(b_j)|^2 \\ + L \sum_{n \in Z} |\beta_j^n|^2 |s(x_3)|^2 |e^{-\mathbf{i}\beta_j^n r_j(x_3)} + e^{\mathbf{i}\beta_j^n r_j(x_3)}|^2 |\zeta_j^n(b_j)|^{-2} |\psi_\alpha^{(n)}(b_j)|^2. \end{aligned} \tag{4.7}$$

Since $\mathrm{Re}\,\beta_j^n \ge 0$ and $\mathrm{Im}\,\beta_j^n \ge 0$, we have

$$\mathrm{Re}\,(-\mathbf{i}\beta_j^n r_j(x_3)) = \mathrm{Im}\,(\beta_j^n) r_j^R(x_3) + \mathrm{Re}\,(\beta_j^n) r_j^I(x_3) \ge 0.$$

Thus

$$\begin{aligned} &|e^{-\mathbf{i}\beta_j^n r_j(x_3)} \pm e^{\mathbf{i}\beta_j^n r_j(x_3)}|^2 \\ &\quad = |e^{2\mathbf{i}\mathrm{Im}\,(-\mathbf{i}\beta_j^n r_j(x_3))} \pm e^{-2\mathrm{Re}\,(-\mathbf{i}\beta_j^n r_j(x_3))}|^2 e^{2\mathrm{Re}\,(-\mathbf{i}\beta_j^n r_j(x_3))} \\ &\quad \le 4 e^{2(\mathrm{Im}\,(\beta_j^n) r_j^R(x_3) + \mathrm{Re}\,(\beta_j^n) r_j^I(s_3))}. \end{aligned}$$

Similarly, we have

$$\begin{aligned} |\zeta_j^n(b_j)|^2 &= |e^{-\mathbf{i}\beta_j^n \sigma_j} - e^{\mathbf{i}\beta_j^n \sigma_j}|^2 \\ &= |e^{2\mathbf{i}\mathrm{Im}\,(-\mathbf{i}\beta_j^n \sigma_j)} - e^{-2\mathrm{Re}\,(-\mathbf{i}\beta_j^n \sigma_j)}|^2 e^{2\mathrm{Re}\,(-\mathbf{i}\beta_j^n \sigma_j)} \\ &\ge e^{2(\sigma_j^R \mathrm{Im}\,\beta_j^n + \sigma_j^I \mathrm{Re}\,\beta_j^n)} |1 - e^{-2(\sigma_j^R \mathrm{Im}\,\beta_j^n + \sigma_j^I \mathrm{Re}\,\beta_j^n)}|^2. \end{aligned}$$

Therefore, by $r_j(b_j) = \sigma_j$, we have

$$\begin{aligned} &|e^{-\mathbf{i}\beta_j^n r_j(x_3)} \pm e^{\mathbf{i}\beta_j^n r_j(x_3)}|^2 |\zeta_j^n(b_j)|^{-2} \\ &\quad \le 4 e^{-2(\mathrm{Im}\,\beta_j^n |\int_{b_j}^{x_3} s_1(\tau) d\tau| + \mathrm{Re}\,\beta_j^n |\int_{b_j}^{x_3} s_2(\tau) d\tau|)} (1 - e^{-2(\sigma_j^R \mathrm{Im}\,\beta_j^n + \sigma_j^I \mathrm{Re}\,\beta_j^n)})^{-2}. \end{aligned} \tag{4.8}$$

For $n \in U_j$, we have $\mathrm{Re}\,\beta_j^n = \Delta_j^n \ge \Delta_j^-$, $\mathrm{Im}\,\beta_j^n = 0$, and thus

$$|e^{-\mathbf{i}\beta_j^n r_j(x_3)} \pm e^{\mathbf{i}\beta_j^n r_j(x_3)}|^2 |\zeta_j^n(b_j)|^{-2} \le 4 e^{-2R_j(x)} (1 - e^{-2\Delta_j^- \sigma_j^I})^{-2}.$$

Similarly, for $n \notin U_j$, we have $\operatorname{Re}\beta_j^n = 0, \operatorname{Im}\beta_j^n = \Delta_j^n \geq \Delta_j^+$, and thus

$$\begin{aligned}
&|e^{-\mathbf{i}\beta_j^n r_j(x_3)} \pm e^{\mathbf{i}\beta_j^n r_j(x_3)}|^2 |\zeta_j^n(b_j)|^{-2} \\
&\qquad \leq 4e^{-2(\Delta_j^n-\Delta_j^+)|\int_{b_j}^{x_3} s_1(\tau)d\tau|} e^{-2R_j(x_3)}(1-e^{-2\Delta_j^+\sigma_j^R})^{-2} \\
&\qquad \leq 4e^{-2R_j(x_3)} e^{-2(\Delta_j^n-\Delta_j^+)|x_3-b_j|}(1-e^{-2\Delta_j^+\sigma_j^R})^{-2},
\end{aligned}$$

where we have used the fact $s_1(\tau) \geq 1$ (see (2.11)). Define $I_1 = (b_1, b_1+\delta_1)$ and $I_2 = (b_2-\delta_2, b_2)$. It is easy to see that

$$\int_{I_j} e^{-2(\Delta_j^n-\Delta_j^+)|x_3-b_j|} dx_3 \leq \min\left(\delta_j, \frac{1}{2(\Delta_j^n-\Delta_j^+)}\right).$$

Then, by substituting the above estimates into (4.7), we obtain that

$$\begin{aligned}
&\|s^{-1}e^{R_j}\nabla\psi\|^2_{L^2(\Omega_j^{\mathrm{PML}})} = \int_{I_j} |s(x_3)|^{-2} e^{2R_j(x_3)} \int_0^L |\nabla\psi(x_1,x_3)|^2 dx_1 dx_3 \\
&\quad \leq 4L\delta_j \sum_{n\in U_j} \frac{|\alpha_n+\alpha|^2+|\beta_j^n|^2}{(1-e^{-2\Delta_j^-\sigma_j^I})^2} |\psi_\alpha^{(n)}(b_j)|^2 \\
&\qquad + 4L \sum_{n\notin U_j} \frac{|\alpha_n+\alpha|^2+|\beta_j^n|^2}{(1-e^{-2\Delta_j^+\sigma_j^R})^2} |\psi_\alpha^{(n)}(b_j)|^2 \min\left(\delta_j, \frac{1}{2(\Delta_j^n-\Delta_j^+)}\right) := \mathrm{I} + \mathrm{II}.
\end{aligned}$$

If $n \in U_j$, then $|\alpha_n+\alpha|^2 + |\beta_j^n|^2 = k_j^2$ and we get

$$\mathrm{I} \leq C_j^2 \hat{C}^{-2} L \sum_{n\in U_j} |\psi_\alpha^{(n)}(b_j)|^2.$$

If $n \notin U_j$, then $|\alpha_n+\alpha|^2 - k_j^2 = |\beta_j^n|^2 = |\Delta_j^n|^2$ and we have $|\alpha_n+\alpha|^2 + |\beta_j^n|^2 \leq k_j^2 + 2|\Delta_j^n|^2 \leq 2(k_j+\Delta_j^n)^2$ and $|\alpha_n+\alpha|^2 + |\beta_j^n|^2 \leq 2|\alpha_n+\alpha|^2$. Hence

$$|\alpha_n+\alpha|^2 + |\beta_j^n|^2 \leq 2|\alpha_n+\alpha|(k_j+\Delta_j^n).$$

Therefore

$$(|\alpha_n+\alpha|^2 + |\beta_j^n|^2) \min\left(\delta_j, \frac{1}{2(\Delta_j^n-\Delta_j^+)}\right) \leq |\alpha_n+\alpha|(1+2\delta_j(\Delta_j^+ + k_j)),$$

which yields

$$\mathrm{II} \leq C_j^2 \hat{C}^{-2} L \sum_{n\notin U_j} |\alpha_n+\alpha| |\psi_\alpha^{(n)}(b_j)|^2.$$

This completes the proof upon using Lemma 2.3. □

LEMMA 4.4. *For any $\psi \in X(\Omega)$, let ψ be extended to the whole domain D according to* (4.1). *Then we have the following estimate:*

$$\left\| s^{-1}\frac{\partial\psi}{\partial x_3} \right\|_{L^2(\Gamma_1^{\mathrm{PML}})} \leq \hat{C} M_3 \|\, \psi \,\|_{H^1(\Omega)},$$

where

$$M_3 = \max\left(\frac{2\Delta_1^- e^{-\Delta_1^- \sigma_1^I}}{1 - e^{-2\Delta_1^- \sigma_1^I}},\ \ \frac{2\Delta_1^+ e^{-\Delta_1^+ \sigma_1^R}}{1 - e^{-2\Delta_1^+ \sigma_1^R}}\right).$$

Proof. From (4.1) we deduce easily that

$$\frac{\partial \psi}{\partial x_3}(x_1, b_1 + \delta_1) = -2 \sum_{n \in Z} \mathbf{i}\bar{\beta}_1^n \bar{s}(b_1 + \delta_1)\bar{\zeta}_1^n(b_1)^{-1} \psi_\alpha^{(n)}(b_1) e^{\mathbf{i}(\alpha_n + \alpha)x_1}.$$

Thus

$$\left\| s^{-1} \frac{\partial \psi}{\partial x_3} \right\|_{L^2(\Gamma_1^{\mathrm{PML}})} = 2\left(\sum_{n \in Z} |\beta_1^n|^2 |\zeta_1^n(b_1)|^{-2} |\psi_\alpha^{(n)}(b_1)|^2\right)^{1/2}.$$

Moreover,

$$|\beta_1^n||\zeta_1^n(b_1)|^{-1} = \left|\frac{\beta_1^n e^{\mathbf{i}\beta_1^n \sigma_1}}{1 - e^{2\mathbf{i}\beta_1^n \sigma_1}}\right|.$$

The proof now follows by using an argument similar to that used in Lemma 2.2 and using Lemma 2.3. □

4.3. Proof of Theorem 3.1. Since we are going to interpolate nonsmooth functions satisfying quasi-periodic boundary conditions, we resort to an interpolation operator $\Pi_h : \overset{\circ}{X}(D) \to \overset{\circ}{V}_h(D)$ of Scott and Zhang [21]. Let $N_h = \{a_i\}_{i=1}^N$ be the set of all nodes of $\mathcal{M}_h$, and $\{\phi_i\}_{i=1}^N$ be the corresponding nodal basis of $V_h(D)$. For any node a_i that is interior to D, we take $\sigma_i = e$, any side in $\mathcal{B}_h$ having a_i as one of its vertices. For any node a_i that is in the interior of the left boundary, that is, $a_i = (0, z_i)$ for some $z_i \in (b_2 - \delta_2, b_1 + \delta_1)$, we take σ_i as any side on the left boundary with one vertex a_i. Now for the corresponding node $a_k = (L, z_i)$ on the right boundary, we choose σ_k as the corresponding side of σ_i on the right boundary. For the nodes a_i lying on $\overline{\Gamma_1^{\mathrm{PML}}} \cup \overline{\Gamma_2^{\mathrm{PML}}}$, we can choose σ_i as any side on $\overline{\Gamma_1^{\mathrm{PML}}}$ or $\overline{\Gamma_2^{\mathrm{PML}}}$ which has a_i as one vertex. Let $a_{i,1} = a_i$, and let $\{a_{i,j}\}_{j=1}^2$ be the set of nodal points in σ_i with nodal basis $\{\phi_{i,j}\}_{j=1}^2$. Define $\{\psi_{i,j}\}_{j=1}^2$ as the $L^2(\sigma_i)$ dual basis:

$$\int_{\sigma_i} \psi_{i,j}(x)\phi_{i,k}(x)ds = \delta_{jk}, \quad j, k = 1, 2,$$

where δ_{jk} is the Kronecker delta. We let $\psi_i = \psi_{i,1}$. Then the interpolation operator $\Pi_h : H^1(D) \to W_h(D)$, the conforming linear finite element space, is defined by

$$\Pi_h v(x) = \sum_{i=1}^N \phi_i(x) \int_{\sigma_i} \psi_i(x) v(x) ds.$$

This operator enjoys the following interpolation estimates (see [21]):

$$(4.9)\quad \| v - \Pi_h v \|_{L^2(T)} \le C h_T \| \nabla v \|_{L^2(\tilde{T})}, \qquad \| v - \Pi_h v \|_{L^2(e)} \le C h_e^{1/2} \| \nabla v \|_{L^2(\tilde{e})},$$

where $\tilde{T}$ and $\tilde{e}$ are the union of all elements in $\mathcal{M}_h$ having nonempty intersection with $T \in \mathcal{M}_h$ and the side e, respectively.

It remains to check whether Π_h keeps the boundary condition. It is clear that $\Pi_h v = 0$ on $\Gamma_1^{\rm PML}\cup\Gamma_2^{\rm PML}$ since for any $a_i \in \overline{\Gamma_1^{\rm PML}\cup\Gamma_2^{\rm PML}}, \sigma_i \subset \overline{\Gamma_1^{\rm PML}}$ or $\overline{\Gamma_2^{\rm PML}}$ and $v=0$ on $\Gamma_1^{\rm PML}\cup\Gamma_2^{\rm PML}$. Now let $a_i=(0,z_i)\in\Gamma_{\rm left}$ and $a_k=(L,z_i)\in\Gamma_{\rm right}$. Without loss of generality, we assume $\sigma_i=\{x\in\mathbf{R}^2 : x_1=0, z_i\le x_3\le z_{i+1}\}$. Then by construction we have $\sigma_k=\{x\in\mathbf{R}^2 : x_1=L, z_i\le x_3\le z_{i+1}\}$. The nodal basis $\phi_{i,1}=(z_{i+1}-x_3)/(z_{i+1}-z_i)$, $\phi_{i,2}=(x_3-z_i)/(z_{i+1}-z_i)$ in σ_i, and simple calculation yields the dual basis

$$\psi_{i,1}=\frac{4}{d_i}\phi_{i,1}-\frac{2}{d_i}\phi_{i,2},\qquad \psi_{i,2}=-\frac{2}{d_i}\phi_{i,1}+\frac{4}{d_i}\phi_{i,2}\quad\text{in}\quad\sigma_i,$$

where $d_i=z_{i+1}-z_i$. Similar computation implies that

$$\psi_k(L,x_3)=\psi_i(0,x_3)=\frac{4}{d_i}\phi_{i,1}(x_3)-\frac{2}{d_i}\phi_{i,2}(x_3).$$

Thus for any $v\in X(D)$, that is, $v(0,x_3)=e^{-\mathbf{i}\alpha L}v(L,x_3)$, we have

$$\Pi_h v(a_i)=\int_{\sigma_i}\psi_i(0,x_3)v(0,x_3)dx_3=e^{-\mathbf{i}\alpha L}\int_{\sigma_k}\psi_k(L,x_3)v(L,x_3)dx_3=e^{-\mathbf{i}\alpha L}\Pi_h v(a_k).$$

This shows that $\Pi_h v\in\overset{\circ}{V}_h(D)$ if $v\in\overset{\circ}{X}(D)$.

Now we take $\psi_h=\Pi_h\psi\in\overset{\circ}{V}_h(D)$ in the error representation formula (4.4) to get

$$\begin{aligned}b(u-\hat u_h,\psi) &= \int_D g(\overline{\psi-\Pi_h\psi})dx - a_{_D}(\hat u_h,\psi-\Pi_h\psi)\\ &\quad+\int_{\Gamma_1}(T_1-T_1^{\rm PML})(\hat u_h-u_{\rm I})\bar\psi dx_1+\int_{\Gamma_2}(T_2-T_2^{\rm PML})\hat u_h\bar\psi dx_1\\ &\quad+\int_{\Gamma_1^{\rm PML}}\frac{1}{s(x_3)}\frac{\partial\bar\psi}{\partial x_3}(I_h u_{\rm I}-u_{\rm I})dx_1\\ &:= {\rm III}+\cdots+{\rm VII}.\end{aligned}\tag{4.10}$$

We observe that, by integration by parts and using (3.3)–(3.5),

$${\rm III}+{\rm IV}=\sum_{T\in\mathcal{M}_h}\left(\int_T R_T(\overline{\psi-\Pi_h\psi})dx+\sum_{e\subset\partial T}\frac{1}{2}\int_e J_e(\overline{\psi-\Pi_h\psi})ds\right).$$

Standard argument in the a posteriori error analysis using (4.9) and Lemma 4.3 implies

$$\begin{aligned}|{\rm III}+{\rm IV}| &\le C\sum_{T\in\mathcal{M}_h}\eta_{_T}\|\,\rho^{-1}\nabla\psi\,\|_{L^2(\tilde T)}\\ &\le C(1+C_1+C_2)\left(\sum_{T\in\mathcal{M}_h}\eta_{_T}^2\right)^{1/2}\|\,\psi\,\|_{H^1(\Omega)}.\end{aligned}\tag{4.11}$$

By Lemmas 2.2 and 2.3, we obtain

$$\begin{aligned}|{\rm V}+{\rm VI}| &\le M_1\|\,\hat u_h-u_{\rm I}\,\|_{L^2(\Gamma_1)}\|\,\psi\,\|_{L^2(\Gamma_1)}+M_2\|\,\hat u_h\,\|_{L^2(\Gamma_2)}\|\,\psi\,\|_{L^2(\Gamma_2)}\\ &\le(\hat C M_1\|\,\hat u_h-u_{\rm I}\,\|_{L^2(\Gamma_1)}+\hat C M_2\|\,\hat u_h\,\|_{L^2(\Gamma_2)})\|\,\psi\,\|_{H^1(\Omega)}.\end{aligned}\tag{4.12}$$

Finally, by Lemmas 4.4 and 2.3, we get

$$\begin{aligned}|{\rm VII}| &\le M_3\|\,I_h u_{\rm I}-u_{\rm I}\,\|_{L^2(\Gamma_1^{\rm PML})}\|\,\psi\,\|_{L^2(\Gamma_1)}\\ &\le\hat C M_3\|\,I_h u_{\rm I}-u_{\rm I}\,\|_{L^2(\Gamma_1^{\rm PML})}\|\,\psi\,\|_{H^1(\Omega)}.\end{aligned}\tag{4.13}$$

Combining (4.10)–(4.13), we obtain the desired estimate. □

5. The $\mathrm{Im}\,\varepsilon_2 > 0$ case. In this section we consider briefly the case in which $\mathrm{Im}\,\varepsilon_2 > 0$. Let $k_2^2 = \omega^2\varepsilon_2\mu$ satisfy $\mathrm{Im}\,k_2 > 0$. The constants β_2^n in the Rayleigh expansion (2.5) in Ω_2 satisfy

$$(\beta_2^n)^2 = k_2^2 - (\alpha_n + \alpha)^2, \qquad \mathrm{Im}\,\beta_2^n \geq 0.$$

Define

$$\xi_n = \frac{1}{2}(\mathrm{Re}\,(k_2^2) - (\alpha_n + \alpha)^2), \qquad \eta = \frac{1}{2}\mathrm{Im}\,(k_2^2);$$

then $(\beta_2^n)^2 = 2(\xi_n + \mathbf{i}\eta)$. Since $\mathrm{Im}\,(k_2^2) = \omega^2\mathrm{Im}\,(\varepsilon_2)\mu > 0$, we obtain that

$$\mathrm{Re}\,\beta_2^n = (\sqrt{\xi_n^2 + \eta^2} + \xi_n)^{1/2}, \qquad \mathrm{Im}\,\beta_2^n = (\sqrt{\xi_n^2 + \eta^2} - \xi_n)^{1/2}. \tag{5.1}$$

Thus $\mathrm{Re}\,(-2\mathbf{i}\beta_2^n\sigma_2) = 2(\sigma_2^R\mathrm{Im}\,\beta_2^n + \sigma_2^I\mathrm{Re}\,\beta_2^n) \geq 2\sigma_2^R\mathrm{Im}\,\beta_2^n$, which yields

$$\begin{aligned}|\mathbf{i}\beta_2^n(1 - \coth(-\mathbf{i}\beta_2^n\sigma_2))| = \left|\frac{2\beta_2^n}{e^{-2\mathbf{i}\beta_2^n\sigma_2} - 1}\right| &\leq \frac{2|\beta_2^n|}{e^{2\sigma_2^R\mathrm{Im}\,\beta_2^n} - 1} \\ &= \frac{2\sqrt{2}(\xi_n^2 + \eta^2)^{1/4}}{e^{2\sigma_2^R(\sqrt{\xi_n^2+\eta^2}-\xi_n)^{1/2}} - 1}.\end{aligned}$$

Simple but tedious calculation shows that the function on the right-hand side is increasing with respect to $\xi_n \in \mathbf{R}$. Since $\xi_n \leq \frac{1}{2}\mathrm{Re}\,(k_2^2)$, we deduce after some algebraic manipulations that

$$|\mathbf{i}\beta_2^n(1 - \coth(-\mathbf{i}\beta_2^n\sigma_2))| \leq \mathcal{M}_2 := \frac{2|k_2|}{e^{2\sigma_2^R\mathrm{Im}\,k_2} - 1}.$$

Therefore, we conclude that Lemma 2.2 and thus Theorems 2.4 and 2.5 are also valid in the case $\mathrm{Im}\,\varepsilon_2 > 0$, with the definition of M_2 being replaced by $M_2 = \mathcal{M}_2$.

Moreover, we deduce from (4.8) that

$$|e^{-\mathbf{i}\beta_2^n r_2(x_3)} \pm e^{\mathbf{i}\beta_2^n r_2(x_3)}|^2|\zeta_2^n(b_2)|^{-2} \leq 4e^{-2\mathrm{Im}\,\beta_2^n\int_{x_3}^{b_2} s_1(\tau)d\tau}(1 - e^{-2\sigma_2^R\mathrm{Im}\,\beta_2^n})^{-2}.$$

Since $\xi_n \leq \frac{1}{2}\mathrm{Re}\,(k_2^2)$, we see easily from (5.1) that $\mathrm{Im}\,\beta_2^n \geq \mathrm{Im}\,k_2$. Now the argument in the proof of Lemma 4.3 implies that

$$\|\, s^{-1}e^{\mathcal{R}_2}\nabla\psi\,\|_{L^2(\Omega_2^{\mathrm{PML}})} \leq \mathcal{C}_2\|\,\psi\,\|_{H^1(\Omega)},$$

with $\mathcal{R}_2(x_3) = \mathrm{Im}\,(k_2)\int_{x_3}^{b_2} s_1(\tau)d\tau$ and

$$\mathcal{C}_2 = \hat{C}\frac{2[\max(1,|k_2|)(1 + 2\delta_2(\mathrm{Im}\,k_2 + |k_2|))]^{1/2}}{1 - e^{-2\sigma_2^R\mathrm{Im}\,k_2}}.$$

Therefore, we know that Theorems 3.1 and 3.2 are also valid, with the definitions of $R_2(x_3)$, C_2, and M_2 being replaced by $R_2 = \mathcal{R}_2(x_3)$, $C_2 = \mathcal{C}_2$, and $M_2 = \mathcal{M}_2$, respectively.

6. Implementation and numerical examples. The implementation of the adaptive algorithm in this section is based on the PDE toolbox of MATLAB. We use the a posteriori error estimate from Theorem 3.1 in the TE case and from Theorem 3.2 in the TM case to determine the PML parameters. According to the discussion in section 2, we choose the PML medium property as the power function, and thus we need to specify only the thickness δ_j of the layers and the medium parameters σ_j^m (see (2.26)). Recall from Theorem 3.1 that the a posteriori error estimate consists of two parts: the PML error $\mathcal{E}_{\mathrm{PML}}$ and the finite element discretization error $\mathcal{E}_{\mathrm{FEM}}$, where

$$\mathcal{E}_{\mathrm{PML}} = M_1\| \hat{u}_h - u_{\mathrm{I}} \|_{L^2(\Gamma_1)} + M_2\| \hat{u}_h \|_{L^2(\Gamma_2)}, \tag{6.1}$$

$$\mathcal{E}_{\mathrm{FEM}} = M_3\| \hat{u}_h - u_{\mathrm{I}} \|_{L^2(\Gamma_1^{\mathrm{PML}})} + \left(\sum_{T\in\mathcal{M}_h} \eta_T^2 \right)^{1/2}. \tag{6.2}$$

$\mathcal{E}_{\mathrm{PML}}$ and $\mathcal{E}_{\mathrm{FEM}}$ should be changed accordingly in the TM case. In our implementation we first choose δ_j and σ_j^m such that $M_j L^{1/2} \le 10^{-8}$, which makes the PML error negligible compared with the finite element discretization errors. Once the PML region and the medium property are fixed, we use the standard finite element adaptive strategy to modify the mesh according to the a posteriori error estimate (6.2). For any $T \in \mathcal{M}_h$, we define the local a posteriori error estimator as follows:

$$\tilde{\eta}_T = \eta_T + M_3\| I_h u_{\mathrm{I}} - u_{\mathrm{I}} \|_{L^2(\Gamma_1^{\mathrm{PML}}\cap\partial T)}.$$

Now we describe the adaptive algorithm we have used in this paper.

ALGORITHM 6.1. Given tolerance TOL > 0. Let $m = 2$, $\delta_1 = \delta_2 = \delta$.

```
• Choose $\delta$ and $\sigma_j^m$ such that $M_j L^{1/2} \le 10^{-8}$ for $j = 1, 2$;
• Set the computational domain $D = \Omega_2^{\mathrm{PML}} \cup \Gamma_2 \cup \Omega \cup \Gamma_1 \cup \Omega_1^{\mathrm{PML}}$ and generate
  an initial mesh $\mathcal{M}_h$ over $D$;
• While $\mathcal{E}_{\mathrm{FEM}} >$ TOL do
    – refine the mesh $\mathcal{M}_h$ according to the strategy
          if $\tilde{\eta}_T > \frac{1}{2}\max_{T\in\mathcal{M}_h} \tilde{\eta}_T$, refine the element $T \in \mathcal{M}_h$,
    – solve the discrete problem (3.1) or (3.7) on $\mathcal{M}_h$,
    – compute error estimators on $\mathcal{M}_h$,
  end while.
```

In the following, we report several numerical examples to demonstrate the competitive behavior of the proposed algorithm. In all the experiments we document only the value δ of the thickness of the PML layers. The medium parameters σ_j^m are determined accordingly through the relation $M_j L^{1/2} \le 10^{-8}$ for $j = 1, 2$. We normalize the space variables so that $\mu = 1$. We also scale the error estimator by a factor of 0.15 as in the PDE toolbox of MATLAB.

Example 1. We consider the simplest grating structure, a straight line. Assume that a plane wave $u_{\mathrm{I}} = e^{\mathbf{i}\alpha x_1 - \mathbf{i}\beta x_3}$ is incident on the straight line $\{x_3 = 0\}$, which separates two homogeneous media whose dielectric coefficients are ε_1 and ε_2, respectively. The exact solution is known (see [7]):

$$u = \begin{cases} u_{\mathrm{I}} + r e^{\mathbf{i}\alpha x_1 + \mathbf{i}\beta x_3} & \text{if } x_3 > 0, \\ t e^{\mathbf{i}\alpha x_1 - \mathbf{i}\hat{\beta} x_3} & \text{if } x_3 < 0, \end{cases}$$

Table 6.1

Comparison of numerical results using adaptive and uniform mesh refinements. N_k is the number of nodal points of mesh $\mathcal{M}_k$. $\mathcal{E}_k = (\sum_{T\in\mathcal{M}_k} \tilde{\eta}_T^2)^{1/2}$ and $e_k = \|\nabla(u-\hat{u}_k)\|_{L^2(\Omega)}$.

Adaptive mesh				Uniform mesh			
k	N_k	e_k	$\mathcal{E}_k/e_k$	k	N_k	e_k	$\mathcal{E}_k/e_k$
0	29	8.7043	0.7830	0	29	8.7043	0.7830
2	62	3.4572	0.9866	1	71	3.7531	1.0117
4	133	2.1563	1.1251	2	143	2.6595	1.1021
5	195	1.6772	1.1143	3	283	1.7305	1.2505
6	302	1.3618	1.1088	4	575	1.3448	1.1709
7	477	0.9488	1.2212	5	1145	0.9556	1.2237
9	1135	0.5874	1.2348	6	2321	0.6984	1.1734
10	1838	0.4592	1.2371	7	4639	0.5065	1.1982
11	3379	0.3503	1.1864	8	9359	0.3549	1.1812

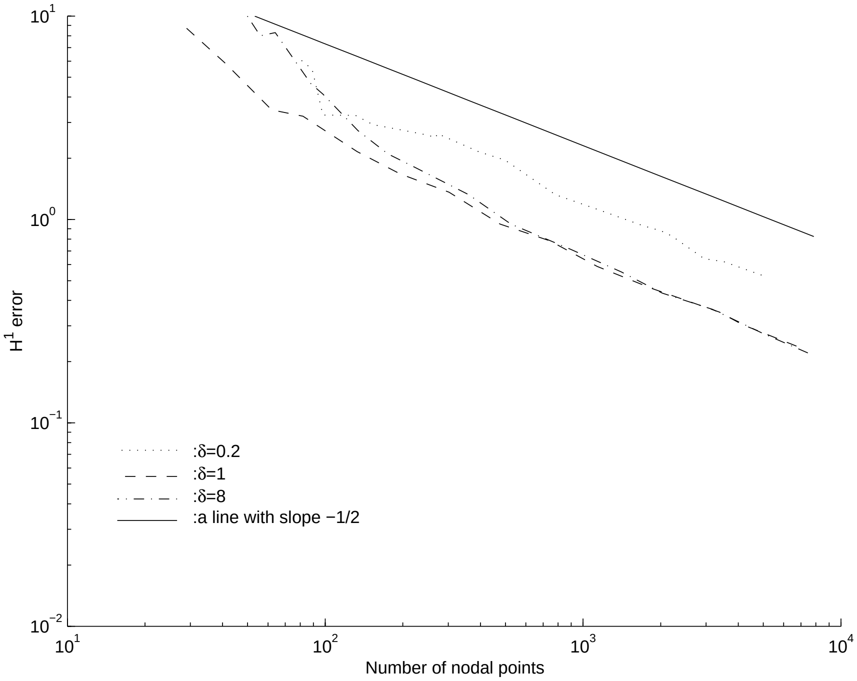


Fig. 6.1. *Quasioptimality of the adaptive mesh refinements.*

where $\hat{\beta} = (k_2^2 - \alpha^2)^{1/2}$, $t = 2\beta/(\beta+\hat{\beta})$, and $r = (\beta-\hat{\beta})/(\beta+\hat{\beta})$. The domain $\Omega = (0,L)\times(-b,b)$, $b>0$.

In our experiment, the parameters are chosen as $\varepsilon_1 = 1$, $\varepsilon_2 = (0.22+6.71\mathbf{i})^2$, $\theta = \pi/6$, $\omega = \pi$, and $L = 2$. Table 6.1 compares the results using adaptive and uniform mesh refinements when $b = \delta = 1$. It clearly shows the advantage of using adaptive mesh refinements. Moreover, our a posteriori error estimate $\mathcal{E}_k = (\sum_{T\in\mathcal{M}_k}\tilde{\eta}_T^2)^{1/2}$ provides a rather good estimate of the interested error $e_k = \|\nabla(u-\hat{u}_k)\|_{L^2(\Omega)}$.

Figure 6.1 shows the curves of $\log N_k$ versus $\log\|\nabla(u-\hat{u}_k)\|_{L^2(\Omega)}$, where N_k is the number of nodes of the mesh $\mathcal{M}_k$. It indicates that for the proposed method,

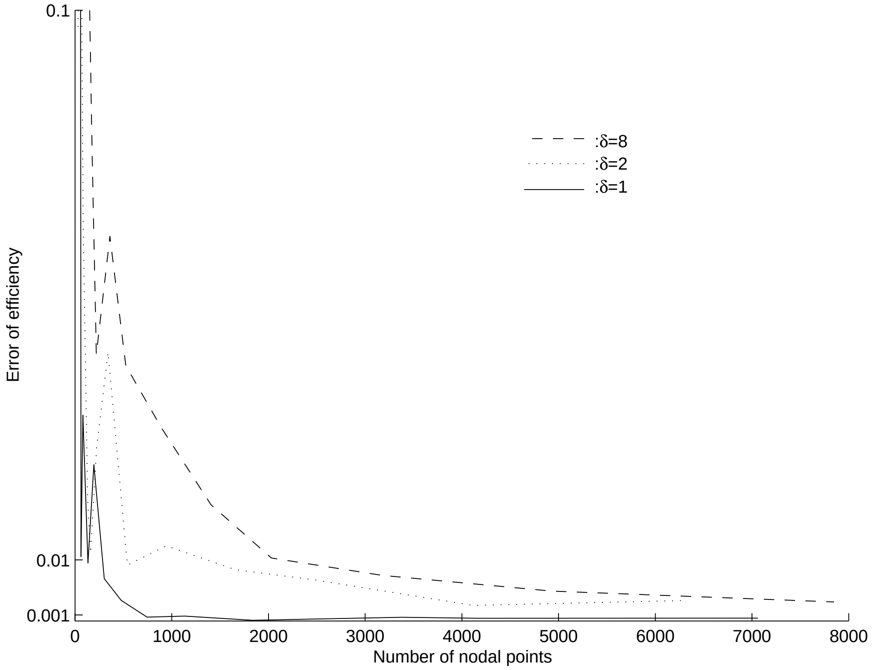


FIG. 6.2. *Robustness of the grating efficiency with respect to the thickness of PML layers.*

the meshes and the associated numerical complexity are quasi-optimal: $\| \nabla(u - \hat{u}_k) \|_{L^2(\Omega)} = CN_k^{-1/2}$ is valid asymptotically.

Figure 6.2 shows the robustness of the proposed method with respect to the choice of the thickness of PML layers: The error of the grating efficiency is insensitive to the thickness δ. The definition of the grating efficiency can be found, for example, in [7]. The reason for this robustness is the exponential decay factor in our a posteriori error estimator in the PML region. Our experiences indicate that the choice of δ in the range of $0.5L$ to $1.5L$ usually produces satisfactory results.

The parameter b determines the position of the PML layers, which, in the traditional PML technique, should be sufficiently far away from the grating surface to allow the evanescent waves to be sufficiently decayed. Figure 6.3 shows the robustness of our method with respect to the position of the PML layers, which is the purpose of our extended PML technique for attenuating both the outgoing and evanescent waves.

Example 2. This example concerns the TM polarization on a grating surface with a sharp angle, indicated in Figure 6.4. The parameters are the same as those in Example 1. There are two reflected outgoing waves. The grating efficiency of the reflected waves as well as the total grating efficiency are displayed in Figure 6.5. Figure 6.6 shows the mesh and the amplitude of the associated solution after 10 adaptive iterations when the grating efficiency is stabilized. The mesh has 3585 nodes, and the a posteriori error estimate over the mesh is 0.6468. The initial a posteriori error estimate is 3.7838. This example shows clearly the ability of the proposed method to capture the singularities of the problem. The meshes near the upper PML boundary are rather coarse, as a result of the exponential decay factor in our a posteriori error estimator.

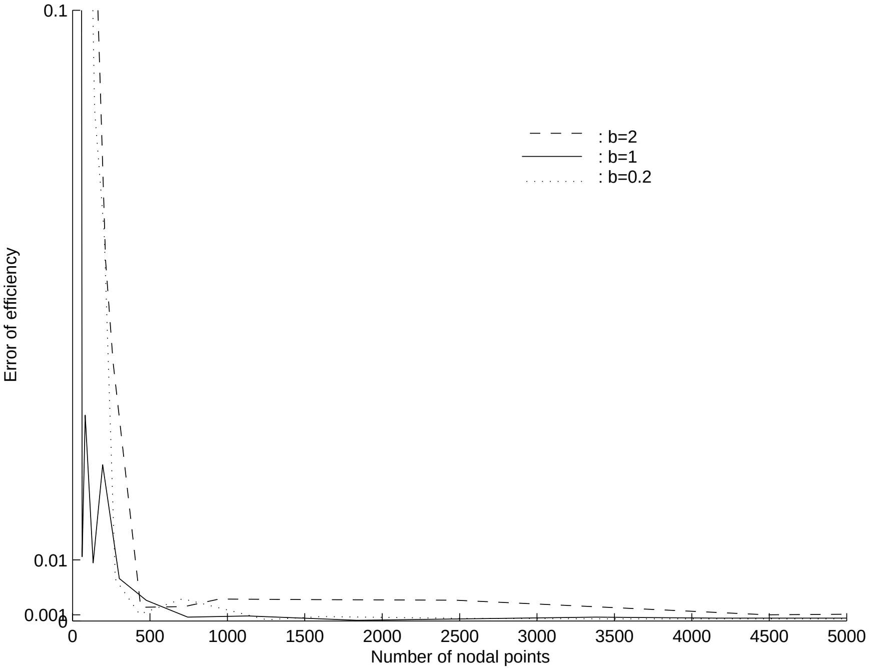


FIG. 6.3. *Robustness of the grating efficiency with respect to the position of PML layers.*

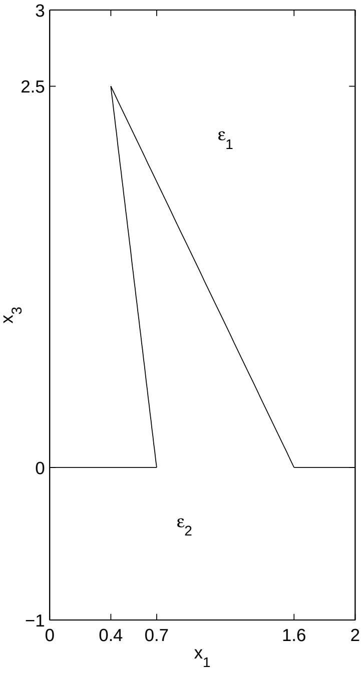


FIG. 6.4. *Geometry of the domain in Example* 2.

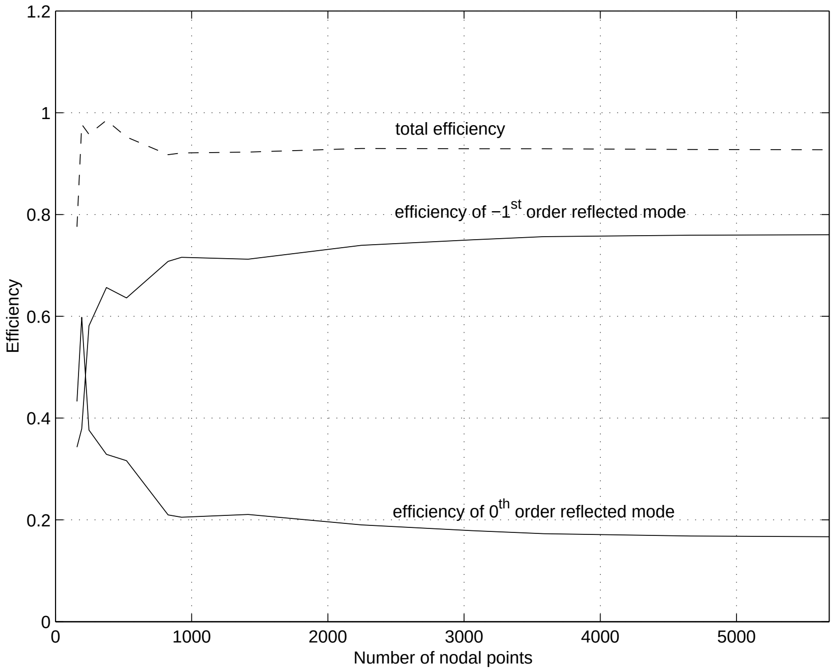


FIG. 6.5. *Grating efficiency of Example* 2.

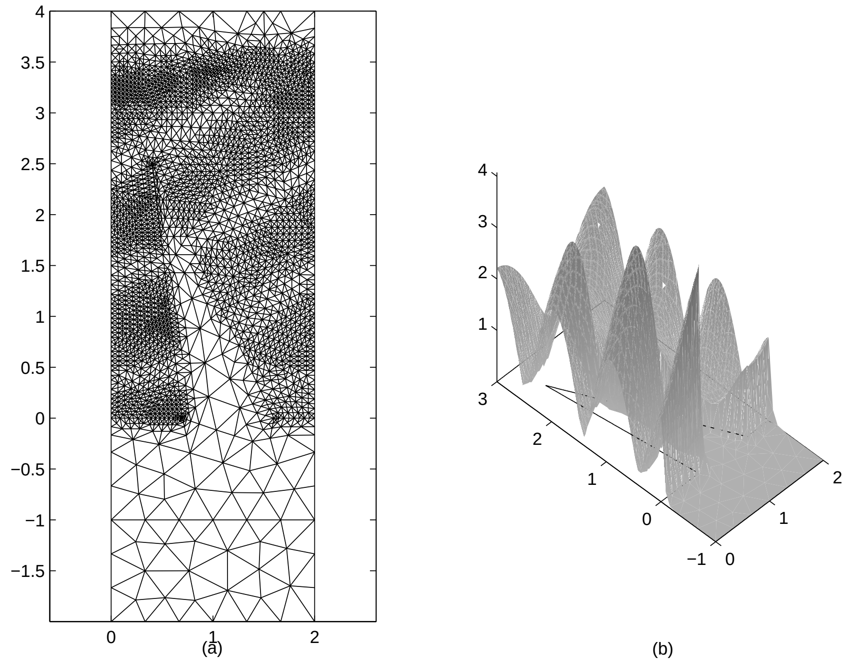


FIG. 6.6. *The mesh* (a) *and the surface plot of the amplitude of the associated solution* (b) *after* 10 *adaptive iterations. The mesh has* 3585 *nodes.*

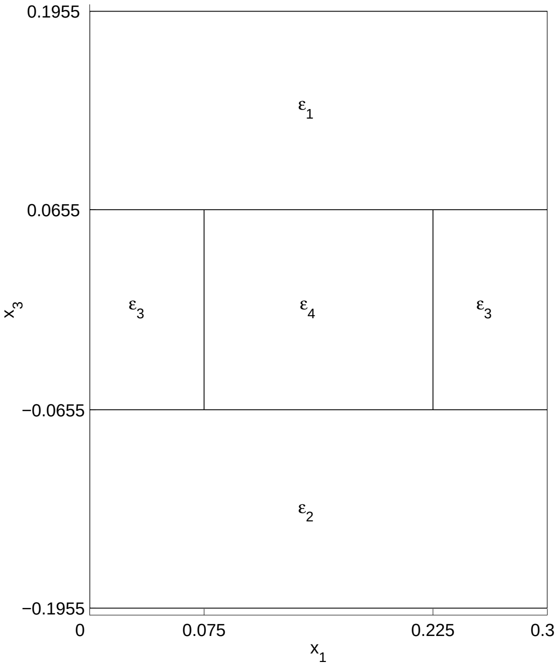


Fig. 6.7. *Geometry of the domain in Example* 3.

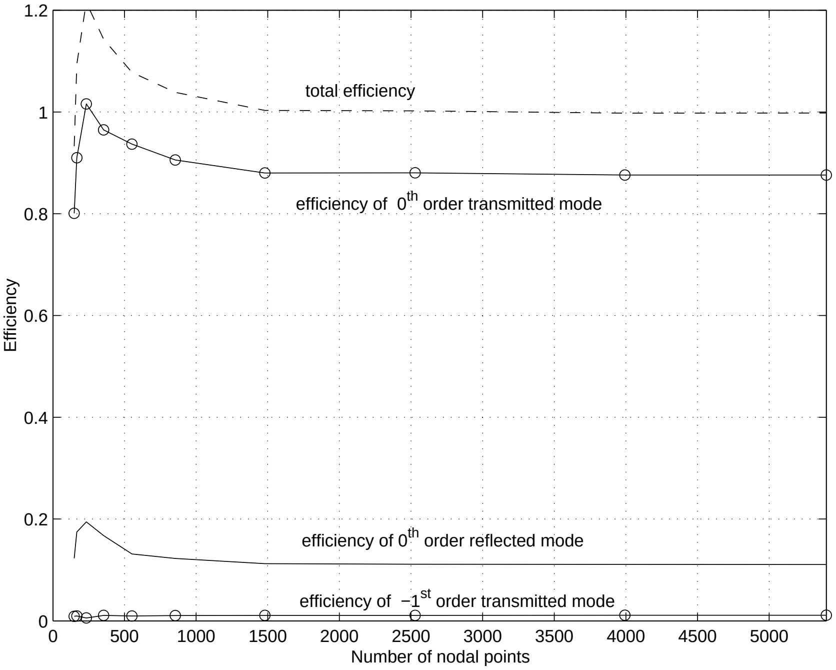


Fig. 6.8. *Grating efficiency of Example* 3.

Example 3. The final example is taken from [7] for TE polarization. The grating structure consists of multiple interfaces, as shown in Figure 6.7. This type of grating

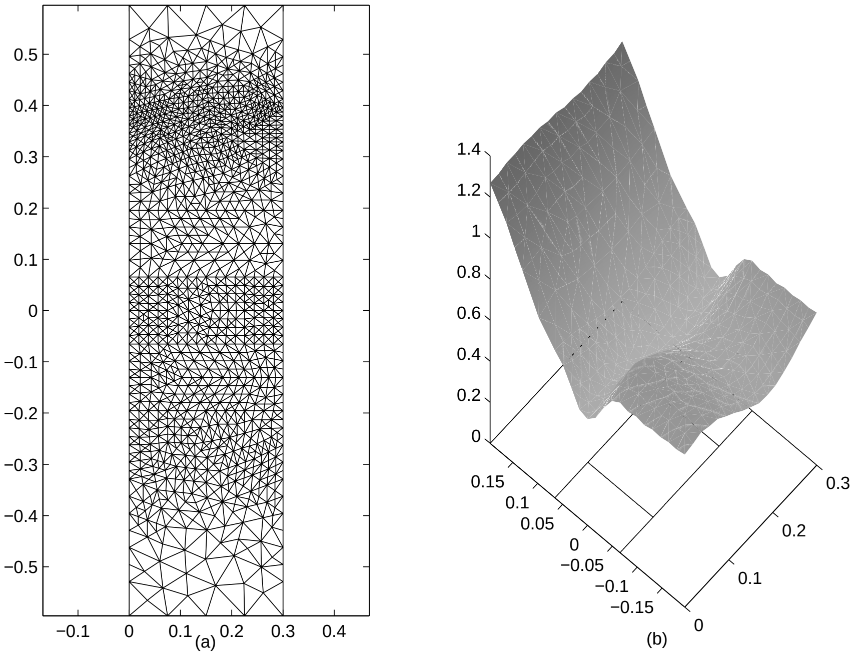


Fig. 6.9. *The mesh* (a) *and the surface plot of the amplitude of the associated solution* (b) *after* 7 *adaptive iterations. The mesh has* 1480 *nodes.*

structure has applications in optical filters and guided mode resonance devices. The parameters are taken as follows: $\varepsilon_1 = 1$, $\varepsilon_2 = 2.31$, $\varepsilon_3 = 4.4$, $\varepsilon_4 = 3.6$, $\theta = \pi/4$, $L = 0.3$, and $\omega = 2\pi/0.526$. The thickness of the PML layers $\delta = 0.4$. There are two transmitted waves and one reflected outgoing wave. The grating efficiency of the reflected and transmitted waves as well as the total grating efficiency are displayed in Figure 6.8. Figure 6.9 shows the mesh and the amplitude of the associated solution after 7 adaptive iterations when the grating efficiency is stabilized. The mesh has 1480 nodes, and the a posteriori error estimate over the mesh is 0.8189. The initial a posteriori error estimate is 3.0701. Again the meshes near the upper PML boundary are rather coarse, as a result of the exponential decay factor in our a posteriori error estimator.

Acknowledgment. The authors wish to thank Gang Bao for many inspiring discussions on the physical background and mathematical modeling of the grating problems.

REFERENCES

[1] T. Abboud, *Electromagnetic waves in periodic media*, in Proceedings of the Second International Conference on Mathematical and Numerical Aspects of Wave Propagation, R. Kleinman, T. Angell, D. Colton, F. Santosa, and I. Stakgold, eds., SIAM, Philadelphia, 1993, pp. 1–9.

[2] I. Babuška and A. Aziz, *Survey lectures on mathematical foundations of the finite element method*, in The Mathematical Foundations of the Finite Element Method with Application to Partial Differential Equations, A. Aziz, ed., Academic Press, New York, 1973, pp. 5–359.

[3] I. Babuška and W.C. Rheinboldt, *Error estimates for adaptive finite element computations*, SIAM J. Numer. Anal., 15 (1978), pp. 736–754.

[4] G. Bao, D.C. Dobson, and J.A. Cox, *Mathematical studies in rigorous grating theory*, J. Opt. Soc. Amer. A, 12 (1995), pp. 1029–1042.

[5] G. Bao, *Variational approximation of Maxwell's equations in biperiodic structures*, SIAM J. Appl. Math., 57 (1997), pp. 364–381.

[6] G. Bao, L. Cowsar, and W. Masters, eds., *Mathematical Modeling in Optical Science*, Frontiers Appl. Math. 22, SIAM, Philadelphia, 2001.

[7] G. Bao, Y. Cao, and H. Yang, *Numerical solution of diffraction problems by a least-square finite element method*, Math. Methods Appl. Sci., 23 (2000), pp. 1073–1092.

[8] J.-P. Berenger, *A perfectly matched layer for the absorption of electromagnetic waves*, J. Comput. Phys., 114 (1994), pp. 185–200.

[9] Z. Chen and S. Dai, *Adaptive Galerkin methods with error control for a dynamical Ginzburg–Landau model in superconductivity*, SIAM J. Numer. Anal., 38 (2001), pp. 1961–1985.

[10] Z. Chen and S. Dai, *On the efficiency of adaptive finite element methods for elliptic problems with discontinuous coefficients*, SIAM J. Sci. Comput., 24 (2002), pp. 443–462.

[11] Z. Chen, R.H. Nochetto, and A. Schmidt, *A characteristic Galerkin method with adaptive error control for continuous casting problem*, Comput. Methods Appl. Mech. Engrg., 189 (2000), pp. 249–276.

[12] F. Collino and P.B. Monk, *Optimizing the perfectly matched layer*, Comput. Methods Appl. Mech. Engrg., 164 (1998), pp. 157–171.

[13] D. Dobson, *A variational method for electromagnetic diffraction in biperiodic structures*, Math. Model. Numer. Anal., 28 (1994), pp. 419–439.

[14] D. Dobson and A. Friedman, *The time-harmonic Maxwell equations in a doubly periodic structure*, J. Math. Anal. Appl., 166 (1992), pp. 507–528.

[15] M. Lassas and E. Somersalo, *On the existence and convergence of the solution of PML equations*, Computing, 60 (1998), pp. 229–241.

[16] P. Monk, *A posteriori error indicators for Maxwell's equations*, J. Comput. Appl. Math., 100 (1998), pp. 173–190.

[17] P. Monk and E. Süli, *The adaptive computation of far-field patterns by a posteriori error estimation of linear functionals*, SIAM J. Numer. Anal., 36 (1998), pp. 251–274.

[18] P. Morin, R.H. Nochetto, and K.G. Siebert, *Data oscillation and convergence of adaptive FEM*, SIAM J. Numer. Anal., 38 (2000), pp. 466–488.

[19] R. Petit, ed., *Electromagnetic Theory of Gratings*, Topics in Current Physics 22, Springer-Verlag, Heidelberg, 1980.

[20] A.H. Schatz, *An observation concerning Ritz–Galerkin methods with indefinite bilinear forms*, Math. Comp., 28 (1974), pp. 959–962.

[21] L.R. Scott and S. Zhang, *Finite element interpolation of nonsmooth functions satisfying boundary conditions*, Math. Comp., 54 (1990), pp. 483–493.

[22] E. Turkel and A. Yefet, *Absorbing PML boundary layers for wave-like equations*, Appl. Numer. Math., 27 (1998), pp. 533–557.

[23] R. Verfürth, *A Review of A Posteriori Error Estimation and Adaptive Mesh Refinement Techniques*, Teubner, Stuttgart, 1996.

SIAM J. Numer. Anal.
Vol. 41, No. 3, pp. 827–855

COMPUTING THE ZEROS AND TURNING POINTS OF SOLUTIONS OF SECOND ORDER HOMOGENEOUS LINEAR ODES*

AMPARO GIL† AND JAVIER SEGURA‡

Abstract. Algorithms to compute the zeros and turning points of the solutions of second order ODEs $y'' + B(x)y' + A(x)y = 0$ are developed. Two fixed point methods are introduced. The first method consists of fixed point iterations that are built from the first order differential system relating the problem function with a contrast function w; the contrast function w has zeros interlaced with those of y, and it is a solution of a different second order ODE. This method, which generalizes previous findings [J. Segura, *SIAM J. Numer. Anal.*, 40 (2002), pp. 114–133], requires the evaluation of the ratio of functions y/w. The second method is based on fixed point iterations stemming from the second order ODE; it requires the computation of the logarithmic derivative y'/y. Both are quadratically convergent methods; error bounds are provided. The particular case of second order ODEs depending on one parameter, $y_n'' + B_n(x)y_n' + A_n(x)y_n(x) = 0$, with applications to the computation of the zeros and turning points of special functions, is discussed in detail. The combination of both methods provides algorithms for the efficient computation of the zeros and turning points of a broad family of special functions, including hypergeometric and confluent hypergeometric functions of real parameters and variables (Jacobi, Laguerre, and Hermite polynomials are particular cases), Bessel, Airy, Coulomb, and conical functions, among others. We provide numerical examples showing the efficiency of the methods.

Key words. second order ODEs, zeros and turning points, special functions, fixed point method

AMS subject classifications. 65H05, 33XX

PII. S0036142901392754

1. Introduction.

The study of the properties of zeros and turning points of special functions and orthogonal polynomials has become an active field in applied mathematics, which is not surprising, because of the importance of these topics in approximation theory, the theory of differential equations, and in many fields of physics and engineering.

In [1] a method of computation of zeros of special functions was presented whose applicability was illustrated for first kind Bessel functions and regular Coulomb wave functions. This article applies matrix eigenvalue methods, first put forward by Grad and Zakrajsek [10] and later enhanced by Ikebe and colleagues [11, 12, 15], who also applied the method for regular Bessel and Coulomb functions.

In [22] a fixed point method was introduced which covers the problems solved by matrix methods and can be applied to more general situations like, for instance, evaluating the zeros of *any* cylinder (Bessel) function or of *any* Coulomb function, and not necessarily the regular solutions. This method generalizes previous findings regarding Bessel functions [19, 21], and the fixed point iterations improve the convergence rate of the global Newton method considered in [19, 20]. The case of second order ODEs depending on one parameter, $y_n'' + B(x)y_n' + A_n(x)y_n(x) = 0$ (B not depending on n),

*Received by the editors July 23, 2001; accepted for publication (in revised form) November 21, 2002; published electronically May 22, 2003.

http://www.siam.org/journals/sinum/41-3/39275.html

†Departamento de Matemáticas, Universidad Autónoma de Madrid, 48039-Madrid, Spain, and Mathematics Department, Kassel University, 34132-Kassel, Germany (amparo.gil@uam.es). The research of this author was supported by the A. von Humboldt foundation.

‡Departamento de Matemáticas, Universidad Carlos III de Madrid, 28911-Leganés, Madrid, Spain, and Mathematics Department, Kassel University, 34132-Kassel, Germany. Current address: Departamento de Matemáticas, Estadistica y Computacion, Universidad de Cantabria, 39005 Santander, Spain (javier.segura@unican.es). The research of this author was supported by DAAD.

was described in detail. The global convergence of the resulting fixed point iteration was proved, and algorithms to compute with certainty the zeros inside an interval were given.

In this paper, the fixed point method described in [22], based on the first order systems associated with a set of second order ODEs, is established in its most general form (including the case of dependence of both A and B on the parameter n). In the particular case of hypergeometric functions, these fixed point methods can be directly applied for the computation of both the zeros and turning points (TPs). However, cases will be described for which these methods are difficult to apply to the computation of TPs, the main difficulty being that the change of variables is not analytically invertible. An alternative method for the computation of TPs is built. It is based on fixed point iterations stemming from the second order ODE. These methods require that an algorithm to evaluate the logarithmic derivative of the function be available. An explicit algorithm for the method based on the second order ODE is provided (section 3.3); this algorithm, combined with those described in [22], can be applied to the computation of the zeros and TPs of a broad family of special functions including hypergeometric and confluent hypergeometric functions of real parameters and variables (which includes classical orthogonal polynomials: Jacobi, Laguerre, Hermite), Bessel and Airy functions, Coulomb wave functions, and conical functions, among others.

Both methods are quadratically convergent, and error bounds are given. The methods are illustrated with three examples that combine both methods (Bessel, Coulomb, and conical functions). These examples show the efficiency of the proposed computational schemes. Comparison with other methods is also provided.

2. Method I: Fixed point methods based on first order systems.

In [22] a method for the computation of the zeros of the solutions of second order ODEs

$$y'' + B_y(x)y' + A_y(x)y = 0, \tag{2.1}$$

with continuous coefficients B_y and A_y in an interval I, was described. This method can be applied when contrast functions w exist which are solutions of a differential equation with continuous coefficients in I,

$$w'' + B_w(x)w' + A_w(x)w = 0,$$

such that, given two fundamental systems of solutions of these ODEs, $\{y^{(1)}, y^{(2)}\}$ and $\{w^{(1)}, w^{(2)}\}$, the functions y and w (twice continuously differentiable) are related by a first order system

$$\begin{aligned} y' &= \alpha(x)y + \delta(x)w, \\ w' &= \beta(x)w + \gamma(x)y \end{aligned} \tag{2.2}$$

both for $\{y^{(1)}, w^{(1)}\}$ and $\{y^{(2)}, w^{(2)}\}$, with continuous coefficients α, β, δ, γ in I. Such coefficients exist and are unique [14, Theorem 1], and therefore the restriction in the contrast function is given by the continuity of the coefficients. We restrict the analysis to real intervals I where all the coefficients are continuous and the solutions are twice continuously differentiable, which implies the differentiability of the coefficients in (2.2).

In [22], the case of uniparametric families of differential equations ($B_w = B_{n-1}$, $A_w = A_{n-1}$, $B_y = B_n$, $A_y = A_n$, with n a parameter) was considered; however, the theory applies to more general situations. Also, the restriction $B_y = B_w$ was considered; as remarked in this same reference, this is not an essential restriction because the B_y and B_w terms can be eliminated by proper changes of the dependent variable.

An essential property is the fact that the δ and γ coefficients are different from zero for any x in I [22, Lemma 2.1]; this is so because the first order system is simultaneously satisfied by $\{y^{(1)}, w^{(1)}\}$ and $\{y^{(2)}, w^{(2)}\}$ and, therefore, the first equation in (2.2) implies $\delta(x) = W[y^{(1)}, y^{(2)}]/Z(x)$, where W is the Wronskian of two independent solutions of a second order ODE (and therefore never vanishes) and

$$Z(x) = \begin{vmatrix} y^{(1)} & w^{(1)} \\ y^{(2)} & w^{(2)} \end{vmatrix}. \tag{2.3}$$

$Z(x)$ is never zero in I because we are assuming that the coefficients of the first order system are continuous in I.

Another essential ingredient of the method is the fact that $\delta(x)\gamma(x) < 0 \ \forall x \in I$ if there is a solution y of the original ODE or w of the contrast ODE with at least two zeros in I [22, Lemma 2.1].

2.1. First order systems in normal form and oscillating functions. Let us briefly summarize the transformations which must be performed over the system (2.2) in order to build global fixed point iterations to compute the zeros of $y(x)$. First, a change of the dependent functions is considered

$$y(x) = \lambda_y(x)\bar{y}(x), \qquad w(x) = \lambda_w(x)\bar{w}(x), \tag{2.4}$$

with $\lambda_y(x) \neq 0$, $\lambda_w(x) \neq 0 \ \forall x \in I$ in such a way that $\bar{y}$ and $\bar{w}$ satisfy

$$\begin{aligned} \bar{y}' &= \bar{\alpha}\,\bar{y} + \bar{\delta}\,\bar{w}, \\ \bar{w}' &= \bar{\beta}\,\bar{w} + \bar{\gamma}\,\bar{y} \end{aligned} \tag{2.5}$$

with $\bar{\delta} > 0$ and $\bar{\delta} = -\bar{\gamma}$. This is accomplished by choosing

$$\lambda_y = \operatorname{sign}(\delta)\lambda_w\sqrt{-\frac{\delta}{\gamma}}, \tag{2.6}$$

and the argument of the square root is positive if y and/or w have at least two zeros in I. The new functions $\bar{y}$ and $\bar{w}$ obviously have the same zeros as y and w; it is easy to verify that $\bar{y}$ and $\bar{w}$ satisfy second order ODEs with the same B-coefficient accompanying the first derivative term. This is the situation described in more detail in [22].

Considering now a change of variables

$$z(x) = \int \bar{\delta}(x)dx, \tag{2.7}$$

the system reads

$$\begin{aligned} \dot{\bar{y}} &= \bar{a}\,\bar{y} + \bar{w}, \\ \dot{\bar{w}} &= \bar{b}\,\bar{w} - \bar{y}, \end{aligned} \qquad \text{where } \bar{a} = \frac{\bar{\alpha}}{\bar{\delta}} \text{ and } \bar{b} = \frac{\bar{\beta}}{\bar{\delta}}, \tag{2.8}$$

where dots mean derivatives with respect to z.

From (2.8) we see that the ratio $H(z) = \bar{y}/\bar{w}$ satisfies the first order nonlinear ODE

$$\dot{H} = 1 + H^2 - 2\eta H, \qquad \text{where } \eta = \frac{\bar{b} - \bar{a}}{2}, \tag{2.9}$$

which resembles the behavior of the tangent function, at least for small η. The zeros and singularities of H are interlaced, as happens with the tangent function; this is a consequence of the continuity of the coefficients in (2.2).

For later convenience, let us consider an additional change of the dependent variables; namely, we can take $\bar{y}(z) = \nu(z)\widetilde{y}(z)$ and $\bar{w}(z) = \nu(z)\widetilde{w}(z)$, with $\nu(z) = \exp\left(\int \frac{1}{2}(\bar{a} + \bar{b})dz\right)$. Then, we have

$$\begin{aligned} \dot{\widetilde{y}} &= -\eta\,\widetilde{y} + \widetilde{w}, \\ \dot{\widetilde{w}} &= \eta\,\widetilde{w} - \widetilde{y}, \end{aligned} \tag{2.10}$$

and $\widetilde{y}$ and $\widetilde{w}$ satisfy second order ODEs in normal form; for instance,

$$\ddot{\widetilde{y}} + \widetilde{A}(z)\widetilde{y} = 0 \qquad \text{with } \widetilde{A}(z) = 1 + \dot{\eta} - \eta^2, \tag{2.11}$$

and similarly for w with $\widetilde{A}(z) = 1 - \dot{\eta} - \eta^2$.

We will say that the system (2.10) is in normal form. The function η, together with an initial value $H(z_0)$, determines the location of the zeros; η measures the deviation of $H(z)$ from a tangent function. Generally, we will have $|\eta(x)| < 1$ ($\eta(x) = \eta(z(x))$) in the interval I where our problem function oscillates. (This is an important feature when we later analyze the rate of convergence of the method.) This is so because of the following.

THEOREM 2.1. *If $|\eta| > 1$ in an interval I, then $H(z(x))$ has at most one zero and one singularity in this interval*

Proof. If $|\eta| > 1$, then (2.9) implies that $\dot{H} < 0$ for $|H - \eta| < \sqrt{\eta^2 - 1}$. Taking into account that H is differentiable except at its singularities, that the singularities and zeros are interlaced, and that H is increasing at their zeros, then if z_s is a singularity of H in I, necessarily $\lim_{z \to z_s^+} H(z) = -\infty$ and there cannot be any other singularity $z_{s'} > z_s$ (such that $\lim_{z \to z_{s'}^-} H(z) = +\infty$) because $H(z)$ becomes decreasing in a band-shaped region (above the z axis for $\eta > 0$ and below for $\eta < 0$). On the other hand, due to interlacing, if there can be no more than one singularity in I, then there can be no more than one zero in I. □

Therefore, we expect that $|\eta| < 1$ in an interval where the problem function oscillates and has several zeros. A related result is the following sufficient condition for the existence of infinitely many real zeros.

THEOREM 2.2. *If $\int_{z_0}^{+\infty}(1 - \eta(z)^2)dz = +\infty$ with η bounded in $[z_0, +\infty)$, then $H(z)$ has infinitely many interlaced zeros and singularities for $z > z_0$.*

Proof. This is a direct consequence of a well known result: the solutions of an equation in normal form $\ddot{y}(z) + A(z)y(z) = 0$ that satisfies $\int_{z_0}^{+\infty} A(z)dz = +\infty$ have infinitely many zeros for $z > z_0$. Considering (2.11), we see that $H(z) = \widetilde{y}/\widetilde{w}$ has infinitely many zeros because $\widetilde{y}$ satisfies $\ddot{\widetilde{y}} + \widetilde{A}(z)\widetilde{y} = 0$ with $\widetilde{A}(z) = 1 + \dot{\eta}(z) - \eta(z)^2$ and

$$\int_{z_0}^{+\infty} \left[1 + \dot{\eta} - \eta^2\right] dz = \lim_{z \to +\infty} \left[\eta(z) - \eta(z_0) + \int_{z_0}^{z} (1 - \eta(z)^2) dz\right] = +\infty,$$

and because the zeros and singularities are interlaced, there is also an infinite number of singularities. □

COROLLARY 2.3. *If $|\eta(x)| < 1-\epsilon$, $0<\epsilon<1$, $\forall x \geq x_0$, then $H(z(x))$ has infinitely many interlaced zeros and singularities for $x > x_0$.*

2.2. Fixed point methods for first order systems. In [22], it was shown that, given the ratio $H(z)$, with zeros and singularities interlaced and satisfying the first order nonlinear ODE (2.9), the fixed point iteration

$$T(z) = z - \arctan(H(z)) \tag{2.12}$$

is globally convergent to the zeros of the problem function $y(x(z))$ in intervals where η does not change sign. More specifically, denoting the zeros of the problem function $y(x(z))$ by $z_y^{(j)}$ (zeros of $H(z)$) and those of the contrast function by $z_w^{(j)}$ (singularities of $H(z)$), j being integer numbers such that

$$\cdots < z_w^{(j)} < z_y^{(j)} < z_w^{(j+1)} < z_y^{(j+1)} < \cdots, \tag{2.13}$$

then

$$\lim_{n\to\infty} T^{(n)}(z_0) = z_y^{(j)} \qquad \forall z_0 \in (z_w^{(j)}, z_w^{(j+1)}) \tag{2.14}$$

and the convergence to $z_y^{(j)}$ is monotonic if $\eta < 0$ and $z_0 \in (z_w^{(j)}, z_y^{(j)})$ or $\eta > 0$ and $z_0 \in (z_y^{(j)}, z_w^{(j+1)})$.

Global bounds for the distance between the zeros and singularities of $H(z)$ when η does not change sign can be established (see [22, Corollary 4.4]):

$$z_y^{(j)} - z_w^{(j)} > \frac{\pi}{2} \quad \text{and} \quad z_w^{(j+1)} - z_y^{(j)} < \frac{\pi}{2} \qquad \text{for} \;\; \eta < 0, \tag{2.15}$$

and the contrary if $\eta > 0$.

These global bounds stem from the fact that, when $\eta < 0$, the graph of $H(z)$ lies above the graph of $\tan(z - z_y^{(j)})$ around $z_y^{(j)}$; this feature also explains the global convergence of the method (see [22]). Using these global bounds, iterative schemes to compute with certainty all the zeros inside a given interval I were developed; for instance, when $\eta < 0$, we have that $z_w^{(j+1)} < z_y^{(j)} + \pi/2 < z_y^{(j+1)} < z_w^{(j+1)}$ and therefore

$$z_y^{(j+1)} = \lim_{n\to\infty} T^{(n)}(z_y^{(j)} + \Delta z^{(j)}) \qquad \text{with} \;\; \Delta z^{(j)} = \frac{\pi}{2}, \tag{2.16}$$

and the iteration converges monotonically. Hence, the successive zeros inside I can be found by using this forward iterative scheme. Obviously, when $\eta > 0$, a backward scheme is the option:

$$z_y^{(j-1)} = \lim_{n\to\infty} T^{(n)}(z_y^{(j)} + \Delta z^{(j)}) \qquad \text{with} \;\; \Delta z^{(j)} = -\frac{\pi}{2}. \tag{2.17}$$

When η changes sign, forward and backward sweeps can be combined. For the usual situation in which η changes sign only once, at $z = z_\eta$, explicit algorithms were described in [22]. These algorithms, depending on the sign of $S = (z - z_\eta)\eta$, compute zeros by an expansive sweep when $S \leq 0$ (forward sweep for $z > z_\eta$ and backward sweep for $z < z_\eta$) or by a contractive sweep when $S \geq 0$.

2.3. Improvement of the iteration step. The iteration steps $\Delta z^{(j)} = \pm\pi/2$ can be improved under certain monotonicity conditions of the $\widetilde{A}(z)$ coefficient in (2.11). For instance, when $\eta < 0$ in I, a forward sweep (2.16) can be used to compute the zeros of $y(x(z))$ in I, the convergence to each zero being monotonic because $z_y^{(j)} + \pi/2 \in (z_w^{(j+1)}, z_y^{(j+1)})$. (See [22], where the intervals $(z_w^{(j+1)}, z_y^{(j)})$ are called subintervals of monotonic convergence, or SMCs.) In this case ($\eta < 0$), the step $\Delta z^{(j)} = \pi/2$ can be improved when $\widetilde{A}(z)$ is a decreasing function; indeed, this means that $\widetilde{y}(z)$ (which has the same zeros as the problem function $y(x(z))$) oscillates more slowly as z increases, and therefore $z_y^{(j+1)} - z_y^{(j)} > z_y^{(j)} - z_y^{(j-1)}$. Hence $z_y^{(j)} + [z_y^{(j)} - z_y^{(j-1)}] < z_y^{(j+1)}$. On the other hand, because of the global bounds between zeros and singularities, $\widetilde{\Delta} z^{(j)} \equiv z_y^{(j)} - z_y^{(j-1)} > \pi/2$, and therefore

$$z_w^{(j+1)} < z_y^{(j)} + \frac{\pi}{2} < z_y^{(j)} + \widetilde{\Delta} z^{(j)} < z_y^{(j+1)}.$$

Hence, the starting value $z_y^{(j)} + \widetilde{\Delta} z^{(j)}$ guarantees (monotonic) convergence to $z_y^{(j+1)}$, and it is a better iteration step than $\Delta z^{(j)}$ because $z_y^{(j)} + \widetilde{\Delta} z^{(j)}$ lies closer to $z_y^{(j+1)}$ than $z_y^{(j)} + \Delta z^{(j)}$ (the convergence being monotonic in both cases).

In other words, when $\widetilde{A}(z)$ is decreasing and η is negative, the difference between the two previously evaluated zeros ($z_y^{(j)}$ and $z_y^{(j-1)}$) can be used as the iteration step to generate a starting value to compute the next zero ($z_y^{(j+1)}$). Obviously, a similar scheme can be devised for the case $\eta > 0$ when $\widetilde{A}(z)$ is increasing. These arguments lead to the following result.

THEOREM 2.4. *If* $\eta(z)\dot{\widetilde{A}}(z) > 0$ *in* $(z_y^{(j-1)}, z_y^{(j+1)})$*, then*

$$z_y^{(j\pm1)} = \lim_{n\to\infty} T^{(n)}(z_y^{(j)} + \Delta z^{(j)}) \qquad \textit{with } \Delta z^{(j)} = z_y^{(j)} - z_y^{(j\mp1)},$$

where the upper sign is for the case $\eta < 0$*, and the lower sign for* $\eta > 0$*. The convergence is monotonic.*

These iteration steps, as mentioned, improve the steps $\Delta z^{(j)} = \pm\pi/2$. Notice, however, that the improvement cannot be considered until two zeros have been evaluated. (The step $\Delta z^{(j)} = \pi/2$ must be considered in order to compute the second zero.)

As we will discuss next, the condition $\eta\dot{\widetilde{A}} > 0$ can be achieved with great generality for the case of families of second order ODEs depending continuously on one parameter (typical case of special functions).

2.3.1. Difference-differential equations and three term recurrence relations. Let us consider now the case, described in [22], of second order ODEs depending continuously on one parameter k,

$$y_k''(x) + B_k(x)y_k'(x) + A_k(x)y_k(x) = 0, \tag{2.18}$$

with general difference-differential equations (DDEs)

$$\begin{aligned} y_k'(x) &= a_k(x)y_k(x) + d_k(x)y_{k-1}(x), \\ y_{k-1}'(x) &= b_k(x)y_{k-1}(x) + e_k(x)y_k(x) \end{aligned} \tag{2.19}$$

satisfied by two families of independent solutions $\{y_k^{(1)}\}$ and $\{y_k^{(2)}\}$; we take integer differences of k. These general DDEs will allow us, if permitted by the range of k, to consider two different contrast functions (y_{n-1} or y_{n+1}) in order to compute the zeros of a problem function y_n; in other words, both equations in (2.19), taking $k = n$ and $k = n+1$, can be used to build the fixed point iterations to compute the zeros of y_n.

Explicitly and following the steps described in section 2.1, the following two fixed point iterations can be built:

$$T_i(z) = z - \arctan(H_i(z)), \qquad i = \pm 1, \tag{2.20}$$

where

$$H_i(z) = -i\,\mathrm{sign}(d_{n_i}) K_i \frac{y_n(x(z))}{y_{n+i}(x(z))}, \qquad n_i = \begin{cases} n, & i = -1, \\ n+1, & i = +1, \end{cases} \tag{2.21}$$

and

$$z(x) = \int \sqrt{-d_{n_i} e_{n_i}}\, dx, \qquad K_i = \left(-\frac{d_{n_i}}{e_{n_i}}\right)^{i/2}, \tag{2.22}$$

and $x(z)$ is the inverse function of $z(x)$.

The functions $H_i(z)$ satisfy $\dot{H}_i = 1 + H_i^2 - 2\eta H_i$, where the function η is given by $\eta_i(z) = \eta_i(x(z))$ and

$$\eta_i(x) = i \frac{1}{2\sqrt{-d_{n_i} e_{n_i}}} \left(a_{n_i} - b_{n_i} + \frac{1}{2}\left(\frac{e'_{n_i}}{e_{n_i}} - \frac{d'_{n_i}}{d_{n_i}} \right) \right). \tag{2.23}$$

It is interesting to note that η_{+1} is equal to η_{-1} with reversed sign and with the substitution $n \to n+1$. Therefore, quite generally, η_{-1} and η_{+1} have opposite signs. This means, as described above, that two different types of sweep are possible for computing the zeros of y_n: forward and backward when the η functions do not change sign, and expansive and contractive when the η functions change sign once. This is an important feature for improving the iteration step $\Delta z^{(j)} = \pm\pi/2$ by $\widetilde{\Delta} z^{(j)} = z_y^{(j)} - z_y^{(j\mp 1)}$ as described in the previous section, given that the sign of η can be chosen in such a way that the hypothesis of Theorem 2.4 is met.

On the other hand, (2.19), satisfied by two independent sets of solutions ($\{y_k^{(1)}\}$ and $\{y_k^{(2)}\}$), implies the existence of a three term recurrence relation between the functions $Y_k = \alpha y_k^{(1)} + \beta y_k^{(2)}$ for any α, β, namely,

$$Y_{n+1} = r_n\, Y_n + s_n\, Y_{n-1} \qquad \text{with} \quad r_n = \frac{a_n - b_{n+1}}{e_{n+1}}, \quad s_n = \frac{d_n}{e_{n+1}}, \tag{2.24}$$

with $s_n \neq 0\, \forall x \in I$.

Given that, by hypothesis, the coefficients of the DDEs are continuous in I, then $\{y_k^{(1)}\}$ and $\{y_k^{(2)}\}$ are necessarily independent solutions of the three term recurrence relation because the determinant $Z(x)$ of (2.3) (with $y \equiv y_n^{(k)}$, $k = 1, 2$, and $w = y_{n-1}^{(k)}$ or $w = y_{n+1}^{(k)}$) can never vanish in I. This is an important fact from a computational point of view because recurrence relations are useful tools for computing the ratios $y_n/y_{n\pm 1}$, which are at the heart of the fixed point method. In section 4, we will discuss these recursive computations.

2.4. Order of convergence of the method. The fixed point iteration $T(z) = z - \arctan(H(z))$ with $H(z)$ satisfying (2.9) is, as shown in [22], a globally convergent iteration. Next, we will show that the order of convergence is 2, and the speed of convergence increases as $|\eta|$ is smaller. As discussed before, generally $|\eta| < 1$ in the region where the problem function y oscillates.

Defining $\epsilon_n = z^{(n)} - z_y^{(j)}$, where $z^{(n)} = T^{(n)}(z_0)$, $z_0 \in (z_w^{(j)}, z_w^{(j+1)})$, we have, expanding in Taylor series, that

$$\epsilon_{n+1} = T(z_y^{(j)} + \epsilon_n) - z_y^{(j)} = \eta(z_y^{(j)})\epsilon_n^2 + \mathcal{O}(\epsilon_n^3)$$

because $\dot{T}(z) = 1 - \dot{H}(z)/(1 + H(z)^2) = 2\eta H(z)/(1 + H(z)^2)$ and then $\dot{T}(z_y^{(j)}) = 0$, $\ddot{T}(z_y^{(j)}) = 2\eta(z_y^{(j)})$.

Therefore the fixed point iteration is quadratically convergent. The convergence will be faster as η is smaller.

The error after the nth iteration can be bounded by standard methods when $|\dot{T}(z)| \leq M < 1$. It is well known that in this case $|\epsilon_n| \leq |z^{(1)} - z^{(0)}| M^n/(1-M)$, and given that $|\dot{T}(z)| = \left|2\eta H/(1 + H^2)\right| \leq |\eta(z)|$, the absolute error in the computation of $z_y^{(j)}$ after the nth iteration can be bounded by

$$|\epsilon_n| \leq \frac{M_\eta^n}{1 - M_\eta} |\arctan(H(z^{(0)}))| \leq \frac{\pi}{2} \frac{M_\eta^n}{1 - M_\eta}, \tag{2.25}$$

where $|\eta| \leq M_\eta < 1$ in $(z_w^{(j)}, z_w^{(j+1)})$.

As we know from Theorem 2.1 and Corollary 2.3, we can expect that $|\eta| < 1$ in the oscillating region I, while $|\eta|$ is expected to become larger than 1 outside this region. This is, for instance, the case of Bessel functions: as $x \to 0$ the oscillations end and $|\eta|$ becomes larger than one. According to (2.25) we can expect that the convergence will improve as larger zeros are considered since $|\eta|$ decreases with increasing x. (For Bessel functions the associated change of variables is trivial $z(x) = x$.) Later, we will illustrate this behavior with numerical examples.

2.5. Examples of application: Hypergeometric functions and Bessel functions. The method described will be particularly indicated for those systems for which the associated change of variables (2.7) is a simple transformation which can be analytically inverted; otherwise, we would need to develop an additional root finding scheme to invert the function $z(x)$ in order to obtain the zeros of the problem function $y(x)$.

In the particular case of the hypergeometric and confluent hypergeometric equations of real parameters and real variables, it is easy to build global fixed point iterations with simple changes of variables. On the other hand, the method described can also be applied to compute the TPs of these functions, or, more generally, the zeros of any derivative, given that the derivatives of (confluent) hypergeometric functions are (confluent) hypergeometric functions. This is an important category of functions, which, as a subset, includes classical orthogonal polynomials (Hermite, Laguerre, Jacobi).

We will illustrate the method for (confluent) hypergeometric functions. For imaginary arguments (Bessel functions), we will see how difficulties arise in the computation of the TPs, which suggest the development of an alternative method, to be developed in section 3. This alternative method will also be the method of choice for computing the turning points of other functions related to hypergeometric functions of complex parameters or variables (Coulomb and conical functions).

2.5.1. Hypergeometric functions. Let us consider the confluent hypergeometric equation

$$xy'' + (\beta - n - x)y' + ny = 0, \tag{2.26}$$

with n and β real numbers. One of the solutions of this ODE is Kummer's function ${}_1\mathrm{F}_1(-n;\beta-n;x)$. (We modify the standard notation for the parameters ${}_1\mathrm{F}_1(a;b;x)$.) We denote by y_n our problem function, which is a solution of (2.26). Taking the derivative of (2.26), we observe that y_n' satisfies the same equation with n replaced by $n-1$; we define $y_{n-1} \equiv y_n'$ and, considering (2.26), we can write the first order system relating these functions:

$$\begin{aligned} y_n' &= y_{n-1}, \\ y_{n-1}' &= \left(1 - \frac{\beta - n}{x}\right) y_{n-1} - \frac{n}{x} y_n. \end{aligned} \tag{2.27}$$

For instance, in the case of Kummer functions, $y_{n-1} = y_n' = \frac{n}{n-\beta}\, {}_1\mathrm{F}_1(-n+1;\beta-n+1;x)$; we can also take as a contrast function a confluent hypergeometric function y_{n+1} such that $y_{n+1}' = y_n$; in terms of the Kummer function we could choose $y_{n+1} = \frac{n+1-\beta}{n+1}\, {}_1\mathrm{F}_1(-n-1,\beta-n-1,x)$. In fact, it is not difficult to see that the selection of y_{n+1} as our contrast function is more appropriate since improved iteration steps (Theorem 2.4) can be implemented for this selection (contrary to the case when y_{n-1} is the contrast function).

Similarly one can build fixed point iterations for hypergeometric functions starting with the differential equation

$$x(1-x)y'' + [c - (a+b+1)x]y' - aby = 0; \tag{2.28}$$

as before, the derivative of a solution y is also a solution of the hypergeometric equation with the parameters a, b, c incremented by one. Taking y' as the contrast function for y, we have that the associated change of variables is $z(x) = \int \sqrt{ab}/\sqrt{x(1-x)}dx$. Considering, for instance, the case $I = (0,1)$, we see that, if y has at least two zeros in I, then $ab > 0$ and the change of variables is $z(x) = \sqrt{ab}\arcsin(2x-1)$. The change of variables is simple and analytically invertible, and we can proceed similarly to confluent hypergeometric functions.

As for confluent hypergeometric functions, the first order system built from the differential equation is equivalent to the first order DDEs for uniparametric families of functions with parameters $a = -n$, $b = \beta - n$, $c = \gamma - n$; this means that to compute the zeros of a solution $y(a,b;c;x)$ of (2.28), we choose as contrast functions $y(a\pm1, b\pm1; c\pm1; x)$. This is not the only possibility since, for instance, we could also choose $y(a\pm1, b; c; x)$ (and for the confluent case, $y(a\pm1; b; x)$); this was the choice considered in [22] for the case of Legendre, Hermite, and Laguerre polynomials. The selection of the most appropriate contrast functions depending on the ranges of the parameters and the dependent variables lies outside the scope of the present work [7].

As mentioned before, for (confluent) hypergeometric functions, the same methods apply for the computation of the TPs, and because the derivatives of (confluent) hypergeometric functions are (confluent) hypergeometric functions, also satisfy the (confluent) hypergeometric equation. These methods therefore enable us to compute the zeros of classical orthogonal polynomials (Hermite, Laguerre, Legendre, Jacobi, Gegenbauer) and their derivatives, or the zeros and TPs of any other solution of the same differential equation.

2.5.2. Bessel functions. Bessel functions are Kummer functions of imaginary variable [25]; they are a simple example of an important type of function for which the method described above can be used to compute the zeros [21]; however, it is not so simple to employ the same methods for the computation of their turning points. Let us consider Ricatti–Bessel functions y_n (Bessel functions multiplied by $\sqrt{x}$), which satisfy second order ODEs in normal form,

$$y_n'' + A_n(x)y_n = 0, \qquad \text{where } A_n(x) = 1 - \frac{n^2 - 1/4}{x^2}, \tag{2.29}$$

and first order DDEs in normal form

$$\begin{aligned} y_n' &= -\eta y_n + y_{n-1}, \\ y_{n-1}' &= \eta y_{n-1} - y_n, \end{aligned} \qquad \text{with } \eta = \frac{n - 1/2}{x}, \tag{2.30}$$

which can be used to build the corresponding fixed point methods. In this case, the transformations considered in section 2.1 are not required because both the ODE and the DDEs are in normal form; therefore, the change of variables is trivial $z(x) = x$. Since $\eta(x)A_n'(x) > 0$ for $n > 0$, these DDEs are suitable for applying improved iteration steps (Theorem 2.4) for positive orders. (Positive and negative orders are related by reflection formulas; (see [19, equation (24)]).)

Let us notice that we could have considered the first order system associated with the second order ODE (as we did for hypergeometric functions of real variables); however, the associated change of variables

$$z(x) = \int \sqrt{A_n(x)}dx \tag{2.31}$$

is not simple enough to allow for an analytical inversion. For this reason, the first order DDEs (2.30) are preferred. A second essential difference with respect to hypergeometric functions of real variables and parameters is that the computation of TPs is not so simple as the computation of the zeros through (2.30); the derivatives of Bessel functions (or Ricatti–Bessel functions) are not Bessel functions, and they do not satisfy the same type of DDEs. We could consider the first order system associated to the second order ODE to compute the zeros of y_n' using as a contrast function the same y_n; however, as commented, we are faced with the numerical inversion of the change of variables (2.31). Also, the possibility of considering y_n' as the problem function with contrast function y_{n-1}' (as we can do for hypergeometric functions) is also possible, but the associated change of variables is even more involved than (2.31) and is related to elliptic integrals.

In the next section, we propose an alternative method, which is based on the direct use of second order ODEs in normal form.

3. Method II: Fixed point methods based on second order ODEs in normal form.

Much as did the method based on first order systems, the method that we are next describing converges with certainty. Also, as we will show, the method is quadratically convergent. It can be used to compute zeros and TPs of any solution of a given second order ODE in normal form,

$$y''(x) + A(x)y(x) = 0, \tag{3.1}$$

where we assume that $A(x)$ is continuous and that $y(x)$ has continuous second derivative (and is not a trivial solution).

Differently from Method I, Method II does not provide iteration steps in order to find with certainty all the zeros (or TPs) of a solution of the second order ODE in a given interval I. However, we can use this method to compute bracketed zeros or TPs.

In particular, we will describe the computation of TPs of solutions of second order ODEs in normal form, which are bracketed by the zeros of these solutions. This new method will solve a considerable number of cases for which Method I, based on first order systems, failed to provide a simple solution.

This method can also be efficiently applied to second order ODEs, which can be transformed to normal form by a simple (analytically invertible) change of variables. It is important to realize that for calculating the TPs of solutions of a second order ODE

$$y'' + \tilde{B}(x)y' + \tilde{A}(x)y = 0 \tag{3.2}$$

we cannot consider changes of the dependent variable in order to write down an equation in normal form, because the extrema of $\nu(x)y(x)$ are not the extrema of $y(x)$. However, considering a change of variable $z = z(x)$, we arrive at

$$(z')^2\ddot{y} + [z'' + \tilde{B}z']\dot{y} + \tilde{A}y = 0, \tag{3.3}$$

where dots mean derivatives with respect to z, and the primes are derivatives with respect to x. And, choosing

$$z(x) = \int \exp\left(-\int \tilde{B}dx\right)dx, \tag{3.4}$$

we have

$$\ddot{y}(z) + A(z)y(z) = 0, \quad A(z) = A(x(z))\dot{x}^2. \tag{3.5}$$

The TPs of $y(x)$ will be $z^{-1}(z_{y'}^{(j)})$, with $z_{y'}^{(j)}$ being the turning points of $y(x(z))$. Thus, the problem of finding the TPs of any second order ODE with continuous coefficients $\widetilde{B}(x)$, $\widetilde{A}(x)$ can be reduced to the problem of finding the turning points of a second order ODE in normal form. We will require that this change of the independent variable be a simple change (i.e., analytically invertible in terms of elementary functions).

Of course, similarly to Method I, Method II can be applied in general, regardless of whether the change of variable is elementary. However, the cases when the change is simple are the most indicated for the method. Combining Methods I and II, the spectrum of functions whose zeros and TPs can be easily computed (with simple changes of variable) becomes very wide. As we know, the zeros and TPs of (confluent) hypergeometric functions of real parameters and variables can be obtained with Method I. Combining Methods I and II, we can add to the list of satisfactorily solved problems, among others, the computation of the zeros and TPs of Airy, Bessel ($\mathcal{C}_\nu = \cos\alpha J_\nu(x) - \sin\alpha Y_\nu(x)$), and Coulomb wave functions; Whittaker functions of real parameters; associated Legendre functions (ALFs) of real parameters and conical functions (ALFs of complex degrees); and the zeros of Bessel-related functions like $x^\alpha C_\nu(x)$. Both Methods I and II deal with any solution of the corresponding second order ODE (and not only the regular solution or the polynomial solutions, as is the case of matrix methods [1]).

3.1. Construction of the method for ODEs in normal form. In what follows, we consider that the zeros of a function $y(x)$ inside an interval $[x_1, x_2]$ have already been evaluated by using Method I (or any other method) and that this function satisfies a second order ODE in normal form (3.1). We further assume that $A(x)$ has no more than one zero in the interval of interest $[x_1, x_2]$. If this were not so, we would subdivide the search for TPs for different subintervals where this property holds.

If there is a value $x_a \in [x_1, x_2]$ for which $A(x_a) = 0$, we will take $A(x) < 0$ in $[x_1, x_a)$ and $A(x) > 0 \in (x_a, x_2]$, making the change of independent variable $x \to -x$, if necessary.

It is straightforward to prove that y can have no more than one zero in $[x_1, x_a)$; this is so because $y''y \geq 0$ in $[x_1, x_a]$. It is also a simple matter to check that the zeros of y and y' are interlaced in $(x_a, x_2]$; see the following.

LEMMA 3.1. *The zeros of y and y' are interlaced in $(x_a, x_2]$.*

Proof. y and $v \equiv y'$ can not vanish simultaneously, given that we assume that y is not a trivial solution. Interlacing is an immediate consequence of the fact that the Wronskian of y and v is always different from zero. Indeed, $W[y, v] = y'v - yv' = (y')^2 - y''y = (y')^2 + A(x)(y)^2 > 0$. □

The first step towards an algorithm to evaluate the TPs of y is computing each TP between two consecutive zeros of y in $(x_a, x_2]$. For this, given that the TPs are bracketed by the zeros of y (because they are interlaced) except for one possible exception (see the discussion after Lemma 3.12), one can, for instance, use bisection. Instead, we introduce fixed point iterations which are quadratically convergent.

Given the interlacing of the zeros of y and y', it is expected that the logarithmic derivative of y will behave roughly like the tangent function, which can be used to set globally convergent fixed point iterations. Namely, defining

$$h(x) = -\frac{y'(x)}{y(x)}, \tag{3.6}$$

taking the derivative, and using $y''(x) + A(x)y(x) = 0$, we get

$$h' = A + h^2; \tag{3.7}$$

thus h is monotonically increasing when $A(x) > 0$ except at the zeros of y. This equation is the starting point of the method. In order to implement the method we require only that an algorithm to evaluate the logarithmic derivative be available. Assuming that the function $y \equiv y_n$ satisfies relations of the form (2.19), we need only to evaluate ratios y_n/y_{n-1} (or y_n/y_{n+1}) for computing both the zeros (Method I) and TPs (Method II).

Let us first discuss the calculation of TPs when $A(x) > 0$, that is, for $x > x_a$. The next theorem provides globally convergent fixed point iterations on subintervals $(x^{(j)}, x^{(j+1)})$, with $x^{(k)}$ being the zeros of $y(x)$.

THEOREM 3.2. *Let D be the logarithmic derivative of y. Let $x^{(j)}$, $x^{(j+1)}$ be two consecutive zeros of y, and $x' \in (\bar{x}_1, \bar{x}_2) \subseteq (x^{(j)}, x^{(j+1)})$ a turning point of y. If $0 < A(x) < K$ $\forall x \in (\bar{x}_1, \bar{x}_2)$, then*

$$\lim_{p \to \infty} T^{(p)}(x) = x' \qquad \forall x \in (\bar{x}_1, \bar{x}_2),$$

where

$$T(x) = x + \frac{1}{\sqrt{K}} \arctan\left(\frac{D}{\sqrt{K}}\right) \tag{3.8}$$

and the convergence is monotonic.

Proof. Let $h = -D = -\frac{y'}{y}$; then $h' < K + h^2$, that is, $\frac{h'}{K+h^2} - 1 < 0$, and then

$$\operatorname{sign}(x - x') \int_{x'}^{x} \left[\frac{h'}{K + h^2} - 1 \right] < 0.$$

Performing the integral and taking into account that $h(x') = 0$, we get

$$\operatorname{sign}(x - x') \left[\frac{1}{\sqrt{K}} \arctan\left(\frac{h}{\sqrt{K}} \right) - (x - x') \right] < 0. \tag{3.9}$$

Then, if $x > x'$ $(x \in (\bar{x}_1, \bar{x}_2))$,

$$x' < x - \frac{1}{\sqrt{K}} \arctan\left(\frac{h}{\sqrt{K}} \right) \equiv T(x)\,,$$

and given that $h' > 0$ in $(x_a, x_2]$, then $h > 0$ in $(x', \bar{x}_2)$ (and $h < 0$ in $(\bar{x}_1, x')$). Therefore

$$x' < T(x) = x + \frac{1}{\sqrt{K}} \arctan\left(\frac{D}{\sqrt{K}} \right) < x. \tag{3.10}$$

Similarly, if $x < x'$,

$$x' > T(x) > x, \tag{3.11}$$

and given that x' is the only fixed point of $T(x)$ in $(\bar{x}_1, \bar{x}_2)$ and considering that, according to (3.10) and (3.11), the successive iterations of T form monotonic sequences, we have that

$$\lim_{p\to\infty} T^{(p)}(x) = x' \qquad \forall x \in (\bar{x}_1, \bar{x}_2)$$

and convergence is monotonic. □

Corollary 3.3. *If x is a zero adjacent to a turning point x' of y and $0 < A < K$ between x and x', then $|x - x'| > \pi/(2\sqrt{K})$.*

Proof. Take x in (3.9) to be a zero of y adjacent to x'. □

An immediate consequence of this corollary is the following.

Corollary 3.4. *If $x^{(j)}$, $x^{(j+1)}$ are two consecutive zeros of y and $0 < A < K$ between $x^{(j)}$ and $x^{(j+1)}$, then $|x^{(j+1)} - x^{(j)}| > \pi/\sqrt{K}$.*

This is nothing but the classical result from Sturm's comparison theorem [25].

Similarly, it is easy to prove the following result.

Proposition 3.5. *If x is a zero adjacent to a turning point x' of y and $0 < k < A$ between x and x', then $|x - x'| < \pi/(2\sqrt{k})$.*

Corollary 3.6. *Similarly, if $x^{(j)}$, $x^{(j+1)}$ are two consecutive zeros of y and $0 < k < A$ between $x^{(j)}$ and $x^{(j+1)}$, then $|x^{(j+1)} - x^{(j)}| < \pi/\sqrt{k}$.*

Corollaries 3.4 and 3.6 also apply to consecutive turning points of y.

The bounds from Corollary 3.3 and Proposition 3.5 can be sufficient when A varies slowly to obtain accurate values for the turning points, as in the following.

Proposition 3.7. *Let x' be a turning point between two consecutive zeros of y such that $x' \in (x^{(j)}, x^{(j+1)})$, and let $0 < k < A(x) < K$ in $(x^{(j)}, x^{(j+1)})$. Then $x' \in J \equiv (x_m - \lambda, x_m + \lambda)$ with $x_m = (x^{(j)} + x^{(j+1)})/2$ and $\lambda = \left[k^{-1/2} - K^{-1/2}\right]\pi/4$. Also, $J \subset (x^{(j)}, x^{(j+1)})$ if $K/k < 9$.*

Then, if the upper and lower bounds of A are similar, x_m will be a good approximation to x' with relative accuracy better than λ/x_m. Even if x_m is not a good enough approximation, we can use it as a starting value to evaluate x' by means of the fixed point iteration previously discussed.

When the function $A(x)$ is monotonic in some interval, the speed of convergence of the fixed point iteration can be improved by adjusting the upper bound K in each iteration. For instance, if $A(x)$ is positive and increasing inside an interval $\bar{I} = (\bar{x}_1, \bar{x}_2)$ which contains a TP of $y(x)$, x', and $y(x) \neq 0 \ \forall x \in \bar{I}$, then $x' < T(x) < x$ for any $x \in \bar{I}$ such that $h(x) > 0$. $T(x)$ is the fixed point iteration of (3.8) with K the upper bound of $A(x)$ in $\bar{I}$. But, because $x' < T(x) < x$, the same holds if K is replaced by the upper bound of $A(x)$ in $(x', x]$, which is $A(x)$. Therefore we have the following.

THEOREM 3.8. *Let D be the logarithmic derivative of y. Let $\bar{I} = (\bar{x}_1, \bar{x}_2)$ be an interval such that $A(x) > 0$ in $\bar{I}$ and $y(x) \neq 0 \, \forall x \in \bar{I}$; let $x' \in \bar{I}$ be a turning point of $y(x)$. If A is monotonic in $\bar{I}$ and $x_0 \in \bar{I}$ is such that $A'(x_0)h(x_0) \geq 0$, then*

$$\lim_{p \to \infty} T^{(p)}(x_0) = x',$$

where

$$T(x) = x + \frac{1}{\sqrt{A(x)}} \arctan\left(\frac{D(x)}{\sqrt{A(x)}}\right) \tag{3.12}$$

and the convergence is monotonic.

Later we will show that the fixed point method of the previous theorem is quadratically convergent.

If $A(x)$ is monotonic, we can find a guideline for choosing a starting value for which Theorem 3.8 holds. Although $x_m = (x^{(j)} + x^{(j+1)})/2$ could be a good estimation, it is not the best choice, given that for this initial value one can not apply Theorem 3.8, as the next result shows.

PROPOSITION 3.9. *Let $A(x)$ be positive ($x > x_a$) and x' the turning point of y in $(x^{(j)}, x^{(j+1)})$.*

1. *If $A(x)$ is increasing in $(x^{(j)}, x^{(j+1)})$, then $x_m \equiv (x^{(j)} + x^{(j+1)})/2 < x'$ and $h_m(x_m) < 0$.*
2. *If $A(x)$ is decreasing in $(x^{(j)}, x^{(j+1)})$, then $x_m \equiv (x^{(j)} + x^{(j+1)})/2 > x'$ and $h(x_m) > 0$.*

Proof.

1. $A(x') \equiv k$ is the upper bound of A in $(x^{(j)}, x']$, and then $x' - x^{(j)} > \pi/(2\sqrt{k})$ (Corollary 3.3). On the other hand, k is the lower bound of A in $[x', x^{(j+1)})$, and then $x^{(j+1)} - x' < \pi/(2\sqrt{k})$ (Proposition 3.5). Then $x_m < x'$ and $h(x_m) < 0$ because h is monotonically increasing for $x > x_a$.
2. This is shown similarly as in 1. □

However, Proposition 3.7 gives a recipe for choosing a starting value for which Theorem 3.8 will hold, whenever A does not show very strong variations and $K/k < 9$.

(a) If A is increasing in $(x^{(j)}, x^{(j+1)})$, take as a starting value $x_m + \lambda$.

(b) If A is decreasing in $(x^{(j)}, x^{(j+1)})$, take as a starting value $x_m - \lambda$.

If $K/k > 9$, we could, for instance, choose x_m as a starting value and apply Theorem 3.2. (We have never encountered such a situation in the examples we will later discuss.) An alternative, which is generally more efficient, is locating a value of x for which Theorem 3.8 applies; this can be done by applying one or two bisection steps. This possibility is generally better because, as mentioned, the fixed point iteration of Theorem 3.8 is quadratically convergent.

If $A(x)$ has a turning point in $(x^{(j)}, x^{(j+1)})$, the previous two rules are substituted with the following:

(c) If $A(x)$ has a maximum in $x_{Max} \in (x^{(j)}, x^{(j+1)})$, take x_{Max} as a starting value;

(d) If $A(x)$ has a minimum in $x_{Min} \in (x^{(j)}, x^{(j+1)})$, take $x_{Min} - \operatorname{sign}(h(x_{Min}))\lambda$ (provided $K/k < 9$).

Of course, in principle several extrema of $A(x)$ could be located between two consecutive zeros of y. However, rules (a), (c), and (d) will suffice in most circumstances. With this choice of starting values the improved iteration in Theorem 3.8 converges monotonically.

We have all the ingredients for computing the TPs between two consecutive zeros which are greater than x_a. The only thing left is the TPs in $[x_1, x^{(a+)})$, with $x^{(a+)}$ being the smallest zero larger than x_a (in the case that $x_1 < x_a$). First, it is important to note the following results.

Lemma 3.10. *If x_a is a TP, x_a is a double root of $y' = 0$ and there are no other TPs in $[x_1, x^{(a+)})$.*

Lemma 3.11. *If there are no TPs in $(x_a, x^{(a+)})$, then there are no TPs in $[x_1, x_a)$.*

Lemma 3.12. *If there is a zero of y in $[x_1, x_a]$, then there are no TPs in $[x_1, x_a]$.*

Lemmas 3.11 and 3.12 are consequences of the different convexity properties of y for $x > x_a$ and $x < x_a$ and the continuity of y''.

One should always test the existence of a TP in $[x_a, x^{(a+)})$ and compute it when it exists. If $h(x_a) = 0$, there is a turning point at x_a and no more TPs in $[x_1, x_a)$ (Lemma 3.10). If this is not so and there are no TPs in $(x_a, x^{(a+)})$, then there are no TPs in $[x_1, x_a]$ (Lemma 3.11). On the contrary, we must compute the TP in $(x_a, x^{(a+)})$ and later check the existence of a TP in (x_1, x_a) in case there is no zero of y in this subinterval (Lemma 3.12).

To compute the possible TP in $(x_a, x^{(a+)})$, bisection can be used; the convergence is accelerated by using the iteration of Theorem 3.8 when $A(x)$ is monotonic or has at most one extremum in $(x_a, x^{(a+)})$. If this is not so, Theorem 3.2 can be applied.

For finding the possible TP in (x_1, x_a) one cannot use the fixed point iterations discussed so far. Bisection is a safe choice but not the fastest method. We can use a fixed point iteration, which converges for values close enough to the TP, in order to "polish" the roots as follows.

Theorem 3.13. *Let x' be a TP in $[x_1, x_a]$ and let $-C < A(x) < 0$ in an interval I such that $x' \in I \subseteq [x_1, x_a)$, with C a positive constant; then*

$$\lim_{p\to\infty} \bar{T}^{(p)}(x) = x' \qquad \forall x \in I \textit{ such that } h^2 < C \textit{ between } x \textit{ and } x',$$

where

$$\bar{T}(x) = x - \frac{1}{\sqrt{C}} \tanh^{-1}\left(\frac{D}{\sqrt{C}}\right). \tag{3.13}$$

Proof. $h' = A + h^2 > -C + h^2$, but $h^2 < C$ and therefore $\frac{h'}{-C+h^2} - 1 < 0$. Integrating, we get

$$\operatorname{sign}(x - x') \int_{x'}^{x} \left(\frac{h'}{-C + h^2} - 1\right) dx < 0$$

and

$$\operatorname{sign}(x - x')[R(x) - (x - x')] < 0, \tag{3.14}$$

where

$$R(x) = \frac{1}{2\sqrt{C}} \log \frac{\sqrt{C} - h}{\sqrt{C} + h} = \frac{1}{\sqrt{C}} \tanh^{-1}\left(\frac{D}{\sqrt{C}}\right),$$

and similarly as in Theorem 3.2, monotonic convergence holds from (3.14). □

In the common case in which A is increasing in $[x_1, x_a]$, the following alternative to Theorem 3.13 can be considered, with the advantage that the iteration is quadratically convergent.

THEOREM 3.14. *Let x' be a TP in $[x_1, x_a)$ and let $A(x)$ be increasing in $[x_1, x_a)$. Let $x_0 < x'$ such that $0 < h(x_0) < \sqrt{-A(x_0)}$; then*

$$\lim_{p\to\infty} \bar{T}^{(p)}(x_0) = x',$$

where

$$\bar{T}(x) = x - \frac{1}{\sqrt{-A(x)}} \tanh^{-1}\left(\frac{D}{\sqrt{-A(x)}}\right), \tag{3.15}$$

and the convergence is monotonic.

If $A(x)$ may become negative on two subintervals $I_1 = [x_1, x_a)$ and $I_2 = (x_b, x_2]$ contained in the interval of interest $[x_1, x_2]$, the following Corollary can be applied to compute the possible TPs in I_1 (case of Theorem 3.14) and I_2.

COROLLARY 3.15. *Let x' be a TP inside an interval B where $A(x) < 0$, $A(x)$ being monotonic in B. Let $x_0 \in B$ such that $0 < h(x_0)\mathrm{sign}(A(x_0)) < \sqrt{-A(x_0)}$; then $\lim_{p\to\infty} \bar{T}^{(p)}(x_0) = x'$ with $T(x)$ as in Theorem 3.14. The convergence is monotonic.*

3.2. Order of convergence of the method. A good reason to consider the improved iterations from Theorems 3.8 and Corollary 3.15 instead of the iterations from Theorems 3.2 and 3.13 is that the first two iterations are quadratically convergent. Given that we can easily choose starting values such that the hypothesis of Theorems 3.8 and Corollary 3.15 are met, our algorithms (see the next section) will be based on these fixed point iterations.

It is straightforward to check that the fixed point iterations (3.12) and (3.15) are quadratically convergent. For instance, taking the derivative of the iteration (3.12) and considering the second order ODE (2.1), we have

$$T'(x) = -\frac{A'}{2A}\left[\frac{y'/y}{A + (y'/y)^2} + \frac{1}{\sqrt{A}} \arctan\left(\frac{y'/y}{\sqrt{A}}\right)\right]. \tag{3.16}$$

Therefore, if x' is a turning point of y for $x > x_a$ (a fixed point of $T(x)$), we have that $T'(x') = 0$ and then the convergence is quadratical.

Besides, we see that for $x > x_a$, $|T'(x)| \leq \frac{|A'(x)|}{4A^{3/2}(x)}\pi = |\frac{\pi}{2}\frac{d}{dx}A^{-1/2}(x)|$, which, similarly as for Method I, can be used to bound the absolute error after n iterations. The convergence is faster as $A^{-1/2}(x)$ varies more slowly.

A similar analysis can be made with the iteration (3.15).

3.3. Algorithm for the computation of TPs. Let us now describe the algorithm (using a FORTRAN-like style) for the computation of the TPs of a function $y(x)$ in an interval $I = [x_1, x_2]$. As noted in the beginning of section 3.1, we need to consider only the evaluation of TPs in an interval $I = [x_1, x_2]$, where $A(x)$ changes sign at most once, at x_a, and $A(x) < 0$ if $x < x_a$, $A(x) > 0$ if $x > x_a$. However, we have also considered the possibility that $A(x)$ may have two zeros $x_a^{(-1)}$ and $x_a^{(+1)}$ ($A(x)$ being positive in $(x_a^{(-1)}, x_a^{(+1)})$), which is an easy-to-carry generalization; we consider only the possibility of $A(x)$ having one maximum in I or $A(x)$ being monotonic in I; this is enough for a vast number of special functions.

The algorithm consists of the implementation of Theorem 3.8 for the computation of TPs for $x \in (x_a^{(-1)}, x_a^{(+1)})$ and Theorem 3.14 for $x < x_a^{(-1)}$ (which should be replaced by Theorem 3.13 if $A(x)$ is not increasing in $(x_1, x_a^{(-1)})$). When $x > x_a^{(+1)}$, according to the fact that we consider that $A(x)$ has at most one maximum, we suppose that $A(x)$ is decreasing. For the computation of the possible TP in this interval, we apply Corollary 3.15.

Algorithm: Computation of Turning Points.

Let $I \equiv [x_1, x_2]$ be the interval for the computation of TPs of $y(x)$, which is the solution of a second order ODE $y'' + A(x)y = 0$. $A(x)$ is a differentiable function with at most one extremum ($x = x_e$) in I, such possible extremum being a maximum.

Let L be such that $-L < x_1$, $L > x_2$. We define the quantities $x_a^{(-1)}$ and $x_a^{(+1)}$ as follows: $x_a^{(s)} \in I$ are such that $A(x_a^{(s)}) = 0$ and $A'(x_a^{(s)}) * s < 0$; if a value $x_a^{(s)}$ does not exist, we set $x_a^{(s)} = s * L$.

x_e is the maximum of $A(x)$. We set $x_e = -L$ if $A'(x_1) \leq 0$, and $x_e = L$ if $A'(x_2) \geq 0$. (There can be no maximum in these cases.)

Let $x^{(j)}$ be the zeros of $y(x)$, $x^{(n)} \in I$ being the smallest zero in I greater than $x_a^{(-1)}$, and $x^{(m)}$ ($m \geq n$) the largest zero in I smaller than $x_a^{(+1)}$.

Also, we will use the function $h(x) = -y'(x)/y(x)$ and, finally, given three real values x_l, x_u ($x_l < x_u$) and x_p, we define

$$\mathrm{isg}(x_l, x_u, x_p) = \begin{cases} -1 & \text{if } x_p \leq x_l, \\ 0 & \text{if } x_p \in]x_l, x_u[, \\ +1 & \text{if } x_p \geq x_u. \end{cases}$$

Then, the following algorithm computes with certainty all the zeros of $y'(x)$ in I.

Input: x_1; x_2; $x_a^{(\pm 1)}$; x_e, $x^{(n)} \dots x^{(m)}$; $\epsilon \equiv$ relative precision
Output: $x'(j)$: turning points in I. $j = ((n-2), n-1), n, \dots, m-1, (m, (m+1))$
Common parameters and functions for the routines: x_e, ϵ, $A(x)$, $h(x)$

```
DO $i = n, m-1$
    IC=isg($x^{(i)}$,$x^{(i+1)}$,$x_e$)
    IF (IC=0) THEN
        $x_o = x_e$
    ELSEIF ($1/9 < A(x^{(i+1)})/A(x^{(i)}) < 9$) THEN
        $x_0 = (x^{(i+1)} + x^{(i)})/2 + IC * \frac{\pi}{4}[A(x^{(i+1)})^{-1/2} - A(x^{(i)})^{-1/2}]$
    ELSE
        CALL SX(+1,IC,$x^{(i)}$,$x^{(i+1)}$,$x_0$)
    ENDIF
    CALL FP(+1,$x_0$,$x$)
    $x'(i) = x$
ENDDO
$X_{out}$ =MIN($X_a^{(-1)}$,$x_1$); $X_{A0}$ = MAX($x_a^{(-1)}$,$x_1$); $X_{in} = x^{(n)}$
CALL EXTR(−1,$X_{out}$,$X_{A0}$,$X_{in}$, $x'_{ext}(1)$, $x'_{ext}(2)$)
IF ($x_{ext}(2) \geq x_1$) THEN
    $x'(n-1) = x'_{ext}(2)$; IF ($x'_{ext}(1) \geq x_1$) THEN $x'(n-2) = x'_{ext}(1)$ ENDIF
ENDIF
$X_{out}$ = MAX($x_a^{(+1)}$,$x_2$); $X_{A0}$ = MIN($x_a^{(+1)}$,$x_2$); $X_{in} = x^{(m)}$
CALL EXTR(+1,$X_{out}$,$X_{A0}$,$X_{in}$, $x'_{ext}(1)$, $x'_{ext}(2)$)
```

```
(20)  IF ($x_{ext}(2) \le x_2$), THEN
(21)      $x'(m) = x'_{ext}(2)$; IF ($x'_{ext}(1) \le x_2$) THEN $x'(m+1) = x'_{ext}(1)$ ENDIF
(22)  ENDIF
(23)  END
```

SUBROUTINE SX(par,sign,x_l,x_u,x_o)

Input: par, sign, x_l, x_u. **Output:** x_0

```
$(1)_1$   IF (sign=0), THEN
$(2)_1$       $x_0 = x_e$
$(3)_1$   ELSE
$(4)_1$       $x_m = (x_l + x_u)/2$; $h_m = h(x_m)$; $h_l = -$par
$(5)_1$       DOWHILE ($h_m *$ sign $< 0$ or $-$par $* h_m^2 > -0.5 * (1 -$ par$) * A(x_m)$)
$(6)_1$           IF ($h_m * h_l < 0$), THEN
$(7)_1$               $x_u = x_m$
$(8)_1$           ELSE
$(9)_1$               $x_l = x_m$
$(10)_1$          ENDIF
$(11)_1$          $x_m = (x_l + x_u)/2$; $h_m = h(x_m)$
$(12)_1$      ENDDO
$(13)_1$      $x_0 = x_m$
$(14)_1$  ENDIF
$(15)_1$  END
```

SUBROUTINE FP(par,x_0,x)

Input: par, x_0. **Output:** x

```
$(1)_2$   $x = x_0$
$(2)_2$   DO WHILE (Err$> \epsilon$)
$(3)_2$       $x_p = x$
$(4)_2$       IF (par=1), THEN
$(5)_2$           $x = x + \frac{1}{\sqrt{A(x)}} \arctan\left(\frac{y'(x)/y(x)}{\sqrt{A(x)}}\right)$
$(6)_2$       ELSE
$(7)_2$           $x = x - \frac{1}{\sqrt{-A(x)}} \mathrm{atanh}\left(\frac{y'(x)/y(x)}{\sqrt{-A(x)}}\right)$
$(8)_2$       ENDIF
$(9)_2$       Err$= |1 - x/x_p|$
$(10)_2$  ENDDO
$(11)_2$  END
```

SUBROUTINE EXTR(side,X_{out},X_{A0},X_{in}, $x'_{ext}(1)$,$x'_{ext}(2)$)

Input: side, X_{out}, X_{A0}, X_{in}. **Output:** $x'_{ext}(1), x'_{ext}(2)$

```
$(1)_3$   IF ($h(X_{A0}) = 0$), THEN
$(2)_3$       $x'_{ext}(2) = X_{A0}$; $x'_{ext}(1) =$side*L
$(3)_3$   ELSEIF ($h(X_{A0}) *$ side $> 0$), THEN
$(4)_3$       CALL ZE(+1,MIN($X_{A0}$,$X_{in}$), MAX($X_{A0}$,$X_{in}$),x); $x'_{ext}(2) = x$
$(5)_3$       IF (L-side*$X_{out} < 0$) and ($h(X_{out}) *$ side $< 0$), THEN
$(6)_3$           CALL ZE($-1$, MIN($X_{A0}$,$X_{out}$), MAX($X_{A0}$,$X_{out}$),x); $x'_{ext}(1) = x$
$(7)_3$       ELSE
$(8)_3$           $x'_{ext}(1) =$L*side
$(9)_3$       ENDIF
$(10)_3$  ELSE
$(11)_3$      $x'_{ext}(1) = x'_{ext}(2) =$side*L
$(12)_3$  ENDIF
$(13)_3$  END
```

SUBROUTINE ZE(par,x_l,x_u,x)

Input: par, x_l, x_u. **Output:** x

```
$(1)_4$   IC=isg($x_l$,$x_u$,$x_e$); CALL SX(par,IC,$x_l$,$x_u$,$x_0$); CALL FP(par,$x_0$,x)
$(2)_4$   END
```

The structure of the algorithm is as follows.

1. Lines (1)–(23) are the main program.
    (i) (1)–(12). The TPs of y, which are bracketed by their zeros (inside the interval where $A(x)$ is positive), are computed. Line (4) implements rule

(c) (one maximum of $A(x)$ between the two consecutive zeros). In line (5) the rules (a) or (b) (after Proposition 3.9) to obtain the starting values are implemented (case of monotonic $A(x)$), in the case when the upper (K) and lower bounds (k) for $A(x)$ verify $K/k < 9$; if this condition is not met, rules (a) or (b) cannot be applied. Then, the program calls SX, which computes a starting value (using a few bisection steps) for which the conditions of Theorem 3.8 are met. Then, the subroutine FP, which implements this theorem (as well as Corollary 3.15), is applied.

(ii) (13)–(17). The TPs in $[x_a^{(-1)}, x^{(n)})$ and $[x_1, x_a^{(-1)})$ are computed; if they exist, they are stored in the positions: $n-1$ and $n-2$ (respectively) of the array x'. Such computation is performed by the routine EXTR with first parameter -1.

(iii) (17)–(21). Same as before, but for the intervals $(x^{(m)}, x_a^{(+1)}]$ and $(x_a^{(-1)}, x_2]$. The TPs, if they exist, are stored in the positions m and $m+1$. Such computation is performed by the routine EXTR with first parameter $+1$.

2. Lines $(1)_1$–$(15)_1$. As described above, the routine SX performs a few bisection steps until either Theorem 3.8 or Corollary 3.15 can be applied.
3. Lines $(1)_2$–$(11)_2$. The fixed point iterations of Theorem 3.8 (line $(5)_2$) and Corollary 3.15 (line $(7)_2$) are implemented.
4. Lines $(1)_3$–$(13)_3$. This routine checks the existence of extreme TPs (those described above in items 1(ii) and 1(iii)). If they exist, starting values are computed which guarantee that Theorem 3.8 and/or Corollary 3.15 can be applied.
5. Lines $(1)_4$–$(2)_4$. Auxiliary routine called by EXTR (lines $(1)_3$–$(13)_3$).

4. Computation of the ratios y/w and the logarithmic derivatives y'/y.

In the methods described in the two previous sections, it is assumed that algorithms are available to compute the ratio between the problem function and the contrast function y/w for Method I and the logarithmic derivative for Method II. Although the numerical analysis of the computation of these ratios lies beyond the scope of the present paper, we give here some hints for the computation of ratios for functions satisfying three term recurrence relations (TTRRs).

We consider the particular case of functions which are solutions of a uniparametric family of second order ODEs, as described in section 2.3.1; this is the case of a great number of special functions. In this case, the ratios for both methods are related. The contrast function for the computation of the zeros of y_n (Method I) will be y_{n-1} or y_{n+1}. Using (2.19), we have that the logarithmic derivative for the problem function can be written as $y_n'/y_n = d_n(x) + a_n(x)y_n(x)/y_{n-1}(x)$; on the other hand, the ratio y_n/y_{n+1} is related to y_n/y_{n-1} through the TTRR (2.24), which can be used to compute the ratios $R_n = y_n/y_{n-1}$ from recursion:

$$\begin{aligned} R_n &= r_{n-1} + \frac{s_{n-1}}{R_{n-1}} \qquad \text{(forward recurrence)}, \\ R_n &= \frac{s_n}{R_{n+1} - r_n} \qquad \text{(backward recurrence)}. \end{aligned} \tag{4.1}$$

The problem of computing ratios of special functions of consecutive orders, $R_n = y_n/y_{n-1}$, satisfying a TTRR has been broadly studied in the literature. A classical reference for condition and stability issues is the article by Gautschi [6]; see also [30, 4].

Important results are Perron's and Pincherle's theorems. Perron's theorem is a classical result for determining when a TTRR admits a minimal solution $y_n^{\downarrow}$ (see [6] and [25]) and what its asymptotic behavior is as $n \to \infty$. A solution $y_n^{\downarrow}$ of a TTRR is said to be minimal if, for any other solution $y_n^{\uparrow}$ (dominant solution), we have that $\lim_{n\to\infty} y_n^{\downarrow}/y_n^{\uparrow} = 0$. (Of course, the minimal solution, if it exists, is unique.) Perron's theorem does not always provide an answer to the problem of the existence of a minimal solution; there are cases which need to be analyzed by other means. Of course, there is also the possibility that a TTRR has no minimal solution (and hence no dominant ones).

When a solution $y_n^{\uparrow}$ is known to be dominant, forward recursion is well conditioned, and the first equation in (4.1) can be safely iterated to compute the values of $R_n = y_n^{\uparrow}/y_{n-1}^{\uparrow}$ starting from a ratio $y_k^{\uparrow}/y_{k-1}^{\uparrow}$ $(k < n)$; for a minimal solution the same happens for the backward recursion (that is, starting from $k > n$ and using the second equation in (4.1)). A related result is Pincherle's theorem, which guarantees the convergence of the infinite continued fraction for the ratio of consecutive minimal solutions $y_n^{\downarrow}/y_{n-1}^{\downarrow}$ obtained by iterating infinitely many times the second recurrence of (4.1).

Then, if the function under consideration is a minimal solution of the corresponding TTRR, an infinite continued fraction can be used to compute the ratios $y_n^{\downarrow}/y_{n\pm1}^{\downarrow}$. It should be noted that only the coefficients r_n and s_n of the TTRR are needed for such a computation. In order to compute the continued fraction, the use of the Lentz–Thompson algorithm [2, 26, 17] has the great advantage that overflow problems caused by cancellation of the denominators are under control; this is an important feature when computing ratios of functions both having zeros in the interval under consideration. Minimal solutions are, for instance, the Bessel functions J_ν, the conical functions $P_{-1/2+i\tau}^{m}$, and the regular Coulomb functions $F_{L,\rho}(x)$. Of course, depending on the range of the parameters and the variables, the continued fraction will converge at different speeds. For instance, for the case of first kind Bessel functions $J_\nu(x)$ that we will later describe, the convergence of the continued fraction slows down as x increases. The same is true for regular Coulomb wave functions and for conical functions.

Apart from the condition issues previously described, it is impossible to provide a general recipe for the computation of these ratios of special functions. For instance, for dominant solutions (or more generally nonminimal solutions), even if the forward recurrence relation can be used to compute the ratios y_n/y_{n-1} starting from, say, y_1/y_0, the difficulty in the computation of this initial ratio is strongly dependent on the function under consideration: it is trivial to compute for orthogonal polynomials (understanding that n is the degree of the polynomial), but it is not so for irregular Coulomb wave functions, among other examples. (For more trivial and nontrivial examples, see, for instance, [18].) In fact, for the case of Coulomb functions that we will later analyze, the computation of the zeros for the regular Coulomb wave function can be made by using its continued fraction representation, while for any other solution combining the regular and the irregular solutions, we will directly compute the numerator and the denominator using Barnett's code [2].

We will not insist on the analysis of the computation of special function values. For the theory of the computation through TTRRs, see [4, 6, 25, 30]; numerical details on the computation of special function ratios can be found, for instance, in [2, 21, 20].

As previously noted, there are a great number of functions for which stable methods to compute the ratios are available. This has enabled the authors to develop algorithms which are able to compute the zeros and TPs of the classical orthogonal polynomials and of several important special functions: Jacobi and related polynomials (Gegenbauer, Legendre), Laguerre polynomials, Hermite polynomials, Bessel functions, Airy functions, and Coulomb wave functions.

5. Numerical examples.

In order to illustrate the performance of the methods developed, we will consider three different examples for which Methods I and II are combined to compute the zeros and TPs: Bessel functions, Coulomb wave functions, and conical functions. Of these three examples, the cases of the zeros of Bessel functions [1, 10, 12] and Coulomb wave functions [1, 11, 12] have been also considered as illustrations of the matrix methods.

5.1. Bessel functions.

The Bessel equation reads

$$y_n'' + \frac{1}{x} y_n' + \left(1 - \frac{n^2}{x^2}\right) y_n = 0 \tag{5.1}$$

with real n. The zeros and TPs of the regular Bessel function can be computed by means of matrix eigenvalue methods [1, 10, 12] (being the minimal solution of a TTRR). The zeros and TPs of irregular solutions $Y_n(x)$ can not be computed using these methods, but methods based on asymptotics [24] or more general-purpose methods [13, 27, 28] have been used. The methods discussed here cover not only these cases but also general cylinder functions $\mathcal{C}_n \equiv \cos\alpha J_n - \sin\alpha Y_n$ (and thus, for instance, also the case of Airy functions).

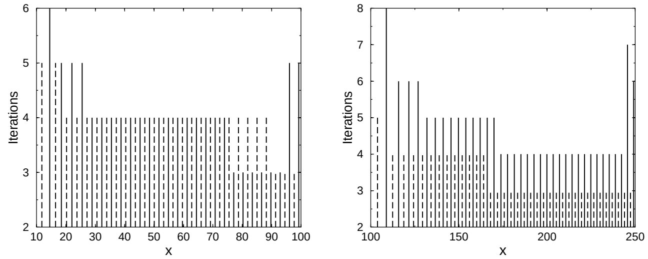
Ricatti–Bessel functions are the solutions of the second order ODE

$$y_n'' + \left(1 - \frac{n^2 - 1/4}{x^2}\right) y_n = 0, \tag{5.2}$$

with solutions related to Bessel functions by $y_n(x) = \sqrt{x}\mathcal{C}_n(x)$. Their zeros, of course, coincide with the zeros of Bessel functions but not their TPs. We discuss the evaluation of the zeros and TPs of Bessel and Ricatti–Bessel functions (appearing, for instance, in quantum electromagnetism [3]) and, in general, of TPs of $x^\alpha \mathcal{C}_n(x)$, that is, of zeros of $\alpha\mathcal{C}_n(x) + x\mathcal{C}_n'(x)$.

The evaluation of the zeros of Bessel or Ricatti–Bessel functions is performed as discussed in section 2.5.2. Figure 1 shows the performance of the fixed point method (Method I) for the evaluation of the zeros of Bessel functions (backward sweep). FORTRAN language was used, and the ratios J_n/J_{n-1} were computed using the continued fraction representation. The number of iterations needed to achieve a double precision ($\sim$ 15 digits) computation for each zero is shown. Notice that the number of iterations needed for the evaluation of the smallest zero increases with increasing order. This is as expected since the function $|\eta(x)|$ increases with decreasing x, which means that the convergence of the fixed point method must be slower for low x; besides, the spacing between the first zeros becomes larger as the order increases. However, the number of iterations increases mildly. Notice also that the two largest zeros evaluated need more iterations than the immediately smaller zeros; this is so because we need to compute two zeros before improved iteration steps can be considered. This figure thus illustrates the effect of improving the iteration step.

We have also implemented the algorithm in Maple, with the advantage that one can select almost arbitrarily the number of digits for the computation. The description of a Maple procedure to compute zeros of special functions using Method I can be found in [8]. With this, one can check that with few additional iterations the achieved



(a) (b)

FIG. 1. *Number of iterations needed for the evaluation of the zeros of the first kind Bessel function $J_n(x)$ and its derivative, plotted as a function of the values of the zeros of J_n and its derivative. Results for two different orders are shown:* (a) $n = 10$ *and* (b) $n = 100$. *The solid vertical bars correspond to the number of iterations needed in order to compute the zeros with* 15 *exact digits; the dashed lines represent the number of iterations for the computation of TPs.*

accuracy can be greatly improved. For instance, in order to achieve 10^{-100} accuracy only two or three additional iterations of the function $T(z)$ are needed for $n = 10$ and similarly for $n = 100$.

Regarding the TPs, the zeros of the derivative of Bessel–Ricatti functions can be computed directly using Theorem 3.8, given that these functions satisfy a second order ODE in normal form. For the case of Bessel functions, defined by (5.1), we have to consider a change of variables to transform the equation into normal form (section 3):

$$z(x) = \int \exp\left(-\int B dx\right) dx, \qquad B(x) = \frac{1}{x} \to z(x) = \ln x. \tag{5.3}$$

In the new variable z the differential equation reads

$$\ddot{y}_n + [e^{2z} - n^2] y_n = 0, \tag{5.4}$$

and $\tilde{A}(z) = e^{2z} - n^2$ ($\tilde{A}(z(x)) = x^2 - n^2$) is increasing. After this change of variables, one can apply the methods described in sections 3.1 and 3.3. Notice that the x variable in these sections corresponds to the z variable of (5.4). We can undo this change in the fixed point iteration, first rewriting the logarithmic derivative as

$$D(z(x)) = \frac{\dot{y}_n(x(z))}{y_n(x(z))} = x\frac{y_n'(x)}{y_n(x)}. \tag{5.5}$$

With this, the fixed point iteration of Theorem 3.12 reads

$$T(z) = z + \frac{1}{\sqrt{x^2 - n}} \arctan\left(\frac{x y_n'(x)}{y_n(x)\sqrt{x^2 - n}}\right), \qquad x = e^z. \tag{5.6}$$

For seeking the TPs inside intervals $(x_n^{(j)}, x_n^{(j+1)})$ the previous iteration is used, taking as a starting value $z_0 = z_m + \lambda$, with

$$z_m = \frac{\ln(x_n^{(j)}) + \ln(x_n^{(j+1)})}{2} = \ln\left(\sqrt{x_n^{(j)} x_n^{(j+1)}}\right) \tag{5.7}$$

and

$$\lambda = \frac{\pi}{4}\left(k_n^{-1/2} - K_n^{-1/2}\right), \quad k_n = (x_n^{(j)})^2 - n^2, \quad K_n = (x_n^{(j+1)})^2 - n^2. \tag{5.8}$$

Here z_0 is larger than $z(x'_n)$ and lies inside $(z(x_n^{(j)}), z(x_n^{(j+1)}))$ whenever $K_n/k_n < 9$. The corresponding starting value x_0 is therefore $x_0 = \exp(\lambda)\sqrt{x_n^{(j)} x_n^{(j+1)}}$.

For TPs outside intervals $(x_n^{(j)}, x_n^{(j+1)})$ (or in the rare situations for which $K_n/k_n \geq 9$) we proceed as discussed in sections 3.1 and 3.3.

With this, it turns out that the computation of the TPs is very efficient and that typically 3-4 iterations are enough to attain double precision values (15 digits) for the turning points (Figure 1). Similarly, as happens with the computation of the zeros, only two additional iterations are required when the accuracy demanded is 10^{-100}.

As a spin-off of the method, it is interesting to note that the fact that $z_m = z(x_m) < z(x'_n)$ leads to the inequality $\sqrt{x_n^{(j)} x_n^{(j+1)}} < x'_n$ relating two consecutive zeros of a cylinder function and the zero of the derivative between such zeros. This is a general inequality for the solutions of Bessel equations which proves a conjecture by Elbert [5, 23].

THEOREM 5.1. *Let $j_{\nu,\kappa}$ and $j_{\nu,\kappa+1}$ be two consecutive zeros of a solution of Bessel differential equations and $j'_{\nu,\kappa}$ the TP between them; then*

$$j'_{\nu,\kappa} > \sqrt{j_{\nu,\kappa} j_{\nu,\kappa+1}}.$$

Similar methods can be used to compute the zeros of other functions related to Bessel functions. For instance, the turning points of $x^\alpha \mathcal{C}_n(x)$ are the roots of

$$\alpha \mathcal{C}_n(x) + x\mathcal{C}'_n(x) = 0, \tag{5.9}$$

which are important roots in many applications. Using (5.1), the differential equation satisfied by the functions $y_n = x^\alpha \mathcal{C}_n(x)$ is

$$y''_n + \frac{1-2\alpha}{x} y'_n + \left(1 - \frac{n^2 - \alpha^2}{x^2}\right) y_n = 0, \tag{5.10}$$

which is transformed into the normal form by the change of variables $z(x) = x^{2\alpha}/2\alpha$ ($\alpha \neq 0$, which is the case previously discussed). After changing variables, we have that the coefficient of the second order ODE in normal form reads

$$A(z(x)) = (x^2 - (n^2 - \alpha^2))x^{-4\alpha}, \tag{5.11}$$

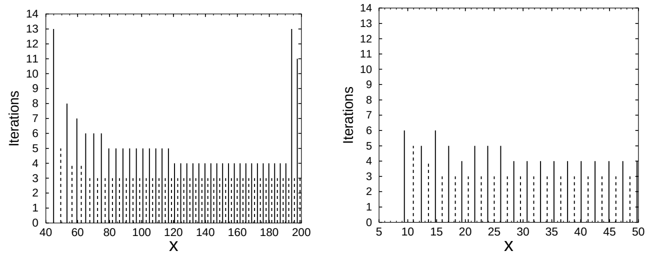
and the interlacing between the zeros of $\mathcal{C}_n$ and $\alpha\mathcal{C}_n(x) + x\mathcal{C}'_n(x)$ for $x > \sqrt{n^2 - \alpha^2}$ follows immediately. Methods similar to those used before can therefore be applied. Additional properties of the zeros of $x^\alpha \mathcal{C}_\nu(x)$ have been explored in [23] using these ideas.

5.2. Coulomb wave functions.

Coulomb functions are the solutions of the second order ODE

$$y''_n + A_n(x) y_n = 0, \qquad A_n(x) = 1 - \frac{2\gamma}{x} - \frac{n(n+1)}{x^2} \tag{5.12}$$

and satisfy the DDEs (2.19) with coefficients $a_n = -b_n = -(n^2/x + \gamma)/n$, $d_n = -e_n = \sqrt{n^2 + \gamma^2}/n$. Ricatti–Bessel functions are Coulomb functions with $\gamma = 0$.



(a) (b)

FIG. 2. *Number of iterations needed for the evaluation of the zeros and TPs with* 15 *exact digits for two different values of n and the parameter γ:* (a) $n = 10$*,* $\gamma = 20$ *and* (b) $n = 20$*,* $\gamma = -20$*. The zeros and TPs correspond to the combination of regular* (F) *and irregular* (G) *solutions* $0.5F - \sqrt{3}G/2$*.*

The main distinction of Coulomb functions with respect to Bessel functions is that $A_n(x)$ is not a monotonic function when $\gamma < 0$, which results in new phenomena with respect to the evaluation of zeros and TPs. Similarly to Bessel functions, the change of variables is trivial in this case (a factor times x). We will consider the evaluation of the positive zeros. (The negative zeros are the positive zeros of a Coulomb function with the opposite sign of γ.)

The functions η_i (see (2.23)) which determine the type of sweep to be applied are $\eta_i(x) = -i(\gamma + n_i^2/x)/\sqrt{n_i^2+\gamma^2}$, where $i = \pm 1$ and $n_{+1} = n + 1$, $n_{-1} = n$. The iteration corresponding to $i = -1$ is the most appropriate since the improved iteration step (Theorem 2.4) can be considered; indeed, $A'(x)\eta_{-1}(x) > 0$ except when $\gamma < 0$ and $x \in [x_{\eta_{-1}}, x_{Max}]$, where $x_{Max} = -n(n+1)/\gamma$ is the maximum of $A(x)$ and $\eta_{-1}(x_{\eta_{-1}}) = 0$ ($x_{\eta_{-1}} = -n^2/\gamma$). In that interval, one could always use the iteration step without improvement ($\Delta z = \pm\pi/2$). For $\gamma \geq 0$, $A(x)$ is always increasing, and the method of computation of zeros is a backward sweep ($\eta_{-1} > 0$). For $\gamma < 0$, as remarked, $A_n(x)$ has a maximum, and the zeros are computed by an expansive sweep starting from $x_{\eta_{-1}}$ (see [22]).

The monotonicity properties of $A_n(x)$ also have consequences for the evaluation of TPs. For instance, for TPs inside intervals $(x_n^{(j)}, x_n^{(j+1)})$, when $A_n(x)$ is increasing we should use starting values $x_0 = x_m + \lambda$, while we must use $x_0 = x_m - \lambda$ when $A_n(x)$ is decreasing; in this way, Theorem 3.8 can be applied. In the interval $(x_n^{(j)}, x_n^{(j+1)})$ where x_m lies we can also use Theorem 3.8 by proceeding as described after Proposition 3.9.

Figure 2 shows the performance of the method for the following choice of parameters: $n = 10$, $\gamma = 20$ (Figure 2(a)) and $n = 20$, $\gamma = -20$ (Figure 2(b)). The improved iteration step is used for computing the zeros. We use Barnett's code [2] for computing the F and G functions and obtaining the ratios appearing in the fixed point iterations.

The Maple implementation of the algorithms (for the computation of zeros of F by means of the continued fraction representation for the ratios) shows that two or three extra iterations are required to compute the zeros (and the TPs) for an accuracy of 10^{-100} (in addition to those required for 10^{-15} accuracy).

5.3. Conical functions. Conical functions are Legendre functions of degrees $-1/2 + i\tau$ with real τ ($y_n \equiv P^n_{-1/2+i\tau}(x)$ or $y_n \equiv Q^n_{-1/2+i\tau}(x)$). They satisfy general

DDEs (see (2.19)) with coefficients

$$\begin{aligned} &y_n''(x) + B(x)y_n'(x) + A_n(x)y_n(x) = 0, \\ &B(x) = \frac{2x}{x^2-1}, \qquad A_n(x) = \frac{1/4 + \tau^2 - n^2/(x^2-1)}{x^2-1}. \end{aligned} \tag{5.13}$$

We consider $n > 0$ (the differential equation is invariant under the replacement $n \to -n$), and we restrict our study to $x > 1$ (where the functions oscillate).

The DDEs (2.19) have the coefficients

$$a_n = \frac{-nx}{x^2-1}, \quad b_n = \frac{(n-1)x}{x^2-1}, \quad d_n = -\frac{\lambda_n^2}{\sqrt{x^2-1}}, \quad e_n = \frac{1}{\sqrt{x^2-1}}. \tag{5.14}$$

The η_i functions are (see (2.23)) $\eta_i(x) = i(-n_i + 1/2)x/(\lambda_{n_i}\sqrt{x^2-1})$, where, as usual, $n_{+1} = n+1$ and $n_{-1} = n$. With this, we see that $i = -1$ is the appropriate iteration in order to apply the improved iteration. Indeed, choosing $i = -1$, the associated change of variables (see (2.22)) is

$$z(x) = \lambda_n \cosh^{-1} x, \tag{5.15}$$

with $\lambda_n = \sqrt{(n-1/2)^2 + \tau^2}$, and the coefficient $\tilde{A}(z)$ of the equation in normal form (see (2.11)) is

$$\tilde{A}(z) = 1 + \dot{\eta}_{-1} - \eta_{-1}^2 = \frac{1}{\lambda_n^2}\left[\tau^2 - \frac{n^2 - 1/4}{\sinh^2(z/\lambda_n)}\right] \tag{5.16}$$

so that $\dot{\tilde{A}}(z)\eta_{-1}(z) > 0$ for $n > 0$, and therefore Theorem 2.4 can always be applied. The computation of zeros will be performed in the backward direction for $n > 1/2$ and in the forward direction of $n < 1/2$ because $\text{sign}(\eta_{-1}) = \text{sign}(n - 1/2)$.

Notice that from the expressions for both η_{-1} (Theorem 2.2 and Corollary 2.3) and $A(z)$ one deduces the following.

THEOREM 5.2. *Conical functions have infinitely many zeros for $x > 1$ and $\tau \neq 0$.*

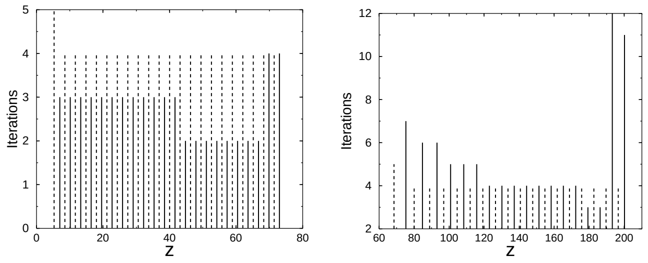
Conical functions illustrate how the change of variables associated to the DDE tends to make uniform the distribution of zeros. In fact, we see that $\tilde{A}(z) \to (\tau/\lambda_n)^2$ as $z \to +\infty$, which means that the difference between zeros in the z variable for large z tends to

$$z_y^{(j+1)} - z_y^{(j)} = \pi\sqrt{1 + \left(\frac{n-1/2}{\tau}\right)^2}, \qquad j \to +\infty.$$

In fact, for the particular case $n = 1/2$ we observe that the zeros are equally spaced in the z variable.

Regarding the TPs and going back to (5.13), the equation can be transformed to the normal form by taking a change of variables so that $dz/dx = \exp(-\int B dx)$. Let us again insist (as we did for Bessel functions) that the z variable here corresponds to the x variable in sections 3.1–3.2 and that this z variable is unrelated to (5.15). With this,

$$z(x) = -\tanh^{-1}\left(\frac{1}{x}\right), \qquad D(z) = \frac{\dot{y}_n(z)}{y_n(z)} = (x^2-1)\frac{y_n'(x)}{y_n(x)}, \tag{5.17}$$



(a) (b)

FIG. 3. *Number of iterations required for the computation of the zeros of* $P^n_{-1/2+i\tau}$ *(solid bars) and those of its derivative (dashed bars) with* 15 *exact digits as a function of its location (in the z variable of* (5.15)*). The left figure is for* $\tau = 20$*,* $n = 1$ *while the figure on the right corresponds to* $\tau = 20$*,* $n = 40$.

and the coefficient $\tilde{A}_n(z)$ appearing in the ODE in normal form (see (3.5)) is

$$\tilde{A}_n(x(z)) = (x(z)^2 - 1)A_n(x(z)) = \left(\frac{1}{4} + \tau^2\right)(x(z)^2 - 1) - n^2, \tag{5.18}$$

which is increasing in x. A method similar to the one applied for Bessel functions is the choice for conical functions.

Figure 3 shows the number of iterations required for the computation of the zeros and TPs with 15 exact digits. The effect of the improved iteration step becomes apparent from Figure 3(b), where the two largest zeros are computed by the plain iterative sweep (with $\Delta z = \pi/2$) and require more iterations than for the computation of the rest of zeros, which profit from the improved iteration step. This effect is also observed in Figure 3(a), but the improvement is not so noticeable.

6. Comparison with other methods. Other methods for computing the real zeros of special functions are the general-purpose methods described, for instance, in [13, 27, 28] (which have been applied to Bessel and Airy functions), or the most specialized methods which are based on previous approximations to the roots, mainly asymptotic approximations [16, 24]. Of course, when accurate a priori approximations to the roots are available, one can build efficient codes which use these approximations as starting points of, for instance, a Newton method; the algorithm presented in [24] is an example of an efficient algorithm based on asymptotic approximations.

Differently from these specific methods, the fixed point methods that we have described are general for a considerable number of special functions. Generally speaking, and starting from the least general to the most general methods (attending to the number of cases that they cover), the least general (but very efficient) methods are those based on specific approximations [16, 24], followed by matrix methods [1, 10, 11], our fixed point methods and the more general purpose methods [13, 27, 28].

Our computational scheme can be considered to be more general than matrix methods for the computation of real zeros in the sense that, given a second order ODE to which we can apply our methods, they are valid for any solution of the differential equation. In contrast, matrix methods can be applied only to minimal solutions with respect to a TTRR or to orthogonal polynomials (Golub–Welsch algorithm [9], [29, Problem 9, p. 80]). For example, matrix methods apply to the zeros of the regular Bessel function $J_\nu(x)$ and the regular Coulomb wave functions $F_L(\eta, \rho)$, while fixed point methods apply to any solution of the Bessel equation $\mathcal{C}_\nu = \cos\alpha J_\nu(x) -$

Table 1

Timing comparison for the evaluation of the zeros of Bessel functions $J_\nu(x)$ between the matrix method and the fixed point method (FPM) for 15 *exact digits. The first column represents the order of the function; the first row is the number of computed zeros. In each entry of the table the first number is the ratio between the time spent by the fixed point method and the time spent by the matrix diagonalization (if smaller than one FPM is faster) and the second number is the ratio between the size of the truncated matrix and the number of zeros N.*

ν\N	10	20	50	100
0	0.64|2.5	0.58|2.2	0.48|1.9	0.43|1.8
1	0.71|2.5	0.61|2.2	0.52|1.9	0.44|1.8
2	0.67|2.5	0.67|2.2	0.61|1.9	0.48|1.8
5	0.98|2.8	0.72|2.3	0.60|2.0	0.59|1.8
10	1.1|2.9	0.94|2.3	0.67|2.0	0.65|1.8
20	1.4|3.1	1.0|2.5	0.77|2.1	0.63|1.9
50	2.1|3.7	1.4|2.8	0.85|2.2	0.66|2.0

$\sin\alpha Y_\nu(x)$ or any solution of the Coulomb equation. On the other hand, matrix methods can also be applied to compute complex zeros. There are other advantages of fixed point methods with respect to matrix methods, as discussed in [22], like, for instance the fact that the interval $[x_1, x_2]$ for computing the zeros can be freely chosen (while for matrix methods one has to choose to compute a given number of zeros in increasing order). Besides, our methods can be mixed with more specific methods using a priori approximations to the roots; this can be used to improve the performance.

For the case of orthogonal polynomials, the matrix eigenvalue methods are exact in the sense that the N zeros of an orthogonal polynomial of degree N can be obtained by diagonalizing a real symmetric matrix of size N. However, for the case of functions with infinitely many zeros, of course this is not so; in this case, if the function is the minimal solution of a TTRR, the problem of computing the first N zeros can be approximated by the problem of diagonalizing a truncated matrix, the exact problem being the diagonalization of a matrix of infinite size. The size of the matrix must be greater than the number of zeros that are needed to a prescribed accuracy. The main difficulty consists in estimating the size of the truncated matrix, which is function- and parameter dependent. (See [12] for estimations for the regular Bessel function, and [15] for the regular Coulomb wave function.)

Regarding the efficiency of the different methods, the most general methods tend to be slower than the most specific ones; however, fixed point methods are an exception to this rule. Indeed, the fixed point methods, though being applicable to a wider set of functions than matrix methods, are not slower. On the other hand, the comparison between our method and the more general purpose method [27] based on the concept of topological degree favors the fixed point method: in [20] such a method was compared against a global Newton method which proved to be faster and, as shown in [22], the fixed point method developed here converges faster than the Newton method [20].

In Table 1, we show the relative comparison of CPU times spent by Method I and matrix methods to compute a given number of zeros of Bessel functions for different orders. The comparison in Table 1 is performed in the following way: we compute the first N zeros of the Bessel function with our method for a relative precision better than 10^{-15}. Then we select the size of the matrix to be diagonalized by testing that the N zeros are computed with a relative precision better than 10^{-15}. We use an efficient algorithm for tridiagonal matrices [17]. The comparison in Table 1 is rather unfair because we are selecting the optimal size of the matrix for a given precision and number of zeros by comparing with our method; a stand-alone algorithm based on matrix methods should provide its own estimations of the size of the matrix, which are function dependent and not so easy to compute (see [12, 15]). As mentioned before, the truncation problem is one of the main limitations of the matrix methods for functions with infinitely many zeros.

Our methods admit simple improvements when accurate a priori approximations are available. For instance, for the particular case of Bessel functions, the performance for low numbers of zeros and high orders can be improved by considering the asymptotic approximations of [24] for high orders. And when a high number of zeros is demanded, asymptotic approximations for large zeros are also available. In any case, it is apparent that the fixed point methods are efficient methods by themselves.

7. Conclusions. Two methods for the computation of zeros and turning points of solutions of second order homogeneous linear ODEs have been presented. They are based on fixed point methods which compute with certainty all the zeros and TPs; these procedures have been applied to the computation of zeros and TPs of ODEs depending on one parameter n (like, for instance, hypergeometric functions and Bessel functions). These methods require that algorithms for the computation of ratios of solutions y_n/y_{n-1} be available. For minimal solutions of the corresponding TTRR, this ratio can be computed by using the continued fraction associated with this recurrence; the coefficients of the recurrences are then the only information necessary for computing the zeros and TPs (as happens with matrix methods [1, 11]). This is the case of the Bessel function $J_\nu(x)$, the regular Coulomb wave function $F_L(\eta, \rho)$, and the conical function $P^n_{-1/2+i\tau}(x)$, among others. Differently from matrix methods, the fixed point methods here presented can be applied to general solutions of the second order ODE; it is not required that the solutions be minimal with respect to a three-term recurrence.

The fixed point methods prove to be efficient schemes of computation, as illustrated with Bessel, Coulomb, and conical functions, and they can be systematically and easily implemented. The analytical steps prior to the application of the method are simple to perform and can be made automatic by using symbolic computation packages such as Maple or Mathematica. The methods here presented lead to efficient and portable algorithms for the computation of zeros and TPs of special functions. A Maple algorithm to compute the zeros and the TPs of the three families of functions discussed in detail in this article (Bessel, Coulomb, and conical functions) is available upon request to the authors; in the near future, the Maple algorithm will be expanded to include, among others, the cases of hypergeometric and confluent hypergeometric functions of real parameters and variables (which, as a subset, include classical orthogonal polynomials).

REFERENCES

[1] J. S. Ball, *Automatic computation of zeros of Bessel functions and other special functions*, SIAM J. Sci. Comput., 21 (2000), pp. 1458–1464.

[2] A. R. Barnett, *Coulomb functions for real lambda and positive energy to high accuracy*, Comput. Phys. Comm., 24 (1981), pp. 141–159.

[3] T. H. Boyer, *Concerning the zeros of some functions related to Bessel functions*, J. Math. Phys., 10 (1969), pp. 1729–1744.

[4] P. Deuflhard and P. Hohmann, *Numerical Analysis. A First Course in Scientific Computation*, Walter de Gruyter, Berlin, 1995.

[5] Á. Elbert, *Some recent results on the zeros of Bessel functions and orthogonal polynomials*, J. Comput. Appl. Math., 133 (2001), pp. 65–83.

[6] W. Gautschi, *Computational aspects of three-term recurrence relations*, SIAM Rev., 9 (1967), pp. 24–82.

[7] A. Gil, W. Koepf, and J. Segura, *Numerical algorithms for the computation of the zeros of hypergeometric functions*, in preparation.

[8] A. Gil and J. Segura, *A combined symbolic and numerical algorithm for the computation of the zeros of special functions*, J. Symbolic Comput., 35 (2003), pp. 465–485.

[9] G. H. Golub and J. H. Welsch, *Calculation of Gauss quadrature rules*, Math. Comput., 23 (1969), pp. 221–230.

[10] J. Grad and E. Zakrajsek, *Method for evaluation of zeros of Bessel functions*, J. Inst. Math. Appl., 11 (1973), pp. 57–72.

[11] Y. Ikebe, *The zeros of regular Coulomb wave functions and of their derivatives*, Math. Comput., 29 (1975), pp. 878–887.

[12] Y. Ikebe, Y. Kikuchi, and I. Fujishiro, *Computing zeros and orders of Bessel functions*, J. Comput. Appl. Math., 38 (1991), pp. 169–184.

[13] D. J. Kavvadias and M. N. Vrahatis, *Locating and computing all the simple roots and extrema of a function*, SIAM J. Sci. Comput., 17 (1996), pp. 1232–1248.

[14] I. Marx, *On the structure of recurrence relations* II, Michigan Math. J., 2 (1953), pp. 99–103.

[15] Y. Miyazaki, Y. Kikuchi, D. Cai, and Y. Ikebe, *Error analysis for the computation of zeros of regular Coulomb wave function and its first derivative*, Math. Comp., 70 (2001), pp 1195–1204.

[16] R. Piessens, *On the computation of zeros and turning points of Bessel functions*, Bull. Greek Math. Soc., 31 (1990), pp. 117–222.

[17] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in Fortran* 77, Cambridge University Press, Cambridge, UK, 1992.

[18] J. Segura and A. Gil, *Evaluation of associated Legendre functions off the cut and parabolic cylinder functions*, Electron. Trans. Numer. Anal., 9 (1999), pp. 137–146.

[19] J. Segura, *A global Newton method for the zeros of cylinder functions*, Numer. Algorithms, 18 (1999), pp. 259–276.

[20] J. Segura and A. Gil, *ELF and GNOME: Two tiny codes to evaluate the real zeros of the Bessel functions of the first kind for real orders*, Comput. Phys. Comm., 117 (1999), pp. 250–262.

[21] J. Segura, *Bounds on differences of adjacent zeros of Bessel functions and iterative relations between consecutive zeros*, Math. Comput., 70 (2001), pp. 1205–1220.

[22] J. Segura, *The zeros of special functions from a fixed point method*, SIAM J. Numer. Anal., 40 (2002), pp. 114–133.

[23] J. Segura, *On a conjecture regarding the extrema of Bessel functions and its generalization*, J. Math. Anal. Appl., to appear.

[24] N. M. Temme, *An algorithm with ALGOL* 60 *program for the computation of the zeros of ordinary Bessel functions and those of their derivatives*, J. Comput. Phys., 32 (1979), pp. 270–279.

[25] N. M. Temme, *Special Functions. An Introduction to the Classical Functions of Mathematical Physics*, John Wiley and Sons, New York, 1996.

[26] I. J. Thompson and A. R. Barnett, *Coulomb and Bessel functions of complex arguments and order*, J. Comput. Phys., 64 (1986) pp. 490–509.

[27] N. M. Vrahatis, O. Ragos, T. Skiniotis, F. A. Zafiropoulos, and T. N. Grapsa, *RFSFNS: A portable package for the numerical determination of the number and the calculation of roots of Bessel functions*, Comput. Phys. Comm., 92 (1995), pp. 252–266.

[28] M. N. Vrahatis, T. N. Grapsa, O. Ragos, and F. A. Zafiropoulos, *On the location and computation of zeros of Bessel functions*, Z. Angew. Math. Mech., 77 (1997), pp. 467–475.

[29] H. S. Wilf, *Mathematics for the Physical Sciences*, John Wiley and Sons, New York, 1962.

[30] J. Wimp, *Computation with recurrence relations. Applicable Mathematics Series*, Pitman (Advanced Publishing Program), Boston, 1984.

SIAM J. Numer. Anal.
Vol. 41, No. 3, pp. 856–868

A LEIBNIZ FORMULA FOR MULTIVARIATE DIVIDED DIFFERENCES*

CARL DE BOOR[†]

Abstract. The Leibniz formula, for the divided difference of a product, and Opitz's formula, for the divided difference table of a function as the result of evaluating that function at a certain matrix, are shown to be special cases of a formula available for the coefficients, with respect to any basis, of an "ideal" or "Hermite" polynomial interpolant, in any number of variables.

Key words. polynomials, interpolation, ideal, multivariate, Opitz formula

AMS subject classifications. 41A05, 41A10, 41A63, 65D05, 65D15

PII. S0036142902406818

1. Introduction.

The so-called Leibniz formula

$$\Delta(x_i,\dots,x_j)(fg)=\sum_{k=i}^{j}\Delta(x_i,\dots,x_k)f\ \Delta(x_k,\dots,x_j)g, \tag{1.1}$$

for the divided difference of a product in terms of the divided differences of the factors, has played a major role in the development of spline theory; it was an essential tool in the derivation of the B-spline recurrence relations. My earliest reference for it now is Popoviciu [16, p. 12], who refers, for the case of uniform spacing, to [9] where, on page 105, that special case of the formula is referred to as "bekannt". Nevertheless, the formula is generally credited (see, e.g., [15]) to Steffensen, because of his paper [20].

In this paper, the algebraic background of the Leibniz formula is explored, showing the formula to be equivalent to the Opitz formula (from [15]; see (2.1) below) that gives the divided difference table of any polynomial as the result of applying that polynomial to a certain matrix. This, in turn, is shown to be a particular consequence of the fact that, in G. Birkhoff's [2] terminology, polynomial interpolation is an "ideal" interpolation scheme. This insight is used to explore Leibniz (and Opitz) formulas for certain *multivariate* polynomial interpolation schemes and their associated divided differences.

The paper is laid out as follows. In section 2, the connection between the Leibniz formula and the Opitz formula is recalled, along with Opitz's way of deriving them. The next section brings a brief discussion of the basic features of "ideal" interpolation, i.e., linear projectors on the space of polynomials (in one or several variables, real or complex) whose kernel is a polynomial ideal. Section 4 provides the Opitz formula in the general setting of "ideal" interpolation, and the truncated Taylor series serves as a trivial illustration. The nontrivial details for both the Opitz and the Leibniz formulas are fully worked out for Chung–Yao interpolation in section 6. Such formulas for other divided differences are outlined in section 7. The final section points out that this paper's restriction to interpolation to polynomials is easily removed.

*Received by the editors May 2, 2002; accepted for publication (in revised form) October 15, 2002; published electronically May 22, 2003. This work was supported in part by a Steenbock Professorship.

http://www.siam.org/journals/sinum/41-3/40681.html

[†]Department of Computer Sciences, University of Wisconsin-Madison, 1210 W. Dayton St., Madison, WI 53706-1685 (deboor@cs.wisc.edu).

For ready reference, here is the (mostly, but not entirely, standard) notation used in this paper. $\alpha \in \mathbb{Z}_+^d$ denotes a **multi-index** or, more precisely, a d-index, i.e., a d-vector with nonnegative integer entries; $|\alpha| := \sum_j \alpha(j)$ is its **length** (or "**degree**"); also, $\alpha! := \prod_j \alpha(j)!$. There being no standard notation for it, I use

$$()^\alpha : \mathbb{F}^d \to \mathbb{F} : x \mapsto x^\alpha := \prod_j x(j)^{\alpha(j)}$$

for the monomial of **multidegree** α. Here, $\mathbb{F}$ is either $\mathbb{R}$ or $\mathbb{C}$, though usually it is $\mathbb{C}$. With this,

$$\Pi_{\mathbb{I}} := \operatorname{span}(()^\alpha : \alpha \in \mathbb{I}), \quad \mathbb{I} \subset \mathbb{Z}_+^d,$$

with the special cases

$$\Pi := \Pi(\mathbb{F}^d) := \Pi_{\mathbb{Z}_+^d}, \quad \Pi_k := \operatorname{span}(()^\alpha : |\alpha| \le k).$$

The *ad hoc* abbreviation

$$\widehat{p}(\alpha) := (D^\alpha p)(0)/\alpha!, \qquad p \in \Pi,\ \alpha \in \mathbb{Z}_+^d,$$

with

$$D^\alpha := \prod_j D_j^{\alpha(j)}$$

and D_j differentiation with respect to the jth argument, is convenient. Analogously,

$$()_j : x \mapsto x(j), \quad j = 1{:}d,$$

while

$$()_0 : x \mapsto 1.$$

In the dual, Π', of Π, *evaluation at some point* $v \in \mathbb{F}^d$ is singled out; i.e., the linear functional

$$\epsilon_v : \Pi \to \mathbb{F} : p \mapsto p(v),$$

and, more generally, $\epsilon_v q(D) : p \mapsto (q(D)p)(v)$ for $q \in \Pi$, with

$$q(D) := \sum_\alpha \widehat{q}(\alpha) D^\alpha.$$

Also,

$$Q(D) := \{q(D) : q \in Q\}, \qquad Q \subset \Pi,$$

and

$$\Lambda_\perp := \ker \Lambda := \bigcap_{\lambda \in \Lambda} \ker \lambda, \qquad \Lambda \subset \Pi'.$$

2. The Opitz formula. In his short note [15], describing a talk submitted but not given, Opitz introduces “Steigungsmatrizen” (literally “divided difference matrices”) as matrices of the form

$$S[f;X] := f(A_X),$$

with f a (univariate) polynomial or rational function or, more generally, a suitable limit of such functions, and, correspondingly, $f(A_X)$ the “value” of f at the matrix A_X, with

$$A_X := \begin{bmatrix} x_1 & 1 & & & \\ & x_2 & 1 & & \\ & & x_3 & \ddots & \\ & & & \ddots & 1 \\ & & & & x_n \end{bmatrix},$$

and with $X := (x_1, \ldots, x_n)$ a sequence of pairwise distinct complex numbers. Using the (obvious) eigenstructure of A_X, Opitz readily concludes that, for each i, j,

$$S[f;X](i,j) = \Delta(x_i, \ldots, x_j)f, \tag{2.1}$$

i.e., the divided difference of f at $(x_i, \ldots, x_j)$ (in W. Kahan’s felicitous notation[1]), and hence the name “Steigungsmatrix”. Here, as is customary, $\Delta(x_i, \ldots, x_j) := 0$ for $i > j$.

In other words, $f(A_X)$ is (or, the upper triangular part of $f(A_X)$ provides) the divided difference table for f with respect to the sequence X, and, as Opitz points out, its calculation in this fashion from A_X is less affected by loss of significance than is the direct construction of the divided difference table by the repeated formation of divided differences. In fact, it can be used for the symbolic calculation of divided differences; see, e.g., [10], and, most recently, [18].

Further, Opitz observes that the map

$$f \mapsto S[f;X]$$

is linear as well as multiplicative, and hence a ring homomorphism, from the ring of functions under pointwise addition and multiplication into the ring of matrices of order n. In particular,

$$(fg)(A_X) = f(A_X)g(A_X).$$

Because of (2.1), this is equivalent to the Leibniz formula, (1.1), i.e., to

$$\Delta(x_i, \ldots, x_j)(fg) = \sum_{k=i}^{j} \Delta(x_i, \ldots, x_k)f \ \Delta(x_k, \ldots, x_j)g.$$

Further, if we take (2.1) as the *definition* of $S[f;X]$, then the Leibniz formula implies that $f \mapsto S[f;X]$ is a ring homomorphism and so, in particular, $S[f;X] = f(A_X)$.

[1] I am using here Kahan’s notation not only because it is quite literal, but because the standard notation, $[x_i, \ldots, x_j]$, has already other uses, e.g., the matrix with columns $x_i, \ldots, x_j$ or, in the case $j = i + 1$, the closed interval with endpoints x_i, x_j.

3. Ideal interpolation. If P is a linear projector of finite rank on the linear space F with algebraic dual F', then we can think of P as providing a *linear interpolation scheme* on F: For each $g \in F$, $f = Pg$ is the unique element of $\operatorname{ran} P := P(F)$ for which

$$\lambda f = \lambda g, \quad \lambda \in \operatorname{ran} P' = \{\lambda \in F' : \lambda P = \lambda\}$$

(with $P' : F' \to F' : \lambda \mapsto \lambda P$ the dual of P). In other words, given that $\ker P = \operatorname{ran}(\mathrm{id} - P)$, we have

$$\operatorname{ran} P' = (\ker P)^{\perp} := \{\lambda \in F' : \ker P \subset \ker \lambda\}.$$

In this way, $\operatorname{ran} P'$ provides the *interpolation conditions* matched by P. Not surprisingly, there are exactly as many independent conditions as there are degrees of freedom, i.e.,

$$\dim \operatorname{ran} P = \dim \operatorname{ran} P'.$$

Now we take

$$F = \Pi,$$

the ring of polynomials in d (complex) variables. In [2], Birkhoff defined *ideal interpolation* as any linear projector P on Π whose nullspace or kernel is an ideal. In the interest of brevity, and without passing judgement, we will call such a projector **ideal**. However, Birkhoff seemed not to have been aware of the fact that such projectors had already been looked at carefully before that, by Möller in [12], who called them "Hermite interpolation", for the following reason.

As is well known (and, in this formulation, probably due to Gröbner; see [8, p. 176]), a nonempty subset I of Π is an ideal of finite codimension if and only if

$$I = \bigcap_{v \in \mathcal{V}} \ker(\epsilon_v Q_v(D))$$

for some finite subset $\mathcal{V}$ of $\mathbb{C}^d$ (necessarily the ideal's variety) and some nontrivial D-invariant finite-dimensional polynomial subspaces Q_v, necessarily given by

$$Q_v := \{q \in \Pi : ((D^{\alpha} q)(D)p)(v) = 0, \alpha \in \mathbb{Z}_+^d, p \in I\}.$$

In other words, as Möller rightly stresses, ideal interpolation is characterized by the fact that its interpolation conditions involve values and, possibly, also derivatives at certain sites, subject only to the condition that if the linear functional $\epsilon_v q(D)$ is matched, then so are all "lower" derivatives, i.e., every $\epsilon_v (D^{\alpha} q)(D)$ for $\alpha \in \mathbb{Z}_+^d$.

Since an ideal projector is, in a sense, aware of the multiplicative structure of Π, we would expect insights from considering its interaction with multiplication. The following lemma gives this interaction a handy formulation.

LEMMA 3.1. *A linear projector P on Π is ideal if and only if*

$$P(pq) = P(pPq), \quad p, q \in \Pi. \tag{3.1}$$

Proof. The condition (3.1) is equivalent to having

$$P(\Pi(\mathrm{id} - P)(\Pi)) = \{0\},$$

and, since P is a linear projector, and hence $(\mathrm{id}-P)(\Pi)=\ker P$, this is equivalent to

$$\Pi \ker P \subset \ker P,$$

and hence, given that $\ker P$ is a linear subspace, to $\ker P$ being an ideal. □

It is standard in algebraic geometry (see, e.g., [7, pp. 51ff.]) to consider, on the quotient ring

$$\Pi/\mathcal{I} := \{f+\mathcal{I} : f \in \Pi\}$$

of the polynomials over the ideal $\mathcal{I}$ and for an arbitrary polynomial p, the map

$$\Pi/\mathcal{I} \to \Pi/\mathcal{I} : f+\mathcal{I} \mapsto pf+\mathcal{I}.$$

In our context, it is more convenient to consider, equivalently, the map

$$M_p : \operatorname{ran} P \to \operatorname{ran} P : f \mapsto P(pf). \tag{3.2}$$

Evidently,

$$M_p \in L(\operatorname{ran} P);$$

i.e., M_p is a linear map on $\operatorname{ran} P$. Further, (3.1) implies that, for arbitrary $p, q \in \Pi$ and $f \in \operatorname{ran} P$,

$$M_q M_p f - M_{qp} f = P(qP(pf)) - P(qpf) = 0.$$

It follows that the map

$$m : \Pi \to L(\operatorname{ran} P) : p \mapsto M_p \tag{3.3}$$

is a ring homomorphism onto the commutative algebra generated by the specific linear maps

$$M_j : \operatorname{ran} P \to \operatorname{ran} P : f \mapsto P(()_j f), \quad j = 0{:}d,$$

in terms of which

$$M_p = p(M) := \sum_\alpha \widehat{p}(\alpha)\, M^\alpha, \quad p \in \Pi,$$

with

$$M^\alpha := \prod_j (M_j)^{\alpha(j)} = M_{()^\alpha}$$

independent of the order in which this product is formed from its factors.

It follows, directly from (3.1), that

$$p(M)P()_0 = P(p\,P()_0) = Pp, \quad p \in \Pi. \tag{3.4}$$

Such a formula plays a major role in Mourrain's intriguing paper [14], though it is proved there, consistent with that paper's setting, only for P whose range, $B := \operatorname{ran} P$, is **connected to** 1, meaning that each $b \in B$ can be written in the form $\sum_{j=0}^d ()_j b_j$ with each b_j in $B \cap \Pi_{<\deg b}$; hence, in particular, $()_0 \in B$, and (3.4) simplifies to $p(M)()_0 = Pp$.

(3.4) implies that $\ker m \subset \ker P$, while, if $p \in \ker P$, then $p(M)f = P(pf) = P(fPp) = P0 = 0$ for all $f \in \operatorname{dom} p(M) = \operatorname{ran} P$, i.e., $p(M) = 0$. Thus, altogether,

$$\ker m = \ker P. \tag{3.5}$$

4. A general Opitz formula. If now

$$V : \mathbb{F}^n \to \operatorname{ran} P : a \mapsto \sum_j v_j a(j) =: [v_1, \ldots, v_n]a$$

is any basis for $\operatorname{ran} P$, i.e., $V = [v_1, \ldots, v_n]$ is an invertible linear map, then the matrix representation for $M_p = p(M)$ with respect to this basis is

$$\widehat{M_p} = V^{-1} M_p V = p(\widehat{M}), \tag{4.1}$$

with

$$\widehat{M_j} = V^{-1} M_j V, \quad j = 1{:}d.$$

Consequently,

$$P(pv_j) = p(M)v_j = Vp(\widehat{M})(:, j), \quad p \in \Pi. \tag{4.2}$$

In particular,

$$Pp = p(M)P()_0 = Vp(\widehat{M})a_0, \quad p \in \Pi,$$

with $a_0 := V^{-1}P()_0$ the coordinates of $P()_0$ with respect to V.

(4.1), (4.2) is the promised generalization of the Opitz formula.

To make the connection with (2.1), take, in particular, $d = 1$, and let $P = P_n$ be the linear projector of interpolation from polynomials of degree $< n$ to data at the distinct sites $x_1, \ldots, x_n$. Choosing, specifically, for V the Newton basis

$$v_j := \prod_{j<k\le n} (\cdot - x_k), \quad j = 1{:}n,$$

we compute the jth column of $\widehat{M} := \widehat{M_1}$ as the coordinates, with respect to V, of

$$M_1 v_j = P_n(()_1 v_j) = P_n(x_j v_j + (\cdot - x_j)v_j) = x_j v_j + P_n v_{j-1} = x_j v_j + \begin{cases} v_{j-1}, & j > 1; \\ 0, & j = 1, \end{cases}$$

and hence

$$\widehat{M} = \begin{bmatrix} x_1 & 1 & & & \\ & x_2 & 1 & & \\ & & x_3 & \ddots & \\ & & & \ddots & 1 \\ & & & & x_n \end{bmatrix} = A_X.$$

Consider now $p(M)v_j = P_n(pv_j)$. Certainly, $(P_j p)v_j$ is in $\operatorname{ran} P_n$ and matches pv_j at all the x_i, and hence must equal $P_n(pv_j)$. Therefore,

$$p(M)v_j = \left(\sum_{k=1}^{j} \prod_{k<h\le j} (\cdot - x_h) \ \Delta(x_k, \ldots, x_j)p \right) v_j = \sum_{k=1}^{j} v_k \Delta(x_k, \ldots, x_j)p.$$

Consequently,

$$p(\widehat{M})(k, j) = \Delta(x_k, \ldots, x_j)p, \qquad k, j = 1{:}n.$$

Since Opitz [15] bases his derivations on the eigenstructure of the matrix A_X, it seems appropriate to point out that it is standard in algebraic geometry (see, e.g., [7, pp. 54ff.]) to consider the eigenstructure of the linear maps M_p (defined in (3.2)). To be sure, it is their dual, more precisely the matrix $\mathcal{M}_p$, called a **multiplication table** and defined implicitly by

$$\langle ()^\alpha p\rangle =: \sum_{\beta\in\mathbb{I}} \mathcal{M}_p(\alpha,\beta)\langle ()^\beta\rangle, \qquad \alpha\in\mathbb{I}$$

(with $\langle f\rangle := f+\mathcal{I}$ and $\mathbb{I}$ the set of multidegrees that do not occur among the multidegrees of elements of the ideal), whose eigenstructure is given, by Stetter and his collaborators, a major role in the solving of polynomial systems; see, e.g., [1], [13]. But I find it more convenient to deal with the linear maps M_p.

The bare facts are these: For each v in the variety $\mathcal{V} := \mathcal{V}(\ker P)$ of the ideal $\ker P$, $\epsilon_v \in \operatorname{ran} P'$, and hence, for every $f\in B := \operatorname{ran} P$,

$$\epsilon_v M_p f = \epsilon_v P(pf) = \epsilon_v(pf) = p(v)\epsilon_v f,$$

and this shows ϵ_v (or, more precisely, $\epsilon_v|_B$) to be a left eigenvector of M_p, with corresponding eigenvalue $p(v)$. Hence, if we are dealing with Lagrange interpolation (as is the case in [15] at the outset), i.e., if $(\epsilon_v : v\in\mathcal{V})$ spans $\operatorname{ran} P'$, then M_p is diagonalizable, and $\{p(v) : v\in\mathcal{V}\}$ is its spectrum. In that case, a right eigenbasis for M_p is the basis $(\ell_v : v\in\mathcal{V})$ of $\operatorname{ran} P$ dual to $(\epsilon_v : v\in\mathcal{V})$, i.e., $\ell_v(w) = \delta_{vw}$, the Lagrange basis. Further, $\{p(v) : v\in\mathcal{V}\}$ is also the spectrum of M_p in the general case, with each $q\in Q_v$ that is not in $\sum_{j=1}^d D_j Q_v$ giving rise to a (right) eigenvector of M_p for the eigenvalue $p(v)$.

5. An example: The truncated Taylor series. As a first (and trivial) d-variate example with $d>1$, consider $P = T_k$, the linear map on Π that associates with $p\in\Pi$ its Taylor expansion

$$T_k p := \sum_{|\alpha|<k} ()^\alpha D^\alpha p(0)/\alpha!$$

of order k. Evidently,

$$\operatorname{ran} T_k' = \epsilon_0 \Pi_{<k}(D);$$

thus

$$\ker T_k = \operatorname{ideal}(()^\alpha : |\alpha| = k).$$

In particular, with

$$V_{<k} := [()^\alpha : |\alpha| < k]$$

the power basis for $\Pi_{<k} = \operatorname{ran} T_k$, we find $()_j()^\alpha \in \operatorname{ran} T_k$ if and only if $|\alpha| < k-1$, while, for $|\alpha| = k-1$, $P(()_j()^\alpha) = 0$. Hence, with $\iota_j := (\delta_{ij} : i = 1{:}d)$,

$$\widehat{M_j}(\alpha,\beta) = \delta_{\beta+\iota_j-\alpha}, \qquad |\alpha|, |\beta| < k,$$

a strictly lower triangular matrix in any total ordering of $\mathbb{Z}_+^d$ that respects "degree", i.e., for which $|\alpha| < |\beta| \Longrightarrow \alpha < \beta$. It reflects the evident fact that the action of $\widehat{M_j}$ is to shift the coefficient function

$$\widehat{p} : \alpha \to \widehat{p}(\alpha) = D^\alpha p(0)/\alpha!$$

by ι_j, i.e.,

$$\widehat{M_j p} = \widehat{p}(\cdot - \iota_j),$$

dropping off those terms that are, thereby, pushed outside the relevant index set, $\{\alpha : |\alpha| < k\}$.

Correspondingly (or directly by (4.2)), the αth column of $p(\widehat{M})$ is obtained from $\widehat{p}$ by a shift of $\widehat{p}$ by α, again dropping off those terms that are, thereby, pushed outside $\{\alpha : |\alpha| < k\}$, i.e.,

$$p(\widehat{M})(:, \alpha) = \widehat{p}(\cdot - \alpha).$$

In particular, for any $p, q \in \Pi$,

$$\widehat{(pq)}(\alpha) = (pq)(\widehat{M})(\alpha, 0) = p(\widehat{M})q(\widehat{M})(\alpha, 0) = \sum_{\beta \le \alpha} \widehat{p}(\alpha - \beta)\widehat{q}(\beta),$$

the familiar Leibniz formula for the derivative of a product.

6. An example: Chung–Yao interpolation. In [6], Chung and Yao introduced the eponymous multivariate polynomial interpolation scheme. This scheme provides interpolation from Π_k to data at the sites

$$\Theta_{\mathbb{H}} := \{\theta_H : H \in \tbinom{\mathbb{H}}{d}\},$$

with $\mathbb{H}$ a set of $d + k$ hyperplanes in $\mathbb{R}^d$ in general position and θ_H the unique point common to the d hyperplanes in such an $H \in \tbinom{\mathbb{H}}{d}$. Chung and Yao [6] show that such interpolation is possible, and uniquely so, by exhibiting the interpolant $P_{\mathbb{H}}g$ to g in Lagrange form.

[3] (see [4] for details) provides the following Newton form for $P_{\mathbb{H}}g$:

$$P_{\mathbb{H}}g = \sum_{j=0}^{k} \sum_{K \in \tbinom{\mathbb{H}_{j-1}}{d-1}} p_{j-1,K}\, [\Theta_{\mathbb{H}_j,K} \mid n_K, \ldots, n_K]g, \tag{6.1}$$

with the various terms occurring here defined as follows:

$$\mathbb{H}_{-1} \subset \cdots \subset \mathbb{H}_k := \mathbb{H}$$

is any increasing sequence of subsets of $\mathbb{H}$ with $\#\mathbb{H}_j = d + j$, all j. Further,

$$p_{j,K} := \prod_{h \in \mathbb{H}_j \setminus K} \frac{h}{h_\uparrow(n_K)},$$

with h denoting a hyperplane as well as a particular linear polynomial whose zero set coincides with that hyperplane, and $h_\uparrow$ its **leading term**, i.e., its linear homogeneous part. Also,

$$\Theta_{\mathbb{K},K} := \Theta_{\mathbb{K}} \cap l_K,$$

with

$$l_K := \bigcap_{h \in K} h$$

the straight line common to the $d-1$ hyperplanes in K, while

$$n_K$$

is an arbitrary nontrivial vector parallel to that line. Last, but certainly not least,

$$[X \mid \Xi]g := \int_{[X]} D_\Xi g$$

is the multivariate divided difference (notation) introduced in [3]. In this formula, $X = (x_0, \ldots, x_n)$ and $\Xi = (\xi_1, \ldots, \xi_n)$ are arbitrary sequences in $\mathbb{R}^d$, the first one having one more entry than the second, $D_\Xi := D_{\xi_1} \cdots D_{\xi_n}$ is the composition of directional derivatives $D_\xi := \sum_j \xi(j) D_j$, and

$$f \mapsto \int_{[x_0,\ldots,x_n]} f := \int_0^1 \int_0^{s_1} \cdots \int_0^{s_{n-1}} f(x_0 + s_1 \nabla x_1 + \cdots + s_n \nabla x_n)\, \mathrm{d}s_n \cdots \mathrm{d}s_1 \tag{6.2}$$

(with $\nabla x_j := x_j - x_{j-1}$) is termed, by Micchelli in [11], the **divided difference functional on** $\mathbb{R}^d$ and is familiar from the Genocchi–Hermite formula for the univariate divided difference. $[X \mid \Xi]$ is symmetric in the "sites" $x \in X$, is linear and symmetric in the "directions" $\xi \in \Xi$, and satisfies the recurrence

$$[X \mid \Xi][X', \cdot \mid \Xi'] = [X, X' \mid \Xi, \Xi'].$$

Let now

$$V := [p_{j,K} : (j,K) \in \mathbb{I}], \qquad \text{with } \mathbb{I} := \{(j,K) : K \in \tbinom{\mathbb{H}_j}{d-1}, j = -1{:}(k-1)\},$$

be the corresponding "Newton" basis for $\operatorname{ran} P = \Pi_k$. For $j = 0{:}k$, let h_j be the sole element of $\mathbb{H}_j \backslash \mathbb{H}_{j-1}$, pick $K \in \tbinom{\mathbb{H}_{j-1}}{d-1}$, and let $H := K \cup h_j$. Then

$$(x - \theta_H) = \sum_{h \in H} n_{H \backslash h} \frac{h(x)}{h_\uparrow(n_{H\backslash h})}.$$

This implies that

$$\begin{aligned} x p_{j-1,K}(x) &= (\theta_H \ + \ (x - \theta_H)) p_{j-1,K}(x) \\ &= \theta_H p_{j-1,K}(x) + \sum_{h \in H} n_{H\backslash h} \left(\prod_{h' \in \mathbb{H}_j \backslash H} \frac{h'_\uparrow(n_{H\backslash h})}{h'_\uparrow(n_K)} \right) p_{j,H\backslash h}(x). \end{aligned}$$

Notice that each of the $p_{j,H\backslash h}$ in the sum over H vanishes on $\Theta_{\mathbb{H}_j}$. In particular, for $j = k$, the sum over H vanishes for every $x \in \Theta_{\mathbb{H}}$. It follows that, for $i = 1{:}d$, the matrix representation $\widehat{M_i}$ for $M_i : f \mapsto P(()_i f)$ with respect to the "Newton" basis V is "lower triangular" and quite sparse, with the column corresponding to $p_{j-1,K}$ having nonzero entries only on the diagonal, where it has the value $\theta_{h_j \cup K}(i)$, and at the entries, if any, corresponding to $p_{j,h_j \cup K\backslash h}$ for $h \in h_j \cup K$.

Now, what about $f(\widehat{M})$ for arbitrary $f \in \Pi$? The polynomial $f p_{j-1,K}$ vanishes on $\Theta_{\mathbb{H}_{j-1}}$, and hence depends only on f restricted to $\Theta_{\mathbb{H}} \backslash \Theta_{\mathbb{H}_{j-1}}$. However, this dependence is hardly simple. Formally, we have

$$f(\widehat{M})((j,K),(j',K')) = [\Theta_{\mathbb{H}_{j+1},K} \mid n_K, \ldots, n_K](f p_{j',K'}), \quad (j,K),(j',K') \in \mathbb{I}.$$

The fact that $f(\widehat{M})$ is lower triangular, in any ordering of the index set $\mathbb{I}$ that refines the natural partial ordering provided by the first components, is evident.

With this, from the fact that $(fg)(\widehat{M}) = f(\widehat{M})g(\widehat{M})$, we get the following "Leibniz formula":

$$\begin{aligned}&[\Theta_{\mathbb{H}_j,K} \mid n_K,\ldots,n_K](fg)\\ &\quad= \sum_{(j',K')\in\mathbb{I};j'<j} [\Theta_{\mathbb{H}_j,K} \mid n_K,\ldots,n_K](fp_{j',K'})\,[\Theta_{\mathbb{H}_{j'+1},K'} \mid n_{K'},\ldots,n_{K'}]g.\end{aligned} \tag{6.3}$$

Note that the second factor depends only on g on the sites in $\Theta_{\mathbb{H}_{j'+1}}$, while the first factor depends only on f on the sites in $\Theta_{\mathbb{H}_j}\backslash\Theta_{\mathbb{H}_{j'}}$. In particular, the first factor is trivially zero when $j' \geq j$, and hence the sum's restriction to $j' < j$.

Note also, by way of a check, that, for $d = 1$, $\mathbb{H}$ consists of pairwise distinct points, with $\mathbb{H}_j$ containing $j+1$ points, $h_0,\ldots,h_j$, say. Further, $K = \emptyset$ is the sole element of $\binom{\mathbb{H}_j}{d-1}$, and $l_\emptyset = \mathbb{R}$, and hence we may choose ι_1 for $n_\emptyset$ and, with that,

$$[\Theta_{\mathbb{H}_j,K} \mid n_K,\ldots,n_K] = \Delta(h_0,\ldots,h_j),$$

by the Genocchi–Hermite formula, while, as observed earlier,

$$\Delta(h_0,\ldots,h_j)\left(\prod_{i<j'}(\cdot - h_i)f\right) = \Delta(h_{j'},\ldots,h_j)f.$$

This verifies that, indeed, (6.3) reduces to (1.1) when $d = 1$.

7. Other divided differences. Let T be an arbitrary finite subset of $\mathbb{C}^d$, and assume that the polynomial subspace B is correct for it in the sense that

$$\Lambda_{\mathrm{T}}^{\mathrm{t}} : B \to \mathbb{C}^{\mathrm{T}} : b \mapsto b|_{\mathrm{T}}$$

is one-to-one and onto. Then, with

$$W : \mathbb{C}^W \to B : a \mapsto \sum_{w\in W} a(w)w$$

an arbitrary basis for B (using W to denote both the basis and the associated basis map), the Gram matrix

$$\Lambda_{\mathrm{T}}^{\mathrm{t}} W = (w(\tau) : \tau \in \mathrm{T}, w \in W)$$

is invertible; hence, for any particular ordering of the basis W, there is some ordering of T so that

$$\Lambda_{\mathrm{T}}^{\mathrm{t}} W = LU,$$

with L lower triangular and U unit upper triangular (in the chosen orderings of T and W). Then one is free to call

$$\lambda(\tau_1,\ldots,\tau_i) := \sum_k L^{-1}(i,k)\epsilon_{\tau_k} = \sum_{k\leq j} L^{-1}(i,k)\epsilon_{\tau_k}$$

the "divided difference" at the sequence $(\tau_1,\dots,\tau_i)$, and to call, correspondingly, the polynomials

$$v_j := \sum_k w_k U^{-1}(k,j) = \sum_{k\le j} w_k U^{-1}(k,j)$$

"Newton polynomials", and to call

$$\sum_j v_j\; \lambda(\tau_1,\dots,\tau_j)f$$

the "Newton form" of the interpolant from B to f at T. Assuming that B contains the constant function and that, in fact, $v_1 = ()_0$, it then follows that

$$\lambda(\tau_1,\dots,\tau_j)(fg) = \sum_{k=1}^{j} \lambda(\tau_1,\dots,\tau_j)(fv_k)\, \lambda(\tau_1,\dots,\tau_k)g,$$

with $\lambda(\tau_1,\dots,\tau_j)(fv_k)$ depending only on f at $\tau_k,\dots,\tau_j$. The role reversal of f and g here as compared to (1.1) is due to the fact that the "Newton" basis here is ordered differently than there.

It is in this manner, or, perhaps, in a more relaxed block-triangular way, that one could provide some kind of Leibniz formula and even an Opitz formula in the context of more general schemes of multivariate polynomial interpolation, e.g., the least interpolant of [5] or the Sauer–Xu formulation [19].

The divided difference introduced by Rabut in [17] does not quite fit this pattern. While Rabut does define divided differences as the coefficients of the interpolating polynomial, he sticks to the power basis

$$V_k := [()^\alpha : |\alpha| \le k]$$

rather than some kind of multivariate Newton basis. Precisely, with T some pointset in $\mathbb{R}^d$ correct for interpolation from Π_k, and hence

$$P := V_k(\Lambda_\mathrm{T}^\mathrm{t} V_k)^{-1}\Lambda_\mathrm{T}^\mathrm{t}$$

well-defined, he denotes the (T,α)**-divided difference of** f by

$$f[\mathrm{T}]^\alpha$$

and defines it implicitly by

$$Pf =: \sum_\alpha ()^\alpha f[\mathrm{T}]^\alpha.$$

With this definition, it follows from (4.2) that

$$(p[\mathrm{T}]^\alpha : |\alpha| \le k) = p(\widehat{M})(:,0), \quad p \in \Pi,$$

and hence that

$$(pq)[\mathrm{T}]^\alpha = \sum_\beta (p()^\beta)[\mathrm{T}]^\alpha\, q[\mathrm{T}]^\beta = \sum_{\beta\le\alpha} (p()^\beta)[\mathrm{T}]^\alpha\, q[\mathrm{T}]^\beta.$$

However, since $f[\mathrm{T}]^\alpha$ depends on f on all of T, the first factor in each summand still depends, offhand, on p on all of T.

In Rabut's setting, the matrix representation $\widehat{M_j}$ of

$$M_j : \Pi_k \to \Pi_k : p \mapsto P(()_j p)$$

is, in principle, not that hard to work out. For $|\alpha| \le k$, we have $()_j()^\alpha \in \Pi_k$ if and only if $|\alpha| < k$. Therefore

$$\widehat{M_j}(\alpha,\beta) = \begin{cases} \delta_{\beta+\iota_j-\alpha}, & |\beta| < k; \\ ()^{\beta+\iota_j}[\mathrm{T}]^\alpha, & |\beta| = k. \end{cases}$$

However, this still leaves the particular details of the specific divided differences $()^{\beta+\iota_j}[\mathrm{T}]^\alpha$ for $|\beta| = k$ to be supplied. At this point, I do not know whether it would be worthwhile to make that effort.

8. Extensions. In contrast to the standard literature on polynomial interpolation and divided differences, I have restricted here attention to interpolation to polynomials. However, since a polynomial interpolant depends only on the values at the interpolation sites of the function being interpolated, interpolation extends immediately to any function having values at least at the interpolation sites, and this leads to a natural extension, to such functions, of whatever divided difference notion or polynomial interpolation scheme is used.

In the univariate setting, if the interpolation involves "repeated" sites, i.e., matching of certain "consecutive" derivatives, then, correspondingly, the interpolation scheme and the divided differences extend to functions suitably differentiable at the interpolation sites. The same holds for multivariate ideal interpolation, except that, at present, it is not known whether every such Hermite interpolation scheme can be viewed as the limit of suitable Lagrange interpolation schemes, i.e., whether in this sense multivariate Hermite interpolation can be viewed as interpolation involving "repeated" sites.

Acknowledgements. I am grateful to H. Michael Möller for various discussions concerning the algebraic geometry literature and, in particular, for pointing me toward Hans J. Stetter, and to the latter for pointing out to me B. Mourrain's paper. I also thank the two referees for their most careful reading of the manuscript which further strengthened my belief that it is impossible to get anything right.

REFERENCES

[1] W. Auzinger and H. Stetter, *An elimination algorithm for the computation of all zeros of a system of multivariate polynomial equations*, in Numerical Mathematics, Singapore 1988, Internat. Schriftenreihe Numer. Math. 86, Ravi P. Agarwal, Y. M. Chow, and S. J. Wilson, eds., Birkhäuser, Basel, 1988, pp. 11–30.

[2] G. Birkhoff, *The algebra of multivariate interpolation*, in Constructive Approaches to Mathematical Models, C. V. Coffman and G. J. Fix, eds., Academic Press, New York, 1979, pp. 345–363.

[3] C. de Boor, *A multivariate divided difference*, in Approximation Theory VIII, Vol. 1: Approximation and Interpolation, Charles K. Chui and Larry L. Schumaker, eds., World Scientific, Singapore, 1995, pp. 87–96.

[4] C. de Boor, *The error in polynomial tensor-product, and in Chung-Yao, interpolation*, in Surface Fitting and Multiresolution Methods, A. LeMéhauté, C. Rabut, and L. L. Schumaker, eds., Vanderbilt University Press, Nashville, TN, 1997, pp. 35–50.

[5] C. de Boor and A. Ron, *On multivariate polynomial interpolation*, Constr. Approx., 6 (1990), pp. 287–302.

[6] K. C. Chung and T. H. Yao, *On lattices admitting unique Lagrange interpolations*, SIAM J. Numer. Anal., 14 (1977), pp. 735–743.

[7] D. Cox, J. Little, and D. O'Shea, *Using Algebraic Geometry*, Grad. Texts in Math. 185, Springer-Verlag, New York, 1998.

[8] W. Gröbner, *Algebraische Geometrie* II, Bibliographisches Institut Mannheim, B.I.-Hochschultaschenbuch, Germany, 1970.

[9] M. E. Jacobsthal, *Mittelwertbildung und Reihentransformation*, Math. Z., 6 (1920), pp. 100–117.

[10] W. Kahan and R. Fateman, *Symbolic Computation of Divided Differences*, unpublished report, 1985, http://www.cs.berkeley.edu/∼fateman/papers/divdiff.pdf.

[11] C. A. Micchelli, *On a numerically efficient method for computing multivariate B-splines*, in Multivariate Approximation Theory, W. Schempp and K. Zeller, eds., Birkhäuser, Basel, 1979, pp. 211–248.

[12] H. M. Möller, *Mehrdimensionale Hermite-Interpolation und numerische Integration*, Math. Z., 148 (1976), pp. 107–118.

[13] H. M. Möller and H. J. Stetter, *Multivariate polynomial equations with multiple zeros solved by matrix eigenproblems*, Numer. Math., 70 (1995), pp. 311–329.

[14] B. Mourrain, *A new criterion for normal form algorithms*, in Applied Algebra, Algebraic Algorithms and Error-Correcting Codes, Lecture Notes in Comput. Sci. 1719, M. Fossorier, H. Imai, S. Lin, and A. Pol, eds., Springer-Verlag, Heidelberg, 1999, pp. 430–443.

[15] G. Opitz, *Steigungsmatrizen*, Z. Angew. Math. Mech., 44 (1964), pp. T52–T54.

[16] T. Popoviciu, *Sur quelques propriétés des fonctions d'une ou de deux variables réelles*, Institutul de Arte Grafice "Ardealul," Cluj, Romania, 1933.

[17] C. Rabut, *Multivariate divided differences with simple knots*, SIAM J. Numer. Anal., 38 (2000), pp. 1294–1311.

[18] T. W. Reps and L. B. Rall, *Computational divided differencing and divided-difference arithmetics*, Higher-Order and Symbolic Computations, to appear.

[19] T. Sauer and Y. Xu, *On multivariate Lagrange interpolation*, Math. Comp., 64 (1995), pp. 1147–1170.

[20] J. F. Steffensen, *Note on divided differences*, Danske Vid. Selsk. Math.-Fys. Medd., 17 (1939), pp. 1–12.

SIAM J. Numer. Anal.
Vol. 41, No. 3, pp. 869–892

THE SEMIDISCRETE FILTERED BACKPROJECTION ALGORITHM IS OPTIMAL FOR TOMOGRAPHIC INVERSION*

ANDREAS RIEDER† AND ADEL FARIDANI‡

Abstract. The filtered backprojection algorithm is probably the most often used reconstruction algorithm in two-dimensional computerized tomography. For a semidiscrete version in the parallel scanning geometry we prove optimal L^2-convergence rates for density distributions in Sobolev spaces. Additionally we show L^2-convergence without rates when the density distribution is only in L^2. The key to success is a new representation of the filtered backprojection which enables us to apply techniques from approximation theory. Our analysis provides further a modification of the Shepp–Logan reconstruction filter with an improved convergence behavior. Numerical experiments in the fully discrete setting reproduce the theoretical predictions.

Key words. Radon transform, tomography, filtered backprojection algorithm, reconstruction filter

AMS subject classification. 65R20

PII. S0036142902405643

1. Filters in tomography.

Tomographic reconstruction means finding a density distribution f from all its line integrals $g = \mathbf{R}f$. Here, $\mathbf{R}$ denotes the *Radon transform*,

$$\mathbf{R}f(s,\vartheta) := \int_{L(s,\vartheta)\cap\Omega} f(x)\,\mathrm{d}\sigma(x),$$

mapping a function to its integrals over the lines $L(s,\vartheta) = \{\tau\,\omega^{\perp}(\vartheta) + s\,\omega(\vartheta) \mid \tau \in \mathbb{R}\}$, where $s \in \mathbb{R}$, $\omega(\vartheta) = (\cos\vartheta, \sin\vartheta)^t$, and $\omega^{\perp}(\vartheta) = (-\sin\vartheta, \cos\vartheta)^t$ for $\vartheta \in]0,\pi[$. This parameterization of lines gives rise to the *parallel scanning geometry.* The Radon transform $\mathbf{R}$ maps $L^2(\Omega)$ boundedly to $L^2(Z)$, where Ω is the unit ball in $\mathbb{R}^2$ centered about the origin and Z is the rectangle $Z =]-1,1[\times]0,\pi[$.

Analytically, tomographic reconstruction is represented by the inversion formula

$$f = (2\pi)^{-1}\,\mathbf{R}^*\,\Lambda\,g, \tag{1.1}$$

where the *backprojection operator* $\mathbf{R}^* : L^2(Z) \to L^2(\Omega)$ is the adjoint to $\mathbf{R}$,

$$\mathbf{R}^*\Phi(x) := \int_0^{\pi} \Phi(x^t\,\omega(\vartheta),\vartheta)\,\mathrm{d}\vartheta.$$

Formally, Λ is the square root of the one-dimensional Laplacian $-\Delta$: $\Lambda = (-\Delta)^{1/2}$. In (1.1), Λ acts on the variable s of g. For a proof of (1.1) see, e.g., Natterer [14].

Due to the compactness of $\mathbf{R}$ the reconstruction of f from noisy Radon data g by (1.1) is unstable (Λ amplifies high frequencies). A stable algorithm of tomographic reconstruction is therefore based on

$$f \star e_\gamma = \mathbf{R}^*(\upsilon_\gamma \star_s g), \quad e_\gamma = \mathbf{R}^*\upsilon_\gamma, \tag{1.2}$$

*Received by the editors April 15, 2002; accepted for publication (in revised form) October 23, 2002; published electronically May 22, 2003.
http://www.siam.org/journals/sinum/41-3/40564.html

[†]Institut für Wissenschaftliches Rechnen und Mathematische Modellbildung, Universität Karlsruhe, 76128 Karlsruhe, Germany (andreas.rieder@math.uni-karlsruhe.de).

[‡]Department of Mathematics, Oregon State University, Corvallis, OR 97331 (faridani@math.orst.edu). The research of this author was supported by NSF grant DMS-9803352.

where $\star$ denotes convolution and $\star_s$ denotes convolution with respect to the variable s. In (1.2), $e_\gamma(x) = e(x/\gamma)/\gamma^2$, $\gamma > 0$, and $e = e_1$ is a *mollifier*, that is, a smooth function with normalized mean value. Thus, $f \star e_\gamma$ is a smoothed or mollified approximation to f. The function $\upsilon = \upsilon_1$ is called the *reconstruction kernel* or *reconstruction filter* which is independent of the angle ϑ for radially symmetric mollifiers (which we assume in what follows). Note that $\upsilon_\gamma(s) = \upsilon(s/\gamma)/\gamma^2$. By the inversion formula (1.1) we can compute the reconstruction kernel from a mollifier e:

$$\upsilon = \frac{1}{2\pi}\,\Lambda\mathbf{R}e. \tag{1.3}$$

The convolution $\upsilon_\gamma \star_s g$ realizes a low pass filtered version of $\Lambda g/(2\pi)$.

A straightforward discretization of (1.2) together with an interpolation step yields the *filtered backprojection algorithm* (FBA) which is the most frequently used algorithm in computerized tomography; see, e.g., Natterer [14, Chap. V]. In what follows let f be a density distribution compactly supported in Ω. If we assume to know the discrete Radon data $g_{k,j} := \mathbf{R}f(s_k, \vartheta_j)$ for $s_k = k/q$, $k = -q, \dots, q$, and $\vartheta_j = j\,\pi/p$, $j = 0, \dots, p-1$, then the FBA reconstructs f_{FB} by

$$f_{\mathrm{FB}}(x) := \mathbf{R}_p^*\mathrm{I}_h(w \star_q g)(x). \tag{1.4}$$

In the FBA, first the discrete convolution

$$(w \star_q g)_{\ell,j} := \frac{1}{q}\sum_{k\in\mathbb{Z}} w_{\ell-k}\, g_{k,j} \approx \big(\upsilon_\gamma \star_s g(\cdot, \vartheta_j)\big)(s_\ell) \tag{1.5}$$

is performed, where $\{w_k\}$ is a weight sequence associated with the chosen kernel υ_γ. In the second step, an interpolation operator I_h is applied (with respect to ℓ). Finally, the discrete backprojection operator

$$\mathbf{R}_p^*\Phi(x) := \frac{\pi}{p}\sum_{j=0}^{p-1}\Phi(x^t\,\omega(\vartheta_j), \vartheta_j) \tag{1.6}$$

is evaluated.

Except for the interpolation process, the discrete convolution (1.5) is the most delicate step in the FBA: the discrete convolution kernel $\{w_k\}$ has to be chosen carefully from the continuous kernel υ_γ. For instance, a common choice is

$$w_k = \upsilon_\gamma(s_k). \tag{1.7}$$

Here γ has to be adjusted to the *discretization step size* $h = 1/q$. The sensitivity of the reconstructed image to γ has been noticed probably for the first time by Smith in [22, p. 20]. Rules for selecting γ have been suggested by Smith and Keinert [23, Sect. VI], Natterer [14], and Rieder [16]. For local tomography, see Faridani [8] and Rieder, Dietz, and Schuster [17].

Smith [22, pp. 18–19] propagated a different way to define the w_k's. He intended the discrete convolution (1.5) to be exact for a large class of functions. Let $E_h u$ be an approximation to the function u given as the superposition of translated and scaled versions of a function B; that is,

$$E_h u(s) = \sum_{k\in\mathbb{Z}} u(s_k)\, B_h(s - s_k), \qquad \text{where } B_h(s) = B(s/h). \tag{1.8}$$

For instance, E_h could be an interpolation operator. Defining

$$w_k \;:=\; \frac{1}{h}\int \upsilon_\gamma(s)\; B_h(s_k - s)\;\mathrm{d}s \;=\; \frac{1}{h}\,\upsilon_\gamma \star B_h(s_k), \quad k\in\mathbb{Z}, \tag{1.9}$$

we have that

$$(w \star_q u)_\ell \;=\; \upsilon_\gamma \star_s E_h u(s_\ell), \quad \ell\in\mathbb{Z}.$$

Moreover, if E_h is interpolating, then

$$(w \star_q E_h u)_\ell \;=\; \upsilon_\gamma \star_s E_h u(s_\ell), \quad \ell\in\mathbb{Z};$$

that is, the discrete convolution (1.5) is exact for $E_h u$. Numerical as well as theoretical considerations (see [16, 22]) showed that the reconstructed images f_{FB} are less sensitive to changes in γ when working with (1.9) rather than working with (1.7). Indeed, we will show in the next section that the discrete filter $\{w_k\}$ from (1.9) converges for $\gamma \to 0$ and that its limit $\{w_k^\infty\}$ is again a reconstruction filter belonging to a compactly supported mollifier. This limit filter has an interesting feature: computing $\Lambda E_h u(s_\ell)/(2\pi)$ can now be realized by the discrete convolution

$$\frac{1}{2\pi}\;\Lambda E_h u(s_\ell) \;=\; (w^\infty \star_q u)_\ell.$$

The latter equation is the starting point in section 3 for a reformulation of the FBA leading to optimal L^2-convergence rates in a semidiscrete setting (Theorem 3.7) where in (1.4) the discrete backprojection operator $\mathbf{R}_p^*$ is replaced by the continuous one $\mathbf{R}^*$. We see how the reconstruction filter, the interpolation process (I_h in (1.4)), and the Sobolev regularity of the density distribution f influence the convergence rate. As a by-product of our analysis we discover a new reconstruction filter (Example 4.1) with an improved convergence behavior compared to the widely used Shepp–Logan filter [21] (sections 4 and 5). Indeed, our modified Shepp–Logan filter yields optimal convergence for Sobolev orders up to 5/2, whereas the convergence order of the original Shepp–Logan filter saturates at 2 (Example 5.1). Numerical experiments in the fully discrete setting of (1.4) agree completely with our theoretical predictions and are presented in section 6. Auxiliary but new approximation properties of (quasi-) interpolation operators, which we need for the analysis, are proved in several appendices.

2. The limit. We will now investigate the convergence in Sobolev spaces of $\upsilon_\gamma \star B$ as γ tends to zero. We define the Sobolev spaces $H^\alpha(\mathbb{R}^d)$, $\alpha\in\mathbb{R}$, to be the closure of $L^2(\mathbb{R}^d)$ with respect to the norm

$$\|f\|_\alpha^2 \;:=\; \int_{\mathbb{R}^d} \left(1+\|\xi\|^2\right)^\alpha\; |\widehat{f}(\xi)|^2\;\mathrm{d}\xi,$$

where $\widehat{f}(\xi) \;:=\; (2\pi)^{-d/2}\int_{\mathbb{R}^d} f(x)\,\mathrm{e}^{-\imath\,\xi^t x}\,\mathrm{d}x$ is the Fourier transform of a function f in $L^1(\mathbb{R}^d)\cap L^2(\mathbb{R}^d)$. The Fourier transform can be extended to L^2-functions and tempered distributions by continuity and duality, respectively. The Λ-operator,

$$\widehat{\Lambda f}(\xi) \;:=\; \|\xi\|\;\widehat{f}(\xi),$$

maps $H^\alpha(\mathbb{R}^d)$ boundedly to $H^{\alpha-1}(\mathbb{R}^d)$.

The latter mapping property of Λ together with a smoothing effect of $\mathbf{R}$ (see [14, Chap. II, Thm. 5.1]) and the Sobolev embedding theorem shows that υ from (1.3) is continuous whenever e is a radially symmetric compactly supported mollifier in $H^\alpha(\mathbb{R}^2)$, $\alpha > 1$. Furthermore, $\upsilon \in L^1(\mathbb{R})$; see [16, Lem. 3.1]. Thus, $\upsilon_\gamma \star B$ is well defined in $H^t(\mathbb{R})$ for $B \in H^t(\mathbb{R})$, $t \in \mathbb{R}$; see, e.g., Aubin [1, Prop. 9.3.2].

Lemma 2.1. *Let $e \in H^\alpha(\mathbb{R}^2)$, $\alpha > 1$, be a radially symmetric compactly supported mollifier, and let υ be the corresponding reconstruction kernel* (1.3). *Then,*

$$\lim_{\gamma\to 0} \Big\| \upsilon_\gamma - \frac{1}{2\pi}\,\Lambda\delta \Big\|_{-\beta} \;=\; 0 \quad \textit{for any } \beta > 3/2, \tag{2.1}$$

where δ denotes the Dirac generalized function. Moreover, if $B \in H^t(\mathbb{R})$, $t \in \mathbb{R}$, then

$$\lim_{\gamma\to 0} \Big\| \upsilon_\gamma \star B - \frac{1}{2\pi}\,\Lambda B \Big\|_{t-1} \;=\; 0. \tag{2.2}$$

For values of s such that ΛB is continuous near s we have

$$\lim_{\gamma\to 0} \upsilon_\gamma \star B(s) \;=\; \frac{1}{2\pi}\;\Lambda B(s).$$

Proof. We prove (2.2) which then implies (2.1) when setting $B = \delta$ and recalling that $\delta \in H^t(\mathbb{R})$ for $t < -1/2$. With

$$I(\xi,\gamma) \;=\; (1+|\xi|^2)^{t-1}\,\left|\sqrt{2\pi}\;\widehat{\upsilon_\gamma}(\xi)\;\widehat{B}(\xi) - |\xi|\;\widehat{B}(\xi)/(2\pi)\right|^2$$

we obtain that

$$\Big\| \upsilon_\gamma \star B - \frac{1}{2\pi}\,\Lambda B \Big\|_{t-1}^2 \;=\; \int_{\mathbb{R}} I(\xi,\gamma)\;\mathrm{d}\xi.$$

By the projection slice theorem (see, e.g., Natterer [14, Chap. II, Thm. 1.1]) we find (e is a radially symmetric function)

$$\widehat{\upsilon_\gamma}(\xi) \;=\; \frac{1}{2\pi}\;|\xi|\;\widehat{\mathbf{R}e_\gamma}(\xi) \;=\; \frac{1}{\sqrt{2\pi}}\;|\xi|\;\widehat{e_\gamma}(\xi,0) \;=\; \frac{1}{\sqrt{2\pi}}\;|\xi|\;\widehat{e}(\gamma\,\xi,0),$$

which yields

$$I(\xi,\gamma) \;\le\; (1+|\xi|^2)^t\;\big|\widehat{B}(\xi)\big|^2\;\big|\widehat{e}(\gamma\,\xi,0) - 1/(2\pi)\big|^2.$$

The stated convergence follows now from $\widehat{e}(0,0) \,=\, 1/(2\pi)$, the Riemann–Lebesgue lemma, and the dominated convergence theorem. □

Let us look at an example. For χ being the indicator function of the interval $[-1/2, 1/2]$ we are able to compute $\Lambda\chi$ by

$$\Lambda\chi(s) = -\frac{1}{\pi}\int_{\mathbb{R}} |s-t|^{-2}\;\chi(t)\;\mathrm{d}t, \quad |s| > 1/2; \tag{2.3}$$

see Faridani et al. [9, Form. (2.1)]. Evaluating the integral gives

$$\Lambda\chi(s) \;=\; \frac{4}{\pi}\;\frac{1}{1-4\,s^2}. \tag{2.4}$$

The above formula holds for all $s \in \mathbb{R} \setminus \{-1/2, 1/2\}$. This can be verified using the relation $\Lambda(1-\chi) = -\Lambda\chi$ and applying formula (2.1) of [9] to $1-\chi$, the indicator function of $\mathbb{R} \setminus [-1/2, 1/2]$. So we have that

$$\lim_{\gamma\to 0} \upsilon_\gamma \star \chi(s) \;=\; \frac{2}{\pi^2}\,\frac{1}{1-4\,s^2}, \quad |s| \neq 1/2.$$

In weaker Sobolev norms we can even give convergence rates. For formulating the respective result and later in the paper we use the following convenient notation: $A \lesssim B$ indicates the existence of a generic constant c such that $A \le c\,B$ holds uniformly with respect to all parameters A and B may depend on.

COROLLARY 2.2. *Let* $0 \le s \le 2$. *Under the assumptions of Lemma* 2.1 *we have that*

$$\Big\|\upsilon_\gamma \star B - \frac{1}{2\pi}\,\Lambda B\Big\|_{t-1-s} \lesssim \gamma^s\;\|B\|_t.$$

Proof. As in the proof of Lemma 2.1 we obtain that

$$\Big\|\upsilon_\gamma \star B - \frac{1}{2\pi}\,\Lambda B\Big\|^2_{t-1-s} \;\le\; \int_{\mathbb{R}} (1+|\xi|^2)^{t-s}\;\big|\widehat{B}(\xi)\big|^2\;M(\gamma\,\xi, 0)\;\mathrm{d}\xi,$$

where $M(z) := |\widehat{e}(z) - 1/(2\pi)|^2$, $z \in \mathbb{R}^2$. Since e is an even function all its first order moments vanish. Therefore, all first order derivatives of $\widehat{e}$ are zero at the origin. Thus the Taylor expansion of $\widehat{e}$ about the origin becomes

$$\widehat{e}(z) \;=\; \frac{1}{2\pi} + \sum_{\substack{\nu\in\mathbb{N}_0^2\\ \nu_1+\nu_2=2}} \frac{D^\nu \widehat{e}(\tau_z\,z)}{\nu!}\;z^\nu \quad \text{for a } \tau_z \in [0,1],$$

which yields $M(z) \lesssim \|z\|^4$. Now let $s \in [0,2]$. Then,

$$\begin{aligned}\Big\|\upsilon_\gamma \star B - \frac{1}{2\pi}\,\Lambda B\Big\|^2_{t-1-s} &\lesssim \int_{|\xi|\le 1/\gamma} \big|\widehat{B}(\xi)\big|^2\;(1+|\xi|^2)^{t-s}\;M(\gamma\,\xi,0)\;\mathrm{d}\xi\\ &\qquad + \int_{|\xi|>1/\gamma} \big|\widehat{B}(\xi)\big|^2\;(1+|\xi|^2)^{t-s}\;\mathrm{d}\xi\\ &\lesssim \gamma^4 \int_{|\xi|\le 1/\gamma} \big|\widehat{B}(\xi)\big|^2\;(1+|\xi|^2)^{t}\;|\xi|^{4-2s}\;\mathrm{d}\xi\\ &\qquad + \int_{|\xi|>1/\gamma} \big|\widehat{B}(\xi)\big|^2\;(1+|\xi|^2)^{t}\;|\xi|^{-2s}\;\mathrm{d}\xi.\end{aligned}$$

Both latter terms can be bounded by $\gamma^{2s}\;\|B\|_t^2$. □

Remark 2.3. The generalization of Corollary 2.2 to reconstruction kernels υ belonging to mollifiers with higher order vanishing moments is obvious.

3. The FBA is optimal.

We will reformulate the FBA (1.4) for the limit filters considered in the former section; see (3.2) below. This new representation of the FBA allows us to introduce a novel error analysis which shows that the FBA is optimal for tomographic inversion.

3.1. A new representation of the FBA. We start with the following simple observation.

Lemma 3.1. *Let B be in $H^t(\mathbb{R})$ for a $t \in \mathbb{R}$ such that $\Lambda B(s)$ is continuous near integer values of s. For $\psi(s) = \sum_{k\in\mathbb{Z}} c_k\, B_h(s - h\,k)$, where $\{c_k\}$ is a finite sequence and h is positive, we have*

$$\Lambda\psi(h\,\ell) \;=\; h^{-1} \sum_{k\in\mathbb{Z}} c_k \;\Lambda B(\ell - k), \quad \ell \in \mathbb{Z}.$$

Proof. The statement follows directly from the relations $\Lambda B_h(s) = \Lambda B(s/h)/h$ and $\Lambda T^a = T^a \Lambda$, where T^a is the translation operator $T^a u(s) = u(s - a)$. □

Remark 3.2. Relying on Lemma 3.1 we easily derive that

$$\frac{2}{\pi}\;\frac{1}{q + 1/2} \;=\; \sum_{k=-q}^{q} \Lambda\chi(k) \quad \text{for any } q \in \mathbb{N}_0,$$

where χ is as in (2.3). To prove the above identity we mention only that $\sum_{k=-q}^{q} \chi(\cdot - k)$ is the characteristic function of the interval $[-q - 1/2, q + 1/2]$.

Let the operator E_h be given by (1.8) with B as in Lemma 3.1. Define the discrete reconstruction kernel $\{w_k^\infty\}$ by

$$w_k^\infty \;=\; \frac{1}{2\pi\, h^2}\;\Lambda B(k) \;=\; v_h^\infty(h\,k), \tag{3.1}$$

where $v_h^\infty(s) = v^\infty(s/h)/h^2$ and $v^\infty(s) := \Lambda B(s)/(2\pi)$. Then, the discrete convolution (1.5) can be written as the Λ-operator applied *exactly* to a function approximating g from discrete values:

$$(w^\infty \star_q g)_{\ell,j} \;=\; \frac{1}{2\pi}\,\big(\Lambda E_h g(\cdot, \vartheta_j)\big)(h\,\ell), \quad \ell \in \mathbb{Z}.$$

Thus, the reconstructed image f_{FB} may be rewritten as

$$f_{\mathrm{FB}}(x) \;=\; \frac{1}{2\pi}\;\mathbf{R}_p^* \mathrm{I}_h \Lambda E_h g(x); \tag{3.2}$$

see (1.4). Please observe that the three operators E_h, Λ, and I_h act on the first variable of the data $g = \mathbf{R}f$.

Example 3.3. Let $B = \chi$ be the characteristic function of $[-1/2, 1/2]$. Then, the reconstruction kernel w^∞ used for evaluating (3.2) is

$$w_k^\infty \;=\; \frac{2}{\pi^2\, h^2}\;\frac{1}{1 - 4\,k^2},$$

which follows from (2.4) and (3.1). Here, w^∞ is the discrete Shepp–Logan reconstruction filter [21].

Remark 3.4. Let the discrete reconstruction kernel $\{w_k\}$ be given by (1.9). Due to Lemma 2.1 we obtain $\lim_{\gamma\to 0} w_k = w_k^\infty$, implying that

$$\lim_{\gamma\to 0}\,(w \star_q g)_{\ell,j} \;=\; \frac{1}{2\pi}\,\big(\Lambda E_h g(\cdot, \vartheta_j)\big)(h\,\ell), \quad \ell \in \mathbb{Z}.$$

We next ask the question, Which mollifier e^∞ belongs to the reconstruction kernel v^∞ (3.1)? By (1.3) and the projection slice theorem we find that

$$\widehat{e^\infty}(\xi) \;=\; \widehat{B}(\|\xi\|)/\sqrt{2\pi}, \quad \xi \in \mathbb{R}^2, \tag{3.3}$$

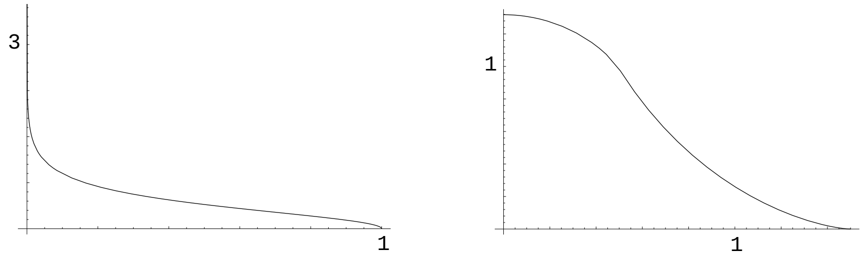


FIG. 1. *Radial part of limit mollifier* e^∞ (3.4) *where* B *is the linear (left) and the quadratic (right) B-spline, respectively.*

which yields (J_0 denoting the Bessel function of the first kind of order 0)

$$e^\infty(x) \,=\, \frac{1}{\sqrt{2\pi}} \int_0^\infty r\ \widehat{B}(r)\ \mathrm{J}_0(\|x\|\, r)\ \mathrm{d}r. \tag{3.4}$$

In view of (3.3) the mollifier e^∞ is in $L^2(\mathbb{R}^2)$ if $\lim_{r\to\infty} r\,|\widehat{B}(r)| = 0$. Further, a compact support of B implies a compact support of e^∞. More precisely, let B be *even* with $\operatorname{supp} B \subset [-R, R]$; then $\operatorname{supp} e^\infty \subset \{x \in \mathbb{R}^2 \mid \|x\| \le R\}$. The latter statement is a consequence from the Paley–Wiener theorems; see, e.g., Rudin [19, Chap. 7].

Example 3.5. Let $B = \chi$ be the characteristic function of the interval $[-1/2, 1/2]$. By formula 6.671.7 from [11] we find that

$$e^\infty(x) \,=\, \begin{cases} \frac{2}{\pi}\ \frac{1}{\sqrt{1-4\,\|x\|^2}} & : \ \ \|x\| < 1/2, \\ \qquad 0 & : \ \ \text{otherwise}, \end{cases}$$

which is the mollifier belonging to $v^\infty(s) = \frac{1}{2\pi}\Lambda\chi(s) = \frac{2}{\pi^2}(1 - 4\,s^2)^{-1}$; see (2.4). The graphs of the radial parts of e^∞ with respect to the linear and quadratic B-splines are plotted in Figure 1.

3.2. A novel error estimate. The new representation (3.2) of the FBA gives us the freedom to provide a novel error analysis based on principles from approximation theory. Indeed, we will be able to prove L^2-convergence of the FBA with optimal rates.

In contrast, the error estimates based on Fourier analysis (see Natterer [14, Chap. V] and Faridani and Ritman [10]) are of qualitative nature in terms of essentially band-limited functions. Since the main tool used is the Poisson summation formula the considered density distributions are required to be continuous functions at least ($\widehat{f} \in L^1$). Convergence has been shown before: Popov [15] established pointwise convergence restricted to a small class of functions (piecewise $\mathcal{C}^\infty$ with jumps across smooth curves). The approach of Rieder and Schuster [18] leads to L^2-convergence for $f \in H_0^\alpha(\Omega)$, $\alpha > 1/2$, however, with suboptimal rates.

In our analysis below we will not take into account the error introduced by discretizing the backprojection $\mathbf{R}^*$; that is, our model of the FBA reconstructs $\widetilde{f}_{\mathrm{FB}}$ by

$$\widetilde{f}_{\mathrm{FB}}(x) \,:=\, \frac{1}{2\pi}\ \mathbf{R}^* \mathrm{I}_h \Lambda E_h \mathbf{R} f(x); \tag{3.5}$$

compare (3.2).

Before bounding the reconstruction error of $\widetilde{f}_{\mathrm{FB}}$ we generalize both operators E_h and I_h. For $u \in H^\alpha(\mathbb{R})$, $\alpha \in \mathbb{R}$, we define

$$E_h u(s) := h^{-1} \sum_{k\in\mathbb{Z}} \big\langle u, \epsilon_h(\cdot - s_k)\big\rangle \, B_h(s - s_k), \tag{3.6}$$

where $\epsilon_h(s) = \epsilon(s/h)$ with $\epsilon \in H^{-\alpha}(\mathbb{R})$ being even and $\widehat{\epsilon}(0) = 1/\sqrt{2\pi}$. Further, $\langle\cdot,\cdot\rangle$ denotes the duality pairing in $H^\alpha(\mathbb{R}) \times H^{-\alpha}(\mathbb{R})$. For $u \in H^\alpha(\mathbb{R})$, $\alpha > 1/2$, we may choose $\epsilon = \delta$ (Dirac distribution). Thus, $h^{-1}\langle u, \epsilon_h(\cdot - s_k)\rangle = u(s_k)$, and the general form (3.6) of E_h coincides with its former definition (1.8). We extended the domain of definition of E_h to cover (generalized) functions in $H^\alpha(\mathbb{R})$ with $\alpha \le 1/2$.

The redefinition of E_h was necessary because we apply E_h to $\mathbf{R}f(\cdot,\vartheta)$ (see (3.5)), and we have only that $\mathbf{R}f(\cdot,\vartheta) \in H_0^{1/2}(-1,1)$ for $f \in L^2(\Omega)$ and almost all ϑ. Moreover, our new model allows for finite width of the rays and detector inhomogeneities in the observed semidiscrete Radon data; see Natterer [14, Chap. V.5.1]. Indeed, for ϵ being a nonnegative function compactly supported in $[-1/2, 1/2]$ with a normalized mean value we obtain

$$h^{-1} \big\langle \mathbf{R}f(\cdot,\vartheta), \epsilon_h(\cdot - s_k)\big\rangle = h^{-1} \int_{s_k - h/2}^{s_k + h/2} \mathbf{R}f(s,\vartheta)\, \epsilon_h(s - s_k)\, \mathrm{d}s. \tag{3.7}$$

Hence, ϵ can be seen as the *sensitivity profile* of the X-ray detectors.

In a very similar way we define I_h by

$$\mathrm{I}_h u(s) := h^{-1} \sum_{k\in\mathbb{Z}} \big\langle u, \eta_h(\cdot - s_k)\big\rangle \, A_h(s - s_k), \tag{3.8}$$

where η and A are like ϵ and B from (3.6), respectively.

Our modifications of E_h and I_h have no effect on the efficient computation of $\mathrm{I}_h \Lambda E_h \mathbf{R}f(\cdot,\vartheta)/(2\pi)$ by discrete convolution. A straightforward calculation reveals that

$$\frac{1}{2\pi}\,\big(\mathrm{I}_h \Lambda E_h \mathbf{R}f(\cdot,\vartheta)\big)(s) = \sum_{\ell\in\mathbb{Z}} \big(w \star_q g^\epsilon(\cdot,\vartheta)\big)_\ell \, A_h(s - s_\ell),$$

where $g^\epsilon(s,\vartheta) = \big(\mathbf{R}f(\cdot,\vartheta) \star_s \epsilon_h\big)(s)/h$ (see (3.7)), and the discrete reconstruction kernel w is given by

$$\begin{aligned} &w_r = v(r)/h^2, \quad r \in \mathbb{Z},\\ &\text{with } v(s) := \frac{1}{\pi} \int_0^\infty \sigma\, \widehat{B}(\sigma)\, \widehat{\eta}(\sigma)\, \cos(s\,\sigma)\, \mathrm{d}\sigma. \end{aligned} \tag{3.9}$$

The above integral exists as a duality pairing whenever $B \in H^t(\mathbb{R})$ and $\eta \in H^{1-t}(\mathbb{R})$.

Example 3.6. We will give the Shepp–Logan reconstruction filter a new interpretation. To this end, let $B(s) = \mathrm{sinc}(\pi\, s)$ be the interpolating function used to define E_h. In I_h let η be the characteristic function of the interval $[-1/2, 1/2]$. We obtain $\widehat{B} = \chi_{[-\pi,\pi]}/\sqrt{2\pi}$ (χ_D characteristic function of interval D) and $\widehat{\eta}(\sigma) = \mathrm{sinc}(\sigma/2)/\sqrt{2\pi}$. Hence,

$$v(s) = \frac{2}{\pi^2}\, \frac{2\,s\,\sin(\pi\,s) - 1}{4\,s^2 - 1} \quad \text{and} \quad w_k = v(k)/h^2 = \frac{2}{\pi^2\,h^2}\, \frac{1}{1 - 4\,k^2};$$

see Example 3.3 and compare formula (1.22) on page 111 in [14].

In estimating the reconstruction error below we will need that the inversion formula (1.1) holds true for functions in $L^2(\Omega)$; that is,

$$f \;=\; (2\pi)^{-1}\,\mathbf{R}^*\,\Lambda\mathbf{R}f \quad \text{for any } f\in L^2(\Omega). \tag{3.10}$$

As far as we know, the most general version of (1.1) is due to Smith, Solomon, and Wagner [24, p. 1257] requiring a compactly supported $f\in H^\alpha(\mathbb{R}^2)$ with $\alpha\ge 1/2$. To verify (3.10) we recall the following mapping property of the Radon transform:

$$\mathbf{R}: H_0^\alpha(\Omega)\to H^{(\alpha+1/2,0)} \quad \text{is bounded for any } \alpha\ge 0, \tag{3.11}$$

which is due to Louis and Natterer [13, Thm. 3.1]; see also [14, Thm. II.5.1]. Above, $H_0^\alpha(\Omega)$ is the closure of $\mathcal{C}_0^\infty(\Omega)$, the space of infinitely differentiable functions compactly supported in Ω, with respect to the norm $\|\cdot\|_\alpha$. Further, $H^{(\beta,0)}$ is the tensor product space $H^\beta(\mathbb{R})\widehat{\otimes}L^2(0,\pi)$.

Now the validity of (3.10) can be seen from the following three facts: 1. The operator $\mathbf{R}^*\,\Lambda\mathbf{R}: L^2(\Omega)\to L^2(\Omega)$ is bounded since all three mappings $\mathbf{R}: L^2(\Omega)\to H^{(1/2,0)}$, $\Lambda : H^{(1/2,0)}\to H^{(-1/2,0)}$, and $\mathbf{R}^* : H^{(-1/2,0)}\to L^2(\Omega)$ are bounded.[1] 2. Formula (3.10) applies to all $f\in\mathcal{C}_0^\infty(\Omega)$; see, e.g., Natterer [14, Thm. II.2.1]. 3. The space $\mathcal{C}_0^\infty(\Omega)$ is dense in $L^2(\Omega)$.

After these preparations we concentrate on the reconstruction error for f in $H_0^\alpha(\Omega)$, $\alpha\ge 0$. Relying on (3.10) we begin with

$$\begin{aligned}\big\|\widetilde{f}_{\mathrm{FB}}-f\big\|_{L^2(\Omega)} &= \frac{1}{2\pi}\,\big\|\mathbf{R}^*\mathrm{I}_h\Lambda E_h\mathbf{R}f-\mathbf{R}^*\Lambda\mathbf{R}f\big\|_{L^2(\Omega)}\\ &\le \big\|\big(\mathbf{R}^*\mathrm{I}_h-\mathbf{R}^*\big)\Lambda E_h\mathbf{R}f\big\|_{L^2(\Omega)}\\ &\qquad\qquad + \big\|\mathbf{R}^*\Lambda\big(E_h\mathbf{R}f-\mathbf{R}f\big)\big\|_{L^2(\Omega)}\end{aligned}$$

and proceed by estimating both norms on the right-hand side.

We saw above that $\mathbf{R}^*\Lambda$ maps $H^{(1/2,0)}$ boundedly to $L^2(\Omega)$. Hence,

$$\big\|\mathbf{R}^*\Lambda\big(E_h\mathbf{R}f-\mathbf{R}f\big)\big\|_{L^2(\Omega)} \lesssim \big\|E_h\mathbf{R}f-\mathbf{R}f\big\|_{H^{(1/2,0)}}.$$

Now we need an approximation property of E_h. Therefore, we assume there are nonnegative constants $\tau_{\max}$ and $\beta_{\min}\le\beta_{\max}$ such that

$$\big\|E_h u-u\big\|_\tau \lesssim h^{\beta-\tau}\,\|u\|_\beta \quad \text{as } h\to 0 \tag{3.12a}$$

$$\text{for } \beta_{\min}\le\beta\le\beta_{\max},\ 0\le\tau\le\beta,\ \ \tau\le\tau_{\max},\ u\in H_0^\beta(-1,1). \tag{3.12b}$$

For instance, if E_h represents piecewise linear interpolation[2], then (3.12) holds with $\beta_{\max}=2$, $\beta_{\min}>1/2$, and $\tau_{\max}<3/2$. For piecewise linear interpolation the approximation property (3.12) is a classical result when $\tau\in\{0,1\}$ and $\beta=2$; see, e.g., Strang and Fix [25, Thm. 1.3]. Also band-limited interpolation[3] yields (3.12) with $\beta_{\min}>1/2$ and any $\beta_{\max}=\tau_{\max}<\infty$. In Appendices A and B we prove (3.12) for more general interpolation-like operators E_h where $\beta_{\min}=0$.

[1] The continuity of $\mathbf{R}^* : H^{(-1/2,0)}\to L^2(\Omega)$ follows from (3.11) by duality.

[2] ϵ is the Dirac distribution and B is the linear B-spline.

[3] ϵ is the Dirac distribution and $B(x)=\mathrm{sinc}(\pi x)$.

Estimates of terms from above by powers of h (like (3.12)) are in what follows always understood asymptotically in the sense of $h \to 0$.

Assume (3.12) to hold with $\beta_{\max} \geq 1/2$ and $\tau_{\max} \geq 1/2$. If $\max\{0, \beta_{\min} - 1/2\} \leq \alpha \leq \beta_{\max} - 1/2$, then

$$\big\|\mathbf{R}^*\Lambda\big(E_h\mathbf{R}f - \mathbf{R}f\big)\big\|_{L^2(\Omega)} \lesssim h^\alpha\ \|\mathbf{R}f\|_{H^{(1/2+\alpha,0)}} \overset{(3.11)}{\lesssim} h^\alpha\ \|f\|_\alpha.$$

Now we turn to $\|(\mathbf{R}^*\mathrm{I}_h - \mathbf{R}^*)\Lambda E_h\mathbf{R}f\|_{L^2(\Omega)}$ which we estimate according to

$$\begin{aligned}\big\|\mathbf{R}^*\big(\mathrm{I}_h - I\big)\Lambda E_h\mathbf{R}f\big\|_{L^2(\Omega)} \\ \leq \|\mathbf{R}^*\|_{H^{(-1/2,0)}\to L^2(\Omega)}\ \|\mathrm{I}_h - I\|_{H^{\alpha-1/2}(\mathbb{R})\to H^{-1/2}(\mathbb{R})} \\ \times\ \|\Lambda\|_{H^{\alpha+1/2}(\mathbb{R})\to H^{\alpha-1/2}(\mathbb{R})}\ \|E_h\mathbf{R}f\|_{H^{(1/2+\alpha,0)}},\end{aligned}$$

where $I : H^{\alpha-1/2}(\mathbb{R}) \hookrightarrow H^{-1/2}(\mathbb{R})$ is the canonical inclusion. Observe that (3.12) implies the boundedness of $E_h : H_0^{1/2+\alpha}(-1,1) \to H^{1/2+\alpha}(\mathbb{R})$ uniformly in h for $0 \leq \alpha \leq \min\{\beta_{\max}, \tau_{\max}\} - 1/2$. Thus,

$$\|E_h\mathbf{R}f\|_{H^{(1/2+\alpha,0)}} \lesssim \|\mathbf{R}f\|_{H^{(1/2+\alpha,0)}} \overset{(3.11)}{\lesssim} \|f\|_\alpha.$$

For the operator I_h we require that

$$\|\mathrm{I}_h - I\|_{H^{\alpha-1/2}(\mathbb{R})\to H^{-1/2}(\mathbb{R})} \lesssim h^\alpha \quad \text{as } h \to 0 \text{ for } 0 \leq \alpha \leq \alpha_{\mathrm{I}}. \tag{3.13}$$

which yields that

$$\big\|\big(\mathbf{R}^*\mathrm{I}_h - \mathbf{R}^*\big)\Lambda E_h\mathbf{R}f\big\|_{L^2(\Omega)} \lesssim h^\alpha\ \|f\|_\alpha.$$

Thus, we have proven the following theorem.

THEOREM 3.7. *Assume* (3.12) *to hold with* $\beta_{\max} \geq 1/2$ *and* $\tau_{\max} \geq 1/2$. *Further, let there exist an* $\alpha_{\mathrm{I}} > 0$ *such that* (3.13) *holds true.*

If $\max\{0, \beta_{\min} - 1/2\} \leq \alpha \leq \min\big\{\alpha_{\mathrm{I}}, \beta_{\max} - 1/2, \tau_{\max} - 1/2\big\}$ *and* $f \in H_0^\alpha(\Omega)$, *then*

$$\Big\|f - \frac{1}{2\pi}\,\mathbf{R}^*\mathrm{I}_h\Lambda E_h\mathbf{R}f\Big\|_{L^2(\Omega)} \lesssim h^\alpha\ \|f\|_\alpha \quad \textit{as } h \to 0. \tag{3.14}$$

The best possible L^2-convergence rate for the reconstruction of $f \in H_0^\alpha(\Omega)$ from Radon data sampled at distance h is h^α as $h \to 0$; see Natterer [14, Chap. IV, Thm. 2.2]. So we just proved that the FBA with an "averaged" limit kernel (3.9) is an optimal reconstruction algorithm (at least for semidiscrete data). The range of Sobolev orders yielding optimal convergence depends on the chosen filter and the used interpolation procedure.

Theorem 3.7 looks similar to Theorem V.1.2 of Natterer [14]. The main difference is that our theorem takes the discretization of the convolution into account. On the other hand, the main result of Popov [15, Thm. 3, p. 35] investigates pointwise convergence utilizing an approach based on asymptotic expansions. It does consider the fully discrete algorithm but is applicable to a smaller class of functions (piecewise $\mathcal{C}^\infty$ with jumps across smooth curves), is stated without a detailed proof, and is not always easily applied to concrete examples.

Example 3.8. Here we provide a simple example for (3.13) which results in a convergence proof of the FBA with the Shepp–Logan filter and nearest-neighbor interpolation.

To this end let both η and A be first order B-splines; that is, $\eta = A = \chi_{[-1/2,1/2[}$. In this situation (3.13) applies with $\alpha_{\mathrm{I}} = 3/2$ as we will demonstrate now. By Theorem A.2,

$$\|\mathrm{I}_h u - u\|_\tau \lesssim h^{\beta-\tau}\, \|u\|_\beta \tag{3.15}$$

for $0 \le \tau \le \beta \le 1$ and $\tau < 1/2$. To estimate $\|\mathrm{I}_h u - u\|_{-1/2}$ we use a duality argument and the symmetry $\mathrm{I}_h = \mathrm{I}_h^*$, where I_h^* is the L^2-adjoint of I_h. We find that, for $0 \le \alpha \le 1/2$,

$$\begin{aligned}\|\mathrm{I}_h u - u\|_{-1/2} &= \sup_{v\in H^{1/2}(\mathbb{R})} \frac{\langle \mathrm{I}_h u - u, v\rangle}{\|v\|_{1/2}} = \sup_{v\in H^{1/2}(\mathbb{R})} \frac{\langle u, \mathrm{I}_h v - v\rangle}{\|v\|_{1/2}}\\ &\le \|u\|_{\alpha-1/2} \sup_{v\in H^{1/2}(\mathbb{R})} \frac{\|\mathrm{I}_h v - v\|_{1/2-\alpha}}{\|v\|_{1/2}} \overset{(3.15)}{\lesssim} h^\alpha\, \|u\|_{\alpha-1/2}.\end{aligned} \tag{3.16}$$

For $1/2 < \alpha \le 3/2$ we estimate similarly, relying on $\mathrm{I}_h^2 = \mathrm{I}_h$,

$$\begin{aligned}\|\mathrm{I}_h u - u\|_{-1/2} &= \sup_{v\in H^{1/2}(\mathbb{R})} \frac{\langle (\mathrm{I}_h - I)^2 u, v\rangle}{\|v\|_{1/2}} = \sup_{v\in H^{1/2}(\mathbb{R})} \frac{\langle \mathrm{I}_h u - u, \mathrm{I}_h v - v\rangle}{\|v\|_{1/2}}\\ &\le \|\mathrm{I}_h u - u\|_{L^2(\mathbb{R})} \sup_{v\in H^{1/2}(\mathbb{R})} \frac{\|\mathrm{I}_h v - v\|_{L^2(\mathbb{R})}}{\|v\|_{1/2}}\\ &\overset{(3.15)}{\lesssim} h^{\alpha-1/2}\, \|u\|_{\alpha-1/2}\, h^{1/2}.\end{aligned} \tag{3.17}$$

Hence, (3.13) holds for $\alpha_{\mathrm{I}} = 3/2$.

Recalling Example 3.6 we observe that the FBA with the Shepp–Logan filter and nearest-neighbor interpolation is represented by $B(s) = \operatorname{sinc}(\pi\, s)$ and $\eta = A = \chi_{[-1/2,1/2[}$ in our framework (3.5). Therefore, our results from Appendix B give that

$$\left\| \widetilde{f}_{\mathrm{FB}} - f \right\|_{L^2(\Omega)} \lesssim h^{\min\{3/2,\,\alpha\}}\, \|f\|_\alpha \quad \text{for } f \in H_0^\alpha(\Omega),\ \alpha > 0,$$

as long as ϵ is either an even, compactly supported, and normalized L^2-function (Theorem B.2) or the Dirac distribution (Theorem B.4).

In the next section we will generalize the above example, covering especially piecewise linear interpolation in I_h.

So far we have not shown L^2-convergence of the FBA when the density distribution f is only in $L^2(\Omega)$. However, we possess all the tools to do this.

COROLLARY 3.9. *Assume* (3.12) *to hold with* $\beta_{\min} \le 1/2$, $\beta_{\max} > 1/2$, *and* $\tau_{\max} > 1/2$. *Further, let there exist an* $\alpha_{\mathrm{I}} > 0$ *such that* (3.13) *holds true. Then,*

$$\lim_{h\to 0} \left\| f - \frac{1}{2\pi}\, \mathbf{R}^* \mathrm{I}_h \Lambda E_h \mathbf{R} f \right\|_{L^2(\Omega)} = 0 \quad \textit{for any } f \in L^2(\Omega). \tag{3.18}$$

Proof. We will use that $\mathbf{R}^*\mathrm{I}_h\Lambda E_h\mathbf{R} : L^2(\Omega) \to L^2(\Omega)$ is uniformly bounded in $h > 0$. This follows by setting $\alpha = 0$ in (3.14) which is allowed since $\beta_{\min} \le 1/2$. Thus, $\|\mathbf{R}^*\mathrm{I}_h\Lambda E_h\mathbf{R}\|_{L^2(\Omega)\to L^2(\Omega)} \lesssim 1$.

Fix an α with $0 < \alpha \le \min\{\alpha_{\mathrm{I}}, \beta_{\max} - 1/2, \tau_{\max} - 1/2\}$. By assumption the upper bound on α is positive. Since $H_0^\alpha(\Omega)$ is dense in $L^2(\Omega)$ there exists a family $\{f_\lambda\}_{\lambda>0} \subset H_0^\alpha(\Omega)$ which converges to f in $L^2(\Omega)$ as $\lambda \to 0$. Without loss of generality we may assume that f is *not* an element of $H_0^\beta(\Omega)$ for any $\beta > 0$ (otherwise we apply Theorem 3.7 to obtain (3.18)). Therefore, the function $\rho(\lambda) := \|f_\lambda\|_\alpha$ explodes: $\rho(\lambda) \to \infty$ as $\lambda \to 0$. Now we choose a family $\{\lambda_h\}_{h>0}$ satisfying

$$\lim_{h\to 0} \lambda_h \;=\; 0 \quad \text{as well as} \quad \lim_{h\to 0} h^\alpha\, \rho(\lambda_h) \;=\; 0.$$

We proceed with

$$\begin{aligned}\Big\| f - \frac{1}{2\pi}\,\mathbf{R}^*\mathrm{I}_h\Lambda E_h\mathbf{R}f\Big\|_{L^2(\Omega)} &\le \|f - f_{\lambda_h}\|_{L^2(\Omega)} \\ &\quad + \Big\| f_{\lambda_h} - \frac{1}{2\pi}\,\mathbf{R}^*\mathrm{I}_h\Lambda E_h\mathbf{R}f_{\lambda_h}\Big\|_{L^2(\Omega)} \\ &\quad + \|\mathbf{R}^*\mathrm{I}_h\Lambda E_h\mathbf{R}(f_{\lambda_h} - f)\|_{L^2(\Omega)} \\ &\lesssim \|f - f_{\lambda_h}\|_{L^2(\Omega)} + h^\alpha\,\rho(\lambda_h),\end{aligned}$$

where we applied Theorem 3.7 in the last step. Finally, the limit $h \to 0$ implies (3.18). □

Example 3.10. We reconsider Example 3.8 in light of Corollary 3.9. The convergence (3.18) holds true when using the Shepp–Logan filter with nearest-neighbor interpolation for I_h and band-limited quasi interpolation for E_h; that is, $\eta = A = \chi_{[-1/2,1/2[}$, $B(s) = \mathrm{sinc}(\pi\, s)$, and ϵ is an even, compactly supported, and normalized L^2-function (Theorem B.2). Please note that band-limited interpolation for E_h (ϵ is the Dirac distribution), which requires $\beta_{\min} > 1/2$, is not covered by Corollary 3.9.

4. Verifying (3.13) for interpolation-like operators $\mathbf{I}_h$ based on orthogonalized B-splines. We consider a special choice for I_h (3.8): let $\widetilde{\eta}$ and A be the B-splines of order $M \ge 1$ and $N \ge 1$, respectively. Define η by

$$\widehat{\eta}(\sigma) \;:=\; \frac{\widehat{\widetilde{\eta}}(\sigma)}{\mathbf{a}(\sigma)} \;=\; \frac{1}{\sqrt{2\pi}}\;\frac{\mathrm{sinc}^M(\sigma/2)}{\mathbf{a}(\sigma)}, \tag{4.1}$$

where

$$\mathbf{a}(\sigma) \;=\; \sum_{\ell\in\mathbb{Z}} a_\ell\; \mathrm{e}^{-\imath\,\ell\sigma} \quad \text{with } a_\ell \;=\; \int_{\mathbb{R}} \widetilde{\eta}(s)\; A(\ell - s)\;\mathrm{d}s. \tag{4.2}$$

Note that $\mathbf{a}$ is a positive even real trigonometric polynomial with $\mathbf{a}(0) = 1$; see Appendix C.1. Further, η and A are dual functions; that is,

$$\big\langle \eta(\cdot - k),\, A(\cdot)\big\rangle \;=\; \delta_{k,0}; \tag{4.3}$$

see Appendix C.2. Especially,

$$(I - \mathrm{I}_h)\,(I - \widetilde{\mathrm{I}}_h) \;=\; I - \mathrm{I}_h, \tag{4.4}$$

where $\widetilde{\mathrm{I}}_h u(s) := h^{-1}\sum_{k\in\mathbb{Z}} \langle u, \widetilde{\eta}_h(\cdot - s_k)\rangle\, A_h(s - s_k)$.

As in (3.16) we obtain

$$\|\mathrm{I}_h u - u\|_{-1/2} \;\le\; \|u\|_{\alpha-1/2} \sup_{v\in H^{1/2}(\mathbb{R})} \frac{\|\mathrm{I}_h^* v - v\|_{1/2-\alpha}}{\|v\|_{1/2}}, \quad 0 \le \alpha \le 1/2.$$

Accordingly, we have to investigate the approximation power of the L^2-adjoint operator $\mathrm{I}_h^* u(s) = h^{-1}\sum_{k\in\mathbb{Z}}\langle u, A_h(\cdot - s_k)\rangle\, \eta_h(s - s_k)$ which is done in Appendix C.3. By (C.3),

$$\|\mathrm{I}_h u - u\|_{-1/2} \;\lesssim\; h^\alpha\; \|u\|_{\alpha-1/2}, \quad 0 \le \alpha \le 1/2.$$

The range $\alpha > 1/2$ we approach as in (3.18) with the help of (4.4):

$$\|\mathrm{I}_h u - u\|_{-1/2} \;\lesssim\; h^{1/2}\; \|\widetilde{\mathrm{I}}_h u - u\|_{L^2(\mathbb{R})}.$$

Applying Theorem A.2 to the above right-hand side implies (3.13) with

$$\alpha_{\mathrm{I}} \;=\; \begin{cases} 3/2 & : \quad N = 1, \\ 5/2 & : \quad N \ge 2. \end{cases}$$

The reconstruction filter belonging to I_h considered in this section is

$$\upsilon(s) \;=\; \frac{1}{\pi\,\sqrt{2\pi}} \int_0^\infty \sigma\; \frac{\mathrm{sinc}^M(\sigma/2)}{\mathbf{a}(\sigma)}\; \widehat{B}(\sigma)\; \cos(s\,\sigma)\; \mathrm{d}\sigma.$$

To find an explicit representation of $\mathbf{a}$ poses no problem since $a_\ell = \mathcal{B}(\ell)$, where $\mathcal{B}$ is the B-spline of order $M + N$. So, $a_\ell \in \mathbb{Q}$ can be found by the B-spline recursion or explicit representations of B-splines. Nevertheless, υ cannot be evaluated explicitly in general. However, the needed values of υ at integers can be computed numerically to any desired accuracy.

Example 4.1. Let $M = 1$, $N = 2$, and $B(s) = \mathrm{sinc}(\pi\, s)$. Then, $\mathbf{a}(\sigma) = \frac{3}{4} + \frac{1}{4}\cos(\sigma)$ and

$$\upsilon(s) \;=\; \frac{4}{\pi^2} \int_0^\pi \frac{\sin(\sigma/2)\,\cos(s\,\sigma)}{3 + \cos(\sigma)}\; \mathrm{d}\sigma.$$

Using this filter in the FBA together with piecewise linear interpolation in I_h yields

$$\big\|\widetilde{f}_{\mathrm{FB}} - f\big\|_{L^2(\Omega)} \;\lesssim\; h^{\min\{5/2,\,\alpha\}}\; \|f\|_\alpha \quad \text{for } f \in H_0^\alpha(\Omega),\; \alpha > 0,$$

since band-limited interpolation (B.4) is considered for E_h ($\beta_{\max} = \tau_{\max} > 3$).

Remark 4.2. The biorthogonalization procedure (4.1) is the same procedure used in the construction of orthogonal spline wavelets; see Lemarié [12]. The connection between wavelets and reconstruction filters can even be extended to increase α_{I}. Choosing A to be a B-spline of order N and η to be a suitable compactly supported dual scaling function (see Cohen, Daubechies, and Feauveau [4]) yields an operator I_h with an α_{I} increasing with N. The needed approximation properties of I_h and $\mathrm{I}_h^\star$ are reported, for instance, by Dahmen [5, Prop. 5.1].

5. Verifying (3.13) for interpolation-like operators $\mathbf{I}_h$ based on B-splines. Our analysis presented so far does not cover operators I_h where η and A are B-splines of order M and N, respectively. We will now investigate this situation.

Let E_h and I_h be defined as earlier with respect to ϵ, B, η, and A. Moreover, let $\mathbf{a}$ be given as in (4.2); however, $\widetilde{\eta}$ is replaced by η. Further, define the operator $\mathbf{A}_h : L^2(\mathbb{R}) \to L^2(\mathbb{R})$, $h > 0$, by $\widehat{\mathbf{A}_h u}(\sigma) := \mathbf{a}(h\,\sigma)\,\widehat{u}(\sigma)$. Note that $\widehat{\mathbf{A}_h^{-1}u}(\sigma) = \widehat{u}(\sigma)/\mathbf{a}(h\,\sigma)$. Now the key observation is that

$$\widetilde{f}_{\mathrm{FB}} = \frac{1}{2\pi}\,\mathbf{R}^*\mathrm{I}_h\Lambda E_h\mathbf{R}f = \frac{1}{2\pi}\,\mathbf{R}^*\mathrm{I}_h\mathbf{A}_h^{-1}\Lambda\mathbf{A}_hE_h\mathbf{R}f.$$

Consequently, we have to study the approximation powers of the products $\mathbf{A}_hE_h$ and $\mathrm{I}_h\mathbf{A}_h^{-1}$. The latter product is exactly the operator I_h studied in the former section. Hence,

$$\|\mathrm{I}_h\mathbf{A}_h^{-1} - I\|_{H^{\alpha-1/2}(\mathbb{R})\to H^{-1/2}(\mathbb{R})} \lesssim h^\alpha, \quad 0 \le \alpha \le \alpha_{\mathrm{I}} = \begin{cases} 3/2 \;:\; N = 1, \\ 5/2 \;:\; N \ge 2. \end{cases}$$

The product $\mathbf{A}_hE_h$ requires a little bit more attention. We begin with

$$\|\mathbf{A}_hE_hu - u\|_\tau \lesssim \|E_hu - u\|_\tau + \|\mathbf{A}_hu - u\|_\tau.$$

In view of (3.12) and (C.2) we obtain

$$\|\mathbf{A}_hE_hu - u\|_\tau \lesssim h^{\beta-\tau}\|u\|_\beta, \quad u \in H_0^\beta(-1,1),$$

for $\beta_{\min} \le \beta \le \min\{\beta_{\max}, 2+\tau\}$, $0 \le \tau \le \beta$, $\tau \le \tau_{\max}$. The parameters $\beta_{\min}$, $\beta_{\max}$, and $\tau_{\max}$ correspond to E_h.

Theorem 3.7 holds accordingly, however, with the following restrictions on α:

$$\max\{0, \beta_{\min} - 1/2\} < \alpha \le \min\big\{\alpha_{\mathrm{I}},\, 2,\, \beta_{\max} - 1/2,\, \tau_{\max} - 1/2\big\};$$

that is, the maximal convergence order cannot exceed 2 which is a tribute to the operator $\mathbf{A}_h$ in front of E_h.

Example 5.1. Using the Shepp–Logan filter ($\eta = \chi_{[-1/2,1/2[}$, $B(s) = \mathrm{sinc}(\pi\,s)$) in the FBA together with piecewise linear interpolation in I_h (A is the linear B-spline) yields

$$\big\|\widetilde{f}_{\mathrm{FB}} - f\big\|_{L^2(\Omega)} \lesssim h^{\min\{2,\,\alpha\}}\,\|f\|_\alpha \quad \text{for } f \in H_0^\alpha(\Omega),\ \alpha > 0,$$

when ϵ is either an even, compactly supported, and normalized L^2-function (Theorem B.2) or the Dirac distribution (Theorem B.4).

6. Numerical illustrations. We provide numerical experiments to illustrate the convergence results proved in the former sections. Especially, we will see that the convergence rates saturate indeed at the given bounds.

To this end we need an $f \in L^2(\Omega)$ with a prescribed Sobolev order and with an analytically computable Radon transform. We favor the following construction. Let p_n be defined by $p_n(x) = (1 - \|x\|^2)^n$, $\|x\| \le 1$, and $p_n(x) = 0$, otherwise. We have that $p_n \in H_0^\alpha(\Omega)$ for any $\alpha < n + 1/2$. The function f for the first numerical experiment is then given by

$$f(x) := \sum_{k=1}^{3} d_k\; p_3\big(U_k(x - b_k)\big) \in H_0^\alpha(\Omega) \text{ for any } \alpha < 7/2, \tag{6.1}$$

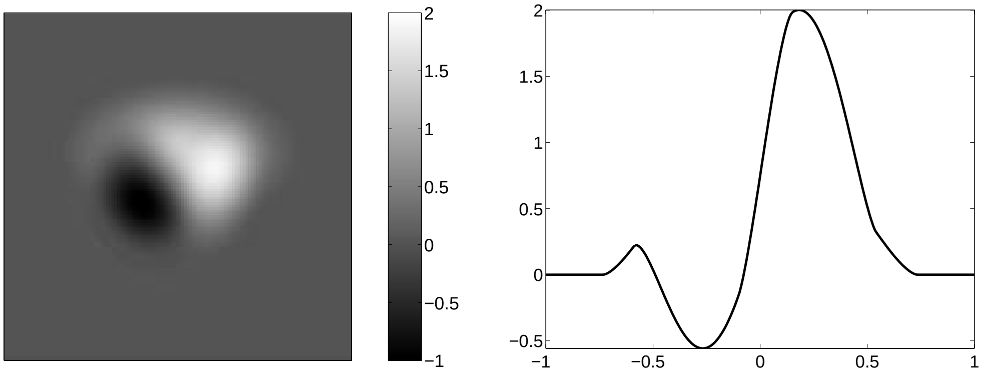


Fig. 2. *The function f from* (6.1) *(left) and its cross section $f(\cdot, 0)$ (right).*

where $d_1 = 1$, $d_2 = -1.5$, $d_3 = 1.5$, and $b_1 = (0.22, 0)^t$, $b_2 = (-0.22, 0)^t$, $b_3 = (0, 0.2)^t$. Further, $U_k = U(\varphi_k, \delta_k, \gamma_k)$, $k = 1, 2, 3$, with

$$U(\varphi, \delta, \gamma) := \begin{pmatrix} \cos(\varphi)/\delta & \sin(\varphi)/\delta \\ -\sin(\varphi)/\gamma & \cos(\varphi)/\gamma \end{pmatrix} \tag{6.2}$$

and

$$\begin{aligned} &\delta_1 = 0.51, && \gamma_1 = 0.31, && \varphi_1 = 72\pi/180, \\ &\delta_2 = 0.51, && \gamma_2 = 0.36, && \varphi_2 = 108\pi/180, \\ &\delta_3 = 0.5, && \gamma_3 = 0.8, && \varphi_3 = \pi/2. \end{aligned}$$

See Figure 2 for a graphical representation of f. We reconstructed f on the grid $\mathfrak{X}_q := \Omega \cap \{(i/q, j/q) \mid -q \leq i, j \leq q\}$ by

$$f_{\mathrm{FB},q}(x) := \frac{1}{2\pi}\, \mathbf{R}^*_{3q} \mathrm{I}_{1/q} \Lambda E_{1/q} \mathbf{R} f(x), \quad x \in \mathfrak{X}_q,$$

where $\mathbf{R}^*_p$ is defined in (1.6). We have chosen the number of directions ($3q$) close to its optimal value; see, e.g., Natterer [14, p. 84].

Now we define the relative ℓ^2-reconstruction error e by

$$e(q) := \Big(\sum_{x \in \mathfrak{X}_q} \big(f_{\mathrm{FB},q}(x) - f(x)\big)^2 \Big/ \sum_{x \in \mathfrak{X}_q} f(x)^2 \Big)^{1/2}. \tag{6.3}$$

In Figure 3 we plotted e as the function of $q \in \{25, 50, 75, 100, 125, 150, 175, 200\}$ on a double logarithmic scale with respect to three different settings in the FBA:

- The Shepp–Logan filter with nearest-neighbor interpolation (Example 3.8). Here, the expected and observed convergence rate is $e(q) \sim q^{-3/2}$; see the dot-dashed line marked with □.
- The Shepp–Logan filter with piecewise linear interpolation (Example 5.1). Here, the expected and observed convergence rate is $e(q) \sim q^{-2}$; see the dashed line marked with △.
- The modified Shepp–Logan filter with piecewise linear interpolation (Example 4.1). Here, the expected and observed convergence rate is $e(q) \sim q^{-5/2}$; see the solid line marked with ○. We also plotted an auxiliary curve decaying exactly like $q^{-5/2}$ (solid line in light gray).

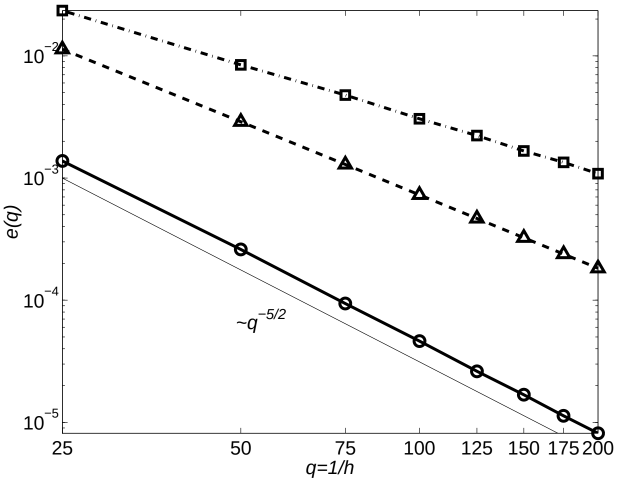


FIG. 3. *The relative ℓ^2-errors e (6.3) for reconstructing f (6.1) by the FBA using the Shepp–Logan filter with nearest-neighbor interpolation (dot-dashed with □), the Shepp–Logan filter with piecewise linear interpolation (dashed with △), and the modified Shepp–Logan filter with piecewise linear interpolation (solid with ○). The auxiliary solid line indicates exact decay $q^{-5/2}$.*

In light of the computational experiments we may conclude that our bounds for the maximal convergence orders cannot be improved (at least for the settings underlying the experiments).

Next, we present the relative ℓ^2-errors in reconstructing the Shepp–Logan head phantom; see Figure 4. The Shepp–Logan head phantom f^{SL} simulates the geometry and the density relations in a human skull. It consists of superimposed indicator functions of ellipses. Hence, $f^{\mathrm{SL}} \in H_0^\alpha(\Omega)$ for any $\alpha < 1/2$.[4] We therefore expect and observe $e(q) \sim q^{-1/2}$ for all three settings from above.

Both experiments agree completely with our theoretical results, although a discretization of the backprojection operator was not investigated. With our last experiment we justify this simplification once more by considering a setting which *might* cause trouble in a convergence analysis including the discrete backprojection operator.

The function to be reconstructed consists of indicator functions of two rectangles $\mathcal{R}_1$ and $\mathcal{R}_2$:

$$f(x) := \chi_{\mathcal{R}_1}(x) + 0.5\,\chi_{\mathcal{R}_2}(x) \tag{6.4}$$

with

$$\mathcal{R}_1 := [-2/5, 2/5] \times [-3/5, 3/5]$$

and (U as in (6.2))

$$\mathcal{R}_2 := \left\{x \in \mathbb{R}^2 \,\middle|\, U(\pi/3, 0.7, 0.4)(x-b) \in [-1,1]^2\right\}, \quad b = (-0.1, -0.1)^t;$$

see Figure 5 (left). Note that f is in $H_0^\alpha(\Omega)$ for any $\alpha < 1/2$. So what is the difference to the Shepp–Logan head phantom? While the fact that $\mathbf{R}f$ as a function

[4] In general, picture densities in medical imaging can be considered elements in $H_0^\alpha(\Omega)$ with $\alpha < 1/2$ but close to 1/2; see Natterer [14, pp. 92ff.].

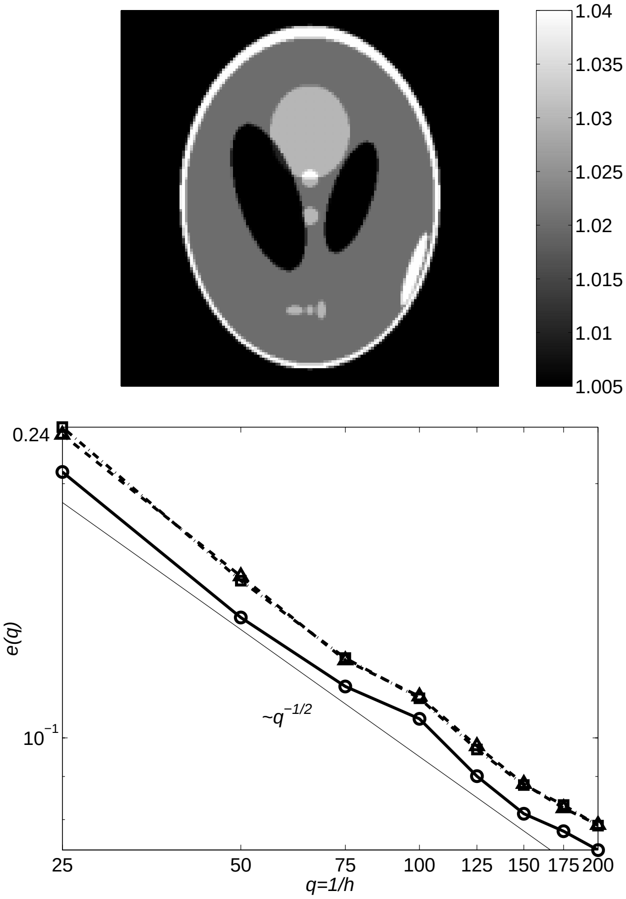


FIG. 4. *Top: head phantom due to Shepp–Logan* [21]*. Bottom: the relative ℓ^2-errors e* (6.3) *for reconstructing the Shepp–Logan phantom by the FBA using the Shepp–Logan filter with nearest-neighbor interpolation (dot-dashed with □), the Shepp–Logan filter with piecewise linear interpolation (dashed with △), and the modified Shepp–Logan filter with piecewise linear interpolation (solid with ○). The auxiliary solid line indicates exact decay $q^{-1/2}$.*

of two variables lies in $H_0^\beta(-1,1)\widehat{\otimes}L^2(0,\pi)$ implies that the functions of one variable $\mathbf{R}f(\cdot,\vartheta)$ lie in $H_0^\beta(-1,1)$ for almost all ϑ, there may be a null set of exceptional angles ϑ, where $\mathbf{R}f(\cdot,\vartheta)$ has less Sobolev regularity. For f given in (6.4) we have that $\mathbf{R}f$ is in $H_0^\beta(-1,1)\widehat{\otimes}L^2(0,\pi)$ for any $\beta<1$, but there exist four angles ϑ, where $\mathbf{R}f(\cdot,\vartheta)$ is less smooth. Indeed,

$$\mathbf{R}f(\cdot,\vartheta)\in H_0^\alpha(-1,1),\ \alpha<1/2,\quad \text{for } \vartheta\in\{0,\pi/3,\pi/2,5\pi/6\};$$

see Figure 5 (right). The bound on α is maximal (there are no such pathological angles for the Shepp–Logan head phantom; however, one expects such angles in real measurements from medical imaging).

In Figure 6 we plotted the relative reconstruction error (6.3) for the same q-values as before. Please note that the discrete Radon data for all q contain integrals over lines which run along the boundary of $\mathcal{R}_1$. Further, all used reconstruction grids $\mathcal{X}_q$ have sufficiently many points on the boundary of $\mathcal{R}_1$; indeed, the cardinality of

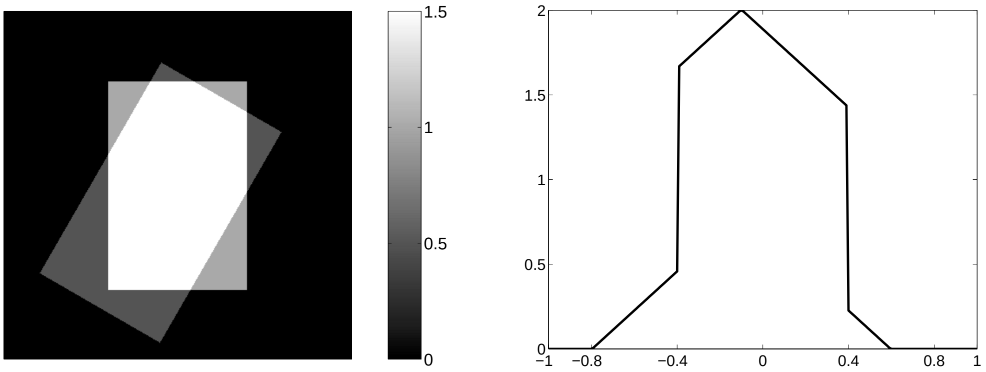


FIG. 5. *The function f from* (6.4) *(left) and its projection* $\mathbf{R}f(\cdot,0)$ *(right). The jumps of* $\mathbf{R}f(\cdot,0)$ *in* $\pm 2/5$ *are clearly visible.*

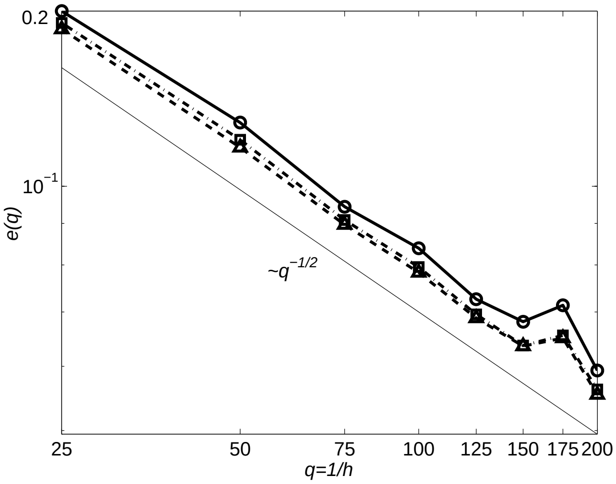


FIG. 6. *The relative ℓ^2-errors e* (6.3) *for reconstructing f* (6.4) *by the FBA using the Shepp–Logan filter with nearest-neighbor interpolation (dot-dashed with □), the Shepp–Logan filter with piecewise linear interpolation (dashed with △), and the modified Shepp–Logan filter with piecewise linear interpolation (solid with ○). The auxiliary solid line indicates exact decay $q^{-1/2}$.*

$\mathcal{X}_q \cap \partial\mathcal{R}_1$ increases like $\mathrm{O}(q)$.

We observe that even “pathological” projections do not deteriorate the convergence rate obtained by using the continuous backprojection operator for the analysis.

Appendix A. Proof of (3.12) for interpolation-like operators $\boldsymbol{E_h}$ based on B-splines. We consider E_h as defined in (3.6) where B is the cardinal B-spline of order $N \geq 1$; that is, B is the N-fold convolution of $\chi_{[-1/2,1/2]}$ with itself. The functional $\epsilon \in H_0^{-\beta_{\min}}(\mathbb{R})$, $\beta_{\min} \geq 0$, is supposed to be even, compactly supported in $\square = [-a, a]$, $a > 0$, and normalized by $\langle 1, \epsilon\rangle = 1$ where $\langle\cdot,\cdot\rangle$ denotes the duality pairing in $H^{\beta_{\min}}(\square) \times H_0^{-\beta_{\min}}(\square)$.

The techniques we use below are standard in approximation theory, yet we are not aware of any reference suitable for our setting; however, see Aubin [1, sect. 8.6].

First, we show that E_h reproduces affine linear functions if $N \geq 2$.

Lemma A.1. *If $N \geq 2$, then $E_h p = p$ for any $p \in \Pi_1$. For $N = 1$ E_h reproduces only constants.*

Proof. Note that the action of E_h on p is well defined since ϵ has compact support. Constants are preserved by $\langle 1, \epsilon(\cdot - k)\rangle = 1$ and $\sum_{k\in\mathbb{Z}} B(s-k) = 1$; see, e.g., Schoenberg [20, p. 16]. Let $p(s) = s$; then $\langle p(\cdot), \epsilon(\cdot - k)\rangle = k$ due to the evenness of ϵ. By $s = \sum_{k\in\mathbb{Z}} k\, B(s-k)$, $N \geq 2$ (see, e.g., Schoenberg [20, p. 16]), we are done. $\square$

Theorem A.2. *Let $\beta_{\min} \leq \beta \leq \min\{2, N\}$, $\tau < N - 1/2$, and $0 \leq \tau \leq \beta$. Then,*

$$\|E_h u - u\|_\tau \lesssim h^{\beta-\tau}\, \|u\|_\beta \quad \textit{as } h \to 0.$$

Proof. We restrict the proof to $N \geq 2$, and we show first a local version of the approximation property. Therefore, let $\square_{h,k} := h\,(\square + 2a\,k)$ for $k \in \mathbb{Z}$. We will rely on the Bramble–Hilbert-like estimate (A.1): there is an affine linear function $P = P(u)$ such that

$$\|u - P\|_{H^\tau(\square_{h,k})} \lesssim h^{\beta-\tau}\, \|u\|_{H^\beta(\square_{h,k})}, \quad 0 \leq \tau \leq \beta \leq 2. \tag{A.1}$$

For $\tau = 0$, (A.1) reduces to the original estimate by Bramble and Hilbert [2]. For positive real τ, see Dupont and Scott [7, Thm. 6.1] or Brenner and Scott [3, Lem. 4.3.8]. By Lemma A.1 and (A.1) we have

$$\|E_h u - u\|_{H^\tau(\square_{h,k})} \lesssim \|E_h(u-p)\|_{H^\tau(\square_{h,k})} + h^{\beta-\tau}\, \|u\|_{H^\beta(\square_{h,k})}.$$

Let $\mathfrak{J}_{h,k} := \{r \in \mathbb{Z} \mid \operatorname{supp} B_h(\cdot - s_r) \cap \square_{h,k} \neq \emptyset\}$. The cardinality of $\mathfrak{J}_{h,k}$ neither depends on h nor on k. We proceed with

$$\begin{aligned}
\|E_h(u-p)\|_{H^\tau(\square_{h,k})} &\lesssim \sum_{r\in\mathfrak{J}_{h,k}} h^{-1}\, \bigl|\bigl\langle u - P, \epsilon_h(\cdot - s_r)\bigr\rangle\bigr|\, \|B_h(\cdot - s_r)\|_{H^\tau(\mathbb{R})} \\
&\lesssim h^{-\tau} \sum_{r\in\mathfrak{J}_{h,k}} h^{-1/2}\, \bigl|\bigl\langle u - P, \epsilon_h(\cdot - s_r)\bigr\rangle\bigr|.
\end{aligned}$$

From the proof of Lemma 5.2 by Dahmen, Prössdorf, and Schneider [6] we know that

$$\bigl|\bigl\langle u - P, h^{-1/2}\, \epsilon_h(\cdot - s_r)\bigr\rangle\bigr|^2 \lesssim \|u - P\|^2_{L^2(\square_{h,r})} + h^{2\,\beta_{\min}}\, \|u - P\|^2_{H^{\beta_{\min}}(\square_{h,r})},$$

which, by (A.1), gives

$$\begin{aligned}
\|E_h u - u\|_{H^\tau(\square_{h,k})} &\lesssim h^{\beta-\tau} \sum_{r\in\mathfrak{J}_{h,k}} \|u\|_{H^\beta(\square_{h,r})} \lesssim h^{\beta-\tau} \Bigl(\sum_{r\in\mathfrak{J}_{h,k}} \|u\|^2_{H^\beta(\square_{h,r})}\Bigr)^{1/2} \\
&\lesssim h^{\beta-\tau}\, \|u\|_{H^\beta(\widetilde{\square}_{h,k})},
\end{aligned}$$

where $\widetilde{\square}_{h,k} := \bigcup_{r\in\mathfrak{J}_{h,k}} \square_{h,r}$. Thus,

$$\|E_h u - u\|_{H^\tau(\square_{h,k})} \lesssim h^{\beta-\tau}\, \|u\|_{H^\beta(\widetilde{\square}_{h,k})}.$$

Squaring both sides of the latter local approximation property and summing over $k \in \mathbb{Z}$ yield finally the stated global approximation property. $\square$

Summary. The above theorem covers especially the cases $\epsilon = \delta$ (Dirac distribution), where $\beta_{\min} > 1/2$, and $\epsilon \in L^2(\square)$ being even with $\int_\square \epsilon(s)\,\mathrm{d}s = 1$, where $\beta_{\min} = 0$. Hence, for both latter cases (3.12) holds with $\beta_{\max} = 2$ and $\tau_{\max} < N - 1/2$.

Appendix B. Proof of (3.12) for interpolation-like operators E_h based on the sinc-function. We consider E_h as defined in (3.6) where B is the *sinus cardinalis*; that is, $B(x) := \operatorname{sinc}(\pi x)$, where $\operatorname{sinc}(x) = \sin(x)/x$, $x \neq 0$, and $\operatorname{sinc}(0) = 1$. Further, $\epsilon \in L^1(\mathbb{R}) \cap L^2(\mathbb{R})$ is an even function with compact support and a normalized mean value, $\int \epsilon(x)\,\mathrm{d}x = 1$. Here, $\langle\cdot,\cdot\rangle$ denotes the $L^2(\mathbb{R})$-inner product.

First we bound E_h uniformly in h.

Lemma B.1. (a) *The operators $E_h : L^2(\mathbb{R}) \to L^2(\mathbb{R})$, $h > 0$, are uniformly bounded in h.*

(b) *Let w be in $L^2(\mathbb{R})$ with* $\operatorname{supp} \widehat{w} \subset [-\pi/h, \pi/h]$. *Then, we have the inverse estimate $\|w\|_\alpha \leq 2^{\alpha/2}\,\pi^\alpha\, h^{-\alpha}\, \|w\|_{L^2(\mathbb{R})}$ for $0 < h \leq \pi$ and any $\alpha \geq 0$.*

Proof. (a) Set $\square_{h,k} = h\,(\operatorname{supp}\epsilon + k)$. The $L^2(\mathbb{R})$-orthogonality of $\{B_h(\cdot - s_k)\}_{k\in\mathbb{Z}}$ gives

$$\|E_h u\|^2_{L^2(\mathbb{R})} = h^{-1}\sum_{k\in\mathbb{Z}} |\langle u, \epsilon_h(\cdot - s_k)\rangle|^2 \leq \sum_{k\in\mathbb{Z}} \|u\|^2_{L^2(\square_{h,k})}\|\epsilon\|^2_{L^2(\mathbb{R})} \lesssim \|u\|^2_{L^2(\mathbb{R})}.$$

(b) The inverse estimate results from a straightforward estimate of $\|w\|^2_\alpha$ taking into account the compact support of $\widehat{w}$. □

After the above preparatory results we are able to prove the claimed convergence estimate.

Theorem B.2. *Let $0 \leq \tau \leq \beta$, $\beta - \tau \leq 2$. Under the assumptions from above we have that*

$$\|E_h u - u\|_\tau \lesssim h^{\beta-\tau}\ \|u\|_\beta \quad \textit{as } h \to 0.$$

Proof. Define an auxiliary operator $P_h : L^2(\mathbb{R}) \to L^2(\mathbb{R})$ by

$$\widehat{P_h w}(\xi) := \chi_{\square_h}(\xi)\ \widehat{w}(\xi).\,^5 \tag{B.1}$$

It is an easy exercise to obtain

$$\|P_h u - u\|_\tau \lesssim h^{\beta-\tau}\ \|u\|_\beta \quad \text{for } 0 \leq \tau \leq \beta < \infty \tag{B.2}$$

whenever the right-hand side is finite.

In a first step we consider $\|E_h P_h u - P_h u\|_\tau$. We have

$$\widehat{E_h P_h u}(\xi) = \frac{1}{\sqrt{2\pi}}\ \chi_{\square_h}(\xi) \sum_{k\in\mathbb{Z}} \langle P_h u,\, \epsilon_h(\cdot - s_k)\rangle\ \mathrm{e}^{-\imath\, hk\xi}$$

and

$$\begin{aligned} h^{1/2}\ \langle P_h u,\, \epsilon_h(\cdot - s_k)\rangle &= h^{1/2}\int_{\square_h} \widehat{P_h u}(\xi)\ \widehat{\epsilon_h}(\xi)\ \mathrm{e}^{\imath\, hk\xi}\ \mathrm{d}\xi \\ &= \Big(\frac{h}{2\pi}\Big)^{1/2} \int_{\square_h} \widehat{P_h u \star \epsilon_h}(\xi)\ \mathrm{e}^{\imath\, hk\xi}\ \mathrm{d}\xi, \end{aligned}$$

which is the kth Fourier coefficient of $\widehat{P_h u \star \epsilon_h}$. Hence,

$$\widehat{E_h P_h u}(\xi) = h^{-1}\ \chi_{\square_h}(\xi)\ \widehat{P_h u \star \epsilon_h}(\xi).$$

[5] Actually, P_h is the orthogonal projector onto the closed subspace of band-limited functions with band-width π/h.

Therefore,

$$\begin{aligned}\|E_h P_h u - P_h u\|_\tau^2 \;&=\; \int_{\square_h} (1+\xi^2)^\tau \,\left|h^{-1}\; \widehat{P_h u \star \epsilon_h}(\xi) - \widehat{P_h u}(\xi)\right|^2 \mathrm{d}\xi \\ &\lesssim\; \int_{\square_h} (1+\xi^2)^\tau \,\left|\widehat{u}(\xi)\right|^2\, M(h\xi)\;\mathrm{d}\xi,\end{aligned}$$

where $M(z) = |\widehat{\epsilon}(z) - 1/\sqrt{2\pi}|^2$, $z \in \mathbb{R}$. As in the proof of Corollary 2.2 one shows that $M(z) \lesssim z^4$ using a Taylor expansion of $\widehat{\epsilon}$ about the origin. Now let $0 \le \beta - \tau \le 2$. Then,

$$\|E_h P_h u - P_h u\|_\tau^2 \lesssim h^4 \int_{\square_h} (1+\xi^2)^\beta \,\left|\widehat{u}(\xi)\right|^2\, \xi^{4-2(\beta-\tau)}\;\mathrm{d}\xi \tag{B.3a}$$

$$\lesssim h^{2(\beta-\tau)}\;\|u\|_\beta^2. \tag{B.3b}$$

In the final step we use both statements from Lemma B.1 as well as (B.2) and (B.3):

$$\begin{aligned}\|E_h u - u\|_\tau \;&\le\; \|E_h u - E_h P_h u\|_\tau \;+\; \|E_h P_h u - P_h u\|_\tau \\ &\qquad + \|P_h u - u\|_\tau \\ &\lesssim\; h^{-\tau}\;\|u - P_h u\|_{L^2(\mathbb{R})} \;+\; h^{\beta-\tau}\;\|u\|_\beta.\end{aligned}$$

Applying (B.2) again we conclude with the proof of Theorem B.2. □

Remark B.3. The upper bound 2 on $\beta - \tau$ in Theorem B.2 may be relaxed by imposing higher order vanishing moments on ϵ.

Now we investigate band-limited interpolation; that is, E_h is defined by

$$E_h u(s) \;=\; \sum_{k\in\mathbb{Z}} u(s_k)\;\mathrm{sinc}\Big(\frac{\pi}{h}(s - s_k)\Big). \tag{B.4}$$

THEOREM B.4. *Let $\beta_{\max} \in \mathbb{N}$. Then, for $1/2 < \beta < \infty$, $0 \le \tau \le \beta$ with $\beta - \tau \le \beta_{\max}$, we have that*

$$\|E_h u - u\|_\tau \;\lesssim\; h^{\beta-\tau}\;\|u\|_\beta \quad \textit{as } h \to 0$$

whenever $u \in H^\beta(\mathbb{R})$ is compactly supported. The constant in the above estimate may *depend on $\beta_{\max}$.*

Proof. Band-limited interpolation is well defined under the assumptions on u. We introduce an auxiliary operator $E_h^{(m)}$. To this end let $\epsilon \in L^2(\mathbb{R})$ be compactly supported with normalized mean value ($\int \epsilon(x)\,\mathrm{d}x = 1$) and vanishing moments up to order $\beta_{\max}$ ($\int x^k\,\epsilon(x)\,\mathrm{d}x = 1$, $k = 1, \dots, \beta_{\max}$). Thus, $\widehat{\epsilon}(0) = 1/\sqrt{2\pi}$ and $\widehat{\epsilon}^{(\nu)}(0) = 0$, $\nu = 1, \dots, \beta_{\max}$. Define $\epsilon^{(m)}(s) := m\,\epsilon(m\,s)$, $m \in \mathbb{N}$, and

$$E_h^{(m)} u(s) \;:=\; h^{-1} \sum_{k\in\mathbb{Z}} \big\langle u, \epsilon_h^{(m)}(\cdot - s_k)\big\rangle\;\mathrm{sinc}\Big(\frac{\pi}{h}(s - s_k)\Big).$$

Observe that $E_h^{(m)} : L^2(\mathbb{R}) \to L^2(\mathbb{R})$ is uniformly bounded in h *and* m; see the proof of Lemma B.1. Hence, we may apply Theorem B.2 to obtain

$$\|E_h^{(m)} u - u\|_\tau \;\lesssim\; h^{\beta-\tau}\;\|u\|_\beta, \tag{B.5}$$

where the constant is bounded in m, as a careful inspection of the proof of Theorem B.2 shows. Moreover, the upper bound on $\beta-\tau$ in (B.5) is $\beta_{\max}$ since all derivatives of $\widehat{\epsilon}$ up to order $\beta_{\max}$ vanish about 0; see Remark B.3. By (B.5),

$$\|E_h u - u\|_\tau \lesssim \|E_h u - E_h^{(m)} u\|_\tau + h^{\beta-\tau}\,\|u\|_\beta. \tag{B.6}$$

Further,

$$\|E_h u - E_h^{(m)} u\|_\tau \le \|\mathrm{sinc}_{h/\pi}\|_\tau \sum_{k\in\mathcal{J}_{m,h}(u)} \bigl|u(s_k) - \bigl\langle u, h^{-1}\epsilon_h^{(m)}(\cdot - s_k)\bigr\rangle\bigr|$$

with $\mathcal{J}_{m,h}(u) = \{k\in\mathbb{Z} \mid s_k \in \operatorname{supp} u\} \cup \{k\in\mathbb{Z} \mid \operatorname{supp}\, u \cap h(m^{-1}\operatorname{supp}\epsilon + k)\}$. The set $\mathcal{J}_{m,h}(u)$ is finite and its cardinality is bounded in m. So we have that $\lim_{m\to\infty}\|E_h u - E_h^{(m)} u\|_\tau = 0$, and the stated estimate is readily seen from (B.6). □

SUMMARY. The band-limited interpolation-like operators considered in Theorem B.2 satisfy (3.12) with $\beta_{\min}=0$, $\beta_{\max}=2+\tau$, and any $\tau_{\max}<\infty$. For the band-limited interpolation (B.4) we have (3.12) with $\beta_{\min}>1/2$ and any positive $\tau_{\max}$ and any fixed $\beta_{\max}>1/2$.

Appendix C. Complement to section 4.

This appendix is devoted to the proof of various auxiliary results from section 4. Throughout this appendix let $\widetilde{\eta}$ and A be B-splines of order $M\ge 1$ and $N\ge 1$, respectively. Further, let η be defined by (4.1).

C.1. The trigonometric polynomial a.

Recall that

$$\mathbf{a}(\sigma) = \sum_{\ell\in\mathbb{Z}} a_\ell\, \mathrm{e}^{-\imath\,\ell\sigma} \quad \text{with } a_\ell = \int_{\mathbb{R}} \widetilde{\eta}(s)\; A(\ell - s)\;\mathrm{d}s.$$

Since $\widetilde{\eta}$ and A are even, so are $\{a_\ell\}_{\ell\in\mathbb{Z}}$ and $\mathbf{a}$. By $\sum_{\ell\in\mathbb{Z}} A(\cdot-\ell)=1$ and $\int\widetilde{\eta}(s)\mathrm{d}s = 1$ (see, e.g., Schoenberg [20, p. 16 and p. 2]), we have that $\mathbf{a}(0)=1$. In the remainder of this appendix we verify that $\mathbf{a}$ has no zeros. Then we have established all properties of $\mathbf{a}$ claimed and needed in section 4.

Straightforward calculations reveal that the a_ℓ's are the Fourier coefficients of the 2π-periodic function $2\pi\sum_{k\in\mathbb{Z}}\widehat{\widetilde{\eta}}(\sigma+2\pi\,k)\,\widehat{A}(\sigma+2\pi\,k)$. Hence,

$$\mathbf{a}(\sigma) = 2\pi\sum_{k\in\mathbb{Z}}\widehat{\widetilde{\eta}}(\sigma+2\pi\,k)\;\widehat{A}(\sigma+2\pi\,k) = \sum_{k\in\mathbb{Z}}\mathrm{sinc}^{M+N}(\sigma/2+\pi\,k). \tag{C.1}$$

If $M+N$ is even, $\mathbf{a}$ clearly has no zeros because there is no σ such that $\mathrm{sinc}^{M+N}(\sigma/2+\pi\,k)=0$ for all $k\in\mathbb{Z}$. It remains to investigate the odd case $M+N=2L+1$, $L\in\mathbb{N}$. We factorize $\mathbf{a}$ according to

$$\mathbf{a}(\sigma) = \sin^{2L}(\sigma/2)\;\mathbf{\Sigma}_{2L+1}(\sigma) \quad \text{with } \mathbf{\Sigma}_{2L+1}(\sigma) := \sum_{k\in\mathbb{Z}}\frac{(-1)^k}{(\sigma/2+\pi\,k)^{2L+1}}.$$

As multiples of 2π are not zeros of $\mathbf{a}$ it suffices to show that $\mathbf{\Sigma}_{2L+1}$ has no zeros in $]0,2\pi[$. Separating even from odd indices we find

$$\mathbf{\Sigma}_{2L+1}(\sigma) = 2^{-(2L+1)}\bigl(S_{2L+1}(\sigma/4) - S_{2L+1}(\sigma/4+\pi/2)\bigr),$$

where $S_l(\sigma) := \sum_{k\in\mathbb{Z}}(\sigma+\pi\,k)^{-l}$, $l\ge 2$. Observe that $S_{2l}(\sigma)>0$, $l\in\mathbb{N}$. Now,

$$\frac{\mathrm{d}}{\mathrm{d}\sigma}S_{2L+1}(\sigma) = -(2L+1)\;S_{2L+2}(\sigma) < 0, \quad \sigma\in]0,2\pi[.$$

Therefore S_{2L+1} is strongly decreasing in $]0,2\pi[$ which gives $\mathbf{\Sigma}_{2L+1}>0$ in $]0,2\pi[$.

C.2. Biorthogonality (4.3). By (4.1) and (C.1) we obtain

$$\begin{aligned}\big\langle \eta(\cdot-k),\, A(\cdot)\big\rangle &= \int_{\mathbb{R}} \widehat{\eta}(\sigma)\ \widehat{A}(\sigma)\ \mathrm{e}^{\imath\, k\sigma}\ \mathrm{d}\sigma\\ &= \int_0^{2\pi} \sum_{n\in\mathbb{Z}} \widehat{\eta}(\sigma+2\pi\, n)\ \widehat{A}(\sigma+2\pi\, n)\ \mathrm{e}^{\imath\, k\sigma}\ \mathrm{d}\sigma\\ &= \int_0^{2\pi} \frac{1}{\mathbf{a}(\sigma)} \ \sum_{n\in\mathbb{Z}} \widehat{\widetilde{\eta}}(\sigma+2\pi\, n)\ \widehat{A}(\sigma+2\pi\, n)\ \mathrm{e}^{\imath\, k\sigma}\ \mathrm{d}\sigma \ = \int_0^{2\pi} \frac{\mathrm{e}^{\imath\, k\sigma}}{2\pi}\ \mathrm{d}\sigma,\end{aligned}$$

which is (4.3).

C.3. Approximation power of $\mathbf{I}_h^*$. We are not able to apply Theorem A.2 directly to I_h^* as η from (4.1) does not have compact support in general. Nevertheless, we will show that the approximation power of $\widetilde{\mathrm{I}}_h^*$ carries over to I_h^* (for the notation see section 4). Since

$$\widehat{\mathrm{I}_h^* u}(\sigma) = \widehat{\widetilde{\mathrm{I}}_h^* u}(\sigma)/\mathbf{a}(h\,\sigma)$$

we have that

$$\|u-\mathrm{I}_h^* u\|_\tau^2 \lesssim \int_{\mathbb{R}} (1+\sigma^2)^\tau\ \big|\mathbf{a}(h\,\sigma)\,\widehat{u}(\sigma) - \widehat{\widetilde{\mathrm{I}}_h^* u}(\sigma)\big|^2\ \mathrm{d}\sigma.$$

Thus,

$$\|u-\mathrm{I}_h^* u\|_\tau \lesssim \|\mathbf{A}_h u - \widetilde{\mathrm{I}}_h^*\mathbf{A}_h u\|_\tau + \|\widetilde{\mathrm{I}}_h^*\mathbf{A}_h u - \widetilde{\mathrm{I}}_h^* u\|_\tau,$$

where $\widehat{\mathbf{A}_h u}(\sigma) = \mathbf{a}(h\,\sigma)\,\widehat{u}(\sigma)$. Theorem A.2 provides

$$\|\mathbf{A}_h u - \widetilde{\mathrm{I}}_h^*\mathbf{A}_h u\|_\tau \lesssim h^{\beta-\tau}\ \|\mathbf{A}_h u\|_\beta \lesssim h^{\beta-\tau}\ \|u\|_\beta$$

for $0\le\beta\le\min\{2,M\}$, $\tau<M-1/2$, and $0\le\tau\le\beta$. Further, also by Theorem A.2,

$$\|\widetilde{\mathrm{I}}_h^*\mathbf{A}_h u - \widetilde{\mathrm{I}}_h^* u\|_\tau \lesssim \|\mathbf{A}_h u - u\|_\tau$$

whenever $0\le\tau<M-1/2$, for $M\le 2$, and $0\le\tau\le 2$, otherwise. A Taylor expansion of $\mathbf{a}$ about 0 proves that $|\mathbf{a}(\sigma)-1|\lesssim\sigma^2$. Now we may copy the proof of Corollary 2.2 to obtain

$$\|\mathbf{A}_h u - u\|_\tau \lesssim h^{\min\{2,\beta-\tau\}}\ \|u\|_\beta, \quad 0\le\tau\le\beta. \tag{C.2}$$

Collecting the pieces we find

$$\|u-\mathrm{I}_h^* u\|_\tau \lesssim h^{\beta-\tau}\ \|u\|_\beta \quad \text{as } h\to 0 \tag{C.3}$$

for $0\le\beta\le\min\{2,M\}$, $\tau<\min\{2,M-1/2\}$, and $0\le\tau\le\beta$.

REFERENCES

[1] J.-P. Aubin, *Applied Functional Analysis*, 2nd ed., Pure Appl. Math., Wiley, New York, 2000.

[2] J. H. Bramble and S. R. Hilbert, *Estimation of linear functionals on Sobolev spaces with application to Fourier transforms and spline interpolation*, SIAM J. Numer. Anal., 7 (1970), pp. 112–124.

[3] S. C. Brenner and L. R. Scott, *The Mathematical Theory of Finite Element Methods*, Texts Appl. Math. 15, Springer, New York, 1994.

[4] A. Cohen, I. Daubechies, and J.-C. Feauveau, *Biorthogonal bases of compactly supported wavelets*, Comm. Pure Appl. Math., 45 (1992), pp. 485–560.

[5] W. Dahmen, *Wavelet and Multiscale Methods for Operator Equations*, Acta Numer. 6, Cambridge University Press, Cambridge, UK, 1997, pp. 55–228.

[6] W. Dahmen, S. Prössdorf, and R. Schneider, *Wavelet approximation methods for pseudodifferential equations* I: *Stability and convergence*, Math. Z., 215 (1994), pp. 583–620.

[7] T. Dupont and L. R. Scott, *Polynomial approximation of functions in Sobolev spaces*, Math. Comp., 34 (1980), pp. 441–463.

[8] A. Faridani, *Praktische Fragen der lokalen Tomographie*, Z. Angew. Math. Mech., 70 (1990), pp. T530–T532.

[9] A. Faridani, D. V. Finch, E. L. Ritman, and K. T. Smith, *Local tomography* II, SIAM J. Appl. Math., 57 (1997), pp. 1095–1127.

[10] A. Faridani and E. L. Ritman, *High-resolution computed tomography from efficient sampling*, Inverse Problems, 16 (2000), pp. 635–650.

[11] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*, Academic Press, San Diego, 1980.

[12] P. G. Lemarié, *Ondelettes à localisation exponentielle*, J. Math. Pures Appl. (9), 67 (1988), pp. 227–236.

[13] A. K. Louis and F. Natterer, *Mathematical problems in computerized tomography*, Proc. IEEE, 71 (1983), pp. 379–389.

[14] F. Natterer, *The Mathematics of Computerized Tomography*, Wiley, Chichester, 1986.

[15] D. A. Popov, *On convergence of a class of algorithms for the inversion of the numerical Radon transform*, in Mathematical Problems of Tomography, Transl. Math. Monogr. 81, L. M. Gelfand and S. G. Gindikin, eds., AMS, Providence, R.I., 1990, pp. 7–65.

[16] A. Rieder, *Principles of reconstruction filter design in* 2*D-computerized tomography*, in Radon Transforms and Tomography, Contemp. Math. 278, T. Quinto, L. Ehrenpreis, A. Faridani, F. Gonzales, and E. Grinberg, eds., AMS, Providence, RI, 2001, pp. 201–226.

[17] A. Rieder, R. Dietz, and Th. Schuster, *Approximate inverse meets local tomography*, Math. Methods Appl. Sci., 23 (2000), pp. 1373–1387.

[18] A. Rieder and Th. Schuster, *The approximate inverse in action* II: *Convergence and stability*, Math. Comp., to appear.

[19] W. Rudin, *Functional Analysis*, 12th ed., Tata McGraw-Hill, New Delhi, India, 1988.

[20] I. J. Schoenberg, *Cardinal Spline Interpolation*, CBMS-NSF Regional Conf. Ser. in Appl. Math. 12, SIAM, Philadelphia, 1973.

[21] L. A. Shepp and B. F. Logan, *The Fourier reconstruction of a head section*, IEEE Trans. Nuc. Sci., 21 (1974), pp. 21–43.

[22] K. T. Smith, *Reconstruction formulas in computed tomography*, Proc. Sympos. Appl. Math., 27 (1982), pp. 7–23.

[23] K. T. Smith and F. Keinert, *Mathematical foundations of computed tomography*, Appl. Optics, 24 (1985), pp. 3950–3957.

[24] K. T. Smith, D. C. Solmon, and S. C. Wagner, *Practical and mathematical aspects of the problem of reconstructing objects from radiographs*, Bull. Amer. Math. Soc., 83 (1977), pp. 1227–1270.

[25] G. Strang and G. J. Fix, *An Analysis of the Finite Element Method*, Prentice-Hall Series in Automatic Computation, Prentice-Hall, Englewood Cliffs, NJ, 1973.

SIAM J. Numer. Anal.
Vol. 41, No. 3, pp. 893–918

h-BOX METHODS FOR THE APPROXIMATION OF HYPERBOLIC CONSERVATION LAWS ON IRREGULAR GRIDS*

MARSHA J. BERGER[†], CHRISTIANE HELZEL[†], AND RANDALL J. LEVEQUE[‡]

Abstract. We study generalizations of the high-resolution wave propagation algorithm for the approximation of hyperbolic conservation laws on irregular grids that have a time step restriction based on a reference grid cell length that can be orders of magnitude larger than the smallest grid cell arising in the discretization. This Godunov-type scheme calculates fluxes at cell interfaces by solving Riemann problems defined over boxes of a reference grid cell length h.

We discuss stability and accuracy of the resulting so-called h-box methods for one-dimensional systems of conservation laws. An extension of the method for the two-dimensional case, which is based on the multidimensional wave propagation algorithm, is also described.

Key words. finite volume methods, conservation laws, nonuniform grids, stability, accuracy

AMS subject classifications. 35L65, 65M12

PII. S0036142902405394

1. Introduction.

We consider the numerical approximation of hyperbolic systems of conservation laws using finite volume schemes on irregular grids. We mainly restrict our considerations to the case of one spatial dimension, although an extension to the two-dimensional case will also be considered. Under appropriate smoothness assumptions the equations can be formulated in the differential form

$$\frac{\partial}{\partial t}q(x,t) + \frac{\partial}{\partial x}f(q(x,t)) = 0, \tag{1.1}$$

where $q(x,t)$ is a vector of conserved quantities and $f(q(x,t))$ is a vector of flux functions. For the numerical approximation we want to use a finite volume method. On an unstructured grid such a scheme can be written in the general form

$$Q_i^{n+1} = Q_i^n - \frac{\triangle t}{\triangle x_i}\left(F_{i+\frac{1}{2}} - F_{i-\frac{1}{2}}\right), \tag{1.2}$$

where Q_i^n is an approximation of the cell average of the conserved quantity over the grid cell $[x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}]$ at time $t = t^n$. The vector valued quantities $F_{i-\frac{1}{2}}$ and $F_{i+\frac{1}{2}}$ are the numerical flux functions at the cell interfaces. We denote the time step by $\triangle t$ and the length of the ith grid cell by $\triangle x_i = x_{i+\frac{1}{2}} - x_{i-\frac{1}{2}}$.

We are particularly interested in the construction of high-resolution schemes for a grid which contains one small grid cell, while all other grid cells have the same length, which will be denoted by $h = \triangle x$. This situation is motivated by a two-dimensional application, namely the construction of Cartesian grid methods with embedded irregular geometry. Away from the boundary one may want to use a

[*] Received by the editors April 10, 2002; accepted for publication (in revised form) November 1, 2002; published electronically May 22, 2003.

http://www.siam.org/journals/sinum/41-3/40539.html

[†] Courant Institute of Mathematical Sciences, 251 Mercer Street, New York, NY 10012-1185 (berger@cims.nyu,edu, helzel@cims.nyu.edu). The work of these authors was supported by DOE grants DE-FG02-88ER25053 and DE-FC02-01ER25472, and AFOSR grant F49620-00-1-0099.

[‡] Department of Applied Mathematics, University of Washington, Box 352420, Seattle, WA 98195-2420 (rjl@amath.washington.edu). The work of this author was supported by DOE grants DE-FG03-00ER25292 and DE-FC02-01ER25474, and NSF grant DMS-0106511.

regular Cartesian grid. Near the boundary one then obtains irregular cut cells, which may be orders of magnitude smaller than the regular grid cells. Our aim in such a situation is to construct a scheme that is stable based on time steps adequate for the regular grid. Such methods were developed by Berger and LeVeque in [4], [5], [6]. The basic idea of these so-called h-box methods is to approximate the numerical fluxes at the interfaces of a small cell based on initial values specified over regions of length h, i.e., of the length of a regular grid cell. If this is done in an appropriate way, then the resulting method remains stable for time steps based on a CFL number appropriate for the regular part of the grid. See also [8], [9], [10], [23], [21], and [25] for other embedded boundary Cartesian grid methods that have this same stability property.

Besides this two-dimensional application, h-box schemes can also offer interesting alternatives to existing irregular grid methods. An extension of h-box methods to a completely irregular grid was considered by Berger, LeVeque, and Stern [7]; see also Stern [28]. We will consider such calculations in section 5. In section 7, we construct a multidimensional h-box method. Other potential applications are the construction of moving mesh or front-tracking algorithms. Stern [28] used an h-box method to construct a conservative finite volume algorithm for a Cartesian grid with an embedded curvilinear grid.

Unsurprisingly, the accuracy of an h-box method depends strongly on the definition of the h-box values. In this paper we develop a one-dimensional as well as a two-dimensional high-resolution h-box method. Our goal here is a systematic study of h-box methods in a relatively simple context to provide fundamental understanding for the more complex applications mentioned above. For the advection equation we show that the one-dimensional scheme leads to a second order accurate approximation of smooth solutions on nonuniform grids (without any restrictions on the grid). We also verify that the resulting method leads to high-resolution approximations for the Euler equations on nonuniform meshes. The approximation of transonic rarefaction waves turns out to require a special treatment. Throughout this paper we will discuss the construction of h-box methods based on LeVeque's high-resolution wave propagation algorithm [18]. This method is implemented in the CLAWPACK software package [13], which provided the basic tool for our test calculations.

The large time step Godunov method of LeVeque described in [14], [15], [16] is related to the h-box method. This scheme allows larger time steps in the approximation of nonlinear systems of conservation laws by increasing the domain of influence of the numerical scheme. This is done in a wave propagation approach, in which waves are allowed to move through more than one mesh cell. The interaction of waves is approximated by linear superposition. At a reflecting boundary this method becomes more difficult than an h-box method, especially in higher dimensions, since the reflection of waves at the boundary has to be considered for waves generated by Riemann problems away from the boundary; see [3]. In [22, Lemma 3.5], Morton showed that high-resolution versions of such a large time step method lead to a second order accurate approximation of the one-dimensional advection equation on a nonuniform grid only if the grid varies smoothly. The high-resolution h-box method described in this paper does not require this smoothness assumption.

2. The wave propagation algorithm.

In this section we describe the basic concept of the high-resolution wave propagation algorithm applied to irregular Cartesian grids; a more general description can be found in LeVeque [18] or [19]. The numerical method for solving (1.1) is a Godunov-type method; i.e., the fluxes at cell interfaces are calculated by solving Riemann problems defined from cell averages of the conserved quantities. This is done by calculating waves that are moving into each grid cell. The first order update of the wave propagation algorithm has the form

$$Q_i^{n+1} = Q_i^n - \frac{\triangle t}{\triangle x_i}\left(\mathcal{A}^+\triangle Q_{i-\frac{1}{2}} + \mathcal{A}^-\triangle Q_{i+\frac{1}{2}}\right).$$

Here the change of the conserved quantities is calculated by taking all waves into account that are moving into the grid cell from the left (respectively, right) cell interface. The solution of Riemann problems at cell interfaces provides a decomposition of the jump $Q_{i+1}^n - Q_i^n$ into waves $\mathcal{W}_{i+\frac{1}{2}}^p$ that are moving with speed $s_{i+\frac{1}{2}}^p$ for $1 \le p \le M_w$,

$$\triangle Q_{i+\frac{1}{2}}^n = Q_{i+1}^n - Q_i^n = \sum_{p=1}^{M_w} \mathcal{W}_{i+\frac{1}{2}}^p.$$

The left- and right-going fluctuations are calculated as

$$\mathcal{A}^+\triangle Q_{i-\frac{1}{2}} = \sum_{p=1}^{M_w} \max(s_{i-\frac{1}{2}}^p, 0)\mathcal{W}_{i-\frac{1}{2}}^p, \quad \mathcal{A}^-\triangle Q_{i+\frac{1}{2}} = \sum_{p=1}^{M_w} \min(s_{i+\frac{1}{2}}^p, 0)\mathcal{W}_{i+\frac{1}{2}}^p.$$

This can be written as a finite volume scheme of the form (1.2) using the relations

$$F_{i+\frac{1}{2}} = f(Q_i) + \mathcal{A}^-\triangle Q_{i+\frac{1}{2}}, \tag{2.1}$$

$$F_{i-\frac{1}{2}} = f(Q_i) - \mathcal{A}^+\triangle Q_{i-\frac{1}{2}}. \tag{2.2}$$

Appropriate waves and speeds for systems of conservation laws can sometimes be calculated by using an exact Riemann solver, but more often an approximative Riemann solver, for instance a Roe–Riemann solver [26], is used.

In the wave propagation algorithm second order correction terms are included by extending the first order method into the form

$$Q_i^{n+1} = Q_i^n - \frac{\triangle t}{\triangle x_i}\left(\mathcal{A}^+\triangle Q_{i-\frac{1}{2}} + \mathcal{A}^-\triangle Q_{i+\frac{1}{2}}\right) - \frac{\triangle t}{\triangle x_i}\left(\tilde{F}_{i+\frac{1}{2}}^2 - \tilde{F}_{i-\frac{1}{2}}^2\right). \tag{2.3}$$

On an irregular grid, the second order correction terms have the form

$$\tilde{F}_{i+\frac{1}{2}}^2 = \frac{1}{2}\sum_{p=1}^{M_w} |s_{i+\frac{1}{2}}^p| \left(\frac{\triangle x_i}{(\triangle x_i + \triangle x_{i+1})/2} - \frac{\triangle t}{(\triangle x_i + \triangle x_{i+1})/2}|s_{i+\frac{1}{2}}^p|\right)\tilde{\mathcal{W}}_{i+\frac{1}{2}}^p. \tag{2.4}$$

In (2.4) the waves $\tilde{\mathcal{W}}^p$ are limited waves—this limiting is necessary in order to avoid oscillations near discontinuities.

The resulting scheme is stable for the approximation of systems of conservation laws (1.1) as long as time steps are restricted such that waves move through at most one mesh cell, which means the Courant number is no larger than one, i.e.,

$$\text{CFL} = \triangle t \max_i \left(\frac{\max(\max_p(s_{i-\frac{1}{2}}^p, 0), |\min_p(s_{i+\frac{1}{2}}^p, 0)|)}{\triangle x_i}\right) \le 1. \tag{2.5}$$

The *h*-box method changes this time step restriction by replacing $\triangle x_i$ in the denominator of (2.5) by h, the width of a reference grid cell. We will use the notation CFL_h if we want to indicate that the Courant number is based on grid cells of width h.

We want to note that some care is necessary in the construction of second order accurate algorithms for irregular grids. There exist versions of the one-dimensional Lax–Wendroff method which lead to second order accurate approximations of the advection equation only if the grid is sufficiently smooth, i.e., if $\triangle x_i/\triangle x_{i-1} = 1 + \mathcal{O}(h)$, where $h = \max_i \triangle x_i$; see, for instance, Wendroff and White [30], [31] and Pike [24]. See also Morton [22] for convergence results of finite volume methods for the approximation of the advection equation on nonuniform grids.

3. The one-dimensional *h*-box method.

First we want to approximate (1.1) on an almost uniform grid that contains one small grid cell in the middle. This example allows simple analytical studies. However, we will show that the results obtained for this simple test case can be extended to more general applications.

We denote the length of a regular grid cell by $h = \triangle x$. The small cell has the length αh, with $0 < \alpha \le 1$. For the small cell the numerical method has to be modified in order to obtain a stable scheme for time steps $\triangle t$ that satisfy the stability condition in the regular part of the grid. The h-box method introduced by Berger and LeVeque [5] defines new left and right states at the edges of the small cell that represent the conserved quantities at these interfaces over boxes of length h; see Figure 1. This guarantees that the domain of dependence of the numerical solution has the size of a regular mesh cell, which is a necessary stability condition.

3.1. First order accurate *h*-box methods.

As a first step we compare the performance of two different h-box schemes applied to the advection equation $q_t(x,t) + aq_x(x,t) = 0$. We will assume that $a > 0$, although analogous considerations can of course be made for the case $a < 0$. In the following we assume that k is the index of the small cell. In order to calculate numerical fluxes at the small cell interfaces new values of the conserved quantity q that represent piecewise constant initial values over boxes of length h will be defined. For the left cell interface of the small cell, these values are denoted by $Q^L_{k-\frac{1}{2}}$ and $Q^R_{k-\frac{1}{2}}$. At the right cell interface of the small cell we have to define values $Q^L_{k+\frac{1}{2}}$ and $Q^R_{k+\frac{1}{2}}$. This is indicated by the shaded boxes at each interface in Figure 1.

The most obvious choice is to define the h-box values via cell averaging over the piecewise constant initial values. (To keep the notation simple we sometimes suppress the time index if it is clear that we mean the values at time t^n.) We obtain

$$\begin{aligned} Q^L_{k-\frac{1}{2}} &= Q_{k-1}, & Q^R_{k-\frac{1}{2}} &= \alpha Q_k + (1-\alpha)Q_{k+1}, \\ Q^L_{k+\frac{1}{2}} &= \alpha Q_k + (1-\alpha)Q_{k-1}, & Q^R_{k+\frac{1}{2}} &= Q_{k+1}. \end{aligned} \tag{3.1}$$

Such h-box values were used in Berger and LeVeque [5] as well as by Forrer and Jeltsch [10]. For the advection equation the update of the small cell value can now be calculated using the upwind method. We obtain

$$\begin{aligned} Q^{n+1}_k &= Q^n_k - \frac{\triangle t}{\alpha h}\left(aQ^L_{k+\frac{1}{2}} - aQ^L_{k-\frac{1}{2}}\right) \\ &= Q^n_k - \frac{a\triangle t}{\alpha h}\left(\alpha Q^n_k + (1-\alpha)Q^n_{k-1} - Q^n_{k-1}\right) \\ &= Q^n_k - \frac{a\triangle t}{h}(Q^n_k - Q^n_{k-1}). \end{aligned} \tag{3.2}$$

Note that the small denominator (that may cause a stability problem) has been removed. One can indeed show TVD stability for this method assuming $\text{CFL}_h \le 1$;

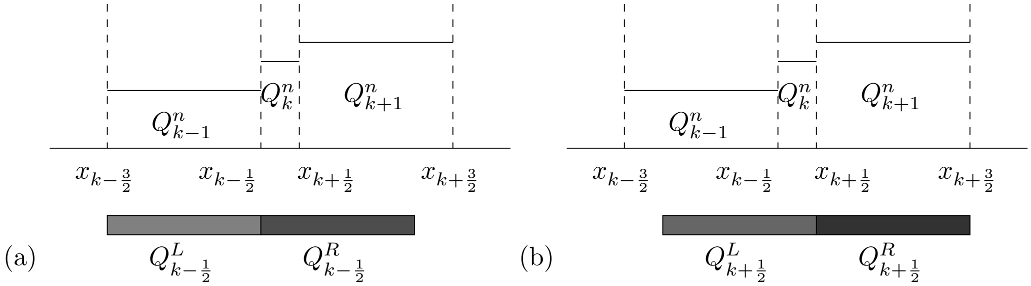


FIG. 1. *Schematic description of h-box values assigned to the left small cell interface (see* (a)*) (respectively, the right small cell interface (see* (b)*)).*

see section 4. However, it is clear that this cannot be a very accurate formula. The truncation error of the scheme (3.2) has the form

$$\begin{aligned} Lq &= \frac{q_k^{n+1} - q_k^n}{\triangle t} + a\frac{q_k^n - q_{k-1}^n}{h} \\ &= q_t(x_k, t^n) + \frac{a}{2}(\alpha + 1)q_x(x_k, t^n) + \mathcal{O}(\triangle t, h) \\ &= \frac{a}{2}(\alpha - 1)q_x(x_k, t^n) + \mathcal{O}(\triangle t, h). \end{aligned}$$

Only for $\alpha = 1$ is the truncation error in cell k of the order $\mathcal{O}(\triangle t, h)$. Note that grid functions for the exact solution q are expressed with lower-case letters, whereas numerical approximations are written in capital letters.

In spite of the apparent inconsistency of the scheme, numerical tests suggest that this h-box method converges with first order. For the advection equation we can indeed prove that under appropriate smoothness assumptions the scheme is first order accurate in the small cell. This so-called *supraconvergence property* can be shown using an idea developed for conservation laws by Wendroff and White [30], [31]. See also [12], [20], where these ideas were introduced for boundary value problems for ODEs.

PROPOSITION 1. *We consider the approximation of the advection equation on an almost uniform grid with mesh width h that contains one small mesh cell of length αh, with $\alpha \leq 1$. The one-dimensional h-box method* (3.2), *based on an upwind discretization with h-box values calculated by averaging over piecewise constant cell average values, leads to a first order accurate approximation for sufficiently smooth solutions of the advection equation, in spite of the fact that the truncation error indicates inconsistency.*

Proof. The basic step of the proof is to calculate the local truncation error for a grid function w, which must be an accurate enough approximation of the grid function of the exact solution q. We want to show that the truncation error for w is first order, i.e., $Lw = \mathcal{O}(h)$. In order to do this we specify the grid function to have the form

$$w_i^n = q_i^n + \frac{1}{2}(1 - \alpha_i)hq_x(x_i, t^n).$$

Here we assume that $\triangle x_i = \alpha_i h$, i.e., $\alpha_i = 1$ for $i \neq k$ and $\alpha_k = \alpha$. The distance between x_k and x_{k-1} is $\frac{1}{2}h(1+\alpha)$. In the simple situation of only one small grid cell we have $w_i^n = q_i^n$ for $i \neq k$ and $w_k^n = q_k^n + \frac{1}{2}(1-\alpha)hq_x(x_k, t^n)$. The truncation error of the grid function w for the scheme (3.2) has in the small cell the form

$$\begin{aligned} Lw &= \frac{w_k^{n+1} - w_k^n}{\triangle t} + a\frac{w_k^n - w_{k-1}^n}{h} \\ &= \frac{q_k^{n+1} + \frac{1}{2}(1-\alpha)hq_x(x_k, t^{n+1}) - q_k^n - \frac{1}{2}(1-\alpha)hq_x(x_k, t^n)}{\triangle t} \\ &\quad + \ a\frac{q_k^n + \frac{1}{2}(1-\alpha)hq_x(x_k, t^n) - q_{k-1}^n}{h} + \mathcal{O}(\triangle t, h) \\ &= \frac{q_k^n + \triangle t q_t(x_k, t^n) + \frac{1}{2}(1-\alpha)hq_x(x_k, t^n) - q_k^n - \frac{1}{2}(1-\alpha)hq_x(x_k, t^n)}{\triangle t} \\ &\quad + \ a\ \frac{q_k^n + \frac{1}{2}(1-\alpha)hq_x(x_k, t^n) - q_k^n + \frac{1}{2}(1+\alpha)hq_x(x_k, t^n)}{h} + \mathcal{O}(\triangle t, h) \\ &= q_t(x_k, t^n) + aq_x(x_k, t^n) + \mathcal{O}(\triangle t, h) \ = \ \mathcal{O}(\triangle t, h). \end{aligned}$$

From the truncation error of w and the stability of the method for $\mathrm{CFL}_h \leq 1$ it follows that $|w_k - Q_k| = \mathcal{O}(\triangle t, h)$. Since $w = q + \mathcal{O}(h)$ we obtain the estimate

$$|q_k - Q_k| = \mathcal{O}(h);$$

i.e., the h-box method (3.2) leads to a first order accurate approximation of the advection equation in the small cell k, in spite of the fact that the scheme is inconsistent in the small cell. Using the same grid function w one can also show that the truncation error Lw in cell $k+1$ is of the order $\mathcal{O}(h)$. In all other regularly spaced grid cells, the method agrees with the upwind scheme for which the truncation error is also $\mathcal{O}(h)$. Therefore, we obtain first order convergence in the whole domain. □

In order to obtain a more accurate small cell scheme, we will now consider the construction of h-box values based on linear interpolation using again grid cell values that are overlapped by the h-boxes. Such h-box values have the general form

$$\begin{aligned} Q^L_{k-\frac{1}{2}} &= Q_{k-1}, & Q^R_{k-\frac{1}{2}} &= \lambda Q_k + (1-\lambda)Q_{k+1}, \\ Q^L_{k+\frac{1}{2}} &= \lambda Q_k + (1-\lambda)Q_{k-1}, & Q^R_{k+\frac{1}{2}} &= Q_{k+1}. \end{aligned}$$

We want to determine λ so that we obtain a consistent h-box scheme, i.e., for which $Lq = \mathcal{O}(h, \triangle t)$. By again using Taylor series expansion we find that only $\lambda = \frac{2\alpha}{1+\alpha}$ leads to an upwind method that satisfies this condition. This suggests that the h-box values should have the form

$$\begin{aligned} Q^L_{k-\frac{1}{2}} &= Q_{k-1}, & Q^R_{k-\frac{1}{2}} &= \frac{2\ \alpha\ Q_k + (1-\alpha)\ Q_{k+1}}{1+\alpha}, \\ Q^L_{k+\frac{1}{2}} &= \frac{2\ \alpha\ Q_k + (1-\alpha)\ Q_{k-1}}{1+\alpha}, & Q^R_{k+\frac{1}{2}} &= Q_{k+1}. \end{aligned} \tag{3.3}$$

Note that this interpolation formula was already given in [4] but not further investigated there. In [7], [28] h-box values were defined in a similar way, and the resulting scheme was shown to give good results for advection and Burgers's equation.

One time step of the h-box method based on the interpolation formula (3.3) again for $a > 0$ has in the small cell the form

$$Q_k^{n+1} = Q_k^n - \frac{a\triangle t}{h} \cdot \frac{Q_k^n - Q_{k-1}^n}{(1+\alpha)/2}. \tag{3.4}$$

We can derive the same method as a finite difference scheme by replacing $q_x(x_k, t^n)$ in the Taylor series expansion of

$$\begin{aligned} q(x_k, t^n + \triangle t) &= q(x_k, t^n) + \triangle t q_t(x_k, t^n) + \mathcal{O}(\triangle t^2) \\ &= q(x_k, t^n) - \triangle t \cdot a\ q_x(x_k, t^n) + \mathcal{O}(\triangle t^2) \end{aligned} \tag{3.5}$$

by an appropriate first order accurate finite difference formula. The *h*-box method (3.4) can be interpreted as a finite difference scheme that approximates the $q_x(x_k)$ terms by one-sided finite differences. This *h*-box method leads to a first order accurate method that approximates linear functions exactly. One can also show that an upwind scheme based on the *h*-box values (3.3) also leads to a consistent first order accurate update in the two neighboring grid cells of the small cell.

If we use the wave propagation algorithm, then the first order update in the small cell can be written in the form

$$Q_k^{n+1} = Q_k^n - \frac{\triangle t}{\alpha h}\left(\mathcal{A}^+\triangle\hat{Q}_{k-\frac{1}{2}} - f(Q^R_{k-\frac{1}{2}}) + \mathcal{A}^-\triangle\hat{Q}_{k+\frac{1}{2}} + f(Q^L_{k+\frac{1}{2}})\right), \tag{3.6}$$

with $\triangle\hat{Q}_{k-\frac{1}{2}} = Q^R_{k-\frac{1}{2}} - Q^L_{k-\frac{1}{2}}$ and $\triangle\hat{Q}_{k+\frac{1}{2}} = Q^R_{k+\frac{1}{2}} - Q^L_{k+\frac{1}{2}}$. In the limit case $\alpha = 1$ we have $Q^R_{k-\frac{1}{2}} = Q^L_{k+\frac{1}{2}}$, and (3.6) reduces to the first order accurate wave propagation algorithm that is valid in the regular parts of the grid. This formula remains valid for nonlinear equations as well as systems of conservation laws, assuming we have a Riemann solver that provides us a decomposition of $Q^R - Q^L$, as described in section 2. We indicate quantities that are calculated from *h*-box values by the "ˆ" symbol.

Numerical results shown in section 5 will demonstrate the superior properties of an *h*-box method with *h*-boxes obtained by linear interpolation.

3.2. A second order accurate *h*-box method. In order to obtain a high-resolution scheme we want to include second order correction terms. This means we want to obtain an update of the small cell that can be written as

$$\begin{aligned} Q_k^{n+1} = Q_k^n &- \frac{\triangle t}{\alpha h}\left(\mathcal{A}^+\triangle\hat{Q}_{k-\frac{1}{2}} - f(Q^R_{k-\frac{1}{2}}) + \mathcal{A}^-\triangle\hat{Q}_{k+\frac{1}{2}} + f(Q^L_{k+\frac{1}{2}})\right) \\ &- \frac{\triangle t}{\alpha h}\left(\hat{F}^2_{k+\frac{1}{2}} - \hat{F}^2_{k-\frac{1}{2}}\right), \end{aligned}$$

where $\hat{F}^2$ denotes the second order correction terms that are implemented in flux differencing form. By analogy to the standard wave propagation algorithm, these second order correction terms should also be calculated by using the waves and speeds obtained from solving Riemann problems at the cell interfaces. For the small cell we again use the waves and speeds from Riemann problems defined by the same *h*-box values used to obtain the first order update. We will restrict our consideration to *h*-box values that are calculated using the interpolation formula (3.3).

The formula (2.4) for the second order correction flux on irregular grids suggests using correction terms of the form

$$\hat{F}^2_{i+\frac{1}{2}} = \frac{1}{2}\sum_{p=1}^{M_w}\left(\frac{1}{(1+\alpha)/2} - \frac{\triangle t}{(1+\alpha)h/2}|\hat{s}^p_{i+\frac{1}{2}}|\right)\cdot|\hat{s}^p_{i+\frac{1}{2}}|\cdot\hat{\mathcal{W}}^p_{i+\frac{1}{2}} \qquad (i = k-1,\ k) \tag{3.7}$$

in the small cell. The waves $\hat{\mathcal{W}}^p_{i+\frac{1}{2}}$ and the speeds $\hat{s}^p_{i+\frac{1}{2}}$ can be obtained by solving Riemann problems defined by the *h*-box values at the small cell interfaces. One can show that the truncation error in the small cell that results from such a high-resolution wave propagation scheme is $Lq_k = \mathcal{O}(h^2, \triangle t^2)$; i.e., assuming the scheme is stable we would obtain a second order accurate approximation in the small cell. However, numerical tests showed that such an approach is not stable for time steps satisfying $\mathrm{CFL}_h \leq 1$.

Instead we use second order correction terms of the form

$$\hat{F}^2_{i+\frac{1}{2}} = \frac{1}{2}\sum_{p=1}^{M_w}\left(1 - \frac{\triangle t}{h}|\hat{s}^p_{i+\frac{1}{2}}|\right) \cdot |\hat{s}^p_{i+\frac{1}{2}}|\hat{\mathcal{W}}^p_{i+\frac{1}{2}} \qquad (i = k-1,\ k). \tag{3.8}$$

The waves are again calculated from Riemann problems defined by the h-box values. The difference from (3.7) is that we do not take the size of the small cell into account in the calculation of the correction fluxes. This reflects the general concept of the h-box method where fluxes are calculated from values defined over regions of length h.

Although the truncation error for the grid cell k now contains first order terms which do not cancel out, the numerical results are very satisfying and indicate second order convergence as well as stability for $\mathrm{CFL}_h \leq 1$. Assuming that the solution is sufficiently smooth we can indeed prove that the resulting method leads to a second order accurate approximation for the advection equation.

PROPOSITION 2. *We consider the approximation of the advection equation on an almost uniform grid with mesh width h that contains one small mesh cell of length αh, with $\alpha \leq 1$. The h-box method consisting of the first order update* (3.4) *and the second order correction terms* (3.8) *(without limiters) leads to a second order accurate approximation for sufficiently smooth solutions of the advection equation.*

Proof. We again use the idea of Wendroff and White and consider the truncation error Lw for a grid function of the form $w^n_i = q^n_i + \frac{1}{8}h^2(\alpha_i+1)(\alpha_i-1)q_{xx}(x_i, t^n)$. Here we assume that $\triangle x_i = \alpha_i h$. We have $\alpha_i = 1$ for $i \neq k$ and $\alpha_k = \alpha$. In regular grid cells $i \neq k$ the grid function w agrees with the exact solution. We want to show only second order convergence in the small cell as well as in the two neighboring cells $k-1$ and $k+1$, since the method reduces to the high-resolution wave propagation algorithm in the other regular grid cells. In the case considered here, the wave propagation algorithm on the regular part of the grid is equivalent to the Lax–Wendroff scheme.

The truncation error for the grid function w has the form

$$\begin{aligned} Lw &= \frac{w^{n+1}_k - w^n_k}{\triangle t} + 2a\frac{w^n_k - w^n_{k-1}}{h(1+\alpha)} + \left(1 - a\frac{\triangle t}{h}\right)a\frac{w^n_{k+1} - 2w^n_k + w^n_{k-1}}{h(1+\alpha)} \\ &= q_t(x_k, t^n) + \frac{\triangle t}{2}q_{tt}(x_k, t^n) + aq_x(x_k, t^n) + \frac{1}{4}h(\alpha - 1)aq_{xx}(x_k, t^n) \\ &\quad - \frac{1}{4}h(1+\alpha)aq_{xx}(x_k, t^n) \\ &\quad + \left(1 - a\frac{\triangle t}{h}\right)a\frac{\frac{1}{4}h^2(1+\alpha)^2 q_{xx}(x_k, t^n) - \frac{1}{4}h^2(\alpha^2 - 1)q_{xx}(x_k, t^n)}{h(1+\alpha)} + \mathcal{O}(\triangle t^2, h^2). \end{aligned}$$

Here we use $h_{k+\frac{1}{2}} = h_{k-\frac{1}{2}} = \frac{1}{2}h(1+\alpha)$ for the distance from the cell center of the small cell k to the cell centers of the neighboring cells. By using the relations $q_t(x_k, t^n) = -aq_x(x_k, t^n)$ and $q_{tt}(x_k, t^n) = a^2 q_{xx}(x_k, t^n)$ all lower order terms in the above equation cancel and we obtain $Lw = \mathcal{O}(\triangle t^2, h^2)$. This shows that $|w_k - Q_k| = \mathcal{O}(\triangle t^2, h^2)$. Since the grid function was chosen to satisfy $w = q + \mathcal{O}(h^2)$, we conclude that

$$|q_k - Q_k| = \mathcal{O}(\triangle t^2, h^2).$$

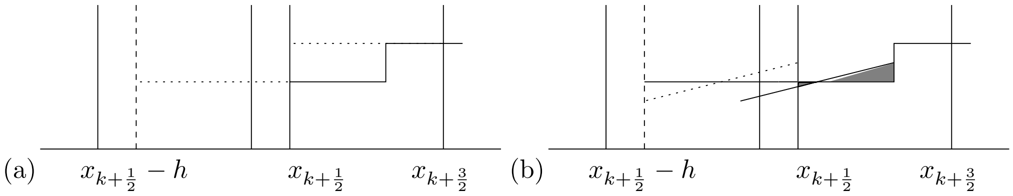


FIG. 2. *h-box values at the interface $x_{k+\frac{1}{2}}$; dotted lines depict the initial values, solid lines the solution after one time step.* (a) *first order update by h-box method;* (b) *second order correction wave of $Q^L_{i+\frac{1}{2}}$.*

Stability of this second order accurate scheme will be shown in the appendix.

Using the same grid function w, one can also show that $Lw = \mathcal{O}(\triangle t^2, h^2)$ in the neighboring grid cells $k-1$ and $k+1$. Therefore, the numerical solution converges with second order accuracy in the whole domain. □

Figure 2 shows a schematic description of the first order update and the high-resolution correction for cell $k+1$. The dotted lines depict the initial values, i.e., $Q^L_{k+\frac{1}{2}}$ and $Q^R_{k+\frac{1}{2}} = Q_{k+1}$. In a first step the piecewise constant values are propagated over a distance $a\triangle t$, as shown in Figure 2(a). In order to increase the accuracy, the piecewise constant initial values are replaced by piecewise linear functions. In Figure 2(b), we show the piecewise linear reconstructed function $Q^L_{k+1}(x, t^n)$ that has the slope $\sigma = (Q^R_{k+\frac{1}{2}} - Q^L_{k+\frac{1}{2}})/h$. Since we already calculated the first order update, the second order correction terms, calculated by propagating piecewise linear initial values $Q^L_{k+\frac{1}{2}}(x, t^n)$ instead of the piecewise constant value $Q^L_{k+\frac{1}{2}}$, take only the shaded region shown in Figure 2(b) into account. Compare with LeVeque [17], where such second order correction terms were described for the approximation of the advection equation on a uniform grid.

3.3. Limiters for the h-box method. In order to have control over unphysical oscillations near discontinuities some kind of limiters must be used in the second order correction terms (2.4). In the wave propagation algorithm this is done by using *wave limiters* that modify the magnitude of the waves $\mathcal{W}^p$ $(p = 1, \dots, M_w)$ in the fluxes that model the second order correction terms. A limited p-wave $\mathcal{W}^p_{i+\frac{1}{2}}$ is obtained by comparison of this wave with the neighboring p-waves $\mathcal{W}^p_{i-\frac{1}{2}}$ or $\mathcal{W}^p_{i+\frac{3}{2}}$, depending on the direction of flow; see LeVeque [18] or [19] for details.

In our high-resolution h-box method we can use the same limiting process in order to obtain limited versions of the waves that were calculated from h-box values. These limited waves can then be used in the second order correction fluxes (3.8). In order to obtain the limiter for waves at a small cell interface, we compare those waves with waves arising from Riemann problems at a distance h away from the cell interface. This can be done by constructing two additional h-boxes at the small cell interface. The waves resulting from the solution of Riemann problems defined by these new h-box values to the left- and right-hand side of a small cell interface can then be used in order to estimate the wave limiter for the waves at the small cell interface. This requires the solution of two additional Riemann problems for each small cell interface; see Figure 3. We used such a limiting process in order to approximate a shock wave solution on an irregular grid shown in section 5.

In addition to the wave limiting process we also include a limiter into the approximation of the h-box values. Note that the h-box values (3.3) can also be obtained by

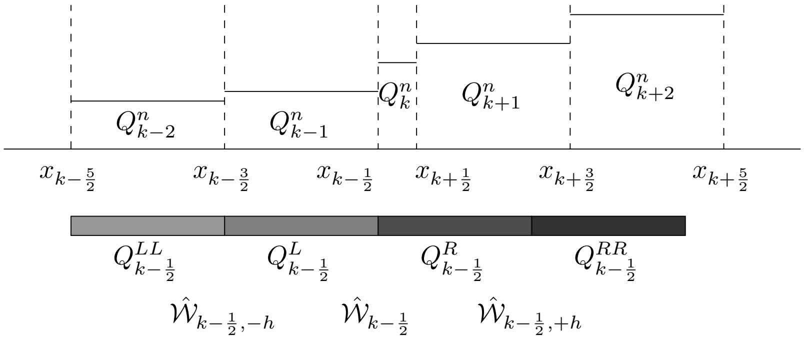


FIG. 3. *Schematic description of h-box values assigned to the left small cell interface. Two additional h-box values are needed to estimate the wave limiter for the second order correction terms.*

reconstructing a piecewise linear function $\overline{Q}(x)$ from the cell averages Q_i for all i and calculating the average value of this piecewise linear function over the same boxes of length h, as indicated in Figure 1. If the reconstructed function has the form

$$\overline{Q}_{k-1}(x) = Q_{k-1} + \frac{Q_k - Q_{k-1}}{\frac{1}{2}h(1+\alpha)}(x - x_{k-1}) \quad \text{for} \quad x \in [x_{k-\frac{3}{2}}, x_{k-\frac{1}{2}}), \tag{3.9}$$

$$\overline{Q}_{k+1}(x) = Q_{k+1} + \frac{Q_{k+1} - Q_k}{\frac{1}{2}h(1+\alpha)}(x - x_{k+1}) \quad \text{for} \quad x \in [x_{k+\frac{1}{2}}, x_{k+\frac{3}{2}}), \tag{3.10}$$

then averaging over boxes of length h leads to h-box values that have the form (3.3). The slopes of the piecewise linear initial values are

$$\sigma_{k-1} = \frac{Q_k - Q_{k-1}}{h(1+\alpha)/2} \quad \text{and} \quad \sigma_{k+1} = \frac{Q_{k+1} - Q_k}{h(1+\alpha)/2}.$$

Near discontinuities such piecewise linear values may not represent a good approximation of the solution. We can use standard *slope limiters* in order to obtain better approximations there. We can, for instance, use a slope limiter proposed by van Leer [29]. Here the slopes are replaced by limited versions that have the form $\hat{\sigma}_i = \sigma_i \phi_i$ for $i \in \{k-1, k+1\}$. For our application the limiter has the form

$$\phi_i(\theta_i) = \min\left(1, \frac{|\theta_i| + \theta_i}{1 + |\theta_i|}\right)$$

with

$$\theta_{k-1} = \frac{Q_{k-1} - Q_{k-2}}{Q_k - Q_{k-1}} \quad \text{and} \quad \theta_{k+1} = \frac{Q_{k+2} - Q_{k+1}}{Q_{k+1} - Q_k}.$$

It may be replaced by other limiter functions. Note that we do not want to use a steeper slope than σ_{k-1} (respectively, σ_{k+1}) for the construction of h-box values, because only those values lead to a second order approximation in smooth regions. However, near discontinuities we want to limit these slopes. The resulting limited h-box values can be calculated using the formulas

$$Q^L_{k+\frac{1}{2}} = \alpha Q_k + (1-\alpha)\left(Q_{k-1} + \frac{\alpha}{1+\alpha}(Q_k - Q_{k-1})\phi_{k-1}\right),$$

$$Q^R_{k-\frac{1}{2}} = \alpha Q_k + (1-\alpha)\left(Q_{k+1} + \frac{\alpha}{1+\alpha}(Q_k - Q_{k+1})\phi_{k+1}\right).$$

4. On the stability of the *h*-box method. The h-box method retains stability by constructing a finite volume scheme for which the flux difference is of the order of the size of the grid cell. For a small grid cell this requires $F_{k+\frac{1}{2}} - F_{k-\frac{1}{2}} = \mathcal{O}(\alpha h)$. In this case the term αh arising in the denominator of the finite volume scheme should not cause a stability problem. In regions where the solution of the conservation law is smooth, the h-box values are constructed to satisfy an analogous property, namely $Q^{L,R}_{k+\frac{1}{2}} - Q^{L,R}_{k-\frac{1}{2}} = \mathcal{O}(\alpha h)$. Since in our applications the flux function is a Lipschitz continuous function of Q^L and Q^R, the flux difference has the required cancellation property; see [5].

For the advection equation Stern [28] proved that the first order accurate h-box methods are TVD. Here we will briefly outline this proof which follows the general concept described above. The first order h-box method can (for $a > 0$) be rewritten in the form

$$\begin{aligned} Q_i^{n+1} &= Q_i^n - \frac{a\triangle t}{\alpha_i h}(Q^L_{i+\frac{1}{2}} - Q^L_{i-\frac{1}{2}}) \\ &= Q_i^n - \frac{a\triangle t}{\alpha_i h}\left(\alpha_i Q_i^n - \alpha_i \underbrace{\frac{1}{\alpha_i h}\int_{x_{i-\frac{1}{2}}-h}^{x_{i-\frac{1}{2}}-h+\alpha_i h} \overline{Q}^n_{i-1}(x)dx}_{\tilde{Q}_i^n}\right). \end{aligned}$$

Here we assume that each grid cell has the size $h_i = \alpha_i h$, with $0 < \alpha_i \leq 1$. $\overline{Q}^n_{i-1}(x)$ is the piecewise linear reconstructed function (3.9). The stability result also holds on an irregular grid with more than one small cell. See also section 5 for a slightly different generalization of the piecewise linear function that has to be used in the construction of h-box values for a completely irregular grid.

Using this notation we now consider the difference $|Q^{n+1}_{i+1} - Q^{n+1}_i|$ and sum over all grid cells. This sum can be estimated as

$$\begin{aligned} TV(Q^{n+1}) &= \sum_i |Q^{n+1}_{i+1} - Q^{n+1}_i| \\ &\leq \left(1 - \frac{a\triangle t}{h}\right)\sum_i |Q^n_{i+1} - Q^n_i| + \frac{a\triangle t}{h}\sum_i |\tilde{Q}^n_{i+1} - \tilde{Q}^n_i|. \end{aligned}$$

We obtain the TVD property $TV(Q^{n+1}) \leq TV(Q^n)$ for time steps $\text{CFL}_h \leq 1$ if

$$\sum_i |\tilde{Q}^n_{i+1} - \tilde{Q}^n_i| \leq TV(Q^n). \tag{4.1}$$

For the h-box method (3.2) using h-box values that were calculated by averaging over piecewise constant values, (4.1) is always satisfied, since $\tilde{Q}_i^n = Q_i^n$. For the more accurate first order h-box method (3.4), the TVD property can be shown if a TVD slope limiter is used in the construction of the h-box values, as discussed in section 3.3.

Note that for the approximation of the advection equation, the first order h-box method based on h-box values (3.1), i.e., defined by averaging over piecewise constant values of the conserved quantities, is also monotone. This property does not carry over to the first order h-box method with h-box values calculated by the interpolation formula (3.3). Note also that none of these two first order accurate h-box methods applied to Burgers's equation leads to a monotone method.

In the appendix we show stability for the second order accurate h-box method applied to the advection equation. This proof is based on the stability theory of Gustafsson, Kreiss, and Sundström [11].

5. Irregular grid calculation. In order to demonstrate the robustness of the high-resolution h-box method we now apply the scheme to a completely arbitrary grid; see Figure 4. By again assigning values to boxes of length h, we obtain a scheme that remains stable for time steps appropriate for a uniform grid with grid cells of length h. In this more general situation more than two grid cells may be overlapped by an h-box. We assume that grid cells have the length $h_i = \alpha_i h$, with $\alpha_i \leq 1$ for all indices i. We will show that a generalization of the h-box method based on averaging over piecewise linear values of the conserved quantities gives accurate results also in this more difficult situation. We will need to use piecewise linear reconstructed values

$$\overline{Q}_i(x) = Q_i + \frac{Q_{i+1} - Q_i}{h(\alpha_i + \alpha_{i+1})/2}(x - x_i) \quad \text{for } x \in [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}), \quad i \in \{m, l\}, \tag{5.1}$$

$$\overline{Q}_j(x) = Q_j + \frac{Q_j - Q_{j-1}}{h(\alpha_j + \alpha_{j-1})/2}(x - x_j) \quad \text{for } x \in [x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}), \quad j \in \{s, t\}. \tag{5.2}$$

The indices m, l and s, t indicate the grid cells that are only partly covered by the left- (respectively, right-) going h-boxes that are constructed at the cell interfaces of grid cell k. Slopes are needed only in these four cells because averaging over an entire cell gives a value that is independent of the slope. Averaging over these piecewise linear functions leads to the h-box values

$$\begin{aligned}
Q^L_{k-\frac{1}{2}} &= \sum_{i=m+1}^{k-1} \alpha_i Q_i + \Big(1 - \sum_{i=m+1}^{k-1} \alpha_i\Big) \cdot \Big[Q_m + \frac{Q_{m+1} - Q_m}{\alpha_m + \alpha_{m+1}} \Big(\sum_{i=m}^{k-1} \alpha_i - 1\Big)\Big], \\
Q^R_{k-\frac{1}{2}} &= \sum_{i=k}^{s-1} \alpha_i Q_i + \Big(1 - \sum_{i=k}^{s-1} \alpha_i\Big) \cdot \Big[Q_s + \frac{Q_s - Q_{s-1}}{\alpha_s + \alpha_{s-1}} \Big(1 - \sum_{i=k}^{s} \alpha_i\Big)\Big], \\
Q^L_{k+\frac{1}{2}} &= \sum_{i=l+1}^{k} \alpha_i Q_i + \Big(1 - \sum_{i=l+1}^{k} \alpha_i\Big) \cdot \Big[Q_l + \frac{Q_{l+1} - Q_l}{\alpha_l + \alpha_{l+1}} \Big(\sum_{i=l}^{k} \alpha_i - 1\Big)\Big], \\
Q^R_{k+\frac{1}{2}} &= \sum_{i=k+1}^{t-1} \alpha_i Q_i + \Big(1 - \sum_{i=k+1}^{t-1} \alpha_i\Big) \cdot \Big[Q_t + \frac{Q_t - Q_{t-1}}{\alpha_t + \alpha_{t-1}} \Big(1 - \sum_{i=k+1}^{t} \alpha_i\Big)\Big].
\end{aligned} \tag{5.3}$$

5.1. Approximation of the advection equation on irregular grids. We can show that these h-box values used in an upwind scheme (which is equivalent to the first order wave propagation algorithm) lead to a consistent approximation of the advection equation.

PROPOSITION 3. *The h-box method $Q_i^{n+1} = Q_i^n - a\frac{\triangle t}{\alpha_i h}(Q^L_{i+\frac{1}{2}} - Q^L_{i-\frac{1}{2}})$ with h-box values defined by* (5.3) *leads to a first order accurate approximation of the advection equation (with advection speed $a > 0$) on an irregular grid.*

The proof is based on Taylor series expansion and may be found in the preprint version of this paper [2]. Together with the stability result mentioned in section 4, we obtain first order convergence of this h-box method on irregular grids using time steps that satisfy $\text{CFL}_h \leq 1$.

Once the h-box values are defined we can apply the same second order correction terms (3.8) at the cell interfaces of a completely irregular grid. With such an approach

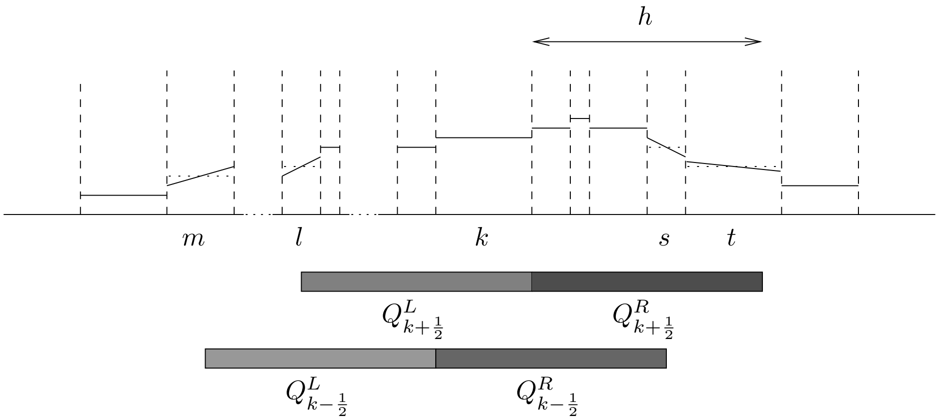


FIG. 4. *Schematic description of the h-box method on a completely irregular grid.*

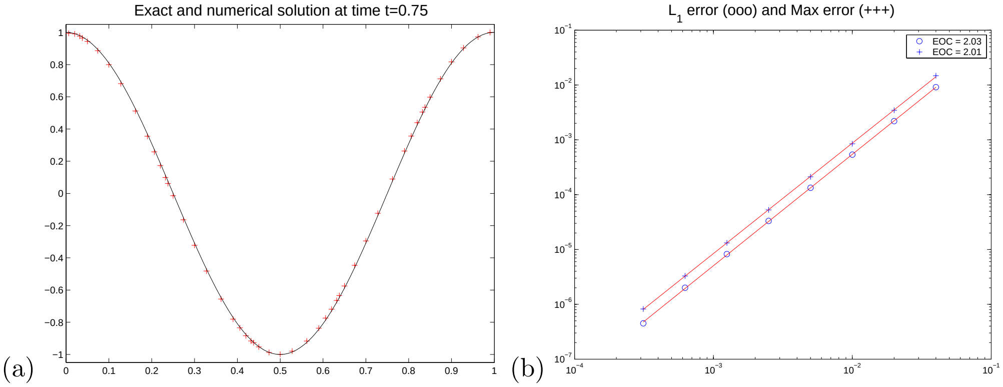


FIG. 5. *Approximation of the advection equation on an irregular grid using the high-resolution h-box method with h-box values calculated by linear interpolation.* (a) *numerical results on an irregular grid with $h = 0.04$;* (b) *log-log-plot of h versus L_1-norm error as well as maximum-norm error shows second order convergence.*

we can expect second order convergence. Figure 5 shows numerical results for the approximation of the advection equation on a sequence of irregular grids. The initial values are set to $q(x, 0) = \sin(2\pi x)$ on the interval $[0, 1]$. Periodic boundary conditions are imposed. A convergence study shows that our new high-resolution h-box method converges with second order accuracy both in the L_1-norm as well as the maximum norm. The accuracy of this calculation compares well with the accuracy of the standard wave propagation algorithm that was briefly described in section 2. However, here we could use much larger time steps. In Figure 6, we show results for the same test case, but here the h-box values were constructed by averaging over piecewise constant values of the conserved quantity, i.e., the formally inconsistent method described in section 3.1. Although we add second order correction terms (which increases the accuracy) the resulting method is only first order accurate. This is analogous to our analytical results for the simpler situation with only one small cell.

5.2. Approximation of the Euler equations on irregular grids. In this section we study the performance of the high-resolution h-box method for one-dimensional

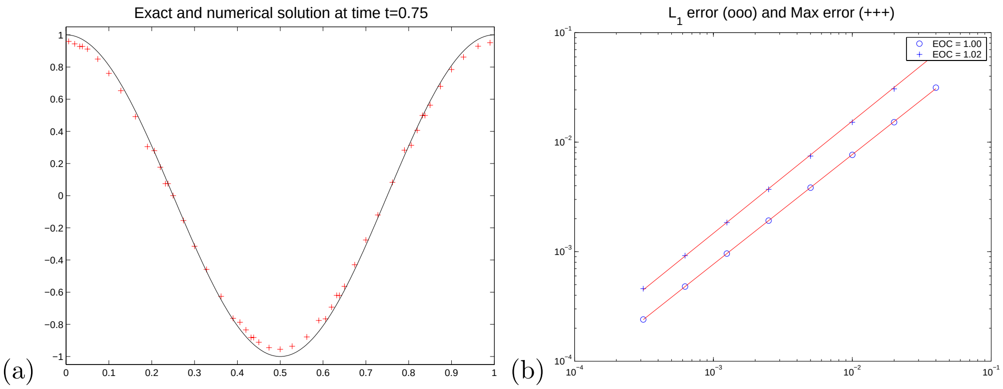


FIG. 6. *Approximation of the advection equation on an irregular grid using an h-box method with second order correction terms, where h-box values are calculated by averaging over piecewise constant values of the conserved quantity.* (a) *numerical results on an irregular grid with $h = 0.04$;* (b) *log-log-plot of h versus L_1-norm error as well as maximum-norm error shows first order convergence.*

Euler equations. The equations can be written in the form (1.1) with

$$q = (\rho, \rho u, E), \quad f(q) = (\rho u, \rho u^2 + p, u(E + p)),$$

where $\rho, p, E,$ and u describe the density, pressure, total energy, and the velocity, respectively. The equation of state has the form

$$E = \frac{p}{\gamma - 1} + \frac{1}{2}\rho u^2.$$

First we consider the approximation of a test problem defined in Example 5.1 on an irregular grid.

Example 5.1. We consider the numerical approximation of the one-dimensional Euler equations on an irregular grid. The grid cells vary in size between $h/10$ and h. The initial values are sufficiently smooth so that the solution does not develop shocks over the time interval considered. Reflecting boundary conditions are imposed on the left and right boundary. The computational domain is the interval $[0, 1]$. Our initial values are

$$\rho(x, 0) = 1 + 0.4 \sin\left(\frac{\pi}{2} + x\pi\right), \quad u(x, 0) = 0.25 - (x - 0.5)^2, \quad p(x, 0) = 1.$$

The ratio of specific heats is set to $\gamma = 1.4$.

In Figure 7 we show numerical results for the approximation of Example 5.1 using our new high-resolution h-box method. A convergence study for density at different time steps is shown in Table 5.1. Here we compare the numerical solution for density on a sequence of irregular grids to a highly resolved reference solution that was calculated on a regular spaced grid. We show results for both the unlimited second order h-box method and a version using the minmod limiter. Next we consider the approximation of a shock wave with the Euler equations.

Example 5.2. We consider the one-dimensional Euler equations with initial values on the interval $[0, 1]$ that have constant density $\rho = 1$ and constant pressure $p = 1$. The velocity is set to $u = 1$ for $x < 0.5$ and $u = -1$ for $x > 0.5$. The ratio of specific heats is $\gamma = 1.05$. The exact solution of this problem consists of two symmetric

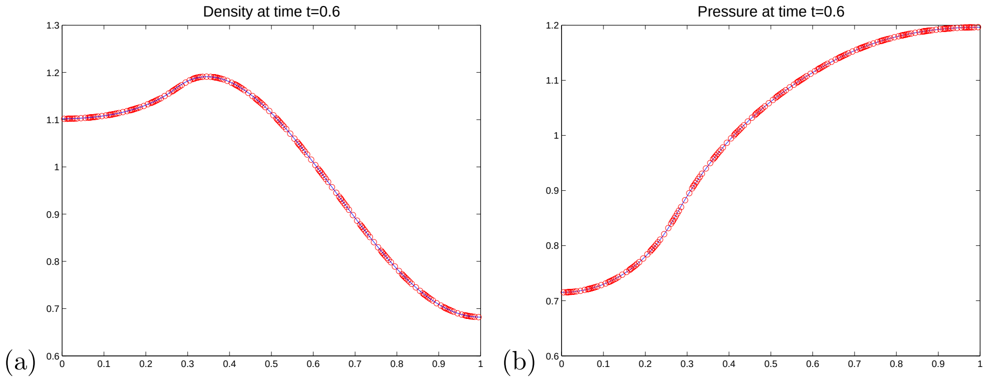


FIG. 7. *Numerical results of density and pressure for Example* 5.1 *on an irregular grid* ($h = 0.02$). *The solid line shows a highly resolved reference solution calculated on a regular grid.*

TABLE 5.1

Convergence study for Example 5.1. L_1 *error of density at different times as well as the experimental order of convergence (EOC) are shown. For this smooth test problem, we show results for the unlimited second order h-box method as well as the limited h-box method using a minmod limiter.*

	t=0.2	t = 0.4	t = 0.6	t = 0.8	t = 1
h/EOC	L_1 *error of density (unlimited method)*				
0.02	1.1229d-4	1.544d-4	3.4573d-4	5.9017d-4	0.0014
0.01	2.9567d-5	4.2550d-5	9.4628d-5	1.7825d-4	4.2628d-4
EOC	**1.92**	**1.86**	**1.87**	**1.73**	**1.72**
0.005	7.7282d-6	1.1786d-5	2.5092d-5	5.1242d-5	1.3381d-4
EOC	**1.94**	**1.89**	**1.91**	**1.80**	**1.67**
h/EOC	L_1 *error of density (using minmod limiter)*				
0.02	1.6893d-4	2.0212d-4	3.1620d-4	5.2083d-4	0.0012
0.01	5.6937d-5	6.5761d-5	1.1105d-4	1.9282d-4	3.6960d-4
EOC	**1.57**	**1.62**	**1.51**	**1.46**	**1.70**
0.005	1.6357d-5	2.1260d-5	4.5036d-5	7.6802d-5	1.2018d-4
EOC	**1.80**	**1.63**	**1.30**	**1.33**	**1.62**

shock waves that are propagating outwards. We use an irregular grid with grid cells that may be smaller than $h = 0.01$ on the left half of the interval. For $x > 0.5$ the grid is regular with mesh length $\triangle x = 0.01$. We use time steps that correspond to $\mathrm{CFL}_h \approx 0.9$.

Figure 8 shows numerical results of Example 5.2 for the high-resolution *h*-box method based on the linear interpolation formula. Our numerical results in Figure 8(a) show that the limiters described in section 3.3 can suppress spurious oscillations near the discontinuity. The approximation of the shock wave that is moving into the region of the irregular grid is in good agreement with the symmetric shock wave that is moving into the regular part of the grid. On the irregular grid the shock is smeared out over more grid cells than on the regular grid. The reason for this more smeared out shock profile is that a jump in the conserved quantities can influence several *h*-box values. In Figure 8(b) we show the results obtained by the second order method without limiters.

6. Approximation of transonic rarefaction waves. In this section we point out that *h*-box methods can cause numerical difficulties in the approximation of tran-

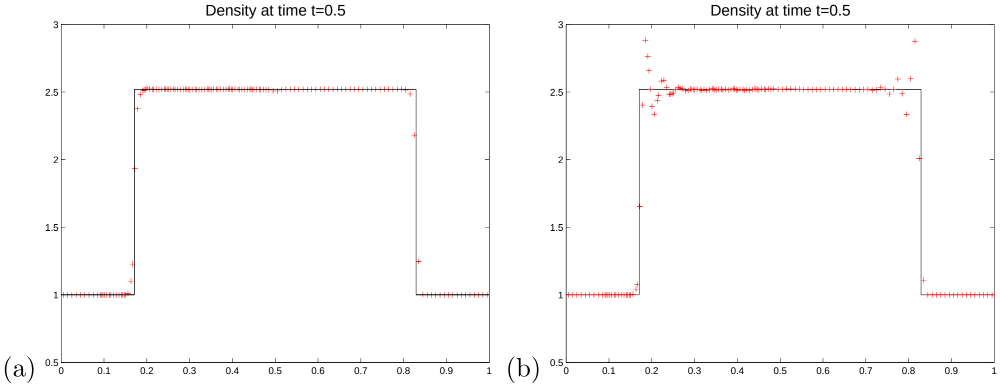


FIG. 8. *Approximation of Example* 5.2 *on a grid that is irregular for $x < 0.5$ and regular for $x > 0.5$.* (a) *plot of density with limiters;* (b) *plot of density without limiters. The solid line indicates the exact solution.*

sonic rarefaction waves that do not appear for standard Godunov-type methods on regular or irregular spaced grids. To see this we first consider the approximation of Burgers's equation $q_t + (q^2/2)_x = 0$ with initial values $q(x, 0) = -0.5$ for $x \leq 0.5$ and $q(x, 0) = 0.5$ for $x > 0.5$ on an irregular grid. The first order accurate fluxes at the cell interface $x_{i-\frac{1}{2}}$ can be calculated by using the exact formula, i.e.,

$$F_{i-\frac{1}{2}} = \begin{cases} \min_{Q^L_{i-\frac{1}{2}} \leq q \leq Q^R_{i-\frac{1}{2}}} f(q) & : \quad Q^L_{i-\frac{1}{2}} \leq Q^R_{i-\frac{1}{2}}, \\ \max_{Q^R_{i-\frac{1}{2}} \leq q \leq Q^L_{i-\frac{1}{2}}} f(q) & : \quad Q^R_{i-\frac{1}{2}} \leq Q^L_{i-\frac{1}{2}}, \end{cases}$$

with the flux $f(q) = \frac{1}{2}q^2$. Figure 9(a) demonstrates that this method produces unphysical oscillations around the sonic point. Note that in this section we use only first order accurate methods to isolate this phenomenon from the flux limiting procedure. The numerical problem can be avoided by using the Lax–Friedrichs flux, which has at the interface $x_{i-\frac{1}{2}}$ the form

$$F_{i-\frac{1}{2}} = \frac{1}{2}\big(f(Q^L_{i-\frac{1}{2}}) + f(Q^R_{i-\frac{1}{2}})\big) + \frac{h}{2\triangle t}\big(Q^L_{i-\frac{1}{2}} - Q^R_{i-\frac{1}{2}}\big).$$

See Figure 9(b) for numerical results.

The same effect can also be observed in the approximation of a transonic rarefaction wave for the Euler equations. To see this we consider a shock tube problem for which the solution consists of a right-moving shock wave, a contact discontinuity, and a left-moving transonic rarefaction wave. The initial values are $\rho = 1$, $u = 0.75$, $p = 1$ for $x \leq 0.3$ and $\rho = 0.125$, $u = 0$, $p = 0.1$ for $x > 0.3$. The ratio of specific heats is $\gamma = 1.4$. For the numerical approximation we used a Roe–Riemann solver with standard entropy fix for transonic rarefaction waves. The results of this calculation are shown in Figure 10. The numerical solution shows some oscillations around the sonic point; see Figure 10(b) for a closer view of the region around the sonic point. If the fluxes at the cell interfaces are again calculated by the Lax–Friedrichs method this numerical problem does not arise; see Figure 11.

In the preprint [2] of this paper, we studied the entropy consistency of the h-box method for the approximation of Burgers's equation. For the h-box method with Godunov flux, we showed that a discrete entropy inequality is satisfied away from sonic

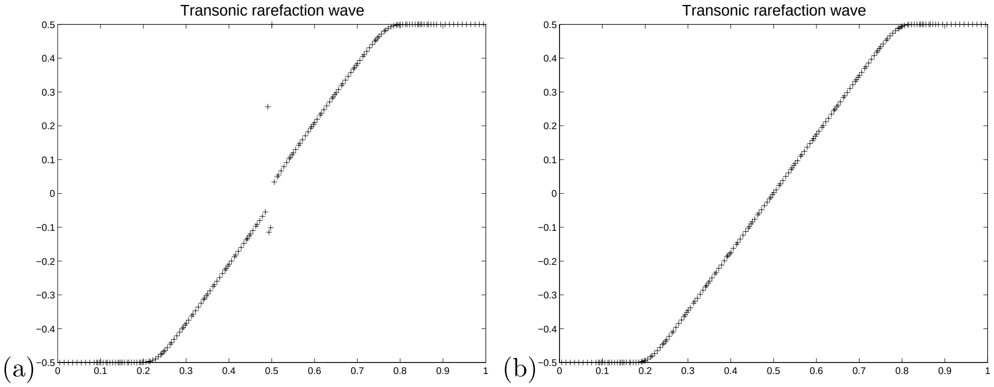


FIG. 9. *Approximation of a transonic rarefaction wave solution for Burgers's equation on an irregular grid.* (a) *h-box method based on Godunov flux;* (b) *h-box method based on Lax–Friedrichs flux.*

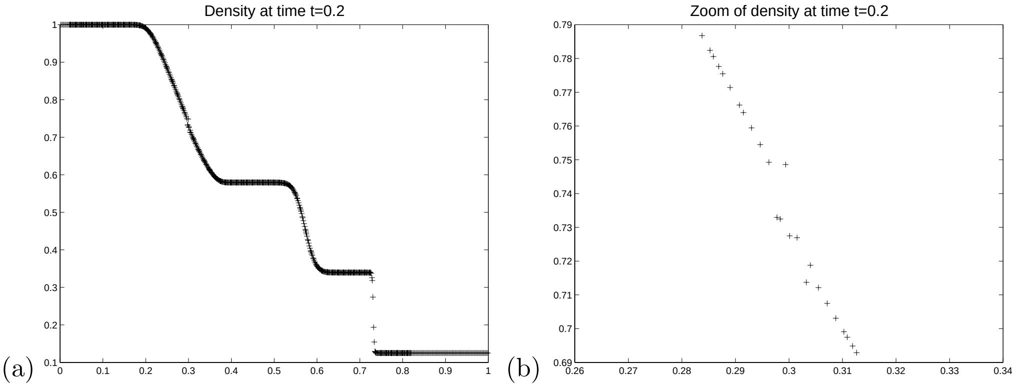


FIG. 10. *Approximation of a shock tube problem for the Euler equations.* (a) *plot of density obtained by first order Roe solver with entropy fix;* (b) *zoom of density around sonic point.*

points. This implies that the numerical solution converges to the entropy consistent weak solution of the conservation law. We showed that this discrete entropy inequality can be violated at a sonic point. While this does not give us any prediction whether or not the method is entropy consistent at the sonic point, it is interesting to note that this is exactly the case where the *h*-box method leads to numerical difficulties. We plan to further investigate the entropy consistency of *h*-box methods in order to develop an entropy fix that is less dissipative than the Lax–Friedrichs method and that can be extended to a high-resolution method.

7. Higher-dimensional irregular grid calculations. Now we will consider two-dimensional systems of conservation laws in the form

$$\frac{\partial}{\partial t}q(x,y,t)+\frac{\partial}{\partial x}f(q(x,y,t))+\frac{\partial}{\partial y}g(q(x,y,t))=0. \tag{7.1}$$

The simplest way to extend a one-dimensional method for conservation laws to multidimensional problems is to use dimension splitting. Equation (7.1) would be approximated by solving one-dimensional subproblems in an alternating way. The high-resolution one-dimensional *h*-box method could be used in each substep.

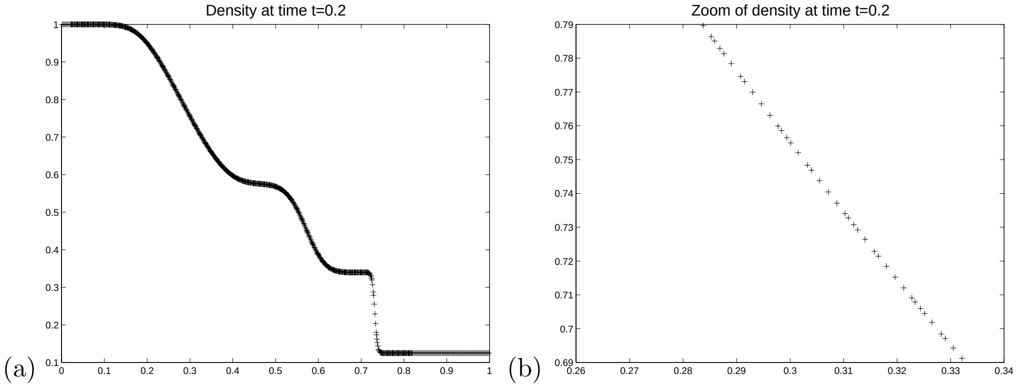


FIG. 11. *Approximation of a shock tube problem for the Euler equations.* (a) *plot of density obtained by Lax–Friedrichs method;* (b) *zoom of density around sonic point.*

Instead of using a dimensional splitting approach, we will here develop a two-dimensional h-box method that is based on the multidimensional wave propagation algorithm [17], [18]. We assume that the reader is familiar with the two-dimensional wave propagation algorithm and with the notation used below. As a first step in this approach we solve one-dimensional Riemann problems normal to each cell interface. Based on formula (3.6), which describes the one-dimensional h-box method, we obtain

$$\begin{aligned} Q_{ij}^{n+1} &= Q_{ij}^n + \triangle_{ij}^{up} \\ &= Q_{ij}^n - \frac{\triangle t}{\triangle x_i}\left(\mathcal{A}^+\triangle\hat{Q}_{i-\frac{1}{2},j} + \mathcal{A}^-\triangle\hat{Q}_{i+\frac{1}{2},j} + f(Q_{i+\frac{1}{2},j}^L) - f(Q_{i-\frac{1}{2},j}^R)\right) \\ &\quad - \frac{\triangle t}{\triangle y_j}\left(\mathcal{B}^+\triangle\hat{Q}_{i,j-\frac{1}{2}} + \mathcal{B}^-\triangle\hat{Q}_{i,j+\frac{1}{2}} + g(Q_{i,j+\frac{1}{2}}^L) - g(Q_{i,j-\frac{1}{2}}^R)\right). \end{aligned} \tag{7.2}$$

The method (7.2) is stable for time steps that satisfy $\text{CFL}_h \leq \frac{1}{2}$. Second order correction terms of the form (3.8) can be included in x as well as in y direction, which leads to a method of the form

$$Q_{ij}^{n+1} = Q_{ij}^n + \triangle_{ij}^{up} - \frac{\triangle t}{\triangle x_i}\left(\hat{F}_{i+\frac{1}{2},j}^2 - \hat{F}_{i-\frac{1}{2},j}^2\right) + \frac{\triangle t}{\triangle y_j}\left(\hat{G}_{i,j+\frac{1}{2}}^2 - \hat{G}_{i,j-\frac{1}{2}}^2\right). \tag{7.3}$$

The second order correction terms are again obtained by using the waves and speeds calculated from solving Riemann problems defined by h-box values. Limiters are used in exactly the same form as described earlier for the one-dimensional case.

In addition to fluxes in the normal direction, the multidimensional wave propagation algorithm also calculates waves that are moving in a transverse direction. For the usual wave propagation scheme one has $Q_{i+\frac{1}{2},j}^L = Q_{i-\frac{1}{2},j}^R$ and $Q_{i,j+\frac{1}{2}}^L = Q_{i,j-\frac{1}{2}}^R$. In this case the transverse propagation of waves can be obtained by a decomposition of the flux differences $\mathcal{A}^\pm\triangle Q$, $\mathcal{B}^\pm\triangle Q$ into transverse fluctuations. For the h-box method this transverse propagation has to be modified. In order to explain the transverse propagation we consider the two-dimensional advection equation

$$\frac{\partial}{\partial t}q(x,y,t) + a\frac{\partial}{\partial x}q(x,y,t) + b\frac{\partial}{\partial y}q(x,y,t) = 0, \qquad a, b > 0.$$

Assuming first that $\triangle x_i = h$ and $\triangle y_j \leq h$, the change of the cell average of the conserved quantity q in grid cell (i,j) due to the first order update in the x-direction

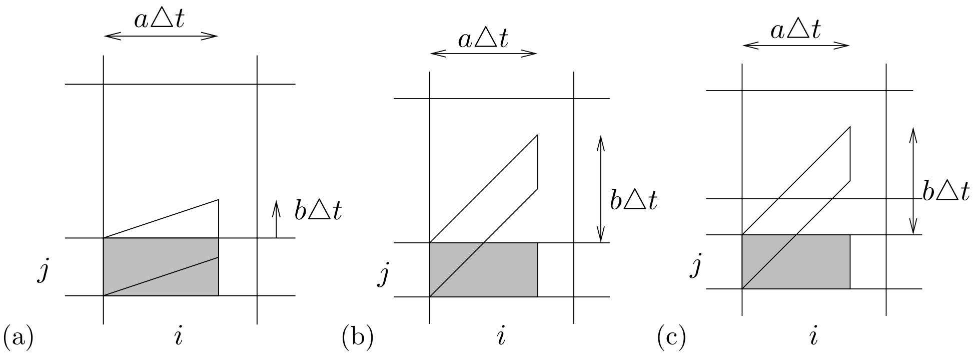


FIG. 12. *Different possibilities for transverse propagation of a right-moving wave for the advection equation on a nonuniform Cartesian grid.*

has the form

$$-\frac{\triangle t}{\triangle x_i}\mathcal{A}^+\triangle Q_{i-\frac{1}{2},j} = -\frac{\triangle t}{h}a(Q^n_{i,j} - Q^n_{i-1,j}). \tag{7.4}$$

Since we assume that the advection speed a in the x-direction is positive, there is no wave that moves into this cell from the right cell interface. Furthermore, the difference $f(Q^L_{i+\frac{1}{2},j}) - f(Q^R_{i-\frac{1}{2},j})$ vanishes in the case $\triangle x_i = h$. In the two-dimensional case a part of the right-moving flux difference $\mathcal{A}^+\triangle Q$ should affect other grid cells. This is indicated in Figure 12. The shaded regions indicate the influence of the jump $Q_{ij} - Q_{i-1,j}$ (initially located at the left cell interface) due to the solution of the Riemann problem in the normal direction. In a multidimensional method the solution of the Riemann problem at the interface $x_{i-\frac{1}{2}}$ should not affect only the cell average of the conserved quantities in the grid cell $(i-1,j)$ and (i,j). It should also have an effect on grid cells in the tangential direction. In the situation shown in Figure 12(a), the triangular portion of the wave describes the fraction that should affect the grid cell $(i,j+1)$. The transverse propagation of the wave considered in Figure 12(a) should change the cell average of the conserved quantity in grid cell (i,j) by the amount

$$\frac{(\triangle t)^2}{\triangle x_i \triangle y_j}\frac{1}{2}b\mathcal{A}^+\triangle Q_{i-\frac{1}{2},j} = \frac{(\triangle t)^2}{\triangle x_i \triangle y_j}\frac{1}{2}\mathcal{B}^+\mathcal{A}^+\triangle Q_{i-\frac{1}{2},j}.$$

The change of the cell average of the conserved quantity in cell $(i,j+1)$ due to the transverse propagation of this wave has the form

$$-\frac{(\triangle t)^2}{\triangle x_i \triangle y_{j+1}}\frac{1}{2}\mathcal{B}^+\mathcal{A}^+\triangle Q_{i-\frac{1}{2},j}.$$

The notation $\mathcal{B}^\pm\mathcal{A}^\pm\triangle Q$ was introduced in [18] to describe transverse propagations of left- and right-moving flux differences. For the wave propagation algorithm with time step restriction CFL ≤ 1 the transverse propagation has always the triangular form depicted in Figure 12(a), even if the grid is irregular. Since the transverse propagation approximates terms that are needed in order to obtain second order accuracy, we include those terms into the second order correction terms used in (7.3). The up-going flux difference $\mathcal{B}^+\mathcal{A}^+\triangle Q_{i-\frac{1}{2}}$ contributes to the $\tilde{G}$ term in the form of an update

$$\hat{G}^2_{i,j+\frac{1}{2}} := \hat{G}^2_{i,j+\frac{1}{2}} - \frac{1}{2}\frac{\triangle t}{\triangle x_i}\mathcal{B}^+\mathcal{A}^+\triangle Q_{i-\frac{1}{2},j}.$$

For our h-box method we have to extend the transverse propagation to allow also wave propagation of other forms, for instance those shown in Figures 12(b) or (c). For the situation shown in Figure 12(b) the update of the flux $\tilde{G}$ due to the transverse propagation has the form

$$\hat{G}^2_{i,j+\frac{1}{2}} := \hat{G}^2_{i,j+\frac{1}{2}} - \frac{\triangle t - \frac{1}{2}\triangle y_j/b}{b\triangle t\triangle x_i}\triangle y_j \mathcal{B}^+\mathcal{A}^+\triangle Q_{i-\frac{1}{2},j}.$$

In the situation shown in Figure 12(c), the transverse propagation of $\mathcal{A}^+\triangle Q_{i-\frac{1}{2},j}$ leads to an update of $\hat{G}^2_{i,j+\frac{1}{2}}$ as well as $\hat{G}^2_{i,j+\frac{3}{2}}$, depending on the fraction of the wave considered. As demonstrated in these examples, simple geometric routines can be used to calculate the fraction of the waves that determine the change of the cell average of the conserved quantity due to the transverse propagation. Note that the wave speed in the normal direction (i.e., a in our example) is present in the fluctuations $\mathcal{A}^\pm\triangle Q$. In order to calculate the transverse propagations no other information from the structure of the Riemann problem in the normal direction is needed. Therefore, even for a system of conservation laws, we have only to decompose the left- and right-moving flux differences, instead of decomposing each wave resulting from the Riemann problem in the normal direction separately.

So far we have assumed that $\triangle x_i = h$. If $\triangle x_i < h$, we want to use the one-dimensional h-box method in order to calculate the fluxes in the normal direction. The transverse propagation will take a very similar form as discussed above. Now we could interpret the grid cells (i,j), $(i,j+1)$ shown in Figure 12 as h-boxes constructed at the interface $x_{i-\frac{1}{2}}$. The transverse propagation of waves should again depend on the fraction of the wave that moves through the h-box considered. This can be calculated in exactly the same way as described above for the case $\triangle x_i = h$. In order to obtain the correct cancellation property needed for a stable update, we have to include the terms $f(Q^L_{i+\frac{1}{2},j})$ and $f(Q^R_{i-\frac{1}{2},j})$ that arise in (7.2) into our transverse propagation. Motivated by (2.1), (2.2) we do this by applying an update of the form

$$\begin{aligned}\mathcal{A}^+\triangle Q_{i-\frac{1}{2},j} &:= \mathcal{A}^+\triangle Q_{i-\frac{1}{2},j} - f(Q^R_{i-\frac{1}{2}}),\\ \mathcal{A}^-\triangle Q_{i-\frac{1}{2},j} &:= \mathcal{A}^-\triangle Q_{i-\frac{1}{2},j} + f(Q^L_{i-\frac{1}{2}})\end{aligned}$$

before we calculate the change of the fluxes $\hat{G}^2$. For our example of the advection equation with positive advection speeds, this update of $\mathcal{A}^\pm\triangle Q$ has the effect that $\mathcal{A}^-\triangle Q$ is no longer equal to zero. Moreover, the fraction of the wave that is propagated in the transverse direction depends only on the size of the grid cells and the speed b. Therefore, our transverse propagation has the effect that a fraction of the update used in (7.3) is propagated in the transverse direction. The update, which describes the wave propagation in the x-direction was already constructed to be of the order $\mathcal{O}(\triangle x)$ with $\triangle x \le h$. Our transverse propagation allows that at most a fraction of magnitude $\mathcal{O}(\triangle y)$ $(\triangle y \le h)$ is propagated in the transverse direction. (See, for instance, Figure 12(c).) Therefore, our transverse propagation satisfies the cancellation property. The transverse propagation of $\mathcal{B}^+\triangle Q$ also has to be included in an analogous way. By including the transverse propagation into our two-dimensional h-box method, we obtain stability for time steps that satisfy the condition $\text{CFL}_h \le 1$.

A transverse propagation of the second order correction (3.8) can be included into the transverse propagation in the same form as it was discussed for the wave propagation algorithm in [18]. This further increases the accuracy of the method. It was used in our test calculations below.

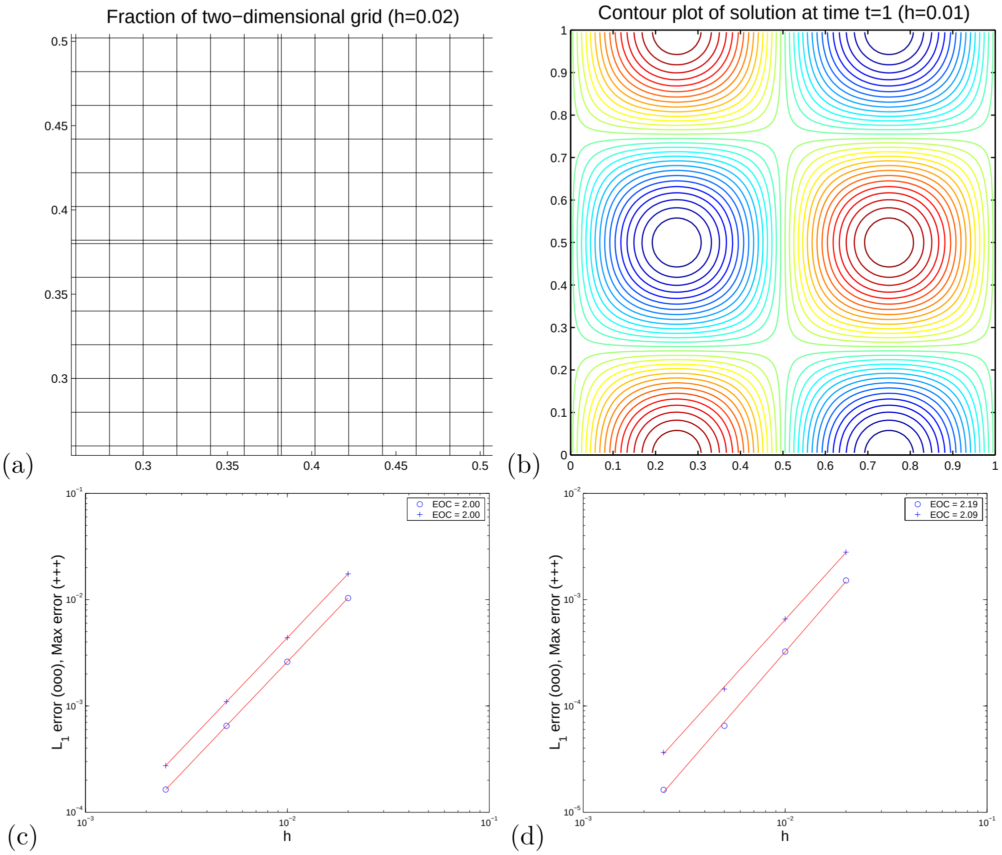


FIG. 13. *Approximation of Example* 7.1. (a) *The grid for a discretization with* $h = 0.02$; (b) *contour plot of the solution using the two-dimensional* h*-box approach with* $h = 0.01$ *and* $CFL_h \approx 0.9$; (c) *convergence study for irregular grid* CLAWPACK *algorithm,* $CFL \approx 0.9$; (d) *convergence study for* h*-box method,* $CFL_h \approx 0.9$ (o*-symbol: error in* L_1*-norm;* $+$*-symbol: error in maximum norm).*

We now demonstrate the performance of our two-dimensional high-resolution h-box method for the approximation of the advection equation. We will compare the numerical results obtained for this h-box scheme with results obtained using the standard high-resolution CLAWPACK algorithm for irregular grid calculations. The latter method requires the time step restriction CFL ≤ 1, while the h-box method is stable for time steps that satisfy $\text{CFL}_h \leq 1$. We first study the accuracy for the two-dimensional advection equation.

Example 7.1. We consider the approximation of the advection equation $q_t + q_x + q_y = 0$, with initial values $q(x, y, 0) = \sin(2\pi x)\cos(2\pi y)$ on the domain $[0, 1] \times [0, 1]$. We impose periodic boundary conditions. The grid contains two lines as well as two columns of grid cells with height (respectively, width) $0.1h$ and $0.9h$. All other grid cells have the size $h \times h$. See Figure 13(a) for a plot of a fraction of the grid.

Test calculations for Example 7.1 confirm that the h-box method leads to second order accurate approximations also in this multidimensional application. In Figure 13(d) we document the experimental order of convergence of the h-box method in both the L_1-norm (depicted by o-symbols) as well as in the maximum norm (+-symbols). For this grid, inaccuracies near the small cells would be displayed in the

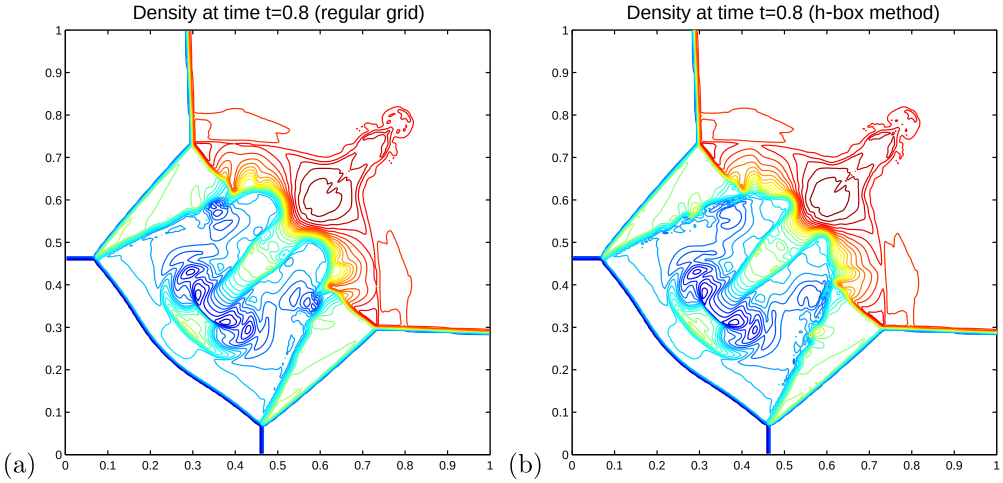


FIG. 14. (a) *Contour plot of density obtained by the high-resolution wave propagation algorithm on a uniform grid with* $h = 0.005$*;* (b) *contour plot of density obtained with high-resolution h-box method,* $h = 0.005$*. We used the monotonized centered limiter.*

maximum norm rather than in the L_1-norm. However, in both norms the experimental order of convergence is about 2. The results for the h-box method compare well with numerical results obtained with the standard wave propagation algorithm with appropriate modifications that allow the approximation on a nonuniform grid. Both schemes converge with second order, but the error is slightly smaller if we use the h-box method. This is due to the numerical viscosity, since the time step restriction CFL ≈ 0.9 for the wave propagation algorithm leads away from the small cell to time steps that correspond to CFL ≤ 0.1.

Our two-dimensional h-box method can be extended to systems of conservation laws in the same way as the standard wave propagation algorithm. The modifications described above now have to be applied to each wave resulting from the decomposition of the left- (respectively, right-) going flux differences into up- and down-going waves. In our last example we consider the approximation of a two-dimensional Riemann problem for the Euler equations, as studied in [27]. This same example was considered in [18], where results of CLAWPACK calculations on a uniform grid are shown. The initial values are piecewise constant in four quadrants, and the solution of each single Riemann problem is a shock wave. Due to the interaction a complex solution structure is obtained. For this calculation we have, in addition to the regular grid cells of the size $h \times h$, 10 lines and 10 columns with height (respectively, width) varying between $0.1h$ and $0.9h$. Our solution on a nonuniform grid calculated by the high-resolution h-box method with $h = 0.005$ compares well with those obtained on a regular grid; see Figure 14. The shock waves are equally well approximated with both methods. Slight differences are visible only at the unstable contact lines, which are very sensitive to the numerical method; see also [18], where it was shown that different limiters have quite a large impact on the approximation.

8. Conclusions. We studied high-resolution h-box methods for the approximation of hyperbolic systems of conservation laws on irregular grids and showed that the definition of the h-box values is important in order to construct accurate schemes. In forthcoming work we will use this to construct a new two-dimensional high-resolution

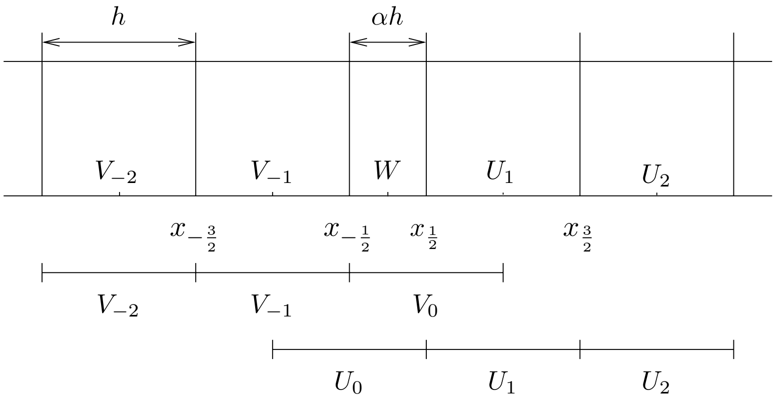


FIG. 15. *Notation for GKS stability, with one small cell in the middle.*

h-box scheme for the approximation of conservation laws with embedded irregular boundaries. So far there is no Cartesian grid embedded boundary method that leads to a second order accurate approximation at boundary cells. Further work will also concentrate on the entropy consistency of *h*-box methods and the approximation of transonic rarefaction waves.

Appendix A. Stability of the second order *h*-box method. In this appendix we prove the stability of the second order *h*-box scheme for $q_t = q_x$ using linear interpolation according to the theory of Gustafsson, Kreiss, and Sundström [11] (henceforth GKS). We treat the small cell with mesh width αh as a boundary condition for the Lax–Wendroff scheme applied on either side of the small cell, using the notation of Figure 15. Here the conserved quantity assigned to the right *h*-box at the interface $x_{-\frac{1}{2}}$ is denoted by V_0. The left *h*-box value at the interface $x_{\frac{1}{2}}$ is U_0. The derivation of the stability condition for the update of the small cell is similar to those used in Berger [1], where stability for schemes with local grid refinement was analyzed.

Both U and V are computed using the second order Lax–Wendroff scheme,

$$\text{(A.1)}\qquad \begin{aligned} U_j^{n+1} &= U_j^n + \lambda/2(U_{j+1}^n - U_{j-1}^n) + \lambda^2/2(U_{j+1}^n - 2U_j^n + U_{j-1}^n), \qquad j \geq 1,\\ V_j^{n+1} &= V_j^n + \lambda/2(V_{j+1}^n - V_{j-1}^n) + \lambda^2/2(V_{j+1}^n - 2V_j^n + V_{j-1}^n), \qquad j \leq -1, \end{aligned}$$

with $\lambda = \frac{\triangle t}{h}$. Using the approach of [1], we look for solutions of the form

$$\text{(A.2)}\qquad \begin{aligned} U_j^n &= \rho\kappa^j z^n, |\kappa| \leq 1, \qquad j = 1, 2, \dots,\\ V_j^n &= \sigma\tau^j z^n, |\tau| \geq 1, \qquad j = -1, -2, \dots. \end{aligned}$$

With this numbering, for l_2 solutions the root κ of the characteristic equation for U on the right side has magnitude less than 1, and τ has magnitude greater than 1. Roughly speaking, the scheme is unstable if and only if there are l_2 solutions satisfying the interpolation conditions with growth in time $|z| > 1$.

The linear interpolation conditions (3.3) give us

$$\text{(A.3)}\qquad U_0 = \frac{1-\alpha}{1+\alpha}V_{-1} + \frac{2\alpha}{1+\alpha}W, \qquad V_0 = \frac{1-\alpha}{1+\alpha}U_1 + \frac{2\alpha}{1+\alpha}W,$$

where the small cell, labeled W above, satisfies the "small cell" version of Lax–Wendroff,

$$W^{n+1} = W^n + \frac{\Delta t}{\alpha h}\left[\frac{U_0+U_1}{2} + \frac{\Delta t}{2h}(U_1 - U_0) - \frac{V_0+V_{-1}}{2} - \frac{\Delta t}{2h}(V_0 - V_{-1})\right]. \tag{A.4}$$

The characteristic equation for W is $W^n = \hat{w}z^n$. We normalize the equations and take $\hat{w} = 1$. Substituting the characteristic roots for U, V into the interpolation conditions (A.3) gives

$$\rho = \frac{1-\alpha}{1+\alpha}\sigma\tau^{-1} + \frac{2\alpha}{1+\alpha}, \qquad \sigma = \frac{1-\alpha}{1+\alpha}\rho\kappa + \frac{2\alpha}{1+\alpha}. \tag{A.5}$$

Equation (A.5) is easily solved for ρ and σ, giving

$$\rho = \frac{2\alpha\left(1+\alpha+(1-\alpha)\tau^{-1}\right)}{(1+\alpha)^2 - (1-\alpha)^2\kappa\tau^{-1}}, \qquad \sigma = \frac{2\alpha\left(1+\alpha+(1-\alpha)\kappa\right)}{(1+\alpha)^2 - (1-\alpha)^2\kappa\tau^{-1}}. \tag{A.6}$$

Substitution of the resolvent equations for U and V into (A.4) gives

$$z = 1 + \frac{\lambda}{2\alpha}\left[\rho(1+\kappa) + \lambda\rho(\kappa-1) - \sigma(1+\tau^{-1}) - \lambda\sigma(1-\tau^{-1})\right]. \tag{A.7}$$

We use (A.5) to replace ρ and σ in terms of $\sigma\tau^{-1}$ and $\rho\kappa$. Also, for a given mesh width h on both the left and right, it is easily seen that the product of the roots κ and τ are $\kappa\tau = \frac{\lambda-1}{\lambda+1}$, so τ^{-1} can be replaced using κ. Thus, (A.7) simplifies to

$$z = 1 - \frac{2\lambda^2}{1+\alpha} + \frac{\lambda(1+\lambda)}{1+\alpha}\kappa(\rho+\sigma). \tag{A.8}$$

We call this *root condition* for the stability of the small cell scheme with Lax–Wendroff. If there are roots z with $|z| > 1$ and κ, τ^{-1} with magnitude less than or equal to 1, satisfying (A.8), then by the GKS theory the scheme is unstable. Conversely, if there are no such roots, the scheme is stable. As in [1], we will use the maximum principle to reduce the range of values we need to check for stability.

To see that the maximum principle applies, we will show that the right-hand side of (A.8), call it $f(z)$, has no singularities for $|z| \geq 1$ and is bounded as $z \to \infty$. First note that $f(z) = (1 - \frac{2\lambda^2}{1+\alpha}) + \frac{\lambda(1+\lambda)}{1+\alpha}\kappa(\rho+\sigma)$ has no branch points for $|z| \geq 1$. This is because the roots κ, τ satisfy the Lax–Wendroff characteristic equation for (A.1),

$$z = 1 + \frac{\lambda}{2}(\eta - \eta^{-1}) + \frac{\lambda^2}{2}(\eta - 2 + \eta^{-1}), \tag{A.9}$$

which lead to a quadratic equation for η with roots

$$\eta_{1,2} = \frac{z - 1 + \lambda^2 \pm \sqrt{(z-1)^2 + \lambda^2(2z-1)}}{\lambda(\lambda+1)}. \tag{A.10}$$

One of the roots is always inside the unit circle, and the other one is outside the unit circle; see [11, Lemma 6.1]. The root inside the unit circle is the root we call κ above, and τ is the root that is outside the unit circle.

The square root term of (A.10) is zero only for $z = 1 - \lambda$, which being inside the unit circle is outside the region of interest, so there are no branch points for $|z| > 1$.

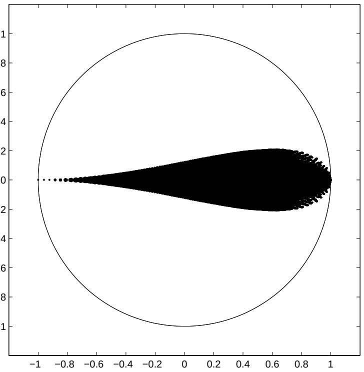


FIG. 16. *Locus of values of $f(z)$ for $|z| = 1$; all values lie inside or on the unit circle.*

Also, note from (A.10) that as $z \to \infty$, the root τ grows like $\frac{2z}{\lambda^2+\lambda}$, so the root κ grows like $\frac{\lambda(\lambda-1)}{2z}$, which is clearly bounded for large z. So the maximum principle applies.

Thus $f(z)$ attains its maximum value on the circle $|z| = 1$. The next step then is to examine the magnitude of $f(z)$ for values of z on the unit circle. Since we can show only analytically that $f(z) \leq 1$ for $\lambda > 0.5$, we instead evaluate $f(z)$ numerically, for $0 \leq \alpha \leq 1$, and $0 < \lambda \leq 1$, on the unit circle for $z = e^{i\theta}, 0 \leq \theta \leq 2\pi$. Figure 16 shows the locus of values of $f(z)$, where the unit circle is also drawn. As the figure and some algebra shows, only for $z = 1, \lambda = 0$, and $z = -1, \lambda = 1$, does $z = f(z)$.

Examining the first value $z = 1 = f(z)$, we have $\lambda = 0$, or equivalently $\Delta t = 0$, so $Q^{n+1} = Q^n$ (with $Q \in \{U, V, W\}$), which is clearly a stable solution. For the other case, we have $z = -1 = f(z)$, whose only solution (again using some numerical evaluation and some algebra) is $\lambda = 1, \alpha = 0$. But $\alpha = 0$ corresponds to the usual Lax–Wendroff scheme without the small cell, and $\lambda = 1$ for this case is straight copying of the solution ($\kappa = 0, \tau = -1$). Again this is stable.

Since Lax–Wendroff is a second order method, the use of linear interpolation with $O(h^2)$ error on a lower-dimensional set of points is reasonable. However, one might consider the use of quadratic interpolation for U_0, V_0. The next question is what stencil to use for the quadratic interpolant. Using the notation of Figure 15, one might consider using the same interpolant based on V_{-1}, W, and U_1 to get both U_0 and V_0. However, this choice reduces the stability region to $\lambda < .5$. If instead the interpolant for U_0 uses the surrounding points V_{-1} and W, and the third point is always the upwind point V_{-2}, full stability for a Courant number of $\lambda \leq 1$ is retained for all small cells with $0 < \alpha < 1$.

REFERENCES

[1] M. J. BERGER, *Stability of interfaces with mesh refinement*, Math. Comp., 45 (1985), pp. 301–318.

[2] M. J. BERGER, C. HELZEL, AND R. J. LEVEQUE, *H-Box Methods for the Approximation of Hyperbolic Conservation Laws on Irregular Grids*, Preprint 2002–022, Conservation law preprint server, http://www.math.ntnu.no/conservation/2002/.

[3] M. J. BERGER AND R. J. LEVEQUE, *An adaptive Cartesian mesh algorithm for the Euler equations in arbitrary geometries*, in Proceedings of the 9th Computational Fluids Conference, AIAA paper AIAA-89-1930, Buffalo, NY, 1989, pp. 305–311.

[4] M. J. BERGER AND R. J. LEVEQUE, *Cartesian meshes and adaptive mesh refinement for hyperbolic partial differential equations*, in Proceedings of the 3rd International Conference on Hyperbolic Problems, Uppsala, Sweden, 1990.

[5] M. J. Berger and R. J. LeVeque, *Stable boundary conditions for Cartesian grid calculations*, Comput. Syst. Engin., 1 (1990), pp. 305–311.

[6] M. J. Berger and R. J. LeVeque, *A rotated difference scheme for Cartesian grids in complex geometries*, in Proceedings of the 10th Computational Fluid Dynamics Conference, AIAA paper CP-91-1602, Honolulu, Hawaii, 1991.

[7] M. J. Berger, R. J. LeVeque, and L. G. Stern, *Finite volume methods for irregular one-dimensional grids*, in Mathematics of Computation 1943–1993: A Half Century of Computational Mathematics, Proc. Sympos. Appl. Math. 48, AMS, Providence, RI, 1994, pp. 255–259.

[8] D. Calhoun and R. J. LeVeque, *A Cartesian grid finite-volume method for the advection diffusion equation in irregular geometries*, J. Comput. Phys., 157 (2000), pp. 143–180.

[9] J. Falcovitz, G. Alfandary, and G. Hanoch, *A two-dimensional conservation laws scheme for compressible flows with moving boundaries*, J. Comput. Phys., 138 (1997), pp. 83–102.

[10] H. Forrer and R. Jeltsch, *A high-order boundary treatment for Cartesian-grid methods*, J. Comput. Phys., 140 (1998), pp. 259–277.

[11] B. Gustafsson, H.-O. Kreiss, and A. Sundström, *Stability theory of difference approximations for mixed initial boundary value problems.* II, Math. Comp., 26 (1972), pp. 649–686.

[12] H. O. Kreiss, T. A. Manteuffel, B. Swartz, B. Wendroff, and A. B. White, *Supra-convergent schemes on irregular grids*, Math. Comp., 47 (1986), pp. 537–554.

[13] R. J. LeVeque, clawpack software, http://www.amath.washington.edu/∼claw/.

[14] R. J. LeVeque, *Large time step shock-capturing techniques for scalar conservation laws*, SIAM J. Numer. Anal., 19 (1982), pp. 1091–1109.

[15] R. J. LeVeque, *Convergence of a large time step generalization of Godunov's method for conservation laws*, Comm. Pure Appl. Math., 37 (1984), pp. 463–477.

[16] R. J. LeVeque, *A large time step generalization of Godunov's method for systems of conservation laws*, SIAM J. Numer. Anal., 22 (1985), pp. 1051–1073.

[17] R. J. LeVeque, *High-resolution conservative algorithms for advection in incompressible flow*, SIAM J. Numer. Anal., 33 (1996), pp. 627–665.

[18] R. J. LeVeque, *Wave propagation algorithms for multidimensional hyperbolic systems*, J. Comput. Phys., 131 (1997), pp. 327–353.

[19] R. J. LeVeque, *Finite Volume Methods for Hyperbolic Problems*, Cambridge University Press, Cambridge, UK, 2002.

[20] T. A. Manteuffel and A. B. White, Jr., *The numerical solution of second-order boundary value problems on nonuniform meshes*, Math. Comp., 47 (1986), pp. 511–535.

[21] D. Modiano and P. Colella, *A higher-order embedded boundary method for time dependent simulation of hyperbolic conservation laws*, in Proceedings of the FEDSM 00 - ASME Fluids Engineering Simulation Meeting, Boston, MA, 2000.

[22] K. W. Morton, *On the analysis of finite volume methods for evolutionary problems*, SIAM J. Numer. Anal., 35 (1998), pp. 2195–2222.

[23] R. B. Pember, J. B. Bell, P. Colella, W. Y. Crutchfield, and M. L. Welcome, *An adaptive Cartesian grid method for unsteady compressible flow in irregular regions*, J. Comput. Phys., 120 (1995), pp. 278–304.

[24] J. Pike, *Grid adaptive algorithms for the solution of the Euler equations on irregular grids*, J. Comput. Phys., 71 (1987), pp. 194–223.

[25] J. J. Quirk, *An alternative to unstructured grids for computing gas dynamic flows around arbitrarily complex two-dimensional bodies*, Comput. & Fluids, 23 (1994), pp. 125–142.

[26] P. L. Roe, *Approximate Riemann solver, parameter vectors, and difference schemes*, J. Comput. Phys., 43 (1981), pp. 357–372.

[27] C. W. Schultz-Rinne, J. P. Collins, and H. M. Glaz, *Numerical solution of the Riemann-problem for two-dimensional gas dynamics*, SIAM J. Sci. Comput., 14 (1993), pp. 1394–1414.

[28] L. G. Stern, *An Explicit Conservative Method for Time-Accurate Solution of Hyperbolic Partial Differential Equations on Embedded Chimera Grids*, Ph.D. thesis, University of Washington, Seattle, WA, 1996. Available from ftp://www.amath.washington.edu/pub/∼rjl/students/stern:thesis.ps.gz.

[29] B. van Leer, *Towards the ultimate conservative difference scheme* II. *Monotonicity and conservation combined in a second order scheme*, J. Comput. Phys., 14 (1974), pp. 361–370.

[30] B. Wendroff and A. B. White, *Some supraconvergent schemes for hyperbolic equations on nonuniform grids*, in Proceedings of the Second International Conference on Hyperbolic Problems, Aachen, Germany, 1988, pp. 671–677.

[31] B. Wendroff and A. B. White, *A supraconvergent scheme for nonlinear hyperbolic systems*, Comput. Math. Appl., 18 (1989), pp. 761–767.

SIAM J. NUMER. ANAL.
Vol. 41, No. 3, pp. 919–944

COUPLING OF FINITE ELEMENTS AND BOUNDARY ELEMENTS IN ELECTROMAGNETIC SCATTERING[*]

R. HIPTMAIR[†]

Abstract. We consider the scattering of monochromatic electromagnetic waves at a dielectric object with a rough surface. We investigate the coupling of a weak formulation of Maxwell's equations inside the scatterer with boundary integral equations that arise from the homogeneous problem in the unbounded region outside the scatterer. The symmetric coupling approach based on the full Calderón projector for Maxwell's equations is employed. By splitting both the electric field inside the scatterer and the surface currents into components of predominantly electric and magnetic nature, we can establish coercivity of the coupled variational problem, provided that the frequency is away from resonant frequencies.

Discretization relies on both **curl**-conforming edge elements inside the scatterer and div_Γ-conforming boundary elements for the surface currents. The splitting idea, adjusted to the discrete setting, permits us to show uniform stability of the discretized problem. We exploit it to come up with a priori convergence estimates.

Key words. electromagnetic scattering, Helmholtz decomposition, Hodge decomposition, Calderón projector, symmetric coupling, edge elements, discrete coercivity

AMS subject classifications. 65N12, 65N38, 78M15

PII. S0036142901397757

1. Introduction.

The simulation of electromagnetic scattering is mainly concerned with approximately solving the homogeneous Maxwell equations in $\mathbb{R}^3$, subject to excitation by some monochromatic incident wave. Outside a bounded object, which is called the scatterer and occupies the bounded domain $\Omega_s \subset \mathbb{R}^3$, the electromagnetic material coefficients ϵ and μ assume the constant values $\epsilon_0 > 0$ and $\mu_0 > 0$, respectively. Inside Ω_s they may display some spatial variation. Commonly faced in applications are scatterers with piecewise smooth, Lipschitz-continuous boundary. For simplicity, we suppose that the surface $\Gamma := \partial\Omega_s$ of the scatterer is connected, but, with slight alterations, all considerations of this article carry over to more general situations.

Let $\mathbf{E}$ denote the complex amplitude of the scattered electric field in $\Omega' := \mathbb{R}^3 \setminus \bar{\Omega}_s$ and the total electric field inside Ω_s. It emerges as the solution of the transmission problem (cf. [45, sect. 5.6.3])

$$
\begin{array}{rcll}
\mathbf{curl}\,\mathbf{curl}\,\mathbf{E} - \kappa^2\mathbf{E} & = & 0 & \text{in } \Omega', \\
\mathbf{curl}\,\mu_r^{-1}\,\mathbf{curl}\,\mathbf{E} - \kappa^2\epsilon_r\mathbf{E} & = & 0 & \text{in } \Omega_s, \\
[\gamma_{\mathbf{t}}\mathbf{E}]_\Gamma = \gamma_{\mathbf{t}}\mathbf{E}_{\mathrm{inc}}, \quad \left[\frac{1}{\mu_r}\gamma_N\mathbf{E}\right]_\Gamma & = & \gamma_N\mathbf{E}_{\mathrm{inc}} & \text{on } \Gamma, \\
\lim\limits_{|\mathbf{x}|\to\infty} \mathbf{curl}\,\mathbf{E}\times\mathbf{x} - i\kappa|\mathbf{x}|\mathbf{E} & = & 0. &
\end{array}
\tag{1.1}
$$

Here, $\kappa := \omega\sqrt{\epsilon_0\mu_0}L$ (with $\omega > 0$ the fixed angular frequency of the excitation, L the characteristic length of the scatterer) stands for the normalized wave number. In what

[*] Received by the editors November 9, 2001; accepted for publication October 17, 2002; published electronically May 22, 2003.

http://www.siam.org/journals/sinum/41-3/39775.html

[†] Seminar for Applied Mathematics, Swiss Federal Institute of Technology Zürich, CH-8092 Zürich, Switzerland (hiptmair@sam.math.ethz.ch).

follows, it should be regarded as merely a nonzero real parameter. $\mathbf{E}_{\text{inc}}$ stands for the complex amplitude of the electric field associated with the incident wave. In addition, we write $\gamma_{\mathbf{t}}\mathbf{E}$ for the tangential components of $\mathbf{E}$ on Γ, and $\gamma_N\mathbf{E}$ for the "magnetic trace" $\mathbf{curl}\,\mathbf{E} \times \mathbf{n}$ on Γ. The exterior unit normal vectorfield $\mathbf{n} \in L^\infty(\Gamma)$ is directed from Ω_s into Ω'. Finally, $[\gamma\phi]_\Gamma$ designates the jump $\gamma\phi_{|\Omega'} - \gamma\phi_{|\Omega_s}$ of some trace γ of a function ϕ across Γ. We remark that a similar, entirely equivalent formulation in terms of the magnetic field $\mathbf{H} := \frac{1}{i\omega\mu_r}\,\mathbf{curl}\,\mathbf{E}$ exists.

Using Rellich's lemma and unique continuation techniques, the following result can be established (cf. [35, Thm. 3.1]).

THEOREM 1.1. *Provided that the relative material parameters μ_r and $\epsilon_r > 0$ are piecewise smooth and bounded away from zero everywhere in Ω_s, the problem* (1.1) *has a unique solution.*

Boundary element methods (BEM) offer the most flexible way to deal with the homogeneous problem in the unbounded exterior domain Ω'. They are based on boundary integral operator equations (BIE) on Γ. Owing to potentially nonconstant material parameters, the field problem inside Ω_s may not be amenable to a treatment by means of boundary elements. Finite element schemes (FEM) founded on a weak variational formulation of the electric wave equation have to be used there. Thus, the topic of the paper comes into focus, namely, how to derive and discretize a suitable coupled problem, and how to analyze the resulting FEM-BEM formulation.

Coupling entails expressing the Dirichlet-to-Neumann map of the exterior problem through boundary equations linking the Cauchy data $\gamma_{\mathbf{t}}\mathbf{E}$ and $\gamma_N\mathbf{E}$ for the electric field. There is a wealth of integral formulations for the exterior electromagnetic boundary value problem. A comprehensive survey is given in Nédélec's recent monograph [45]. In principle, all these methods furnish the Dirichlet-to-Neumann map. However, in many cases, in particular with so-called indirect formulations, the resulting operator lacks structural properties of the Dirichlet-to-Neumann map. This is blatantly obvious in the case of second order elliptic problems [40]. If structure is not preserved, then the linear systems of equations obtained through Galerkin boundary element discretization are adversely affected.

For second order elliptic problems Costabel [23] discovered that the so-called direct boundary integral equations provide a remedy. The key concept is that of the Calderón projector acting on the Cauchy data of the problem. For details and theoretical examinations we refer to [17, sect. 4.5] and [29]. In short, the Calderón projector supplies two sets of boundary integral equations. Judiciously combining them yields a version of the Dirichlet-to-Neumann map that perfectly lends itself to a Galerkin discretization. The realization of Costabel's idea is called the "symmetric coupling approach" to marrying finite elements and boundary elements. It has been applied to a wide range of (strongly elliptic) problems; see, among many others, [15, 37, 41].

Unsurprisingly, the Calderón projector for the Maxwell system has been amply studied (cf. [16, sect. 1.3.2], [32], [45, sect. 5.5], and [39, sect. 3]). The idea of symmetric coupling for the transmission problem was theoretically probed in [1, 2, 3], and in [6] for a related problem involving impedance boundary conditions. All these theoretical results employ compactness arguments and the Fredholm alternative. To this end, most authors have meticulously studied the integral operators on Γ. They have completely succeeded on smooth boundaries, but all efforts to adjust the approach to nonsmooth boundaries have been in vain.

It was fundamental new insights about the trace spaces of electromagnetic fields, presented in [8, 10, 11, 13], that cleared the road to further progress. That progress could finally be achieved by remembering a highly effective policy in the modern treatment of boundary integral equations: The guideline is to stay off the boundary as far as possible by studying variational problems instead of the boundary integral operators directly. This policy has demonstrated its efficacy in the work of Costabel [24]. The recent textbook [43] discusses all nuances of this approach for strongly elliptic boundary value problems. Moving off the boundary helps steer clear of its awkward geometric features. Thus, the foundation for a theory of electromagnetic boundary integral operators on nonsmooth boundaries could be laid in [14].

In addition, in order to harness compactness arguments, we have to employ decompositions of surface vectorfields, on whose components the boundary integral operators will be considered. The classical decomposition is the so-called Hodge decomposition [32], which remains a very effective tool on piecewise smooth boundaries (see [12, 38] and, in particular, [14]). Its counterpart on domains is the Helmholtz decomposition. It is important to realize that there is some leeway in choosing the decomposition, because the exact orthogonality featured by Hodge and Helmholtz decompositions is of minor importance. Instead, we prefer to use related, but simpler, splittings.

All BIE for the exterior Dirichlet problem in electromagnetic scattering and acoustics are haunted by the presence of "forbidden frequencies" [14, 12, 18], for which the equations fail to have a unique solution. Those agree with interior Dirichlet eigenvalues. The symmetrically coupled problem investigated in this paper exhibits the same drawback. Therefore we have to resign ourselves to making the following assumption about the uniqueness of the solutions of the interior Dirichlet problem with constant coefficients.

Assumption 1. If $\mathbf{curl}\,\mathbf{curl}\,\mathbf{U} - \kappa^2\mathbf{U} = 0$ in Ω_s and $\gamma_{\mathbf{t}}\mathbf{U} = 0$ on Γ, then necessarily $\mathbf{U} = 0$.

To discretize the symmetrically coupled problem, we rely on discrete differential forms (edge elements, face elements) on triangulations of both Ω_s and Γ. The Galerkin approach is straightforward, and yet, in the discrete setting another challenge arises, because the continuous decompositions do not directly carry over to the discrete spaces. For pure indirect boundary element formulations (Rumsey's principle) remedies have been explored in [18] and [38]. Direct BIE were successfully tackled in [14]. All these approaches exploit the fact that appropriate discrete splittings can approximate their continuous counterparts reasonably well. In this paper we adapt the ideas pioneered in [14] to the symmetrically coupled FEM-BEM problem. At the heart of the developments will be symmetry and compactness properties of the Calderón projector, which were first established in [14]. We will use variants of these results that do not rely on any sophisticated results on elliptic regularity.

The plan of the paper is as follows. In the next section we discuss the rationale behind the use of decompositions. Then, the theory of tangential traces is reviewed in section 3. After that we introduce the potentials that supply the building blocks for the Stratton–Chu representation formula. These potentials spawn integral operators that are examined in section 5. The coupled variational formulation of the entire scattering problem is derived in section 6. In section 7 we construct decompositions of fields in Ω_s and on Γ. These decompositions give rise to a split variational problem that is shown to be coercive in section 8. Up to this point everything is aimed at the analysis of the continuous variational problem. Then, section 9 is devoted to the spaces of finite elements and boundary elements used for the Galerkin discretization of the coupled problem. Discrete counterparts of the continuous decompositions are presented in section 10. Abstract conditions that the discrete decompositions have to meet are put forth in section 11, which is also devoted to their verification for the concrete decompositions. Finally, in section 12, we give quantitative a priori convergence estimates.

2. Electromagnetic versus acoustic scattering. What is the point of this paper in light of the fairly general theory of symmetric coupling for strongly elliptic boundary value problems? To elucidate this, let us consider the apparently similar problem of time-harmonic acoustic scattering governed by the Helmholtz equation

$$-\Delta\rho - \kappa^2\rho = 0 \quad \text{in } \Omega'.$$

This is a showcase example for the application of the general theory. What accounts for the fundamental difference between electromagnetism and acoustics? Both phenomena, from a physical point of view, are marked by an incessant conversion of energies. In acoustics, potential and kinetic energy of the fluid are converted into each other; in electromagnetism, the same roles are played by the electric and magnetic energy. In acoustics the kinetic energy (with respect to a bounded control volume Ω) is a compact perturbation of the potential energy. Just remember that those are associated with the squared $L^2(\Omega)$- and $H^1(\Omega)$-norms, respectively. Therefore we can clearly single out the Laplacian as the principal part of the Helmholtz operator. This paves the way for proofs of coercivity based on "ignoring" the kinetic energy altogether.

Conversely, for the electric wave equation, the electric energy described by $\|\mathbf{E}\|^2_{\boldsymbol{L}^2(\Omega)}$ is by no means a compact perturbation of the magnetic energy, which is measured by $\|\mathbf{curl}\,\mathbf{E}\|^2_{\boldsymbol{L}^2(\Omega)}$. Both energies are completely symmetric, and none can be given preference at the expense of the other: No term in the differential operator $\mathbf{curl}\,\mathbf{curl}\,\mathbf{E} - \kappa^2\mathbf{E}$ plays the role of a principal part. Formally speaking, the operator of the electric wave equation lacks the essential property of strong ellipticity. However, this is what is required by standard arguments involving compact perturbations.

There is a way to promote one type of energy to dominate the other. This is called *regularization* and is usually done by imposing some constraint on the divergence of the electric field [35, 6, 28]. Regularization is problematic, because $\boldsymbol{H}(\mathbf{curl};\Omega)\cap\boldsymbol{H}(\mathrm{div};\Omega)$ is still not compactly embedded in $\boldsymbol{L}^2(\Omega)$ [4, Prop. 2.7]. In [35] this forces a separation of the surface on which the integral operators are defined and the surface on which they are evaluated in order to salvage coercivity. An alternative consists of tampering with the transmission conditions as in [6, 28, 25], but equivalence to the original problem holds only for smooth Γ.

The superior alternative to regularization is the *splitting* of the fields into two components. One set of components, called the electric, will feature dominant electric energy. With the other set, the magnetic quantities, the situation is reversed. This will promote either **curl curl** or Id to the role of a principal part. As a consequence, on each component the electric wave equation should be amenable to a treatment along the lines of the classical theory.

To find a suitable splitting, we can take our cue from a stationary situation. There, we find that valid electric fields are irrotational. In a nonstationary situation, magnetic induction generates another solenoidal contribution to $\mathbf{E}$. Thus, we have arrived at a Helmholtz decomposition

$$\mathbf{E} = \mathbf{grad}\,\varphi + \mathbf{curl}\,\mathbf{A}$$

of the electric field. The first term has an "electric nature," and the second can be labeled "magnetic." The reader should not be misled by the symbol $\mathbf{E}$ and the parlance "electric field": For positive wave numbers both $\mathbf{E}$ and $\mathbf{H}$ have a twin electric and magnetic nature.

Similar considerations apply to the Cauchy data $\gamma_{\mathbf{t}}\mathbf{E}$ and $\gamma_N\mathbf{E}$. A splitting of $\gamma_{\mathbf{t}}\mathbf{E}$ is instantly induced by the Helmholtz decomposition, and the electric and magnetic parts are also clear from it. In the case of $\gamma_N\mathbf{E}$, which is a rotated tangential trace of the magnetic field, we remember that it is commonly viewed as a "surface current." Hence, its divergence-free components are currents not accompanied by surface charges; they qualify as "magnetic quantities." Currents in a complement space inevitably generate surface charges. Thus, they can be regarded as "electric." This suggests that we start out from the *Hodge decomposition*

$$\gamma_N\mathbf{E} = \mathbf{curl}_\Gamma\varphi + \mathbf{grad}_\Gamma\,\psi$$

of the surface currents, in order to deal with the BIE.

3. Traces and spaces. The natural Hilbert space setting for the analysis of the transmission problem (1.1) is provided by the spaces

$$\boldsymbol{H}_{\mathrm{loc}}(\mathbf{curl};\Omega) := \{\mathbf{V} \in \boldsymbol{L}^2_{\mathrm{loc}}(\Omega),\, \mathbf{curl}\,\mathbf{V} \in \boldsymbol{L}^2_{\mathrm{loc}}(\Omega)\}.$$

Here and in what follows, Ω is a "generic domain," which can either be Ω_s or Ω'. For a thorough examination of these spaces and notations, we refer to [33, Chap. 1].

The Sobolev spaces of scalar functions and related functionals, $H^s(\Gamma)$ and $H^{-s}(\Gamma)$, can be defined invariantly for $0 \le s \le 1$ [34, Sect. 1.3.3]. In addition, we write $\gamma : H^s_{\mathrm{loc}}(\Omega) \mapsto H^{s-\frac{1}{2}}(\Gamma)$, $\frac{1}{2} < s < \frac{3}{2}$, for the usual trace operator [43, Thm. 3.38]. Superscripts $-$ and $+$ will be attached to the trace operators, when it is important whether they act from Ω_s or Ω', respectively.

If Γ is a curvilinear Lipschitz polyhedron (cf. the introduction of [26]) with smooth components Γ_j, $j = 1, \dots, N_\Gamma$, we set

$$\begin{aligned} H^s(\Gamma) :=& \{u \in H^1(\Gamma),\, u_{|\Gamma_j} \in H^s(\Gamma_j),\, j = 1\dots, N_\Gamma\} \quad \text{for } s > 1,\\ \boldsymbol{H}^s_{\mathbf{t}}(\Gamma) :=& \{\mathbf{u} \in \boldsymbol{L}^2_{\mathbf{t}}(\Gamma),\, \mathbf{u}_{|\Gamma_j} \in \boldsymbol{H}^s(\Gamma_j),\, j = 1, \dots, N_\Gamma\} \quad \text{for } s \ge 0, \end{aligned}$$

where $\boldsymbol{L}^2_{\mathbf{t}}(\Gamma) := \{\mathbf{u} \in (L^2(\Gamma))^3, \mathbf{u}\cdot\mathbf{n} = 0\}$. All these spaces are equipped with the natural graph norms.

The tangential components trace $\gamma_{\mathbf{t}}$ and the twisted tangential trace $\gamma_\times$, for $\mathbf{U} \in \boldsymbol{C}^\infty(\bar{\Omega})$ defined by $\gamma_{\mathbf{t}}\mathbf{U}(\mathbf{x}) := \mathbf{n}(\mathbf{x}) \times (\mathbf{U}(\mathbf{x}) \times \mathbf{n}(\mathbf{x}))$ and $\gamma_\times\mathbf{U}(\mathbf{x}) := \mathbf{U}(\mathbf{x}) \times \mathbf{n}(\mathbf{x})$ a.e. on Γ, play a central role in the mathematical treatment of the Maxwell transmission problem. Their extension to $\boldsymbol{H}_{\mathrm{loc}}(\mathbf{curl};\Omega)$ was accomplished for piecewise smooth boundaries in [10, 8] and for Lipschitz domains in [13]. These articles and section 2 of [12] supply the main references for the current section.

THEOREM 3.1. *There are intrinsically defined spaces* $\boldsymbol{H}^{1/2}_{||}(\Gamma), \boldsymbol{H}^{1/2}_{\perp}(\Gamma) \subset \boldsymbol{L}^2_{\mathbf{t}}(\Gamma)$ *such that the tangential trace mappings* $\gamma_{\mathbf{t}}{}^{\pm}/\gamma^{\pm}_{\times} : \boldsymbol{H}^1_{\mathrm{loc}}(\Omega) \mapsto \boldsymbol{H}^{1/2}_{||}(\Gamma)/\boldsymbol{H}^{1/2}_{\perp}(\Gamma)$, $\Omega = \Omega', \Omega_s$, *are continuous, surjective and possess continuous right inverses.*

Proof. See Proposition 1.7 in [10] and [13, sect. 2]. □

The associated dual spaces will be denoted by $\boldsymbol{H}^{-1/2}_{||}(\Gamma)$ and $\boldsymbol{H}^{-1/2}_{\perp}(\Gamma)$, respectively, where the sesqui-linear duality pairings $\langle\cdot,\cdot\rangle_{\frac{1}{2},||,\Gamma} : \boldsymbol{H}^{-1/2}_{||}(\Gamma) \times \boldsymbol{H}^{1/2}_{||}(\Gamma) \mapsto \mathbb{C}$,

$\langle\cdot,\cdot\rangle_{\frac{1}{2},\perp,\Gamma} : \boldsymbol{H}_{\perp}^{-1/2}(\Gamma) \times \boldsymbol{H}_{\perp}^{1/2}(\Gamma) \mapsto \mathbb{C}$ are taken with $\boldsymbol{L}^2_{\mathbf{t}}(\Gamma)$ as pivot space.

The classical Rellich theorem can also be applied to the tangential trace spaces as follows.

Lemma 3.2. *The embeddings* $\boldsymbol{H}_{||}^{\frac{1}{2}}(\Gamma), \boldsymbol{H}_{\perp}^{\frac{1}{2}}(\Gamma) \hookrightarrow \boldsymbol{L}^2_{\mathbf{t}}(\Gamma)$ *are compact.*

Based on surface differential operators (cf. section 3 of [13]) we can define

$$\begin{aligned}\boldsymbol{H}^{-\frac{1}{2}}(\mathbf{curl}_{\Gamma},\Gamma) =&\{\mathbf{v} \in \boldsymbol{H}_{\perp}^{-\frac{1}{2}}(\Gamma),\ \mathrm{curl}_{\Gamma}\,\mathbf{v} \in H^{-\frac{1}{2}}(\Gamma)\},\\ \boldsymbol{H}^{-\frac{1}{2}}(\mathrm{div}_{\Gamma},\Gamma) =&\{\boldsymbol{\zeta} \in \boldsymbol{H}_{||}^{-\frac{1}{2}}(\Gamma),\ \mathrm{div}_{\Gamma}\boldsymbol{\zeta} \in H^{-\frac{1}{2}}(\Gamma)\}.\end{aligned}$$

These spaces are endowed with natural graph norms. They are important as suitable trace spaces for vectorfields in $\boldsymbol{H}(\mathbf{curl};\Omega)$ (see [10, Thms. 2.7, 2.8], [11, Thm. 4.5], [8], [13, sect. 4]).

Theorem 3.3. *The trace mappings* $\gamma_{\mathbf{t}}{}^{+} : \boldsymbol{H}_{\mathrm{loc}}(\mathbf{curl};\Omega') \mapsto \boldsymbol{H}^{-\frac{1}{2}}(\mathbf{curl}_{\Gamma},\Gamma)$, $\gamma_{\mathbf{t}}{}^{-} : \boldsymbol{H}(\mathbf{curl};\Omega_s) \mapsto \boldsymbol{H}^{-\frac{1}{2}}(\mathbf{curl}_{\Gamma},\Gamma)$ *and* $\gamma_{\times}^{+} : \boldsymbol{H}_{\mathrm{loc}}(\mathbf{curl};\Omega') \mapsto \boldsymbol{H}^{-\frac{1}{2}}(\mathrm{div}_{\Gamma},\Gamma)$, $\gamma_{\times}^{-} : \boldsymbol{H}(\mathbf{curl};\Omega_s) \mapsto \boldsymbol{H}^{-\frac{1}{2}}(\mathrm{div}_{\Gamma},\Gamma)$ *are continuous and surjective with continuous right inverses* $\mathsf{F}^{\pm}_{\mathbf{t}}, \mathsf{F}^{\pm}_{\times}$.

We learn that $\boldsymbol{H}^{-\frac{1}{2}}(\mathbf{curl}_{\Gamma},\Gamma)$ is exactly the right space for the Dirichlet data $\gamma_{\mathbf{t}}{}^{\pm}\mathbf{E}$ in (1.1). Hence, we adopt the alternative notation γ_D for $\gamma_{\mathbf{t}}$ to stress that this is the right "Dirichlet trace." As demonstrated in [11, sect. 4], $\boldsymbol{H}^{-\frac{1}{2}}(\mathbf{curl}_{\Gamma},\Gamma)$ and $\boldsymbol{H}^{-\frac{1}{2}}(\mathrm{div}_{\Gamma},\Gamma)$ can be put in duality when $\boldsymbol{L}^2_{\mathbf{t}}(\Gamma)$ is used as pivot space. More precisely, the usual $\boldsymbol{L}^2_{\mathbf{t}}(\Gamma)$-inner product can be extended to a sesqui-linear duality pairing

$$\langle\cdot,\cdot\rangle_{\boldsymbol{\tau}} : \boldsymbol{H}^{-\frac{1}{2}}(\mathrm{div}_{\Gamma},\Gamma) \times \boldsymbol{H}^{-\frac{1}{2}}(\mathbf{curl}_{\Gamma},\Gamma) \mapsto \mathbb{C}$$

by means of Green's formula ("+" for $\Omega = \Omega_s$)

$$\mp \int_{\Omega} \overline{\mathbf{U}} \cdot \mathbf{curl}\,\mathbf{V} - \mathbf{curl}\,\overline{\mathbf{U}} \cdot \mathbf{V}\, d\mathbf{x} = \left\langle \gamma_{\times}^{\pm}\mathbf{V}, \gamma_{\mathbf{t}}{}^{\pm}\mathbf{U}\right\rangle_{\boldsymbol{\tau}} \quad \forall \mathbf{U},\mathbf{V} \in \boldsymbol{H}(\mathbf{curl};\Omega)\ .$$

An overbar designates complex conjugation.[1] A ubiquitous device is the surface twist operator R, for continuous tangential vectorfields given by $(\mathsf{R}\mathbf{u})(\mathbf{x}) := (\mathbf{n}\times\mathbf{u})(\mathbf{x})$ for almost all $\mathbf{x} \in \Gamma$. It gives rise to an isometric mapping $\mathsf{R} : \boldsymbol{H}^{-\frac{1}{2}}(\mathbf{curl}_{\Gamma},\Gamma) \mapsto \boldsymbol{H}^{-\frac{1}{2}}(\mathrm{div}_{\Gamma},\Gamma)$.

We will also need the normal components trace $\gamma_{\mathbf{n}}$ with $\gamma_{\mathbf{n}}\mathbf{U}(\mathbf{x}) = \mathbf{n}(\mathbf{x})\cdot\mathbf{U}(\mathbf{x})$ for almost all $\mathbf{x} \in \Gamma$ and $\mathbf{U} \in C^{\infty}(\bar{\Omega})$. It can be extended to a continuous and surjective mapping $\gamma_{\mathbf{n}} : \boldsymbol{H}(\mathrm{div};\Omega) \mapsto H^{-\frac{1}{2}}(\Gamma)$ [33, Thm. 2.5].

Beside the Dirichlet trace γ_D, the transmission conditions of (1.1) feature a second trace γ_N, aptly called a Neumann trace. It has to be introduced in a weak sense: For

$$\mathbf{U} \in \boldsymbol{H}_{\mathrm{loc}}(\mathbf{curl}^2,\Omega) := \{\mathbf{V} \in \boldsymbol{H}_{\mathrm{loc}}(\mathbf{curl};\Omega),\ \mathbf{curl}\,\mathbf{curl}(\mathbf{v}) \in \boldsymbol{L}^2_{\mathrm{loc}}(\Omega)\}$$

we define $\gamma_N^{\pm}\mathbf{U} \in \boldsymbol{H}^{-\frac{1}{2}}(\mathrm{div}_{\Gamma},\Gamma)$ by

$$\mp \int_{\Omega} \mathbf{curl}\,\mathbf{U} \cdot \mathbf{curl}\,\overline{\mathbf{V}} - \mathbf{curl}\,\mathbf{curl}\,\mathbf{U} \cdot \overline{\mathbf{V}}\, d\mathbf{x} = \left\langle \gamma_N^{\pm}\mathbf{U}, \gamma_D^{\pm}\mathbf{V}\right\rangle_{\boldsymbol{\tau}}$$

[1] We adopt the following convention: Surface vectorfields in $\boldsymbol{H}^{-1/2}(\mathbf{curl}_{\Gamma},\Gamma)$ will be written in small bold Roman print, those in $\boldsymbol{H}^{-1/2}(\mathrm{div}_{\Gamma},\Gamma)$ in small bold Greek characters. Capital bold Roman symbols are used for vectorfields in Ω_s or Ω'. Regular print marks scalar quantities.

for all compactly supported $\mathbf{V} \in \boldsymbol{H}(\mathbf{curl};\Omega)$, where "+" applies to $\Omega = \Omega_s$. Obviously, for smooth fields we recover $\gamma_N \mathbf{U} = \gamma_\times(\mathbf{curl}\,\mathbf{U}) = \mathbf{curl}\,\mathbf{U} \times \mathbf{n}$.

Lemma 3.4 (see [37, Lem. 3.3]). *The traces* $\gamma_N^\pm$ *furnish continuous mappings* $\gamma_N^+ : \boldsymbol{H}_{\mathrm{loc}}(\mathbf{curl}^2, \Omega') \mapsto \boldsymbol{H}^{-\frac{1}{2}}(\mathrm{div}_\Gamma, \Gamma)$ *and* $\gamma_N^- : \boldsymbol{H}(\mathbf{curl}^2, \Omega_s) \mapsto \boldsymbol{H}^{-\frac{1}{2}}(\mathrm{div}_\Gamma, \Gamma)$.

4. Potentials.

In [12, sect. 3], in [16, Chap. 3, sect. 1.3.2], and in [45, sect. 5.5] the *Stratton–Chu representation formula* is derived. It states that any distribution $\mathbf{U} \in \boldsymbol{H}_{\mathrm{loc}}(\mathbf{curl}^2, \Omega')$ that satisfies

$$\mathbf{curl}\,\mathbf{curl}\,\mathbf{U} - \kappa^2 \mathbf{U} = 0 \quad \text{in } \Omega' \tag{4.1}$$

and the Silver–Müller radiation conditions can be written as

$$\mathbf{U} = \boldsymbol{\Psi}_{\mathbf{M}}^\kappa(\gamma_D^+\mathbf{U}) - \boldsymbol{\Psi}_{\mathbf{A}}^\kappa(\gamma_N^+\mathbf{U}) - \mathbf{grad}\,\Psi_V^\kappa(\gamma_{\mathbf{n}}^+\mathbf{U}) \quad \text{in } \Omega'. \tag{4.2}$$

Here, $\boldsymbol{\Psi}_{\mathbf{M}}^\kappa(\cdot)$, $\boldsymbol{\Psi}_{\mathbf{A}}^\kappa(\cdot)$, and $\Psi_V^\kappa(\cdot)$ are *potentials*, i.e, in our parlance, mappings of boundary data to functions defined everywhere off the boundary. In detail, based on the Helmholtz kernel $E_\kappa(\mathbf{x}, \mathbf{y}) := \exp(i\kappa|\mathbf{x} - \mathbf{y}|)/4\pi|\mathbf{x} - \mathbf{y}|$, $\mathbf{x} \neq \mathbf{y}$, Ψ_V^κ is the scalar single layer potential, given by

$$\Psi_V^\kappa(\phi)(\mathbf{x}) := \int_\Gamma E_\kappa(\mathbf{x}, \mathbf{y})\phi(\mathbf{y})\, dS(\mathbf{y}), \quad \mathbf{x} \notin \Gamma, \tag{4.3}$$

and $\boldsymbol{\Psi}_{\mathbf{A}}^\kappa$ its cousin, the vectorial single layer potential

$$\boldsymbol{\Psi}_{\mathbf{A}}^\kappa(\boldsymbol{\mu})(\mathbf{x}) := \int_\Gamma E_\kappa(\mathbf{x}, \mathbf{y})\boldsymbol{\mu}(\mathbf{y})\, dS(\mathbf{y}), \quad \mathbf{x} \notin \Gamma. \tag{4.4}$$

They are joined by the Maxwell double layer potential

$$\boldsymbol{\Psi}_{\mathbf{M}}^\kappa(\mathbf{u})(\mathbf{x}) := \mathbf{curl}_{\mathbf{x}}\, \boldsymbol{\Psi}_{\mathbf{A}}^\kappa(\mathsf{R}\mathbf{u})(\mathbf{x}), \quad \mathbf{x} \notin \Gamma.$$

A simplification of (4.2) is possible by observing that (cf. [13, eq. (26)])

$$\mathrm{div}_\Gamma(\gamma_N^+\mathbf{U}) = \gamma_{\mathbf{n}}^+(\mathbf{curl}\,\mathbf{curl}\,\mathbf{U}) = \kappa^2(\gamma_{\mathbf{n}}^+\mathbf{U}) \quad \text{in } H^{-\frac{1}{2}}(\Gamma) \tag{4.5}$$

for all $\mathbf{U} \in \boldsymbol{H}_{\mathrm{loc}}(\mathbf{curl}^2, \Omega')$ satisfying (4.1). This enables us to get rid of the normal components trace in (4.2). We end up with

$$\mathbf{U} = \boldsymbol{\Psi}_{\mathbf{M}}^\kappa(\gamma_D^+\mathbf{U}) - \boldsymbol{\Psi}_{\mathbf{S}}^\kappa(\gamma_N^+\mathbf{U}) \quad \text{in } \Omega', \tag{4.6}$$

where we introduced the Maxwell single layer potential according to

$$\boldsymbol{\Psi}_{\mathbf{S}}^\kappa(\boldsymbol{\mu})(\mathbf{x}) := \boldsymbol{\Psi}_{\mathbf{A}}^\kappa(\boldsymbol{\mu})(\mathbf{x}) + \frac{1}{\kappa^2}\mathbf{grad}_{\mathbf{x}}\, \Psi_V^\kappa(\mathrm{div}_\Gamma\boldsymbol{\mu})(\mathbf{x}). \tag{4.7}$$

Lemma 4.1 (see [24], [37, Thm. 5.1]). *The single layer potentials* $\boldsymbol{\Psi}_{\mathbf{A}}^\kappa$ *and* Ψ_V^κ *give rise to continuous mappings* $\Psi_V^\kappa : H^{-1/2}(\Gamma) \mapsto H^1_{\mathrm{loc}}(\mathbb{R}^3)$, $\boldsymbol{\Psi}_{\mathbf{A}}^\kappa : \boldsymbol{H}_{||}^{-1/2}(\Gamma) \mapsto \boldsymbol{H}^1_{\mathrm{loc}}(\mathbb{R}^3)$.

Lemma 4.2 (see [42, Lem. 2.3]). *For* $\mathbf{u} \in \boldsymbol{H}^{-1/2}(\mathrm{div}_\Gamma, \Gamma)$ *we have* $\mathrm{div}\, \boldsymbol{\Psi}_{\mathbf{A}}^\kappa(\mathbf{u}) = \Psi_V^\kappa(\mathrm{div}_\Gamma\mathbf{u})$ *in* $\boldsymbol{L}^2(\mathbb{R}^3)$.

From this we get the identities

$$(\mathbf{curl}\,\mathbf{curl} - \kappa^2 Id)\boldsymbol{\Psi}_{\mathbf{A}}^\kappa(\boldsymbol{\mu}) = \mathbf{grad}\,\Psi_V^\kappa(\mathrm{div}_\Gamma\boldsymbol{\mu}) \quad \forall \boldsymbol{\mu} \in \boldsymbol{H}^{-\frac{1}{2}}(\mathrm{div}_\Gamma, \Gamma), \tag{4.8}$$

$$(\mathbf{curl}\,\mathbf{curl} - \kappa^2 Id)\boldsymbol{\Psi}_{\mathbf{M}}^\kappa(\mathbf{u}) = 0 \qquad \forall \mathbf{u} \in \boldsymbol{H}^{-\frac{1}{2}}(\mathbf{curl}_\Gamma, \Gamma), \tag{4.9}$$

off the boundary Γ in the pointwise sense and, globally, in $\boldsymbol{L}^2_{\mathrm{loc}}(\mathbb{R}^3)$. Summing up, both $\boldsymbol{\Psi}^\kappa_{\mathbf{M}}$ and $\boldsymbol{\Psi}^\kappa_{\mathbf{S}}$ are radiating solutions of the homogeneous electric wave equation in $\Omega_s \cup \Omega'$.

From these relationships and Lemma 4.1 we infer the following continuity properties.

Theorem 4.3. *The mappings* $\boldsymbol{\Psi}^\kappa_{\mathbf{S}} : \boldsymbol{H}^{-\frac{1}{2}}(\mathrm{div}_\Gamma, \Gamma) \mapsto \boldsymbol{H}_{\mathrm{loc}}(\mathbf{curl}^2, \Omega_s \cup \Omega')$ *and* $\boldsymbol{\Psi}^\kappa_{\mathbf{M}} : \boldsymbol{H}^{-\frac{1}{2}}(\mathbf{curl}_\Gamma, \Gamma) \mapsto \boldsymbol{H}_{\mathrm{loc}}(\mathbf{curl}^2, \Omega_s \cup \Omega')$ *are continuous.*

The potentials also satisfy fundamental *jump relations* (cf. [22, Thm. 6.11], [45, Thm. 5.5.1], [37, sect. 5]).

Theorem 4.4. *The interior and exterior Dirichlet and Neumann traces of the potentials* $\boldsymbol{\Psi}^\kappa_{\mathbf{S}}$ *and* $\boldsymbol{\Psi}^\kappa_{\mathbf{M}}$ *are well defined and fulfill for* $\boldsymbol{\mu} \in \boldsymbol{H}^{-\frac{1}{2}}(\mathrm{div}_\Gamma, \Gamma)$, $\mathbf{u} \in \boldsymbol{H}^{-\frac{1}{2}}(\mathbf{curl}_\Gamma, \Gamma)$,

$$[\gamma_D \boldsymbol{\Psi}^\kappa_{\mathbf{S}}(\boldsymbol{\mu})]_\Gamma = 0, \quad [\gamma_N \boldsymbol{\Psi}^\kappa_{\mathbf{S}}(\boldsymbol{\mu})]_\Gamma = -\boldsymbol{\mu}, \quad [\gamma_D \boldsymbol{\Psi}^\kappa_{\mathbf{M}}(\mathbf{u})]_\Gamma = \mathbf{u}, \quad [\gamma_N \boldsymbol{\Psi}^\kappa_{\mathbf{M}}(\mathbf{u})]_\Gamma = 0.$$

This theorem in conjunction with Lemma 4.2 and $\boldsymbol{\Psi}^\kappa_{\mathbf{A}}(\mathsf{R}\mathbf{u}) \in \boldsymbol{H}^1_{\mathrm{loc}}(\mathbb{R}^3)$ supplies further jump relations:

$$[\gamma_{\mathbf{n}} \boldsymbol{\Psi}^\kappa_{\mathbf{M}}(\mathbf{u})]_\Gamma = 0, \qquad [\gamma \,\mathrm{div}\, \boldsymbol{\Psi}^\kappa_{\mathbf{A}}(\boldsymbol{\mu})]_\Gamma = 0. \tag{4.10}$$

5. Integral operators. In the usual fashion, we obtain the relevant boundary integral operators by applying Dirichlet and Neumann trace operators to the potentials of the representation formula.

Lemma 5.1. *The integral operators* $\mathbf{A}_\kappa := \gamma_D \boldsymbol{\Psi}^\kappa_{\mathbf{A}} : \boldsymbol{H}^{-1/2}_{||}(\Gamma) \mapsto \boldsymbol{H}^{1/2}_{||}(\Gamma)$, $\tilde{\mathbf{A}}_\kappa := \gamma_\times \boldsymbol{\Psi}^\kappa_{\mathbf{A}} \circ \mathsf{R} : \boldsymbol{H}^{-1/2}_{\perp}(\Gamma) \mapsto \boldsymbol{H}^{1/2}_{\perp}(\Gamma)$, *and* $V_\kappa := \gamma \Psi^\kappa_V : H^{-1/2}(\Gamma) \mapsto H^{1/2}(\Gamma)$ *are continuous.*

Proof. The assertion is immediate by combining Theorem 4.3 with Theorem 3.1 and properties of the standard trace γ. □

Theorem 5.2. *The following integral operators are continuous:*

$$\begin{aligned}
\mathbf{S}_\kappa &:= \gamma_D \boldsymbol{\Psi}^\kappa_{\mathbf{S}} &&: \boldsymbol{H}^{-\frac{1}{2}}(\mathrm{div}_\Gamma, \Gamma) \mapsto \boldsymbol{H}^{-\frac{1}{2}}(\mathbf{curl}_\Gamma, \Gamma),\\
\mathbf{B}_\kappa &:= \frac{1}{2}(\gamma_N^+ + \gamma_N^-) \boldsymbol{\Psi}^\kappa_{\mathbf{A}} &&: \boldsymbol{H}^{-\frac{1}{2}}(\mathrm{div}_\Gamma, \Gamma) \mapsto \boldsymbol{H}^{-\frac{1}{2}}(\mathrm{div}_\Gamma, \Gamma),\\
\mathbf{C}_\kappa &:= \frac{1}{2}(\gamma_D^+ + \gamma_D^-) \boldsymbol{\Psi}^\kappa_{\mathbf{M}} &&: \boldsymbol{H}^{-\frac{1}{2}}(\mathbf{curl}_\Gamma, \Gamma) \mapsto \boldsymbol{H}^{-\frac{1}{2}}(\mathbf{curl}_\Gamma, \Gamma),\\
\mathbf{N}_\kappa &:= \gamma_N \boldsymbol{\Psi}^\kappa_{\mathbf{M}} &&: \boldsymbol{H}^{-\frac{1}{2}}(\mathbf{curl}_\Gamma, \Gamma) \mapsto \boldsymbol{H}^{-\frac{1}{2}}(\mathrm{div}_\Gamma, \Gamma).
\end{aligned}$$

Proof. The continuity properties instantly follow from those of the potentials stated in Theorem 4.3, and the continuity of the trace operators from Theorem 3.3 and Lemma 3.4. □

Beyond continuity, the integral operators possess numerous important properties. In particular, they are closely related as expressed in the next two lemmas.

Lemma 5.3 (see [21, eq. (3.86)]). *The identity* $\mathbf{N}_\kappa = \mathsf{R}^* \circ \mathbf{S}_\kappa \circ \mathsf{R}$ *holds true.*

Lemma 5.4. *There is a compact linear operator* $T_\kappa : \boldsymbol{H}^{-\frac{1}{2}}(\mathrm{div}_\Gamma, \Gamma) \mapsto \boldsymbol{H}^{-\frac{1}{2}}(\mathrm{div}_\Gamma, \Gamma)$ *such that*

$$\langle \mathbf{B}_\kappa \boldsymbol{\zeta}, \mathbf{q} \rangle_{\boldsymbol{\tau}} = \langle \boldsymbol{\zeta}, \mathbf{C}_\kappa \mathbf{q} \rangle_{\boldsymbol{\tau}} - \langle T_\kappa \boldsymbol{\zeta}, \mathbf{q} \rangle_{\boldsymbol{\tau}} \quad \forall \boldsymbol{\zeta} \in \boldsymbol{H}^{-\frac{1}{2}}(\mathrm{div}_\Gamma, \Gamma),\ \mathbf{q} \in \boldsymbol{H}^{-\frac{1}{2}}(\mathbf{curl}_\Gamma, \Gamma).$$

Proof. The proof follows that of [14, Thm. 3.9]. We fix $\rho > 0$ such that $\bar{\Omega}_s$ is contained in the ball $B_\rho := \{\mathbf{x} \in \mathbb{R}^3, |\mathbf{x}| < \rho\}$. Traces onto ∂B_ρ bear a $\widehat{\ }$. Set $\mathbf{U} := \boldsymbol{\Psi}^\kappa_{\mathbf{A}}(\boldsymbol{\zeta})$, $\mathbf{V} := \boldsymbol{\Psi}^\kappa_{\mathbf{M}}(\mathbf{q})$, and observe that (4.9) involves

$$\begin{aligned}\left\langle \gamma_N^+ \mathbf{V}, \gamma_D^+ \mathbf{U} \right\rangle_{\boldsymbol{\tau}} &= -\int_{\Omega^\rho} \mathbf{curl}\,\mathbf{V} \cdot \mathbf{curl}\,\overline{\mathbf{U}} - \mathbf{curl}\,\mathbf{curl}\,\mathbf{V} \cdot \overline{\mathbf{U}}\,d\mathbf{x} + \langle \widehat{\gamma}_N \mathbf{V}, \widehat{\gamma}_D \mathbf{U} \rangle_{\boldsymbol{\tau},\partial B_\rho} \\ &= -\int_{\Omega^\rho} \mathbf{curl}\,\mathbf{V} \cdot \mathbf{curl}\,\overline{\mathbf{U}}\,d\mathbf{x} + \int_{\Omega^\rho} \kappa^2 \mathbf{V} \cdot \overline{\mathbf{U}}\,d\mathbf{x} + \langle \widehat{\gamma}_N \mathbf{V}, \widehat{\gamma}_D \mathbf{U} \rangle_{\boldsymbol{\tau},\partial B_\rho}, \\ \left\langle \gamma_N^- \mathbf{V}, \gamma_D^- \mathbf{U} \right\rangle_{\boldsymbol{\tau}} &= \int_{\Omega_s} \mathbf{curl}\,\mathbf{V} \cdot \mathbf{curl}\,\overline{\mathbf{U}} - \kappa^2 \mathbf{V} \cdot \overline{\mathbf{U}}\,d\mathbf{x}.\end{aligned}$$

The jump relations of Theorem 4.4 reveal that both expressions agree. Moreover, since $\operatorname{div}\mathbf{V} = 0$ in $\Omega_s \cup \Omega'$,

$$\begin{aligned}\left\langle \gamma_N^- \mathbf{U}, \gamma_D^- \mathbf{V} \right\rangle_{\boldsymbol{\tau}} &= \int_{\Omega_s} \mathbf{curl}\,\mathbf{U} \cdot \mathbf{curl}\,\overline{\mathbf{V}} - \mathbf{curl}\,\mathbf{curl}\,\mathbf{U} \cdot \overline{\mathbf{V}}\,d\mathbf{x} \\ &= \int_{\Omega_s} \mathbf{curl}\,\mathbf{U} \cdot \mathbf{curl}\,\overline{\mathbf{V}} - \left(\kappa^2 \mathbf{U} + \mathbf{grad}\,\Psi_V^\kappa(\operatorname{div}_\Gamma \boldsymbol{\zeta})\right) \overline{\mathbf{V}}\,d\mathbf{x} \\ &= \overline{\left\langle \gamma_N^- \mathbf{V}, \gamma_D^- \mathbf{U} \right\rangle_{\boldsymbol{\tau}}} - \overline{\left\langle \gamma_{\mathbf{n}}^- \mathbf{V}, \gamma^- \Psi_V^\kappa(\operatorname{div}_\Gamma \boldsymbol{\zeta}) \right\rangle_{\frac{1}{2},\Gamma}}\end{aligned}$$

and

$$\begin{aligned}\left\langle \gamma_N^+ \mathbf{U}, \gamma_D^+ \mathbf{V} \right\rangle_{\boldsymbol{\tau}} &= -\int_{\Omega^\rho} \mathbf{curl}\,\mathbf{U} \cdot \mathbf{curl}\,\overline{\mathbf{V}} - \mathbf{curl}\,\mathbf{curl}\,\mathbf{U} \cdot \overline{\mathbf{V}}\,d\mathbf{x} + \langle \widehat{\gamma}_N \mathbf{U}, \widehat{\gamma}_D \mathbf{V} \rangle_{\boldsymbol{\tau}} \\ &= -\int_{\Omega^\rho} \mathbf{curl}\,\mathbf{U} \cdot \mathbf{curl}\,\overline{\mathbf{V}}\,d\mathbf{x} + \int_{\Omega^\rho} \left(\kappa^2 \mathbf{U} + \mathbf{grad}\,\Psi_V^\kappa(\operatorname{div}_\Gamma \boldsymbol{\zeta})\right) \overline{\mathbf{V}}\,d\mathbf{x} + \langle \widehat{\gamma}_N \mathbf{U}, \widehat{\gamma}_D \mathbf{V} \rangle_{\boldsymbol{\tau}} \\ &= \overline{\left\langle \gamma_N^+ \mathbf{V}, \gamma_D^+ \mathbf{U} \right\rangle_{\boldsymbol{\tau}}} - \overline{\left\langle \gamma_{\mathbf{n}}^+ \mathbf{V}, \gamma^+ \Psi_V^\kappa(\operatorname{div}_\Gamma \boldsymbol{\zeta}) \right\rangle_{\frac{1}{2},\Gamma}} + \overline{\langle \widehat{\gamma}_{\mathbf{n}} \mathbf{V}, \widehat{\gamma} \Psi_V^\kappa(\operatorname{div}_\Gamma \boldsymbol{\zeta}) \rangle_{1/2,\partial B_\rho}} \\ &\qquad - \overline{\langle \widehat{\gamma}_N \mathbf{V}, \widehat{\gamma}_D \mathbf{U} \rangle_{\boldsymbol{\tau}}} + \langle \widehat{\gamma}_N \mathbf{U}, \widehat{\gamma}_D \mathbf{V} \rangle_{\boldsymbol{\tau}}.\end{aligned}$$

From the extra jump relation (4.10) we infer

$$\left\langle \gamma_N^- \mathbf{U}, \gamma_D^- \mathbf{V} \right\rangle_{\boldsymbol{\tau}} \quad = \quad \left\langle \gamma_N^+ \mathbf{U}, \gamma_D^+ \mathbf{V} \right\rangle_{\boldsymbol{\tau}} \quad + \quad \{\text{Potentials evaluated on } \partial B_\rho\}. \tag{5.1}$$

The potentials, when restricted to domains off the boundary Γ, are C^∞-smoothing (cf. the proof of Theorem 7.6 in [43]). Thus their evaluations on ∂B_ρ will lead to compact operators. Plugging in the definitions of the boundary integral operators, (5.1) gives the assertion. □

Since, ultimately, we aim to resort to a Fredholm alternative argument, compactness properties of the boundary integral operators deserve special attention. It will be crucial that we are able to switch to the "Laplace kernel" E_0 by a compact perturbation only.

LEMMA 5.5 (see [14, Thm. 3.12], [38, Lem. 3.2]). *The following integral operators are compact:*

$$\begin{aligned}\delta\mathbf{A}_\kappa &:= \mathbf{A}_\kappa - \mathbf{A}_0 &&: \boldsymbol{H}_{||}^{-\frac{1}{2}}(\Gamma) \mapsto \boldsymbol{H}_{||}^{\frac{1}{2}}(\Gamma), \\ \delta\tilde{\mathbf{A}}_\kappa &:= \tilde{\mathbf{A}}_\kappa - \tilde{\mathbf{A}}_0 &&: \boldsymbol{H}_{\perp}^{-\frac{1}{2}}(\Gamma) \mapsto \boldsymbol{H}_{\perp}^{\frac{1}{2}}(\Gamma), \\ \delta V_\kappa &:= V_\kappa - V_0 &&: H^{-\frac{1}{2}}(\Gamma) \mapsto H^{\frac{1}{2}}(\Gamma), \\ \delta\mathbf{N}_\kappa &:= \mathbf{N}_\kappa - \mathbf{N}_0 &&: \boldsymbol{H}^{-\frac{1}{2}}(\mathbf{curl}_\Gamma, \Gamma) \mapsto \boldsymbol{H}^{-\frac{1}{2}}(\operatorname{div}_\Gamma, \Gamma).\end{aligned}$$

The significance of the case $\kappa = 0$ is highlighted by the following result (cf. [43, Cor. 8.13], [31, Vol. IV, Chap. XI, sect. 2, Thm. 3], [12, Prop. 4.1]).

LEMMA 5.6 (see [14, Thm. 3.8]). *The operators V_0, $\tilde{\mathbf{A}}_0$, and $\mathbf{A}_0$ are continuous, self-adjoint and fulfill*

$$\begin{aligned}
\langle \mu, V_0\mu\rangle_{\frac{1}{2},\Gamma} &\geq C\,\|\mu\|^2_{H^{-\frac{1}{2}}(\Gamma)} && \forall \mu \in H^{-\frac{1}{2}}(\Gamma),\\
\langle \boldsymbol{\mu}, \mathbf{A}_0\boldsymbol{\mu}\rangle_{\frac{1}{2},\|,\Gamma} &\geq C\,\|\boldsymbol{\mu}\|^2_{\boldsymbol{H}_{||}^{-\frac{1}{2}}(\Gamma)} && \forall \boldsymbol{\mu} \in \boldsymbol{H}_{||}^{-\frac{1}{2}}(\Gamma),\ \mathrm{div}_\Gamma\boldsymbol{\mu} = 0,\\
\langle \mathbf{v}, \tilde{\mathbf{A}}_0\mathbf{v}\rangle_{\frac{1}{2},\perp,\Gamma} &\geq C\,\|\mathbf{v}\|^2_{\boldsymbol{H}_{\perp}^{-\frac{1}{2}}(\Gamma)} && \forall \mathbf{v} \in \boldsymbol{H}^{-\frac{1}{2}}(\mathbf{curl}_\Gamma,\Gamma),\ \mathbf{curl}_\Gamma\mathbf{v} = 0,
\end{aligned}$$

with constants[2] *$C > 0$ depending only on Γ.*

6. Coupled problem. Applying Green's formula to the electric wave equation in Ω_s results in the variational formulation: Seek $\mathbf{E} \in \boldsymbol{H}(\mathbf{curl};\Omega_s)$ such that

$$(6.1)\qquad \left(\mu_r^{-1}\,\mathbf{curl}\,\mathbf{E}, \mathbf{curl}\,\mathbf{V}\right)_{0;\Omega_s} - \kappa^2\left(\epsilon_r\mathbf{E},\mathbf{V}\right)_{0;\Omega_s} - \left\langle \frac{1}{\mu_r}\gamma_N^-\mathbf{E}, \gamma_D^-\mathbf{V}\right\rangle_{\boldsymbol{\tau}} = 0$$

for all $\mathbf{V} \in \boldsymbol{H}(\mathbf{curl};\Omega_s)$. The coupling to the exterior domain is taken into account through the transmission conditions from (1.1):

$$(6.2)\qquad \frac{1}{\mu_r}\gamma_N^-\mathbf{E} = \gamma_N^+\mathbf{E} + \gamma_N\mathbf{E}_{\mathrm{inc}}, \quad \gamma_D^-\mathbf{E} = \gamma_D^+\mathbf{E} + \gamma_D\mathbf{E}_{\mathrm{inc}}.$$

In addition, some realization of the exterior Dirichlet-to-Neumann map has to be provided. It is furnished by the *exterior Calderón projector*, which arises from applying both the exterior Dirichlet and Neumann traces to the representation formula (4.6) (cf. [32, eq. (29)], [45, sect. 5.5], [14, sect. 3.3], [39, eq. (24)]). The resulting identity reads in variational form: For all $\boldsymbol{\mu} \in \boldsymbol{H}^{-\frac{1}{2}}(\mathrm{div}_\Gamma,\Gamma)$, $\mathbf{v} \in \boldsymbol{H}^{-\frac{1}{2}}(\mathbf{curl}_\Gamma,\Gamma)$ the following must hold:

$$(6.3)\qquad \begin{aligned}
\langle \boldsymbol{\mu}, \gamma_D^+\mathbf{E}\rangle_{\boldsymbol{\tau}} &= \left\langle \boldsymbol{\mu}, \left(\frac{1}{2}Id + \mathbf{C}_\kappa\right)(\gamma_D^+\mathbf{E})\right\rangle_{\boldsymbol{\tau}} - \langle \boldsymbol{\mu}, \mathbf{S}_\kappa(\gamma_N^+\mathbf{E})\rangle_{\boldsymbol{\tau}},\\
\langle \gamma_N^+\mathbf{E}, \mathbf{v}\rangle_{\boldsymbol{\tau}} &= \langle \mathbf{N}_\kappa(\gamma_D^+\mathbf{E}), \mathbf{v}\rangle_{\boldsymbol{\tau}} + \left\langle \left(\frac{1}{2}Id - \mathbf{B}_\kappa\right)(\gamma_N^+\mathbf{E}), \mathbf{v}\right\rangle_{\boldsymbol{\tau}}.
\end{aligned}$$

Now we can use the transmission conditions (6.2) and the second equation of the Calderón projector in order to replace the boundary term in (6.1). The trick underlying the symmetric coupling according to Costabel [23] is to retain the first equation of (6.3) in addition (cf. [32, sect. 4] for Maxwell's equations). Adopting the abbreviation $\boldsymbol{\lambda} := \gamma_N^+\mathbf{E}$, we arrive at the formulation: Seek $\mathbf{E} \in \boldsymbol{H}(\mathbf{curl};\Omega_s)$, $\boldsymbol{\lambda} \in \boldsymbol{H}^{-\frac{1}{2}}(\mathrm{div}_\Gamma,\Gamma)$ such that for all $\mathbf{V} \in \boldsymbol{H}(\mathbf{curl};\Omega_s)$, $\boldsymbol{\mu} \in \boldsymbol{H}^{-\frac{1}{2}}(\mathrm{div}_\Gamma,\Gamma)$

$$(6.4)\qquad \begin{aligned}
q_\kappa(\mathbf{E},\mathbf{V}) - \langle \mathbf{N}_\kappa\gamma_D^-\mathbf{E}, \gamma_D^-\mathbf{V}\rangle_{\boldsymbol{\tau}} + \left\langle \left(-\frac{1}{2}Id + \mathbf{B}_\kappa\right)\boldsymbol{\lambda}, \gamma_D^-\mathbf{V}\right\rangle_{\boldsymbol{\tau}} &= f(\mathbf{V}),\\
-\left\langle \boldsymbol{\mu}, \left(-\frac{1}{2}Id + \mathbf{C}_\kappa\right)(\gamma_D^-\mathbf{E})\right\rangle_{\boldsymbol{\tau}} + \langle \boldsymbol{\mu}, \mathbf{S}_\kappa\boldsymbol{\lambda}\rangle_{\boldsymbol{\tau}} &= g(\boldsymbol{\mu}),
\end{aligned}$$

[2] We use C to designate generic constants that may depend only on "fixed quantities" as Γ, κ, and the material parameters ϵ_r, μ_r. Their values might vary between different occurrences.

with right-hand sides $f(\mathbf{V}) := \left\langle \gamma_N \mathbf{E}_{\mathrm{inc}}, \gamma_D^- \mathbf{V} \right\rangle_{\boldsymbol{\tau}} - \left\langle \mathbf{N}_\kappa(\gamma_D \mathbf{E}_{\mathrm{inc}}), \gamma_D^- \mathbf{V} \right\rangle_{\boldsymbol{\tau}}$, $g(\boldsymbol{\mu}) = -\left\langle \boldsymbol{\mu}, (\frac{1}{2} Id + \mathbf{C}_\kappa) \gamma_D \mathbf{E}_{\mathrm{inc}} \right\rangle_{\boldsymbol{\tau}}$, and $q_\kappa(\cdot,\cdot)$ representing the interior sesqui-linear form, that is, $q_\kappa(\mathbf{E},\mathbf{V}) := \left(\mu_r^{-1} \mathbf{curl}\, \mathbf{E}, \mathbf{curl}\, \mathbf{V}\right)_{0;\Omega_s} - \kappa^2 \left(\epsilon_r \mathbf{E}, \mathbf{V}\right)_{0;\Omega_s}$.

LEMMA 6.1. *Provided that Assumption* 1 *holds, a solution of* (6.4) *provides a solution of* (1.1) *by retaining* $\mathbf{E}$ *in* Ω_s *and using the representation formula* (4.6) *for the Cauchy data* $(\gamma_D^- \mathbf{E} + \gamma_D \mathbf{E}_{\mathrm{inc}}, \boldsymbol{\lambda})$ *in* Ω'.

Proof. Our approach is based on [49, sect. 4.3] and [14, sect. 5]. Testing with $\mathbf{V}$ that is compactly supported in Ω_s confirms that $\mathbf{E}$ satisfies (1.1) in Ω_s. We conclude (6.1) for any admissible $\mathbf{V}$. This renders (6.4) equivalent to

$$\begin{aligned} \left\langle \boldsymbol{\xi}, \gamma_D^- \mathbf{V} \right\rangle_{\boldsymbol{\tau}} - \left\langle \mathbf{N}_\kappa \mathbf{u}, \gamma_D^- \mathbf{V} \right\rangle_{\boldsymbol{\tau}} &+ \left\langle \left(-\frac{1}{2} Id + \mathbf{B}_\kappa \right) \boldsymbol{\lambda}, \gamma_D^- \mathbf{V} \right\rangle_{\boldsymbol{\tau}} = 0, \\ \left\langle \boldsymbol{\mu}, \left(-\frac{1}{2} Id + \mathbf{C}_\kappa \right) \mathbf{u} \right\rangle_{\boldsymbol{\tau}} &- \left\langle \boldsymbol{\mu}, \mathbf{S}_\kappa \boldsymbol{\lambda} \right\rangle_{\boldsymbol{\tau}} = 0, \end{aligned}$$

with $\boldsymbol{\xi} := \frac{1}{\mu_r} \gamma_N^- \mathbf{E} - \gamma_N \mathbf{E}_{\mathrm{inc}}$, $\mathbf{u} := \gamma_D^- \mathbf{E} - \gamma_D \mathbf{E}_{\mathrm{inc}}$. In operator notation this means

$$\begin{aligned} -\mathbf{N}_\kappa \mathbf{u} &+ \left(\frac{1}{2} Id + \mathbf{B}_\kappa \right) \boldsymbol{\lambda} = \boldsymbol{\lambda} - \boldsymbol{\xi}, \\ \left(\frac{1}{2} Id - \mathbf{C}_\kappa \right) \mathbf{u} &+ \mathbf{S}_\kappa \boldsymbol{\lambda} = 0. \end{aligned} \tag{6.5}$$

We recognize the operator in (6.5) as an *interior Calderon projector* [14, sect. 3.3]. Its range comprises valid Cauchy data for boundary value problems for $\mathbf{curl}\,\mathbf{curl}\,\mathbf{U} - \kappa^2 \mathbf{U} = 0$ in Ω_s. In particular, $\boldsymbol{\lambda} - \boldsymbol{\xi}$ are seen to be Neumann boundary values of interior Dirichlet eigensolutions. According to Assumption 1 these are trivial, which implies $\boldsymbol{\xi} = \boldsymbol{\lambda}$.

From this we immediately conclude that $(\mathbf{u}, \boldsymbol{\lambda})$ must belong to the range of the exterior Calderon projector. Hence, $(\mathbf{u}, \boldsymbol{\lambda})$ are valid Cauchy data for the exterior field problem. By definition, $(\mathbf{u} + \gamma_D \mathbf{E}_{\mathrm{inc}}, \boldsymbol{\xi})$ are valid Cauchy data for the interior field problem. Taken together with $\boldsymbol{\xi} = \boldsymbol{\lambda}$, this finishes the proof. □

Remark 6.1. If κ^2 coincides with an interior Dirichlet eigenvalue, then the solution of (6.4) is unique only up to a contribution $(0, \boldsymbol{\eta})$, where $\boldsymbol{\eta}$ lies in the span of Neumann data belonging to interior Dirichlet eigensolutions. In particular, $\gamma_D^- \mathbf{E}$ is unique.

To keep notations short, we set $\boldsymbol{\mathcal{V}} := \boldsymbol{H}(\mathbf{curl}; \Omega_s) \times \boldsymbol{H}^{-\frac{1}{2}}(\mathrm{div}_\Gamma, \Gamma)$ and write $a_\kappa : \boldsymbol{\mathcal{V}} \times \boldsymbol{\mathcal{V}} \mapsto \mathbb{C}$ for the sesqui-linear form associated with the variational problem (6.4). Then the latter can be stated as follows: Seek $(\mathbf{E}, \boldsymbol{\lambda}) \in \boldsymbol{\mathcal{V}}$ such that

$$a_\kappa((\mathbf{E}, \boldsymbol{\lambda}), (\mathbf{V}, \boldsymbol{\mu})) = f(\mathbf{V}) + g(\boldsymbol{\mu}) \quad \forall (\mathbf{V}, \boldsymbol{\mu}) \in \boldsymbol{\mathcal{V}}. \tag{6.6}$$

At first glance, the resulting variational problem much resembles those we get in the case of strongly elliptic second order elliptic problems, as they are encountered, for instance, in models for acoustic scattering (cf. section 2). Now, $\mathbf{S}_\kappa$ seems to play the role of a single layer potential V_κ, $\mathbf{B}_\kappa$ and $\mathbf{C}_\kappa$ act as double layer boundary integral operators, and $\mathbf{N}_\kappa$ substitutes for the hypersingular operator D_κ. A closer scrutiny reveals that appearances are deceptive: The key feature of V_κ and D_κ is that they are strongly elliptic operators of order -1 and 1, respectively. Conversely, as is immediate from the formula (4.7) and Lemma 5.3, neither $\mathbf{S}_\kappa$ nor $\mathbf{N}_\kappa$ can be assigned orders, let alone different orders. They both comprise two terms of order 1 and -1, respectively, neither of which can be identified as the principal part. This mirrors the characteristics of the electric wave operator, as discussed in section 2.

7. Decompositions. The considerations of section 2 suggest that we study $\boldsymbol{L}^2(\Omega_s)$-orthogonal Helmholtz decomposition of the electric field into an irrotational component and some complement. However, it turns out that only the energetic stability of the splitting is essential, not its exact orthogonality. Hence, we decided to trade orthogonality for regularity of the magnetic component.

In the case of $\boldsymbol{H}(\mathbf{curl};\Omega_s)$, the construction of such a *Helmholtz-type decomposition* hinges on the existence of vector potentials in $\boldsymbol{H}^1(\Omega_s)$.

LEMMA 7.1 (see [4, Lem. 3.5]). *There is a linear continuous lifting operator* $\mathsf{L} : \boldsymbol{H}(\operatorname{div} 0;\Omega_s) := \{\mathbf{V} \in \boldsymbol{L}^2(\Omega_s), \operatorname{div}\mathbf{V} = 0\} \mapsto \boldsymbol{H}^1(\Omega_s)$ *that satisfies* $\operatorname{div}(\mathsf{L}\mathbf{U}) = 0$ *and* $\mathbf{curl}(\mathsf{L}\mathbf{U}) = \mathbf{U}$ *for all* $\mathbf{U} \in \boldsymbol{H}(\operatorname{div} 0;\Omega_s)$.

Using this device, we introduce the operator

$$\mathsf{P} : \boldsymbol{H}(\mathbf{curl};\Omega_s) \mapsto \boldsymbol{H}^1(\Omega_s), \qquad \mathsf{P}\mathbf{U} := \mathsf{L}(\mathbf{curl}\,\mathbf{U}).$$

From the properties of L we immediately conclude numerous features of P as follows.

LEMMA 7.2. *The operator* P *is a continuous projection that preserves the* $\mathbf{curl}$ *and satisfies* $\operatorname{Ker}(\mathsf{P}) = \operatorname{Ker}(\mathbf{curl}) \cap \boldsymbol{H}(\mathbf{curl};\Omega_s)$.

As $\operatorname{Ker}(\mathsf{P}) = \operatorname{Ker}(\mathbf{curl}) \cap \boldsymbol{H}(\mathbf{curl};\Omega_s)$, it has become evident that the closed subspaces

$$\mathbf{X}(\mathbf{curl},\Omega_s) := \mathsf{P}(\boldsymbol{H}(\mathbf{curl};\Omega_s)) \quad \text{and} \quad \mathbf{N}(\mathbf{curl},\Omega_s) := \operatorname{Ker}(\mathbf{curl}) \subset \boldsymbol{H}(\mathbf{curl};\Omega_s)$$

provide a stable and direct Helmholtz-type splitting

$$\boldsymbol{H}(\mathbf{curl};\Omega_s) = \mathbf{X}(\mathbf{curl},\Omega_s) \oplus \mathbf{N}(\mathbf{curl},\Omega_s). \tag{7.1}$$

For both components we retain the $\boldsymbol{H}(\mathbf{curl};\Omega_s)$-norm. Keeping in mind the discussion of the components of the Helmholtz decomposition in section 2, we easily identify $\mathbf{X}(\mathbf{curl},\Omega_s)$ as the space of "magnetic components" and $\mathbf{N}(\mathbf{curl},\Omega_s)$ as "electric" space. For later compactness arguments the extra regularity of the $\mathbf{X}(\mathbf{curl},\Omega_s)$-component, which is contained in $\boldsymbol{H}^1(\Omega_s)$, is pivotal, since it immediately yields the following compact embedding.

COROLLARY 7.3. *The embedding* $\mathbf{X}(\mathbf{curl},\Omega_s) \hookrightarrow \boldsymbol{L}^2(\Omega_s)$ *is compact.*

In addition, we need splittings of the Neumann trace space $\boldsymbol{H}^{-\frac{1}{2}}(\operatorname{div}_\Gamma,\Gamma)$. Here we could use an $\boldsymbol{L}^2_{\mathbf{t}}(\Gamma)$-orthogonal Hodge decomposition as in [14]. However, as before, we waive orthogonality in favor of enhanced regularity of an algebraic complement of $\operatorname{Ker}(\operatorname{div}_\Gamma)$. The construction is largely parallel to that of $\mathbf{X}(\mathbf{curl},\Omega_s)$: Pick any $\boldsymbol{\lambda} \in \boldsymbol{H}^{-\frac{1}{2}}(\operatorname{div}_\Gamma,\Gamma)$ and set $\omega := \operatorname{div}_\Gamma\boldsymbol{\lambda} \in H^{-\frac{1}{2}}(\Gamma)$. Solve the Neumann problem

$$\Psi \in H^1(\Omega_s)/\mathbb{R} : \quad \Delta\Psi = 0 \quad \text{in } \Omega_s, \quad \gamma_{\mathbf{n}}^{-}\,\mathbf{grad}\,\Psi = w \quad \text{on } \Gamma.$$

We find that $\mathbf{W} := \mathbf{grad}\,\Psi \in \boldsymbol{H}(\operatorname{div} 0;\Omega_s)$ belongs to the domain of the lifting L. Hence, it makes sense to introduce the operator $\mathsf{J} : H^{-\frac{1}{2}}(\Gamma) \mapsto \boldsymbol{H}^1(\Omega_s)$ by $\mathsf{J}\omega := \mathsf{L}\mathbf{W}$. Its continuity is elementary and, thanks to Theorem 3.1, inherited by the mapping $\mathsf{P}^\Gamma := \gamma_\times \circ \mathsf{J} \circ \operatorname{div}_\Gamma : \boldsymbol{H}^{-\frac{1}{2}}(\operatorname{div}_\Gamma,\Gamma) \mapsto \boldsymbol{H}^{\frac{1}{2}}_{\perp}(\Gamma)$. Properties of P^Γ matching those of P can be easily established.

LEMMA 7.4. *The operator* $\mathsf{P}^\Gamma : \boldsymbol{H}^{-\frac{1}{2}}(\operatorname{div}_\Gamma,\Gamma) \mapsto \boldsymbol{H}^{\frac{1}{2}}_{\perp}(\Gamma)$ *is a continuous projection and preserves* $\operatorname{div}_\Gamma$*, and* $\operatorname{Ker}(\mathsf{P}^\Gamma) = \operatorname{Ker}(\operatorname{div}_\Gamma) \cap \boldsymbol{H}^{-\frac{1}{2}}(\operatorname{div}_\Gamma,\Gamma)$.

Through the components

$$\mathbf{X}(\mathrm{div}_\Gamma,\Gamma) := \mathsf{P}^\Gamma(\boldsymbol{H}^{-\frac{1}{2}}(\mathrm{div}_\Gamma,\Gamma)), \quad \mathbf{N}(\mathrm{div}_\Gamma,\Gamma) := \mathrm{Ker}(\mathrm{div}_\Gamma) \cap \boldsymbol{H}^{-\frac{1}{2}}(\mathrm{div}_\Gamma,\Gamma),$$

we arrive at a stable direct decomposition of the space of magnetic traces:

$$\boldsymbol{H}^{-\frac{1}{2}}(\mathrm{div}_\Gamma,\Gamma) := \mathbf{X}(\mathrm{div}_\Gamma,\Gamma) \oplus \mathbf{N}(\mathrm{div}_\Gamma,\Gamma). \tag{7.2}$$

In light of the remarks in section 2, and recalling that the space $\boldsymbol{H}^{-\frac{1}{2}}(\mathrm{div}_\Gamma,\Gamma)$ contains twisted tangential traces of magnetic fields, the two components in the splitting (7.2) correspond to *electric* and *magnetic* field components, respectively.

As before, the extra regularity of $\mathbf{X}(\mathrm{div}_\Gamma,\Gamma)$ rewards us with a valuable compact embedding analogous to [14, Thm. 3.4].

COROLLARY 7.5. *The embedding* $\mathbf{X}(\mathrm{div}_\Gamma,\Gamma) \hookrightarrow \boldsymbol{L}^2_{\mathbf{t}}(\Gamma)$ *is compact.*

8. Coercivity. We decompose the trial and test functions in the variational problem (6.4) according to the splittings provided in the previous section:

$$\begin{aligned}
\mathbf{E} &= \mathbf{E}^\perp + \mathbf{E}^0, \quad \mathbf{E}^\perp \in \mathbf{X}(\mathbf{curl},\Omega_s),\ \mathbf{E}^0 \in \mathbf{N}(\mathbf{curl},\Omega_s),\\
\mathbf{V} &= \mathbf{V}^\perp + \mathbf{V}^0, \quad \mathbf{V}^\perp \in \mathbf{X}(\mathbf{curl},\Omega_s),\ \mathbf{V}^0 \in \mathbf{N}(\mathbf{curl},\Omega_s),\\
\boldsymbol{\lambda} &= \boldsymbol{\lambda}^0 + \boldsymbol{\lambda}^\perp, \quad \boldsymbol{\lambda}^0 \in \mathbf{N}(\mathrm{div}_\Gamma,\Gamma),\ \boldsymbol{\lambda}^\perp \in \mathbf{X}(\mathrm{div}_\Gamma,\Gamma),\\
\boldsymbol{\mu} &= \boldsymbol{\mu}^0 + \boldsymbol{\mu}^\perp, \quad \boldsymbol{\mu}^0 \in \mathbf{N}(\mathrm{div}_\Gamma,\Gamma),\ \boldsymbol{\mu}^\perp \in \mathbf{X}(\mathrm{div}_\Gamma,\Gamma).
\end{aligned}$$

In addition, we sort the unknowns according to their "electric" or "magnetic" nature, grouping them as $(\boldsymbol{\lambda}^\perp, \mathbf{E}^0)$ (electric), $(\boldsymbol{\lambda}^0, \mathbf{E}^\perp)$ (magnetic). Thus we arrive at a variational problem with a distinct block structure. After flipping the signs of the first two equations, it reads: Find $\boldsymbol{\lambda}^\perp \in \mathbf{X}(\mathrm{div}_\Gamma,\Gamma)$, $\mathbf{E}^0 \in \mathbf{N}(\mathbf{curl},\Omega_s)$, $\boldsymbol{\lambda}^0 \in \mathbf{N}(\mathrm{div}_\Gamma,\Gamma)$, $\mathbf{E}^\perp \in \mathbf{X}(\mathbf{curl},\Omega_s)$ such that

$$\begin{array}{ccccccl}
\ast_{11} & + & \ast_{12} & = & g(\boldsymbol{\mu}^\perp) & \forall \boldsymbol{\mu}^\perp \in \mathbf{X}(\mathrm{div}_\Gamma,\Gamma), \\
 & + & & = & f(\mathbf{V}^0) & \forall \mathbf{V}^0 \in \mathbf{N}(\mathbf{curl},\Omega_s), \\
\ast_{21} & + & \ast_{22} & = & g(\boldsymbol{\mu}^0) & \forall \boldsymbol{\mu}^0 \in \mathbf{N}(\mathrm{div}_\Gamma,\Gamma), \\
 & + & & = & f(\mathbf{V}^\perp) & \forall \mathbf{V}^\perp \in \mathbf{X}(\mathbf{curl},\Omega_s),
\end{array} \tag{8.1}$$

where

$$\ast_{11} := \begin{array}{ccc}
\left\langle \boldsymbol{\mu}^\perp, \mathbf{S}_\kappa \boldsymbol{\lambda}^\perp \right\rangle_{\boldsymbol{\tau}} & + & \left\langle \boldsymbol{\mu}^\perp, \left(-\frac{1}{2}Id + \mathbf{C}_\kappa\right) \gamma_D^- \mathbf{E}^0 \right\rangle_{\boldsymbol{\tau}}, \\
\left\langle \left(\frac{1}{2}Id - \mathbf{B}_\kappa\right) \boldsymbol{\lambda}^\perp, \gamma_D^- \mathbf{V}^0 \right\rangle_{\boldsymbol{\tau}} & + & \kappa^2 \left\langle \gamma_D^- \mathbf{V}^0, \tilde{\mathbf{A}}_\kappa \gamma_D^- \mathbf{E}^0 \right\rangle_{\frac{1}{2},\perp,\Gamma} + \kappa^2 \left(\epsilon_r \mathbf{E}^0, \mathbf{V}^0\right)_{0;\Omega_s},
\end{array}$$

$$\ast_{12} := \begin{array}{ccc}
-\left\langle \boldsymbol{\mu}^\perp, \mathbf{A}_\kappa \boldsymbol{\lambda}^0 \right\rangle_{\frac{1}{2},\|,\Gamma} & + & \left\langle \boldsymbol{\mu}^\perp, \left(-\frac{1}{2}Id + \mathbf{C}_\kappa\right) \gamma_D^- \mathbf{E}^\perp \right\rangle_{\boldsymbol{\tau}}, \\
\left\langle \left(\frac{1}{2}Id - \mathbf{B}_\kappa\right) \boldsymbol{\lambda}^0, \gamma_D^- \mathbf{V}^0 \right\rangle_{\boldsymbol{\tau}} & + & \kappa^2 \left\langle \gamma_D^- \mathbf{E}^\perp, \tilde{\mathbf{A}}_\kappa \gamma_D^- \mathbf{V}^0 \right\rangle_{\frac{1}{2},\perp,\Gamma} + \kappa^2 \left(\epsilon_r \mathbf{E}^\perp, \mathbf{V}^0\right)_{0;\Omega^-},
\end{array}$$

$$\ast_{21} := \begin{array}{ccc}
\left\langle \boldsymbol{\mu}^0, \mathbf{A}_\kappa \boldsymbol{\lambda}^\perp \right\rangle_{\frac{1}{2},\|,\Gamma} & - & \left\langle \boldsymbol{\mu}^0, \left(-\frac{1}{2}Id + \mathbf{C}_\kappa\right) \gamma_D^- \mathbf{E}^0 \right\rangle_{\boldsymbol{\tau}}, \\
\left\langle \left(-\frac{1}{2}Id + \mathbf{B}_\kappa\right) \boldsymbol{\lambda}^\perp, \gamma_D^- \mathbf{V}^\perp \right\rangle_{\boldsymbol{\tau}} & - & \kappa^2 \left\langle \gamma_D^- \mathbf{E}^0, \tilde{\mathbf{A}}_\kappa \gamma_D^- \mathbf{V}^\perp \right\rangle_{\frac{1}{2},\perp,\Gamma} - \kappa^2 \left(\epsilon \mathbf{E}^0, \mathbf{V}^\perp\right)_{0;\Omega^-},
\end{array}$$

$$\boldsymbol{\ast}_{22} := \begin{array}{ccc} \left\langle \boldsymbol{\mu}^0, \mathbf{A}_\kappa \boldsymbol{\lambda}^0 \right\rangle_{\frac{1}{2},\|,\Gamma} & - & \left\langle \boldsymbol{\mu}^0, \left(-\frac{1}{2} Id + \mathbf{C}_\kappa\right) \gamma_D^- \mathbf{E}^\perp \right\rangle_{\boldsymbol{\tau}}, \\ \left\langle \left(-\frac{1}{2} Id + \mathbf{B}_\kappa\right) \boldsymbol{\lambda}^0, \gamma_D^- \mathbf{V}^\perp \right\rangle_{\boldsymbol{\tau}} & - & \left\langle \mathbf{N}_\kappa \gamma_D^- \mathbf{E}^\perp, \gamma_D^- \mathbf{V}^\perp \right\rangle_{\boldsymbol{\tau}} + q(\mathbf{E}^\perp, \mathbf{V}^\perp) \; . \end{array}$$

Here, we have used that the first order parts of the operators $\mathbf{S}_\kappa$ and $\mathbf{N}_\kappa$ disappear when those are applied to functions in $\mathbf{N}(\mathrm{div}_\Gamma, \Gamma)$ and $\gamma_D^- \mathbf{N}(\mathbf{curl}, \Omega_s)$, respectively. Evidently, (8.1) and (6.4) produce exactly the same solutions for $\boldsymbol{\lambda} = \boldsymbol{\lambda}^\perp + \boldsymbol{\lambda}^0$ and $\mathbf{E} = \mathbf{E}^\perp + \mathbf{E}^0$, provided that a unique solution exists. It is also clear that (6.4) can be written as a variational problem for a continuous sesqui-linear form $\widehat{a}_\kappa$ on the Hilbert space $\boldsymbol{\mathcal{G}} := \mathbf{X}(\mathrm{div}_\Gamma, \Gamma) \times \mathbf{N}(\mathbf{curl}, \Omega_s) \times \mathbf{N}(\mathrm{div}_\Gamma, \Gamma) \times \mathbf{X}(\mathbf{curl}, \Omega_s)$ that is endowed with the natural graph norm. Also, $\widehat{a}_\kappa$ is given by

$$\begin{aligned} (8.2) \qquad & \widehat{a}_\kappa((\boldsymbol{\zeta}^\perp, \mathbf{U}^0, \boldsymbol{\zeta}^0, \mathbf{U}^\perp), (\boldsymbol{\mu}^\perp, \mathbf{V}^0, \boldsymbol{\mu}^0, \mathbf{V}^\perp)) \\ & \qquad = a_\kappa((\mathbf{U}^\perp + \mathbf{U}^0, \boldsymbol{\zeta}^\perp + \boldsymbol{\zeta}^0), (\mathbf{V}^\perp - \mathbf{V}^0, -\boldsymbol{\mu}^\perp + \boldsymbol{\mu}^0)). \end{aligned}$$

The goal is to resort to compact perturbations and achieve a block structure with elliptic diagonal blocks and off-diagonal blocks that fit a skew-symmetric pattern. To this end we have to identify operators in the above variational problem that can be neglected because they are compact.

LEMMA 8.1. *The operators*

$$(\gamma_D^-)^* \circ \left(-\frac{1}{2} Id + \mathbf{B}_\kappa\right) : \mathbf{N}(\mathrm{div}_\Gamma, \Gamma) \mapsto \mathbf{N}(\mathbf{curl}, \Omega_s)',$$

$$\left(-\frac{1}{2} Id + \mathbf{C}_\kappa\right) \circ \gamma_D^- : \mathbf{N}(\mathbf{curl}, \Omega_s) \mapsto \mathbf{N}(\mathrm{div}_\Gamma, \Gamma)'$$

are compact.

Proof. The proof closely follows that of [14, Prop. 3.13]. Pick $\boldsymbol{\zeta} \in \mathbf{N}(\mathrm{div}_\Gamma, \Gamma)$, $\mathbf{V} \in \mathbf{N}(\mathbf{curl}, \Omega_s)$, and set $\mathbf{v} := \gamma_D^- \mathbf{V}$. Note that there is a direct splitting $\mathbf{N}(\mathrm{div}_\Gamma, \Gamma) = \mathbf{curl}_\Gamma H^{\frac{1}{2}}(\Gamma) \oplus \boldsymbol{\mathcal{H}}_1(\Gamma)$, where $\boldsymbol{\mathcal{H}}_1(\Gamma)$ is some cohomology space of Γ, whose dimension is finite and agrees with the first Betti number $\beta_1(\Gamma)$ (cf. [8]). Hence, we can write

$$\boldsymbol{\zeta} = \mathbf{curl}_\Gamma \phi(\boldsymbol{\zeta}) + \boldsymbol{\eta}(\boldsymbol{\zeta}), \quad \phi(\boldsymbol{\zeta}) \in H^{\frac{1}{2}}(\Gamma), \quad \boldsymbol{\eta}(\boldsymbol{\zeta}) \in \boldsymbol{\mathcal{H}}_1(\Gamma).$$

Further, let $\Phi \in H^1(\Omega_s)$ be some extension of ϕ. The key to the proof is the observation

$$\langle \boldsymbol{\zeta}, \mathbf{v} \rangle_{\boldsymbol{\tau}} = \int_\Gamma \mathbf{curl}\, \mathbf{V} \cdot \overline{\mathbf{grad}\, \Phi} - \mathbf{V} \cdot \mathbf{curl}\, \overline{\mathbf{grad}\, \Phi} \, d\mathbf{x} + \langle \boldsymbol{\eta}(\boldsymbol{\zeta}), \mathbf{v} \rangle_{\boldsymbol{\tau}} = \langle \boldsymbol{\eta}(\boldsymbol{\zeta}), \mathbf{v} \rangle_{\boldsymbol{\tau}} \, .$$

By the jump relations, the identity (4.8), and the weak definition of the Neumann trace γ_N^-,

$$\begin{aligned} \left\langle \gamma_N^+ \boldsymbol{\Psi}_{\mathbf{A}}^\kappa(\boldsymbol{\zeta}), \mathbf{v} \right\rangle_{\boldsymbol{\tau}} &= \left\langle -\boldsymbol{\zeta} + \gamma_N^- \boldsymbol{\Psi}_{\mathbf{A}}^\kappa(\boldsymbol{\zeta}), \mathbf{v} \right\rangle_{\boldsymbol{\tau}} = \left\langle \gamma_N^- \boldsymbol{\Psi}_{\mathbf{A}}^\kappa(\boldsymbol{\zeta}), \mathbf{v} \right\rangle_{\boldsymbol{\tau}} - \langle \boldsymbol{\eta}(\boldsymbol{\zeta}), \mathbf{v} \rangle_{\boldsymbol{\tau}} \\ &= \int_{\Omega_s} \mathbf{curl}\, \boldsymbol{\Psi}_{\mathbf{A}}^\kappa(\boldsymbol{\zeta}) \cdot \mathbf{curl}\, \overline{\mathbf{V}} - \mathbf{curl}\, \mathbf{curl}\, \boldsymbol{\Psi}_{\mathbf{A}}^\kappa(\boldsymbol{\zeta}) \cdot \overline{\mathbf{V}} \, d\mathbf{x} - \langle \boldsymbol{\eta}(\boldsymbol{\zeta}), \mathbf{v} \rangle_{\boldsymbol{\tau}} \\ &= \int_{\Omega_s} \mathbf{curl}\, \boldsymbol{\Psi}_{\mathbf{A}}^\kappa(\boldsymbol{\zeta}) \cdot \mathbf{curl}\, \overline{\mathbf{V}} - \mathbf{grad}\, \Psi_V^\kappa(\mathrm{div}_\Gamma \boldsymbol{\zeta}) \cdot \overline{\mathbf{V}} - \kappa^2 \boldsymbol{\Psi}_{\mathbf{A}}^\kappa(\boldsymbol{\zeta}) \cdot \overline{\mathbf{V}} \, d\mathbf{x} - \langle \boldsymbol{\eta}(\boldsymbol{\zeta}), \mathbf{v} \rangle_{\boldsymbol{\tau}} \, . \end{aligned}$$

The first two terms can be dropped as $\mathbf{V}$ and $\boldsymbol{\zeta}$ are **curl**-free and div_Γ-free, respectively. This leaves us with

$$|\left\langle \gamma_N^+ \boldsymbol{\Psi}_{\mathbf{A}}^{\kappa}(\boldsymbol{\zeta}), \mathbf{v} \right\rangle_{\boldsymbol{\tau}} + \langle \boldsymbol{\eta}(\boldsymbol{\zeta}), \mathbf{v} \rangle_{\boldsymbol{\tau}} | \leq \kappa^2 \left\| \boldsymbol{\Psi}_{\mathbf{A}}^{\kappa}(\boldsymbol{\zeta}) \right\|_{\boldsymbol{L}^2(\Omega_s)} \|\mathbf{V}\|_{\boldsymbol{L}^2(\Omega_s)} .$$

As $\dim \boldsymbol{\mathcal{H}}_1(\Gamma) < \infty$, the mapping $\mathbf{N}(\mathrm{div}_\Gamma, \Gamma) \mapsto \boldsymbol{\mathcal{H}}_1(\Gamma)$, $\boldsymbol{\zeta} \mapsto \boldsymbol{\eta}(\boldsymbol{\zeta})$, is compact. Since $\boldsymbol{\Psi}_{\mathbf{A}}^{\kappa} : \boldsymbol{H}_{||}^{-\frac{1}{2}}(\Gamma) \mapsto \boldsymbol{H}^1(\Omega_s)$ is continuous, the compact embedding $\boldsymbol{H}^1(\Omega_s) \hookrightarrow \boldsymbol{L}^2(\Omega_s)$ confirms the first assertion of the theorem. Thanks to Lemma 5.4, the second is then immediate. $\square$

LEMMA 8.2. *The operators*

$$(\gamma_D^-)^* \circ \left(-\frac{1}{2} Id + \mathbf{B}_\kappa\right) : \mathbf{X}(\mathrm{div}_\Gamma, \Gamma) \mapsto \mathbf{X}(\mathbf{curl}, \Omega_s)',$$

$$\left(-\frac{1}{2} Id + \mathbf{C}_\kappa\right) \circ \gamma_D^- : \mathbf{X}(\mathbf{curl}, \Omega_s) \mapsto \mathbf{X}(\mathrm{div}_\Gamma, \Gamma)'$$

are compact.

Proof. The proof runs parallel to that of [14, Prop. 3.13]. Choose any $\boldsymbol{\zeta} \in \mathbf{X}(\mathrm{div}_\Gamma, \Gamma)$, $\mathbf{V} \in \mathbf{X}(\mathbf{curl}, \Omega_s)$, and recall the definition of $\mathbf{B}_\kappa$ along with the jump conditions

$$\begin{aligned}
&\left\langle \left(-\frac{1}{2} Id + \mathbf{B}_\kappa\right) \boldsymbol{\zeta}, \gamma_D^- \mathbf{V} \right\rangle_{\boldsymbol{\tau}} = \left\langle \gamma_N^- \boldsymbol{\Psi}_{\mathbf{A}}^{\kappa}(\boldsymbol{\zeta}), \gamma_D^+ \mathbf{V} \right\rangle_{\boldsymbol{\tau}} - \left\langle \boldsymbol{\zeta}, \gamma_D^- \mathbf{V} \right\rangle_{\boldsymbol{\tau}} \\
&\quad = \int_{\Omega_s} \mathbf{curl}\, \boldsymbol{\Psi}_{\mathbf{A}}^{\kappa}(\boldsymbol{\zeta}) \cdot \mathbf{curl}\, \overline{\mathbf{V}} - \mathbf{curl}\, \mathbf{curl}\, \boldsymbol{\Psi}_{\mathbf{A}}^{\kappa}(\boldsymbol{\zeta}) \cdot \overline{\mathbf{V}}\, d\mathbf{x} - \left\langle \boldsymbol{\zeta}, \gamma_D^- \mathbf{V} \right\rangle_{\boldsymbol{\tau}} \\
&\quad = \int_{\Omega_s} \mathbf{curl}\, \boldsymbol{\Psi}_{\mathbf{A}}^{\kappa}(\boldsymbol{\zeta}) \cdot \mathbf{curl}\, \overline{\mathbf{V}} - \mathbf{grad}\, \Psi_V^{\kappa}(\mathrm{div}_\Gamma \boldsymbol{\zeta}) \cdot \overline{\mathbf{V}} - \kappa^2 \boldsymbol{\Psi}_{\mathbf{A}}^{\kappa}(\boldsymbol{\zeta}) \overline{\mathbf{V}}\, d\mathbf{x} - \left\langle \boldsymbol{\zeta}, \gamma_D^- \mathbf{V} \right\rangle_{\boldsymbol{\tau}} \\
&\quad = \int_{\Omega_s} \mathbf{curl}\, \boldsymbol{\Psi}_{\mathbf{A}}^{\kappa}(\boldsymbol{\zeta}) \cdot \mathbf{curl}\, \overline{\mathbf{V}} - \kappa^2 \boldsymbol{\Psi}_{\mathbf{A}}^{\kappa}(\boldsymbol{\zeta}) \overline{\mathbf{V}}\, d\mathbf{x} - \overline{\left\langle \gamma_{\mathbf{n}}^- \mathbf{V}, V_\kappa(\mathrm{div}_\Gamma \boldsymbol{\zeta}) \right\rangle_{\frac{1}{2}, \Gamma}} - \left\langle \boldsymbol{\zeta}, \gamma_D^- \mathbf{V} \right\rangle_{\boldsymbol{\tau}} .
\end{aligned}$$

Owing to the construction of $\mathbf{X}(\mathbf{curl}, \Omega_s)$, this means

$$\begin{aligned}
&\left| \left\langle \left(-\frac{1}{2} Id + \mathbf{B}_\kappa\right) (\boldsymbol{\zeta}), \gamma_D^- \mathbf{V} \right\rangle_{\boldsymbol{\tau}} \right| \\
&\leq |\boldsymbol{\Psi}_{\mathbf{A}}^{\kappa}(\boldsymbol{\zeta})|_{\boldsymbol{H}^1(\Omega_s)} \|\mathbf{curl}\, \mathbf{V}\|_{\boldsymbol{L}^2(\Omega_s)} + \kappa^2 \left\| \boldsymbol{\Psi}_{\mathbf{A}}^{\kappa}(\boldsymbol{\zeta}) \right\|_{\boldsymbol{L}^2(\Omega_s)} \|\mathbf{V}\|_{\boldsymbol{L}^2(\Omega_s)} \\
&\qquad + \|V_\kappa(\mathrm{div}_\Gamma \boldsymbol{\zeta})\|_{L^2(\Gamma)} \left\| \gamma_{\mathbf{n}}^- \mathbf{V} \right\|_{L^2(\Gamma)} + \|\boldsymbol{\zeta}\|_{\boldsymbol{L}^2(\Gamma)} \left\| \gamma_D^- \mathbf{V} \right\|_{\boldsymbol{L}^2(\Gamma)} \\
&\leq C \left(\|\boldsymbol{\zeta}\|_{\boldsymbol{H}_{||}^{-\frac{1}{2}}(\Gamma)} + \|V_\kappa(\mathrm{div}_\Gamma \boldsymbol{\zeta})\|_{L^2(\Gamma)} + \|\widehat{\gamma} \Psi_V^{\kappa}(\mathrm{div}_\Gamma \boldsymbol{\zeta})\|_{L^2(\widehat{\Gamma})} + \|\boldsymbol{\zeta}\|_{\boldsymbol{L}^2(\Gamma)} \right) \|\mathbf{V}\|_{\boldsymbol{H}(\mathbf{curl};\Omega_s)} ,
\end{aligned}$$

with some $C = C(\Omega_s) > 0$. It goes without saying that the operators $V_\kappa : H^{-\frac{1}{2}}(\Gamma) \mapsto L^2(\Gamma)$ and $\widehat{\gamma} \Psi_V^{\kappa} : H^{-\frac{1}{2}}(\Gamma) \mapsto L^2(\widehat{\Gamma})$ are compact. Then, the compact embedding of $\mathbf{X}(\mathrm{div}_\Gamma, \Gamma)$ in $\boldsymbol{L}_{\mathbf{t}}^2(\Gamma)$ according to Corollary 7.5 finishes the proof. $\square$

The previous two lemmas reveal that the bilinear forms associated with both $*_{12}$ and $*_{21}$ from (8.1) are compact. This means that the off-diagonal blocks in (8.1) do *not* contribute to the principal part of $\widehat{a}_\kappa$. Rather, in the principal part electric and magnetic field quantities are *completely decoupled*.

To strip compact perturbations off of $*_{11}$ and $*_{22}$, we first deduce from Corollary 7.5, combined with the continuity properties from Theorem 5.3, that the operators $\mathbf{A}_\kappa : \mathbf{X}(\mathrm{div}_\Gamma, \Gamma) \mapsto \boldsymbol{H}_{||}^{1/2}(\Gamma)$ and $\tilde{\mathbf{A}}_\kappa \circ \gamma_D^- : \mathbf{X}(\mathbf{curl}, \Omega_s) \mapsto \boldsymbol{H}_{\perp}^{1/2}(\Gamma)$ are compact. Next, we recall Theorem 5.5 and obtain the following bilinear form through a compact perturbation of $\widehat{a}_\kappa$:

$$\begin{aligned}
&\widehat{b}_\kappa((\boldsymbol{\zeta}^\perp, \mathbf{U}^0, \boldsymbol{\zeta}^0, \mathbf{U}^\perp), (\boldsymbol{\mu}^\perp, \mathbf{V}^0, \boldsymbol{\mu}^0, \mathbf{V}^\perp)) \\
&\qquad = \frac{1}{\kappa^2} \left\langle \mathrm{div}_\Gamma \boldsymbol{\mu}^\perp, V_0 \mathrm{div}_\Gamma \boldsymbol{\zeta}^\perp \right\rangle_{\frac{1}{2},\Gamma} \\
&\qquad + \left\langle \boldsymbol{\mu}^\perp, \left(-\frac{1}{2} Id + \mathbf{C}_\kappa\right) \gamma_D^- \mathbf{U}^0 \right\rangle_{\boldsymbol{\tau}} - \left\langle \left(-\frac{1}{2} Id + \mathbf{B}_\kappa\right) \boldsymbol{\zeta}^\perp, \gamma_D^- \mathbf{V}^0 \right\rangle_{\boldsymbol{\tau}} \\
&\qquad + \kappa^2 \left\langle \gamma_D^- \mathbf{V}^0, \tilde{\mathbf{A}}_0 \gamma_D^- \mathbf{U}^0 \right\rangle_{\frac{1}{2},\perp,\Gamma} + \kappa^2 \left(\epsilon_r \mathbf{U}^0, \mathbf{V}^0\right)_{0;\Omega_s} + \left\langle \boldsymbol{\mu}^0, \mathbf{A}_0 \boldsymbol{\zeta}^0 \right\rangle_{\frac{1}{2},||,\Gamma} \\
&\qquad - \left\langle \boldsymbol{\mu}^0, \left(-\frac{1}{2} Id + \mathbf{C}_\kappa\right) \gamma_D^- \mathbf{U}^\perp \right\rangle_{\boldsymbol{\tau}} + \left\langle \left(-\frac{1}{2} Id + \mathbf{B}_\kappa\right) \boldsymbol{\zeta}^0, \gamma_D^- \mathbf{V}^\perp \right\rangle_{\boldsymbol{\tau}} \\
&\qquad + \left\langle \mathrm{curl}_\Gamma \, \gamma_D^- \mathbf{V}^\perp, V_0 \, \mathrm{curl}_\Gamma \, \gamma_D^- \mathbf{U}^\perp \right\rangle_{\frac{1}{2},\Gamma} + \left(\frac{1}{\mu_r} \mathbf{curl}\, \mathbf{U}^\perp, \mathbf{curl}\, \mathbf{V}^\perp\right)_{0;\Omega_s} .
\end{aligned}$$

Obviously, due to Lemma 5.4, cancellation weeds out all terms that cannot be controlled by compactness. Then the following main result, corresponding to [14, Thm. 3.12], is straightforward.

THEOREM 8.3. *The sesqui-linear form $\widehat{a}_\kappa$ related to the variational problem* (6.4) *is coercive on $\boldsymbol{\mathcal{G}}$; that is, it can be written as a sum $\widehat{a}_\kappa = \widehat{d}_\kappa + \widehat{k}_\kappa$ of a $\boldsymbol{\mathcal{G}}$-elliptic sesqui-linear form $\widehat{d}_\kappa$ and a compact sesqui-linear form $\widehat{k}_\kappa : \boldsymbol{\mathcal{G}} \times \boldsymbol{\mathcal{G}} \mapsto \mathbb{C}$.*

Proof. By the above reasoning, $\widehat{a}_\kappa - \widehat{b}_\kappa$ is compact. Using Lemma 5.4 and the compact operator T_κ introduced there, we readily see that $\widehat{d}_\kappa := \widehat{b}_\kappa - \langle T_\kappa \cdot, \cdot \rangle_{\boldsymbol{\tau}} + \langle T_\kappa \cdot, \cdot \rangle_{\boldsymbol{\tau}}$ is $\boldsymbol{\mathcal{G}}$-elliptic. ∎

Any solution of the scattering problem (1.1) will ultimately turn out to be a solution of (6.6). Consequently, given Assumption 1, Lemma 6.1 combined with Theorem 1.1 guarantees the injectivity of the operator associated with (6.6). Thus, the Fredholm alternative (cf. [43, Thm. 2.33]) instantly vindicates that (6.6) always has a unique solution $(\mathbf{E}, \boldsymbol{\lambda}) \in \boldsymbol{H}(\mathbf{curl}; \Omega_s) \times \boldsymbol{H}^{-\frac{1}{2}}(\mathrm{div}_\Gamma, \Gamma)$. Uniqueness carries over to the split variational problem (8.1). From this we infer the continuous inf-sup condition

$$\sup_{\mathfrak{v} \in \boldsymbol{\mathcal{G}}} \frac{|\widehat{a}_\kappa(\mathfrak{u}, \mathfrak{v})|}{\|\mathfrak{v}\|_{\boldsymbol{\mathcal{G}}}} \geq C \, \|\mathfrak{u}\|_{\boldsymbol{\mathcal{G}}} \quad \forall \mathfrak{u} \in \boldsymbol{\mathcal{G}}. \tag{8.3}$$

9. Finite element spaces. We equip (the curvilinear polyhedron) Ω_s with a family of shape-regular, tetrahedral triangulations $(\Omega_h)_h$. The parameter h designates the meshwidth, that is, the length of the longest edge. Let $\mathbb{H}$ stand for the collection of meshwidths occurring in $(\Omega_h)_h$ and, moreover, assume that $\mathbb{H} \subset \mathbb{R}^+$ forms a decreasing sequence tending to zero. The set $\mathcal{T}_h$ will include all triangles of Ω_h. Restricting Ω_h, $h \in \mathbb{H}$, to Γ gives a sequence $(\Gamma_h)_h$ of surface meshes. They inherit shape-regularity from $(\Omega_h)_h$. We suppose that all Γ_h are aligned with edges of Γ.

Discrete electric fields should be modelled by discrete 1-forms (edge elements). They can be represented by piecewise polynomial vectorfields: For a *fixed* polynomial degree ν, $\nu \in \mathbb{N}_0$, and any tetrahedron $T \in \mathcal{T}_h$ the local spaces are given by (cf. [44])

$$\boldsymbol{\mathcal{E}}^1_{\nu+1}(T) := \{\mathbf{V} \in (\mathcal{P}_{\nu+1}(T))^3,\ \mathbf{V}(\mathbf{x}) \cdot \mathbf{x} = 0\ \forall \mathbf{x} \in T\},$$

where $\mathcal{P}_\iota$ is the space of multivariate polynomials of total degree ν on T. This gives rise to the global finite element space

$$\boldsymbol{\mathcal{E}}^1_{\nu+1}(\Omega_h) := \{\mathbf{U} \in \boldsymbol{H}(\mathbf{curl}; \Omega_s),\ \mathbf{U}_{|T} \in \boldsymbol{\mathcal{E}}^1_{\nu+1}(T)\ \forall T \in \mathcal{T}_h\}.$$

The condition $\boldsymbol{\mathcal{E}}^1_{\nu+1}(\Omega_h) \subset \boldsymbol{H}(\mathbf{curl}; \Omega_s)$ is equivalent to the continuity of tangential components across interelement faces. This renders degrees of freedom based on moments of (tangential components) on edges, faces, and the elements themselves well defined. See [44] and [36] for details and proofs of unisolvence. The discrete 1-forms on $\{\Omega_h\}_h$ form an affine family of finite elements in the sense of [19] with respect to the pullback of 1-forms. Based on the degrees of freedom, we can introduce nodal interpolation operators Π^1_h onto $\boldsymbol{\mathcal{E}}^1_{\nu+1}(\Omega_h)$. To begin with, those are declared for continuous vectorfields. It turns out that this is not enough and that we badly need to apply Π^1_h to less regular vectorfields. The extent to which this is possible is revealed by the following interpolation error estimate.

LEMMA 9.1 (see [20, Lems. 3.2, 3.3]). *If $s > \frac{1}{2}$, then for all $\mathbf{U} \in \boldsymbol{H}^s(\Omega_s)$ such that $\mathbf{curl}\,\mathbf{U} \in \boldsymbol{H}^s(\Omega_s)$*

$$\begin{aligned}\left\|\mathbf{U} - \Pi^1_h\mathbf{U}\right\|_{\boldsymbol{L}^2(\Omega_s)} &\leq \tilde{C} h^{\min\{\nu+1,s\}}(|\mathbf{U}|_{\boldsymbol{H}^s(\Omega_s)} + |\mathbf{curl}\,\mathbf{U}|_{\boldsymbol{H}^s(\Omega_s)}),\\ \left\|\mathbf{curl}(\mathbf{U} - \Pi^1_h\mathbf{U})\right\|_{\boldsymbol{L}^2(\Omega_s)} &\leq \tilde{C} h^{\min\{\nu+1,s\}}\,|\mathbf{curl}\,\mathbf{U}|_{\boldsymbol{H}^s(\Omega_s)},\end{aligned}$$

with constants[3] *$\tilde{C} > 0$ depending only on Ω_s, ν, s, and the shape-regularity of the meshes.*

The reader might be wondering why we want to use the nodal interpolation operator even though it fails to be defined on the entire space $\boldsymbol{H}(\mathbf{curl}; \Omega_s)$. The reason is its exceptional algebraic properties. To explain them, we have to introduce the $\boldsymbol{H}(\mathrm{div}; \Omega_s)$-conforming finite element spaces $\boldsymbol{\mathcal{F}}^1_\nu(\Omega_h)$ of discrete 2-forms of degree ν, also known as Raviart–Thomas elements [7, Chap. 3], [44]. Suitable degrees of freedom for this space are supplied by moments of "face fluxes" and weighted integrals over elements. They induce the nodal interpolation operators Π^2_h onto $\boldsymbol{\mathcal{F}}^1_\nu(\Omega_h)$. Application of the Stokes theorem (cf. [36]) confirms the *commuting diagram property*

$$\mathbf{curl} \circ \Pi^1_h = \Pi^2_h \circ \mathbf{curl}, \tag{9.1}$$

valid for vectorfields in the domain $\mathrm{Dom}(\Pi^1_h)$ of Π^1_h. The relationship (9.1) teaches that Π^1_h leaves the kernel of $\mathbf{curl}$ invariant. This accounts for the pivotal role of the nodal interpolation operator.

To pick a suitable discrete trial space for $\boldsymbol{H}^{-\frac{1}{2}}(\mathrm{div}_\Gamma, \Gamma)$ we also adopt the perspective of differential forms. Be aware that $\boldsymbol{H}^{-\frac{1}{2}}(\mathrm{div}_\Gamma, \Gamma)$ is the trace space for magnetic fields, and keep in mind that those can also be described by 1-forms. This suggests that $\boldsymbol{H}^{-\frac{1}{2}}(\mathrm{div}_\Gamma, \Gamma)$ should be approximated by traces of discrete 1-forms on the surface. In other words, as $\boldsymbol{H}^{-\frac{1}{2}}(\mathrm{div}_\Gamma, \Gamma)$-conforming boundary element space we chose $\gamma_\times \boldsymbol{\mathcal{E}}^1_{\nu+1}(\Omega_h)$. Elementary computations reveal that the procedure generates exactly

[3] A $\tilde{}$ tag for a generic constant indicates that it may also depend on ν and the shape-regularity of the family of meshes.

the two-dimensional Raviart–Thomas elements $\boldsymbol{\mathcal{F}}^1_\nu(\Gamma_h)$ [46] on the surface mesh. The degrees of freedom are also inherited from $\boldsymbol{\mathcal{E}}^1_{\nu+1}(\Omega_h)$. By construction, the induced nodal interpolation operator Π^Γ_h satisfies

$$\Pi^\Gamma_h \circ \gamma_\times = \gamma_\times \circ \Pi^1_h, \tag{9.2}$$

which, due to (9.1), implies another commuting diagram property,

$$\mathrm{div}_\Gamma \circ \Pi^\Gamma_h = \mathsf{Q}^\Gamma_h \circ \mathrm{div}_\Gamma, \tag{9.3}$$

for sufficiently smooth tangential surface vectorfields. Here, Q^Γ_h is the plain $L^2(\Gamma)$-orthogonal projection onto the space $\mathcal{Q}_\nu(\Gamma_h)$ of discontinuous, piecewise polynomials (of degree ν) on Γ_h. Invariance of $\mathrm{Ker}(\mathrm{div}_\Gamma) \cap \mathrm{Dom}(\Pi^\Gamma_h)$ under Π^Γ_h is immediate.

From the results of [46] and [38, sect. 5] we harvest the following interpolation error estimates.

Lemma 9.2. *If $\boldsymbol{\mu} \in \boldsymbol{H}^s_{\mathbf{t}}(\Gamma)$, $\mathrm{div}_\Gamma \boldsymbol{\mu} \in H^s(\Gamma)$ for some $s > 0$, then*

$$\begin{aligned} \left\| \boldsymbol{\mu} - \Pi^\Gamma_h \boldsymbol{\mu} \right\|_{\boldsymbol{L}^2(\Gamma)} &\leq \tilde{C} h^{\min\{s,\nu+1\}} \left(|\boldsymbol{\mu}|_{\boldsymbol{H}^s_{\mathbf{t}}(\Gamma)} + |\mathrm{div}_\Gamma \boldsymbol{\mu}|_{H^s(\Gamma)} \right), \\ \left\| \mathrm{div}_\Gamma (\boldsymbol{\mu} - \Pi^\Gamma_h \boldsymbol{\mu}) \right\|_{L^2(\Gamma)} &\leq \tilde{C} h^{\min\{s,\nu+1\}} \, |\mathrm{div}_\Gamma \boldsymbol{\mu}|_{H^s(\Gamma)} . \end{aligned}$$

Armed with conforming finite element spaces, the Galerkin discretization of the variational problem (6.4), (6.6) is straightforward. Seek $(\mathbf{E}_h, \boldsymbol{\lambda}_h) \in \boldsymbol{\mathcal{E}}^1_{\nu+1}(\Omega_h) \times \boldsymbol{\mathcal{F}}^1_\nu(\Gamma_h)$ such that

$$a_\kappa((\mathbf{E}_h, \boldsymbol{\lambda}_h), (\mathbf{V}_h, \boldsymbol{\mu}_h)) = f(\mathbf{V}_h) + g(\boldsymbol{\mu}_h) \tag{9.4}$$

for all $(\mathbf{V}_h, \boldsymbol{\mu}_h) \in \boldsymbol{\mathcal{E}}^1_{\nu+1}(\Omega_h) \times \boldsymbol{\mathcal{F}}^1_\nu(\Gamma_h)$.

10. Discrete decompositions. The highly effective splitting idea of the continuous setting also has to be adopted for the analysis of the discretized problem (9.4). We follow a simple guideline, which boils down to applying nodal interpolation to the Helmholtz-type splittings of finite element functions. The approach to both $\boldsymbol{\mathcal{E}}^1_{\nu+1}(\Omega_h)$ and $\boldsymbol{\mathcal{F}}^1_\nu(\Gamma_h)$ is completely parallel.

First, we construct a discrete counterpart of $\mathbf{X}(\mathbf{curl}, \Omega_s)$. As in section 7, we rely on a projector. According to the recipe outlined above, it is formally defined as $\mathsf{P}_h := \Pi^1_h \circ \mathsf{P}$. However, even on $\mathsf{P}(\boldsymbol{H}(\mathbf{curl}; \Omega_s)) \subset \boldsymbol{H}^1(\Omega_s)$ the nodal interpolation Π^1_h is not bounded, because the smoothness of the **curl**s is not controlled. Nonetheless, we aim to apply P_h to finite element functions only, which saves the idea.

Lemma 10.1. *If $\mathbf{U} \in \boldsymbol{H}^1(\Omega_s)$ and $\mathbf{curl}\,\mathbf{U} \in \boldsymbol{\mathcal{F}}^1_\nu(\Omega_h)$, then $\mathbf{U} \in \mathrm{Dom}(\Pi^1_h)$ and*

$$\left\| \mathbf{U} - \Pi^1_h \mathbf{U} \right\|_{\boldsymbol{L}^2(\Omega_s)} \leq \tilde{C} h^{\min\{\nu+1,s\}} \, |\mathbf{U}|_{\boldsymbol{H}^s(\Omega_s)} ,$$

with $\tilde{C} > 0$ depending only on Ω_s, ν, and the shape regularity of Ω_h.

Since, by (9.1), $\mathbf{curl}\,\mathsf{P}\mathbf{U}_h \in \boldsymbol{\mathcal{F}}^1_\nu(\Omega_h)$ for $\mathbf{U}_h \in \boldsymbol{\mathcal{E}}^1_{\nu+1}(\Omega_h)$, the same arguments as in the case of P, along with the properties of the latter, give us information on P_h.

Lemma 10.2. *The operator $\mathsf{P}_h : \boldsymbol{\mathcal{E}}^1_{\nu+1}(\Omega_h) \mapsto \boldsymbol{\mathcal{E}}^1_{\nu+1}(\Omega_h)$ is an h-uniformly continuous projection and preserves the* **curl**, *and* $\mathrm{Ker}(\mathsf{P}_h) = \mathrm{Ker}(\mathbf{curl}) \cap \boldsymbol{\mathcal{E}}^1_{\nu+1}(\Omega_h)$.

Setting

$$\mathbf{X}_h(\mathbf{curl}, \Omega_h) := \mathsf{P}_h(\boldsymbol{\mathcal{E}}^1_{\nu+1}(\Omega_h)), \quad \mathbf{N}_h(\mathbf{curl}, \Omega_h) := \mathrm{Ker}(\mathbf{curl}) \cap \boldsymbol{\mathcal{E}}^1_{\nu+1}(\Omega_h),$$

we instantly get an h-uniformly $\boldsymbol{H}(\mathbf{curl};\Omega_s)$-stable direct splitting

$$\boldsymbol{\mathcal{E}}^1_{\nu+1}(\Omega_h) = \mathbf{X}_h(\mathbf{curl},\Omega_h) \oplus \mathbf{N}_h(\mathbf{curl},\Omega_h). \tag{10.1}$$

The following result makes it possible to pursue the same strategy in the case of $\boldsymbol{\mathcal{F}}^1_\nu(\Gamma_h)$.

LEMMA 10.3 (see [38, Lem. 6.2]). *If $\boldsymbol{\lambda} \in \boldsymbol{H}^{\frac{1}{2}}_{\mathbf{t}}(\Gamma)$ and it has its surface divergence in $\mathcal{Q}_\nu(\Gamma_h)$, then $\boldsymbol{\lambda} \in \mathrm{Dom}(\Pi^\Gamma_h)$ and*

$$\left\|\boldsymbol{\lambda} - \Pi^\Gamma_h \boldsymbol{\lambda}\right\|_{\boldsymbol{L}^2(\Gamma)} \leq \tilde{C} h^{\min\{\nu+1,s\}} \left\|\boldsymbol{\lambda}\right\|_{\boldsymbol{H}^s_{\mathbf{t}}(\Gamma)},$$

with $\tilde{C} > 0$ independent of $\boldsymbol{\lambda}$ and the meshwidth h.

Thus, we can define

$$\mathsf{P}^\Gamma_h : \boldsymbol{\mathcal{F}}^1_\nu(\Gamma_h) \mapsto \boldsymbol{\mathcal{F}}^1_\nu(\Gamma_h), \qquad \mathsf{P}^\Gamma_h := \Pi^\Gamma_h \circ \mathsf{P}^\Gamma$$

and find properties corresponding to those of P_h as follows.

LEMMA 10.4. *The mapping P^Γ_h is an h-uniformly continuous projector, which preserves* div_Γ *and fulfills* $\mathrm{Ker}(\mathsf{P}^\Gamma_h) = \mathrm{Ker}(\mathrm{div}_\Gamma) \cap \boldsymbol{\mathcal{F}}^1_\nu(\Gamma_h)$.

The projector furnishes the desired h-uniformly $\boldsymbol{H}^{-\frac{1}{2}}(\mathrm{div}_\Gamma,\Gamma)$-stable splitting of the discrete space of Neumann traces

$$\boldsymbol{\mathcal{F}}^1_\nu(\Gamma_h) = \mathbf{X}_h(\mathrm{div}_\Gamma,\Gamma_h) \oplus \mathbf{N}_h(\mathrm{div}_\Gamma,\Gamma_h), \tag{10.2}$$

with

$$\mathbf{X}_h(\mathrm{div}_\Gamma,\Gamma_h) := \mathsf{P}^\Gamma_h(\boldsymbol{\mathcal{F}}^1_\nu(\Gamma_h)), \quad \mathbf{N}_h(\mathrm{div}_\Gamma,\Gamma_h) := \mathrm{Ker}(\mathbf{curl}) \cap \boldsymbol{\mathcal{F}}^1_\nu(\Gamma_h).$$

We claimed that we were searching for discrete "counterparts" of $\mathbf{X}(\mathbf{curl},\Omega_s)$ and $\mathbf{X}(\mathrm{div}_\Gamma,\Gamma)$. If we had set out to find discrete *subspaces*, the above constructions would not have been at our disposal. Just notice that, in general, the vectorfields in $\boldsymbol{\mathcal{E}}^1_{\nu+1}(\Omega_h)$ are by no means continuous. Conversely, any piecewise polynomial vectorfield in $\mathbf{X}(\mathbf{curl},\Omega_s) \subset \boldsymbol{H}^1(\Omega_s)$ must possess continuous components. Similarly, there are elements in $\mathbf{X}_h(\mathrm{div}_\Gamma,\Gamma_h)$ that cannot occur as twisted tangential traces of continuous vectorfields. In short,

$$\mathbf{X}_h(\mathbf{curl},\Omega_h) \not\subset \mathbf{X}(\mathbf{curl},\Omega_s), \qquad \mathbf{X}_h(\mathrm{div}_\Gamma,\Gamma_h) \not\subset \mathbf{X}(\mathrm{div}_\Gamma,\Gamma).$$

This means that, introducing

$$\boldsymbol{\mathcal{G}}_h := \mathbf{X}_h(\mathrm{div}_\Gamma,\Gamma_h) \times \mathbf{N}(\mathbf{curl},\Omega_s) \times \mathbf{N}_h(\mathrm{div}_\Gamma,\Gamma_h) \times \mathbf{X}_h(\mathbf{curl},\Omega_h)$$

as a discrete approximation space for $\boldsymbol{\mathcal{G}}$, we have made a *nonconforming* choice, as $\boldsymbol{\mathcal{G}}_h \not\subset \boldsymbol{\mathcal{G}}$. This is a special kind of nonconformity, as it is not caused by the choice of finite element spaces, but by the manner in which they are split. Actually, we do not commit any "variational crime" when considering the (split) bilinear form $\widehat{a}_\kappa$ on $\boldsymbol{\mathcal{G}}_h$. However, coercivity of $\widehat{a}_\kappa$ was established only with respect to the split space $\boldsymbol{\mathcal{G}}$. This prevents us from directly applying the known results about the convergence of conforming Galerkin approximations of coercive variational problems [47]. Therefore, coercivity in the discrete setting must be established by a separate argument.

11. Bridge mappings. We recall a variant of the main result from section 4.1 of [14] (also cf. [9, 18]) as follows.

THEOREM 11.1 (see [38, sect. 7]). *Let $a : V \times V \mapsto \mathbb{C}$ be a continuous sesqui-linear form on a Banach space V, whose restriction to a closed subspace $W \subset V$ satisfies the inf-sup condition*

$$\sup_{v \in W} \frac{|a(u,v)|}{\|v\|_V} \geq \underline{c}\, \|u\|_V \quad \forall u \in W.$$

We can write $a := d - k$ with a continuous sesqui-linear form d that is V-elliptic on W, and a compact sesqui-linear form $k : W \times W \mapsto \mathbb{C}$. The family of closed subspaces $W_h \subset V$, $h \in \mathbb{H}$, is to be linked to V by two bridge mappings*: We assume the existence of families of linear, continuous operators $\mathsf{H}_h : W_h \mapsto W$ and $\mathsf{F}_h : W \mapsto W_h$, $h \in \mathbb{H}$, that satisfy*

$$\|Id - \mathsf{H}_h\|_{W_h \to V} \to 0 \quad \forall \mathfrak{u} \in W : \; \|\mathfrak{u} - \mathsf{F}_h \mathfrak{u}\|_V \to 0 \quad \textit{as } h \to 0.$$

Then, there is $h_* > 0$ and a constant $C > 0$ such that

$$\sup_{u_h \in W_h} \frac{|a(v_h, u_h)|}{\|u_h\|_V} \geq C\, \|v_h\|_V \quad \forall v_h \in W_h,\ h < h_*. \tag{11.1}$$

The assumptions of the abstract theory that seem to be most critical concern the existence of appropriate *bridge mappings* $\mathsf{H}_h : \boldsymbol{\mathcal{G}}_h \mapsto \boldsymbol{\mathcal{G}}$ and $\mathsf{F}_h : \boldsymbol{\mathcal{G}} \mapsto \boldsymbol{\mathcal{G}}_h$. For their construction the components of $\boldsymbol{\mathcal{G}}$ will be targeted separately. It turns out that the same tools used in the definition of the decomposition are also very useful for building bridge mappings. We remark that existence of appropriate bridge mappings is equivalent to the assumptions (A1) and (A2) in [14, sect. 4.1].

First, we define $\mathsf{H}_h^\Omega : \mathbf{X}_h(\mathbf{curl}, \Omega_h) \mapsto \mathbf{X}(\mathbf{curl}, \Omega_s)$ by $\mathsf{H}_h^\Omega \mathbf{U}_h := \mathsf{P}\mathbf{U}_h$, $\mathbf{U}_h \in \mathbf{X}_h(\mathbf{curl}, \Omega_h)$. The projection property stated in Lemma 10.2 shows

$$\Pi_h^1 \mathsf{H}_h^\Omega \mathbf{U}_h = \Pi_h^1 \mathsf{P}\mathbf{U}_h = \mathsf{P}_h \mathbf{U}_h = \mathbf{U}_h \quad \forall \mathbf{U}_h \in \mathbf{X}_h(\mathbf{curl}, \Omega_h).$$

As $\mathbf{curl}\, \mathsf{H}_h^\Omega \mathbf{U}_h = \mathbf{curl}\, \mathbf{U}_h \in \boldsymbol{\mathcal{F}}_\nu^1(\Omega_h)$, Lemma 10.1 permits us to estimate

$$\begin{aligned} \left\|\mathbf{U}_h - \mathsf{H}_h^\Omega \mathbf{U}_h\right\|_{\boldsymbol{L}^2(\Omega_s)} &= \left\|(\Pi_h^1 - Id)\mathsf{H}_h^\Omega \mathbf{U}_h\right\|_{\boldsymbol{L}^2(\Omega_s)} \\ &\leq \tilde{C}h \left\|\mathsf{H}_h^\Omega \mathbf{U}_h\right\|_{\boldsymbol{H}^1(\Omega_s)} \leq \tilde{C}h \left\|\mathbf{curl}\, \mathbf{U}_h\right\|_{\boldsymbol{L}^2(\Omega_s)}. \end{aligned} \tag{11.2}$$

The same construction works for $\mathbf{X}_h(\mathrm{div}_\Gamma, \Gamma_h)$. We introduce $\mathsf{H}_h^\Gamma : \mathbf{X}_h(\mathrm{div}_\Gamma, \Gamma_h) \mapsto \mathbf{X}(\mathrm{div}_\Gamma, \Gamma)$ through $\mathsf{H}_h^\Gamma \boldsymbol{\mu}_h := \mathsf{P}^\Gamma \boldsymbol{\mu}_h$, $\boldsymbol{\mu}_h \in \mathbf{X}_h(\mathrm{div}_\Gamma, \Gamma_h)$. As above, now appealing to Lemma 10.4, we get

$$\Pi_h^\Gamma \mathsf{H}_h^\Gamma \boldsymbol{\mu}_h = \Pi_h^\Gamma \mathsf{P}^\Gamma \boldsymbol{\mu}_h = \mathsf{P}_h^\Gamma \boldsymbol{\mu}_h = \boldsymbol{\mu}_h \quad \forall \boldsymbol{\mu}_h \in \mathbf{X}_h(\mathrm{div}_\Gamma, \Gamma_h).$$

Then, $\mathrm{div}_\Gamma \mathsf{H}_h^\Gamma \boldsymbol{\mu}_h = \mathrm{div}_\Gamma \boldsymbol{\mu}_h$, along with Lemma 10.3, shows

$$\begin{aligned} \left\|\boldsymbol{\mu}_h - \mathsf{H}_h^\Gamma \boldsymbol{\mu}_h\right\|_{\boldsymbol{L}^2(\Gamma)} &= \left\|(\Pi_h^\Gamma - Id)\mathsf{H}_h^\Gamma \boldsymbol{\mu}_h\right\|_{\boldsymbol{L}^2(\Gamma)} \\ &\leq \tilde{C}h^{\frac{1}{2}} \left\|\mathsf{H}_h^\Gamma \boldsymbol{\mu}_h\right\|_{\boldsymbol{H}_\perp^{\frac{1}{2}}(\Gamma)} \leq \tilde{C}h^{\frac{1}{2}} \left\|\mathrm{div}_\Gamma \boldsymbol{\mu}_h\right\|_{H^{-\frac{1}{2}}(\Gamma)}. \end{aligned} \tag{11.3}$$

The kernels pose no difficulties, as $\mathbf{N}_h(\mathbf{curl}, \Omega_h) \subset \mathbf{N}(\mathbf{curl}, \Omega_s)$ and $\mathbf{N}_h(\mathrm{div}_\Gamma, \Gamma_h) \subset \mathbf{N}(\mathrm{div}_\Gamma, \Gamma)$. Therefore, we can finally define $\mathsf{H}_h : \boldsymbol{\mathcal{G}}_h \mapsto \boldsymbol{\mathcal{G}}$ through

$$\mathsf{H}_h(\boldsymbol{\mu}_h^\perp, \mathbf{V}_h^0, \boldsymbol{\mu}_h^0, \mathbf{V}_h^\perp) := (\mathsf{H}_h^\Gamma \boldsymbol{\mu}_h^\perp, \mathbf{V}_h^0, \boldsymbol{\mu}_h^0, \mathsf{H}_h^\Omega \mathbf{V}_h^\perp).$$

After replacing W, W_h by $\boldsymbol{\mathcal{G}}$, $\boldsymbol{\mathcal{G}}_h$, the uniform convergence required for Theorem 11.1 holds true.

To define F_h we first consider the $\boldsymbol{L}^2(\Omega_s)$-orthogonal projections $\mathsf{Q}_h^0 : \boldsymbol{L}^2(\Omega_s) \mapsto \mathbf{curl}\,\boldsymbol{\mathcal{E}}_{\nu+1}^1(\Omega_h)$, $h \in \mathbb{H}$. As $\boldsymbol{C}^\infty(\bar{\Omega}_s)$ is dense in $\boldsymbol{H}(\mathbf{curl};\Omega_s)$ and

$$\begin{aligned}\left\|\mathbf{curl}\,\mathbf{U} - \mathsf{Q}_h^0\,\mathbf{curl}\,\mathbf{U}\right\|_{\boldsymbol{L}^2(\Omega_s)} &\leq \left\|\mathbf{curl}\,\mathbf{U} - \Pi_h^2\,\mathbf{curl}\,\mathbf{U}\right\|_{\boldsymbol{L}^2(\Omega_s)} \\ &= \left\|\mathbf{curl}(Id - \Pi_h^1)\mathbf{U}\right\|_{\boldsymbol{L}^2(\Omega_s)} = O(h) \quad \text{as } h \to 0\end{aligned}$$

for fixed $\mathbf{U} \in \boldsymbol{C}^\infty(\bar{\Omega}_s)$, we find

$$\lim_{h\to 0} \left\|(Id - \mathsf{Q}_h^0)\,\mathbf{curl}\,\mathbf{U}\right\|_{\boldsymbol{L}^2(\Omega_s)} = 0 \quad \forall \mathbf{U} \in \boldsymbol{H}(\mathbf{curl};\Omega_s). \tag{11.4}$$

Then we define $\mathsf{F}_h^\Omega := \Pi_h^1 \circ \mathsf{L} \circ \mathsf{Q}_h^0 \circ \mathbf{curl}$. As in section 10, taking into account

$$\mathbf{curl}\,\mathsf{F}_h^\Omega\mathbf{U} = \mathbf{curl}\,\Pi_h^1\mathsf{L}(\mathsf{Q}_h^0\,\mathbf{curl}\,\mathbf{U}) = \Pi_h^2\mathsf{Q}_h^0\,\mathbf{curl}\,\mathbf{U} = \mathsf{Q}_h^0\,\mathbf{curl}\,\mathbf{U}$$

and Lemma 10.1, we see that the definition makes sense. On top of that, setting $\mathbf{U}^* := \mathsf{L}(\mathsf{Q}_h^0\,\mathbf{curl}\,\mathbf{U}) \in \mathbf{X}(\mathbf{curl},\Omega_s)$, we get the estimate

$$\begin{aligned}\left\|\mathbf{U} - \mathsf{F}_h^\Omega\mathbf{U}\right\|_{\boldsymbol{L}^2(\Omega_s)} &\leq \|\mathbf{U} - \mathbf{U}^*\|_{\boldsymbol{L}^2(\Omega_s)} + \left\|\mathbf{U}^* - \Pi_h^1\mathbf{U}^*\right\|_{\boldsymbol{L}^2(\Omega_s)} \\ &\leq C\,\|\mathbf{curl}(\mathbf{U} - \mathbf{U}^*)\|_{\boldsymbol{L}^2(\Omega_s)} + \tilde{C}h\,\|\mathbf{U}^*\|_{\boldsymbol{H}^1(\Omega_s)} \\ &\leq C\left\|(Id - \mathsf{Q}_h^0)\,\mathbf{curl}\,\mathbf{U}\right\|_{\boldsymbol{L}^2(\Omega_s)} + \tilde{C}h\,\|\mathbf{curl}\,\mathbf{U}\|_{\boldsymbol{L}^2(\Omega_s)}\,.\end{aligned} \tag{11.5}$$

Combined with (11.4), pointwise convergence in the $\boldsymbol{H}(\mathbf{curl};\Omega_s)$-norm can be inferred.

On the surface Γ a similar policy succeeds. It is based on the $H^{-\frac{1}{2}}(\Gamma)$-orthogonal projections $\mathsf{Q}_h^{\frac{1}{2}} : H^{-\frac{1}{2}}(\Gamma) \mapsto \mathcal{Q}_\nu(\Gamma_h)$. Density of $\bigcup\{\mathcal{Q}_\nu(\Gamma_h),\ h \in \mathbb{H}\}$ in $H^{-\frac{1}{2}}(\Gamma)$ shows

$$\lim_{h\to 0} \left\|\omega - \mathsf{Q}_h^{\frac{1}{2}}\omega\right\|_{H^{-\frac{1}{2}}(\Gamma)} = 0 \quad \forall \omega \in H^{-\frac{1}{2}}(\Gamma). \tag{11.6}$$

Then, introduce $\mathsf{F}_h^\Gamma := \Pi_h^\Gamma \circ \gamma_\times \circ \mathsf{J} \circ \mathsf{Q}_h^{\frac{1}{2}} \circ \mathrm{div}_\Gamma$ and observe

$$\mathrm{div}_\Gamma\mathsf{F}_h^\Gamma\boldsymbol{\lambda} = \mathsf{Q}_h^\Gamma(\mathrm{div}_\Gamma\gamma_\times\mathsf{J}(\mathsf{Q}_h^{\frac{1}{2}}\mathrm{div}_\Gamma\boldsymbol{\lambda})) = \mathsf{Q}_h^\Gamma\mathsf{Q}_h^{\frac{1}{2}}\mathrm{div}_\Gamma\boldsymbol{\lambda} = \mathsf{Q}_h^{\frac{1}{2}}\boldsymbol{\lambda}.$$

According to Lemma 10.3, the definition of F_h^Γ is meaningful, and as above we derive

$$\left\|\boldsymbol{\lambda} - \mathsf{F}_h^\Gamma\boldsymbol{\lambda}\right\|_{\boldsymbol{L}^2(\Gamma)} \leq C\|(Id - \mathsf{Q}_h^{\frac{1}{2}})\mathrm{div}_\Gamma\boldsymbol{\lambda}\|_{H^{-\frac{1}{2}}(\Gamma)} + \tilde{C}h^{\frac{1}{2}}\,\|\mathrm{div}_\Gamma\boldsymbol{\lambda}\|_{H^{-\frac{1}{2}}(\Gamma)}\,. \tag{11.7}$$

Again, pointwise convergence $\mathsf{F}_h^\Gamma\boldsymbol{\lambda} \to \boldsymbol{\lambda}$ in $\boldsymbol{H}^{-\frac{1}{2}}(\mathrm{div}_\Gamma,\Gamma)$ follows.

The kernels are easier to deal with, because we may just use $\boldsymbol{L}^2(\Omega_s)/H^{-\frac{1}{2}}(\Gamma)$-orthogonal projections $\mathsf{N}_h^0 : \mathbf{N}(\mathbf{curl},\Omega_s) \mapsto \mathbf{N}_h(\mathbf{curl},\Omega_h)$ and $\mathsf{N}_h^{0,\Gamma} : \mathbf{N}(\mathrm{div}_\Gamma,\Gamma) \mapsto \mathbf{N}_h(\mathrm{div}_\Gamma,\Gamma_h)$, respectively. Simple density arguments establish their pointwise convergence in $\boldsymbol{L}^2(\Omega_s)$ and $H^{-\frac{1}{2}}(\Gamma)$, respectively, as $h \to 0$. Eventually, we have found that

$$\mathsf{F}_h(\boldsymbol{\mu}^\perp, \mathbf{V}^0, \boldsymbol{\mu}^0, \mathbf{V}^\perp) := (\mathsf{F}_h^\Gamma\boldsymbol{\mu}^\perp, \mathsf{N}_h^0\mathbf{V}^0, \mathsf{N}_h^{0,\Gamma}\boldsymbol{\mu}^0, \mathsf{F}_h^\Omega\mathbf{V}^\perp) \in \boldsymbol{\mathcal{G}}_h$$

is the right bridge mapping $\mathsf{F}_h : \boldsymbol{\mathcal{G}} \mapsto \boldsymbol{\mathcal{G}}_h$.

Remark 11.1. It is important to realize that the choice of both the continuous and discrete splittings is merely a theoretical tool. It does not affect the discrete problem at all, which remains given by (9.4). Therefore, different splittings may be used to investigate the *same* numerical scheme.

12. Convergence. The discovery of suitable bridge mappings paves the way for a quasi-optimal asymptotic estimate of the discretization error.

THEOREM 12.1. *Under Assumption* 1, *there exists a meshwidth* $h_* \in \mathbb{H}$, *depending only on* Ω_s, κ, ν *and on the shape-regularity of the triangulations* Ω_h, *such that for every* $h < h_*$ *the discrete problem* (9.4) *has a unique solution* $(\mathbf{E}_h, \boldsymbol{\lambda}_h) \in \boldsymbol{\mathcal{E}}^1_{\nu+1}(\Omega_h) \times \boldsymbol{\mathcal{F}}^1_\nu(\Gamma_h)$, *which is quasi-optimal in the sense that*

$$\begin{aligned}\|\mathbf{E}-\mathbf{E}_h\|_{\boldsymbol{H}(\mathbf{curl};\Omega_s)} &+ \|\boldsymbol{\lambda}-\boldsymbol{\lambda}_h\|_{\boldsymbol{H}^{-\frac{1}{2}}(\mathrm{div}_\Gamma,\Gamma)} \\ &\leq \tilde{C}\ \inf\left\{\begin{array}{l}\|\mathbf{E}-\mathbf{V}_h\|_{\boldsymbol{H}(\mathbf{curl};\Omega_s)} + \|\boldsymbol{\lambda}-\boldsymbol{\mu}_h\|_{\boldsymbol{H}^{-\frac{1}{2}}(\mathrm{div}_\Gamma,\Gamma)}, \\ (\mathbf{V}_h,\boldsymbol{\mu}_h) \in \boldsymbol{\mathcal{E}}^1_{\nu+1}(\Omega_h)\times\boldsymbol{\mathcal{F}}^1_\nu(\Gamma_h),\end{array}\right\}\end{aligned}$$

with a constant $\tilde{C} > 0$ *independent of* $(\mathbf{E}, \boldsymbol{\lambda})$ *and* $h \in \mathbb{H}$.

Proof. To begin with, we have to verify the assumptions of the abstract theory of Theorem 11.1: The role of V is now played by $\boldsymbol{H}^{-\frac{1}{2}}(\mathrm{div}_\Gamma,\Gamma) \times \boldsymbol{H}(\mathbf{curl};\Omega_s) \times \boldsymbol{H}^{-\frac{1}{2}}(\mathrm{div}_\Gamma,\Gamma) \times \boldsymbol{H}(\mathbf{curl};\Omega_s)$. The spaces W, W_h have to be replaced by $\boldsymbol{\mathcal{G}}, \boldsymbol{\mathcal{G}}_h$. The bilinear forms $\widehat{a}_\kappa, \widehat{k}_\kappa, \widehat{d}_\kappa$ correspond to a, k, d. Continuity, compactness, and the inf-sup condition are clear from section 8 and, in particular, (8.3). Ultimately, we get h-uniform stability according to (11.1) for $\widehat{a}_\kappa$ on the family $\boldsymbol{\mathcal{G}}_h$, $h \in \mathbb{H}$, provided that h is sufficiently small.

We can now use this insight, the identity (8.2), and the following h-uniform equivalence of norms (deduced from Lemmas 10.2 and 10.4):

$$\left\|(\boldsymbol{\mu}_h^\perp, \mathbf{V}_h^0, \boldsymbol{\mu}_h^0, \mathbf{V}_h^\perp)\right\|_{\boldsymbol{\mathcal{G}}} \asymp \left\|(\mathbf{V}_h^\perp + \mathbf{V}_h^0, \boldsymbol{\mu}_h^\perp + \boldsymbol{\mu}_h^0)\right\|_{\boldsymbol{\mathcal{V}}} \quad \forall (\boldsymbol{\mu}_h^\perp, \mathbf{V}_h^0, \boldsymbol{\mu}_h^0, \mathbf{V}_h^\perp) \in \boldsymbol{\mathcal{G}}_h.$$

Taken together, these directly yield for $h < h_*$

$$\begin{aligned}(12.1)\qquad &\sup_{(\mathbf{U}_h,\boldsymbol{\zeta}_h)\in\boldsymbol{\mathcal{E}}^1_{\nu+1}(\Omega_h)\times\boldsymbol{\mathcal{F}}^1_\nu(\Gamma_h)} \frac{|a_\kappa((\mathbf{V}_h,\boldsymbol{\mu}_h),(\mathbf{U}_h,\boldsymbol{\zeta}_h))|}{\|(\mathbf{U}_h,\boldsymbol{\zeta}_h)\|_{\boldsymbol{\mathcal{V}}}} \\ &\geq \tilde{C} \sup_{(\boldsymbol{\zeta}_h^\perp,\mathbf{U}_h^0,\boldsymbol{\zeta}_h^0,\mathbf{U}_h^\perp)\in\boldsymbol{\mathcal{G}}_h} \frac{|\widehat{a}_\kappa((\boldsymbol{\mu}_h^\perp,\mathbf{V}_h^0,\boldsymbol{\mu}_h^0,\mathbf{V}_h^\perp),(\boldsymbol{\zeta}_h^\perp,\mathbf{U}_h^0,\boldsymbol{\zeta}_h^0,\mathbf{U}_h^\perp))|}{\left\|(\boldsymbol{\zeta}_h^\perp,\mathbf{U}_h^0,\boldsymbol{\zeta}_h^0,\mathbf{U}_h^\perp)\right\|_{\boldsymbol{\mathcal{G}}}} \\ &\geq \tilde{C}\left\|(\boldsymbol{\mu}_h^\perp,\mathbf{V}_h^0,\boldsymbol{\mu}_h^0,\mathbf{V}_h^\perp)\right\|_{\boldsymbol{\mathcal{G}}} \geq \tilde{C}\,\|(\mathbf{V}_h,\boldsymbol{\mu}_h)\|_{\boldsymbol{\mathcal{V}}},\end{aligned}$$

with constants independent of the functions and $h \in \mathbb{H}$. Appealing to Babuška's theory [5], the error estimate of the theorem can be inferred. □

Prerequisite for establishing orders of convergence of best approximations in finite element spaces are assumptions on the smoothness of the continuous solutions. We will take for granted that both the electric and magnetic fields $\mathbf{E}, \mathbf{H} := \frac{1}{i\omega\mu_r}\,\mathbf{curl}\,\mathbf{E}$ belong to $\boldsymbol{H}^\sigma(\Omega_s)$ for some $\sigma > 0$. We point out that the regularity of solutions of Maxwell's equations depends on the discontinuities of the material parameters ϵ_r and μ_r [27]. The investigations in [27] show that we have to brace for very poor regularity with σ slightly larger than zero.

It is reasonable to demand that the discontinuities of μ_r and ϵ_r be resolved by the meshes Ω_h. That is, if Ω_i, $i = 1, \ldots, M$, $M \in \mathbb{M}$, are subdomains of Ω_s on which both material parameters are smooth, then $\Omega_{h|\Omega_i}$ must supply a valid triangulation of Ω_i, $i = 1, \ldots, M$. Then we can exploit $\mathbf{curl}\,\mathbf{E} = i\kappa\mu_r\mathbf{H}$ to see that $\mathbf{curl}\,\mathbf{E}$ is locally in $\boldsymbol{H}^\sigma(\Omega_i)$, $i = 1, \ldots, M$. Globally, $\mathbf{curl}\,\mathbf{E}$ is at least contained in $\boldsymbol{H}^{\min\{\sigma,\frac{1}{4}\}}(\Omega_s)$.

LEMMA 12.2. *If* $\mathbf{E}, \mathbf{H} \in \boldsymbol{H}^\sigma(\Omega_s)$ *for some* $\sigma > 0$, *and if the jumps of* ϵ_r, μ_r *are resolved by all triangulations, we find a constant* $\tilde{C} > 0$ *depending only on* ϵ_r, μ_r, Ω_s, ν, *and the shape-regularity of the meshes* Ω_h *such that*

$$\inf_{\mathbf{V}_h \in \boldsymbol{\mathcal{E}}^1_{\nu+1}(\Omega_h)} \|\mathbf{E} - \mathbf{V}_h\|_{\boldsymbol{H}(\mathbf{curl};\Omega_s)} \leq \tilde{C} h^{\min\{\nu+1,\sigma\}} \left(\|\mathbf{E}\|_{\boldsymbol{H}^\sigma(\Omega_s)} + \sum_{i=1}^M \|H\|_{\boldsymbol{H}^\sigma(\Omega_i)} \right).$$

Proof. First, we restrict our attention to $\sigma > \frac{1}{2}$. Then, according to Lemma 9.1 and thanks to the strict locality of the nodal interpolation operators, we obtain

$$\left\|\mathbf{E} - \Pi^1_h \mathbf{E}\right\|_{\boldsymbol{H}(\mathbf{curl};\Omega_s)} \leq \tilde{C} h^{\min\{\sigma,\nu+1\}} \left(\|\mathbf{E}\|_{\boldsymbol{H}^\sigma(\Omega_s)} + \sum_{i=1}^M \|H\|_{\boldsymbol{H}^\sigma(\Omega_i)} \right).$$

This estimate hinges on the resolution of the jumps of μ_r by the meshes.

Second, in order to cope with $\sigma \leq \frac{1}{2}$, we resort to the Helmholtz-type splitting $\mathbf{E} = \mathbf{E}^\perp + \mathbf{E}^0$ from (7.1). We start with an approximation for $\mathbf{E}^\perp \in \mathbf{X}(\mathbf{curl}, \Omega_s)$. As $\mathbf{X}(\mathbf{curl}, \Omega_s) \subset \boldsymbol{H}^1(\Omega_s)$ and $\mathbf{curl}\,\mathbf{E}^\perp \in \boldsymbol{H}^{\min\{\sigma,\frac{1}{4}\}}(\Omega_s)$, nodal interpolation is an option and it yields

$$\left\|\mathbf{E}^\perp - \Pi^1_h \mathbf{E}^\perp\right\|_{\boldsymbol{H}(\mathbf{curl};\Omega_s)} \leq \tilde{C} h^\sigma \sum_{i=1}^M \|H\|_{\boldsymbol{H}^\sigma(\Omega_i)}. \tag{12.2}$$

Scalar potentials are the key to the treatment of $\mathbf{E}^0$: It is known that the irrotational vectorfield $\mathbf{E}^0$ has a representation

$$\mathbf{E}^0 = \mathbf{grad}\,\Phi + \mathbf{G}, \quad \Phi \in H^1(\Omega_s), \quad \mathbf{G} \in \boldsymbol{\mathcal{H}}(\Omega_s),$$

where $\boldsymbol{\mathcal{H}}(\Omega_s)$ is the space of harmonic Neumann vectorfields in Ω_s (see [4, sect. 3.c]). The dimension of $\boldsymbol{\mathcal{H}}(\Omega_s)$ is finite and agrees with the first Betti number $\beta_1(\Omega_s)$ of Ω_s. Moreover, a basis can be obtained from the solutions of the variational problems [4, Prop. 3.14]: Seek $\eta_j \in H^1_{[]}(\Omega_s \setminus \Sigma_j) := \{\varphi \in H^1(\Omega_s \setminus \Sigma_j), [\varphi]_{\Sigma_j} = \text{const}\}$, $j = 1, \dots, L := \beta_1(\Omega_s)$, such that

$$\int_{\Omega_s \setminus \Sigma_j} \mathbf{grad}\,\eta_j \cdot \mathbf{grad}\,\phi \, d\mathbf{x} = [\phi]_{\Sigma_j} \quad \forall \phi \in H^1_{[]}(\Omega_s \setminus \Sigma_j),$$

where Σ_j, $j = 1, \dots, \beta_1(\Omega_s)$, is a complete set of piecewise smooth Seifert surfaces for Ω_s. Hence, the η_j are fixed as solutions of Neumann problems with a jump condition across Σ_j. From them we construct $\boldsymbol{\mathcal{H}}(\Omega_s) = \text{Span}\,\{\mathbf{grad}\,\eta_1, \dots, \mathbf{grad}\,\eta_L\}$. As the exact position of the Seifert surface does not affect $\mathbf{grad}\,\eta_j$, they can always be assumed to be the union of faces of any mesh Ω_h. Besides, the regularity of $\mathbf{grad}\,\eta_j$ is determined only by the geometry of Ω_s, no matter where the Σ_j are located. More precisely, we find $\mathbf{grad}\,\eta_j \in \boldsymbol{H}^s(\Omega_s)$, where $s+1 < e_N(\Omega_s)$ and $e_N(\Omega_s)$ is the smallest singular exponent for the Neumann problem for Δ in Ω_s. This exponent depends on the angles at reentrant edges and corners of Ω_s. From [30] we know that $e_N > 3/2$ for any Lipschitz-polyhedron, and thus we can choose $s = 1/2$, at worst. At any rate, $\boldsymbol{\mathcal{H}}(\Omega_s) \subset \boldsymbol{H}^{\frac{1}{2}}(\Omega_s)$ and $\boldsymbol{\mathcal{H}}(\Omega_s) \subset \text{Dom}(\Pi^1_h)$ will hold.

To deal with the scalar potential Φ we resort to continuous quasi-interpolation operators $\mathsf{Z}_h : H^1(\Omega_s) \mapsto \mathcal{S}_{\nu+1}(\Omega_h)$ onto the space $\mathcal{S}_{\nu+1}(\Omega_h)$ of continuous, piecewise polynomial (of degree $\nu+1$) scalar functions on Ω_h. Such operators have been introduced, for instance, in [48].

With these powerful tools at our disposal, we set $\mathbf{V}_h^0 := \mathbf{grad}\,\mathsf{Z}_h\Phi + \Pi_h^1\mathbf{G}$. Please note that another commuting diagram property [44] ensures $\mathbf{grad}\,\mathcal{S}_{\nu+1}(\Omega_h) \subset \boldsymbol{\mathcal{E}}_{\nu+1}^1(\Omega_h)$. Using the interpolation properties of the Z_h and that $\mathbf{curl}\,\mathbf{G} = 0$, we end up with

$$\begin{aligned}\left\|\mathbf{E}^0 - \mathbf{V}_h^0\right\|_{\boldsymbol{L}^2(\Omega_s)} &\leq \left\|\mathbf{grad}(Id - \mathsf{Z}_h)\Phi\right\|_{\boldsymbol{L}^2(\Omega_s)} + \left\|(Id - \Pi_h^1)\mathbf{G}\right\|_{\boldsymbol{L}^2(\Omega_s)} \\ &\leq \tilde{C}h^\sigma \left\|\Phi\right\|_{\boldsymbol{H}^{\sigma+1}(\Omega_s)} + \tilde{C}h^s \left\|\mathbf{G}\right\|_{\boldsymbol{H}^s(\Omega_s)} \leq \tilde{C}h^\sigma \left\|\mathbf{E}^0\right\|_{\boldsymbol{H}^\sigma(\Omega_s)} . \qquad (12.3)\end{aligned}$$

Collecting the estimates (12.2) and (12.3), $\Pi_h^1\mathbf{E}^\perp + \mathbf{V}_h^0$ provides the desired approximation of $\mathbf{E}$ of order σ in $\boldsymbol{H}(\mathbf{curl};\Omega_s)$. □

Next, we turn to the approximation of $\boldsymbol{\lambda}$.

LEMMA 12.3. *Assume that the meshes Ω_h resolve the discontinuities of both ϵ_r and μ_r, that $\mathbf{H}, \mathbf{E} \in \boldsymbol{H}^\sigma(\Omega_s)$, $\sigma > 0$, and that $\mathbf{E}_{\mathrm{inc}}$ is smooth. Then*

$$\begin{aligned}\left\|\boldsymbol{\lambda} - \Pi_h^\Gamma\boldsymbol{\lambda}\right\|_{\boldsymbol{H}^{-\frac{1}{2}}(\mathrm{div}_\Gamma,\Gamma)} \leq \tilde{C}h^{\min\{\nu+1,\sigma\}} \Bigg(&\left\|\mathbf{H}\right\|_{\boldsymbol{H}^\sigma(\Omega_s)} + \sum_{i=1}^{M} \left\|\mathbf{E}\right\|_{\boldsymbol{H}^\sigma(\Omega_i)} \\ &+ \left\|\mathbf{E}_{\mathrm{inc}}\right\|_{\boldsymbol{H}^{\sigma+1}(\Omega_s)} \Bigg),\end{aligned}$$

where $\tilde{C} > 0$ depends neither on $\mathbf{E}, \mathbf{H}, \mathbf{E}_{\mathrm{inc}}$ nor on $h \in \mathbb{H}$.

Proof. We exploit that $\boldsymbol{\lambda} = \gamma_{\mathbf{t}}^-\mathbf{H} + \gamma_{\mathbf{t}}\mathbf{H}_{\mathrm{inc}}$. As $\mathbf{E} = i\kappa\epsilon_r\,\mathbf{curl}\,\mathbf{H}$, a complete role reversal of $\mathbf{E}$ and $\mathbf{H}$ is possible in the arguments of the previous proof. The final result is

$$\inf_{\mathbf{V}_h \in \boldsymbol{\mathcal{E}}_{\nu+1}^1(\Omega_h)} \left\|\mathbf{H} - \mathbf{V}_h\right\|_{\boldsymbol{H}(\mathbf{curl};\Omega_s)} \leq \tilde{C}h^{\min\{\nu+1,\sigma\}} \left(\left\|\mathbf{H}\right\|_{\boldsymbol{H}^\sigma(\Omega_s)} + \sum_{i=1}^{M} \left\|\mathbf{E}\right\|_{\boldsymbol{H}^\sigma(\Omega_i)} \right).$$

Then, the trace Theorem 3.3, combined with (9.2), shows the assertion of the lemma. □

Along with Theorem 12.1 this implies convergence of the Galerkin solutions in $\boldsymbol{\mathcal{E}}_{\nu+1}^1(\Omega_h) \times \boldsymbol{\mathcal{F}}_\nu^1(\Gamma_h)$ of the order $O(h^{\min\{\nu+1,\sigma\}})$ in the natural (energy) norms as $h \to 0$.

Acknowledgement. The author is grateful to Dr. A. Buffa, Prof. T. von Petersdorff, and Prof. C. Schwab from ETH Zürich for a very fruitful collaboration that much inspired the present paper.

REFERENCES

[1] H. AMMARI AND J. NÉDÉLEC, *Couplage éléments finis-équations intégrales pour la résolution des équations de Maxwell en milieu hétérogene*, in Equations aux Derivees Pertielles et Applications. Articles dedies a Jacques-Luois Lions, Gauthier-Villars, Paris, 1998, pp. 19–33.

[2] H. AMMARI AND J.-C. NÉDÉLEC, *Coupling of finite and boundary element methods for the time-harmonic Maxwell equations.* II: *A symmetric formulation*, in The Maz'ya Anniversary Collection. Vol. 2, J. Rossmann, ed., Oper. Theory Adv. Appl. 110, Birkhäuser, Basel, 1999, pp. 23–32.

[3] H. AMMARI AND J.-C. NÉDĹEC, *Low-frequency electromagnetic scattering*, SIAM J. Math. Anal., 31 (2000), pp. 836–861.

[4] C. AMROUCHE, C. BERNARDI, M. DAUGE, AND V. GIRAULT, *Vector potentials in three-dimensional nonsmooth domains*, Math. Methods Appl. Sci., 21 (1998), pp. 823–864.

[5] I. Babuška, *Error bounds for the finite element method*, Numer. Math., 16 (1971), pp. 322–333.

[6] A. Bendali, *Boundary element solution of scattering problems relative to a generalized impedance boundary condition*, in Partial Differential Equations: Theory and Numerical Solution (Proceedings of the ICM'98 Satellite Conference, Prague, Czech Republic, 1998), W. Jäger, ed., Chapman & Hall, CRC Res. Notes Math. 406, Boca Raton, FL, 2000, pp. 10–24.

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

[8] A. Buffa, *Hodge decompositions on the boundary of a polyhedron: The multiconnected case*, Math. Models Methods Appl. Sci., 11 (2001), pp. 1491–1504.

[9] A. Buffa and S. Christiansen, *The electric field integral equation on Lipschitz screens: Definition and numerical approximation*, Numer. Math., (2002), to appear.

[10] A. Buffa and P. Ciarlet, *On traces for functional spaces related to Maxwell's equations. Part* I: *An integration-by-parts formula in Lipschitz polyhedra.*, Math. Methods Appl. Sci., 24 (2001), pp. 9–30.

[11] A. Buffa and P. Ciarlet, *On traces for functional spaces related to Maxwell's equations. Part* II: *Hodge decompositions on the boundary of Lipschitz polyhedra and applications*, Math. Methods Appl. Sci., 24 (2001), pp. 31–48.

[12] A. Buffa, M. Costabel, and C. Schwab, *Boundary element methods for Maxwell's equations on non-smooth domains*, Numer. Math., 92 (2002), pp. 679–710.

[13] A. Buffa, M. Costabel, and D. Sheen, *On traces for* $\mathbf{H}(\mathbf{curl}, \Omega)$ *in Lipschitz domains*, J. Math. Anal. Appl., 276 (2002), pp. 845–867.

[14] A. Buffa, R. Hiptmair, T. von Petersdorff, and C. Schwab, *Boundary element methods for Maxwell equations on Lipschitz domains*, Numer. Math., (2002), to appear.

[15] C. Carstensen and P. Wriggers, *On the symmetric boundary element method and the symmetric coupling of boundary elements and finite elements*, IMA J. Numer. Anal., 17 (1997), pp. 201–238.

[16] M. Cessenat, *Mathematical Methods in Electromagnetism*, Adv. Math. Appl. Sci. 41, World Scientific, Singapore, 1996.

[17] G. Chen and J. Zhou, *Boundary Element Methods*, Academic Press, New York, 1992.

[18] S. Christiansen, *Discrete Fredholm Properties and Convergence Estimates for the EFIE*, Technical Report 453, Centre Mathematiques Appliques (CMAP), Ecole Polytechique, Paris, France, 2000.

[19] P. Ciarlet, *The Finite Element Method for Elliptic Problems*, Stud. Math. Appl. 4, North–Holland, Amsterdam, 1978.

[20] P. Ciarlet, Jr., and J. Zou, *Fully discrete finite element approaches for time-dependent Maxwell equations*, Numer. Math., 82 (1999), pp. 193–219.

[21] D. Colton and R. Kress, *Integral Equation Methods In Scattering Theory*, John Wiley & Sons, New York, 1983.

[22] D. Colton and R. Kress, *Inverse Acoustic and Electromagnetic Scattering Theory*, 2nd ed., Appl. Math. Sci. 93, Springer-Verlag, Heidelberg, 1998.

[23] M. Costabel, *Symmetric methods for the coupling of finite elements and boundary elements*, in Boundary Elements IX, C. Brebbia, W. Wendland, and G. Kuhn, eds., Springer-Verlag, Berlin, 1987, pp. 411–420.

[24] M. Costabel, *Boundary integral operators on Lipschitz domains: Elementary results*, SIAM J. Math. Anal., 19 (1988), pp. 613–626.

[25] M. Costabel, *A coercive bilinear form for Maxwell's equations*, J. Math. Anal. Appl., 157 (1991), pp. 527–541.

[26] M. Costabel and M. Dauge, *Maxwell and Lamé eigenvalues on polyhedra*, Math. Methods Appl. Sci., 22 (1999), pp. 243–258.

[27] M. Costabel, M. Dauge, and S. Nicaise, *Singularities of Maxwell interface problems*, Math. Model. Numer. Anal., 33 (1999), pp. 627–649.

[28] M. Costabel and E. Stephan, *Strongly elliptic boundary integral equations for electromagnetic transmission problems*, Proc. Roy. Soc. Edinborough, Sect. A, 109 (1988), pp. 271–296.

[29] M. Costabel and W. Wendland, *Strong ellipticity of boundary integral operators*, J. Reine Angew. Math., 372 (1986), pp. 39–63.

[30] M. Dauge, *Elliptic Boundary Value Problems on Corner Domains*, Lecture Notes in Math. 1341, Springer, Berlin, 1988.

[31] R. Dautray and J.-L. Lions, *Mathematical Analysis and Numerical Methods for Science and Technology*, Vol. 4, Springer, Berlin, 1990.

[32] A. de La Bourdonnaye, *Some formulations coupling finite element and integral equation methods for Helmholtz equation and electromagnetism*, Numer. Math., 69 (1995), pp. 257–268.

[33] V. Girault and P. Raviart, *Finite Element Methods for Navier–Stokes Equations*, Springer, Berlin, 1986.

[34] P. Grisvard, *Elliptic Problems in Nonsmooth Domains*, Pitman, Boston, 1985.

[35] C. Hazard and M. Lenoir, *On the solution of time-harmonic scattering problems for Maxwell's equation*, SIAM J. Math. Anal., 27 (1996), pp. 1597–1630.

[36] R. Hiptmair, *Canonical construction of finite elements*, Math. Comp., 68 (1999), pp. 1325–1346.

[37] R. Hiptmair, *Symmetric coupling for eddy current problems*, SIAM J. Numer. Anal., 40 (2002), pp. 41–65.

[38] R. Hiptmair and C. Schwab, *Natural boundary element methods for the electric field integral equation on polyhedra*, SIAM J. Numer. Anal., 40 (2002), pp. 66–86.

[39] G. Hsiao, *Mathematical foundations for the boundary field equation methods in acoustic and electromagnetic scattering*, in Analysis and Computational Methods in Scattering and Applied Mathematics. A volume in the memory of Ralph Ellis Kleinman, F. Santosa and I. Stakgold, eds., Chapman & Hall/CRC, Res. Notes Math. 147, Boca Raton, FL, 2000, pp. 149–163.

[40] C. Johnson and J. Nédélec, *On the coupling of boundary integral and finite element methods*, Math. Comp., 35 (1980), pp. 1063–1079.

[41] M. Kuhn and O. Steinbach, *FEM-BEM coupling for 3D exterior magnetic field problems*, Math. Methods Appl. Sci., 25 (2002), pp. 357–371.

[42] R. McCamy and E. Stephan, *Solution procedures for three-dimensional eddy-current problems*, J. Math. Anal. Appl., 101 (1984), pp. 348–379.

[43] W. McLean, *Strongly Elliptic Systems and Boundary Integral Equations*, Cambridge University Press, Cambridge, UK, 2000.

[44] J. Nédélec, *Mixed finite elements in* $\mathbb{R}^3$, Numer. Math., 35 (1980), pp. 315–341.

[45] J.-C. Nédélec, *Acoustic and Electromagnetic Equations: Integral Representations for Harmonic Problems*, Appl. Math. Sci. 44, Springer-Verlag, Berlin, 2001.

[46] P. A. Raviart and J. M. Thomas, *A Mixed Finite Element Method for Second Order Elliptic Problems*, Lecture Notes in Math. 606, Springer, New York, 1977, pp. 292–315.

[47] A. Schatz, *An observation concerning Ritz–Galerkin methods with indefinite bilinear forms*, Math. Comp., 28 (1974), pp. 959–962.

[48] L. R. Scott and Z. Zhang, *Finite element interpolation of nonsmooth functions satisfying boundary conditions*, Math. Comp., 54 (1990), pp. 483–493.

[49] T. von Petersdorff, *Boundary integral equations for mixed Dirichlet, Neumann and transmission problems*, Math. Methods Appl. Sci., 11 (1989), pp. 185–213.

SIAM J. NUMER. ANAL.
Vol. 41, No. 3, pp. 945–963

ON MAGNUS INTEGRATORS FOR TIME-DEPENDENT SCHRÖDINGER EQUATIONS[*]

MARLIS HOCHBRUCK[†] AND CHRISTIAN LUBICH[‡]

Abstract. Numerical methods based on the Magnus expansion are an efficient class of integrators for Schrödinger equations with time-dependent Hamiltonian. Though their derivation assumes an unreasonably small time step size, as would be required for a standard explicit integrator, the methods perform well even for much larger step sizes. This favorable behavior is explained, and optimal-order error bounds are derived that require no or only mild restrictions of the step size. In contrast to standard integrators, the error does not depend on higher time derivatives of the solution, which is in general highly oscillatory.

Key words. Magnus integrators, time-dependent Schrödinger equation, commutator bounds, error bounds

AMS subject classifications. 65L05, 65L70, 65M12, 65M20

DOI. 10.1137/S0036142902403875

1. Introduction.

We study numerical integrators for Schrödinger equations with time-dependent Hamiltonian,

$$i\,\frac{d\psi}{dt} = H(t)\psi, \qquad \psi(t_0) = \psi_0. \tag{1.1}$$

The computational Hamiltonian $H(t)$, which is a finite-dimensional Hermitian operator, is typically the sum of a discretized negative Laplacian and a time-dependent potential. As the discretization of an unbounded operator, $H(t)$ can be of arbitrarily large norm.

Magnus integrators are an interesting class of numerical methods for such problems [3, 12]. Though the error behavior of such methods is well understood in the case of moderately bounded $H(t)$ (see [6, 7]), no results are so far available when $\|H(t)\|$ becomes large. The present paper gives optimal-order estimates for situations in which the product of the time step h with $\|H(t)\|$ can be of arbitrary size. Even more interesting than the error bounds themselves are the mechanisms which lead to these bounds and which make Magnus methods perform so well for Schrödinger equations, as compared to standard explicit or implicit numerical integrators.

In section 2 we recall the concepts underlying the construction of Magnus integrators. Section 3 states the main results, which give asymptotically sharp error bounds for Magnus integrators, in a framework that applies to time-dependent Schrödinger equations requiring neither smallness nor bounds of $h\|H(t)\|$. The general procedure for obtaining such estimates is outlined in section 4 and is carried out in detail in sections 5 and 6 for methods of order 2 and 4, respectively. The extension to methods of arbitrary order is done in section 7. Numerical experiments illustrating the theoretical results are given in section 8. A basic assumption for the results of this

[*] Received by the editors March 12, 2002; accepted for publication (in revised form) December 18, 2002; published electronically June 18, 2003.

http://www.siam.org/journals/sinum/41-3/40387.html

[†] Mathematisches Institut, Heinrich-Heine Universität Düsseldorf, Universitätsstr. 1, D–40225 Düsseldorf, Germany (marlis@am.uni-duesseldorf.de).

[‡] Mathematisches Institut, Universität Tübingen, Auf der Morgenstelle 10, D–72076 Tübingen, Germany (lubich@na.uni-tuebingen.de).

paper is commutator bounds; their validity for a spectral discretization is shown in the appendix.

Magnus integrators require computing a matrix exponential multiplying a vector in every time step. For the large matrices (or rather, operators of large dimension) arising from the spatial discretization of Schrödinger equations, this can be done efficiently using operator splitting or Chebyshev or Lanczos approximations. These techniques are well documented in the literature and are not considered here. Because of the stable error propagation, errors arising from the approximation of the matrix exponentials could be straightforwardly included in the error analysis.

2. Magnus integrators. For the linear differential equation

$$\dot{y} = A(t)y, \qquad y(0) = y_0, \tag{2.1}$$

with a time-dependent matrix $A(t)$, the approach of Magnus [9] aims at writing the solution as

$$y(t) = \exp(\Omega(t))y_0 \tag{2.2}$$

for a suitable matrix $\Omega(t)$. An expression for $\Omega(t)$ is obtained by using the ansatz (2.2) and differentiating. This gives

$$\dot{y}(t) = \operatorname{dexp}_{\Omega(t)}(\dot{\Omega}(t))\, y(t),$$

where the dexp operator can be expressed as

$$\operatorname{dexp}_{\Omega}(B) = \varphi(\operatorname{ad}_\Omega)(B) = \sum_{k\ge 0} \frac{1}{(k+1)!}\operatorname{ad}_\Omega^k(B), \tag{2.3}$$

with $\varphi(z) = (e^z - 1)/z$ and $\operatorname{ad}_\Omega(B) = [\Omega, B] = \Omega B - B\Omega$. Hence, (2.2) solves (2.1) if

$$A(t) = \operatorname{dexp}_{\Omega(t)}(\dot{\Omega}(t)), \qquad \Omega(0) = 0. \tag{2.4}$$

As long as $\|\Omega(t)\| < \pi$ (which is *not* the situation of interest in this article), the operator $\operatorname{dexp}_{\Omega(t)}$ is invertible, and the series

$$\operatorname{dexp}^{-1}_{\Omega(t)}(A(t)) = \sum_{k\ge 0} \frac{\beta_k}{k!}\operatorname{ad}^k_{\Omega(t)}(A(t)) \tag{2.5}$$

converges. Here β_k is the kth Bernoulli number appearing in the series $z/(e^z-1) = \sum_0^\infty (\beta_k/k!)z^k$, which converges for $|z| < 2\pi$. (Note that $\|\operatorname{ad}_\Omega(B)\| \le 2\|\Omega\|\cdot\|B\|$, which shows that (2.5) indeed converges for $\|\Omega(t)\| < \pi$.) This gives an explicit differential equation for $\Omega(t)$:

$$\dot{\Omega} = A(t) - \frac{1}{2}[\Omega, A(t)] + \frac{1}{12}[\Omega,[\Omega, A(t)]] + \cdots .$$

Picard iteration yields the *Magnus expansion*

$$\begin{aligned}\Omega(t) = \int_0^t A(\tau)d\tau &- \frac{1}{2}\int_0^t \left[\int_0^\tau A(\sigma)d\sigma, A(\tau)\right] d\tau \\ &+ \frac{1}{4}\int_0^t \left[\int_0^\tau \left[\int_0^\sigma A(\mu)d\mu, A(\sigma)\right] d\sigma, A(\tau)\right] d\tau \\ &+ \frac{1}{12}\int_0^t \left[\int_0^\tau A(\sigma)d\sigma, \left[\int_0^\tau A(\mu)d\mu, A(\tau)\right]\right] d\tau + \cdots .\end{aligned} \tag{2.6}$$

Numerical methods based on this expansion are reviewed by Iserles, et al. [6]. They are of the form

$$y_{n+1} = \exp(\Omega_n) y_n \tag{2.7}$$

to give an approximation to $y(t_{n+1})$ at $t_{n+1} = t_n + h$. Here Ω_n is a suitable approximation of $\Omega(h)$ given by (2.6), with $A(t_n + \tau)$ instead of $A(\tau)$. This approximation involves first truncating the expansion, and second approximating the integrals, e.g., by replacing $A(t)$ locally by an interpolation polynomial $\widehat{A}(t)$ for the nodes $t_n + c_j h$, so that the integrals in the Magnus expansion can be computed analytically. If Ω_n is built up in this way, then we speak of an *interpolatory Magnus integrator.* A method of order p is obtained by combining a pth order truncation of the Magnus series and interpolation of $A(t)$ at the nodes of a pth order quadrature formula. A natural choice is Gaussian quadrature.

For example, the midpoint rule yields a second-order scheme with

$$\Omega_n = hA\left(t_n + \frac{h}{2}\right). \tag{2.8}$$

The two-point Gauss quadrature rule has nodes $c_{1,2} = 1/2 \mp \sqrt{3}/6$. This yields a fourth-order scheme with

$$\Omega_n = \frac{h}{2}(A_1 + A_2) + \frac{\sqrt{3}h^2}{12}[A_2, A_1], \tag{2.9}$$

where $A_j = A(t_n + c_j h)$, $j = 1, 2$.

High-order interpolatory Magnus integrators require the computation of many commutators per step. Their number can be significantly reduced in specially constructed (noninterpolatory) Magnus integrators as given by Blanes, Casas, and Ros [2].

For the purpose of this paper, the Magnus series approach is described only for motivation, since we are interested in the case of large $\|hA(t)\|$, for which $\operatorname{dexp}_{\Omega_n}$ need not be invertible and the Magnus expansion need not converge. The known convergence proofs of the Magnus series (see [1, 10]) require that the time interval be restricted to $\int_0^t \|A(\tau)\|\, d\tau \le r$ with $r \approx 1$, and there are actually examples of matrix functions with divergent Magnus series for $\int_0^t \|A(\tau)\|\, d\tau = \pi$. (The example of [10, p. 30] is, admittedly, not of the type studied in this paper.) In any case, the question of convergence of the Magnus series is irrelevant for the problem of obtaining error bounds, much in the same way as the possible convergence or divergence of Taylor series is of no importance for finite-order error bounds elsewhere in numerical analysis. The possible noninvertibility of the dexp operator and even the nonexistence of a representation (2.2) of the exact solution would appear to be more serious obstacles, but we will show how this problem can be circumvented, using a modified differential equation satisfied by the approximate solution instead of directly estimating the difference between the Magnus expansion and its truncation.

The results of Iserles and Nørsett [7] on the order of Magnus integrators are for $\|hA(t)\| \to 0$ and are obtained by studying the remainder of the truncated Magnus series (2.6). The constants in those estimates depend on norms of commutators of $A(t)$ for different values of t, which all become large with growing $\|A(t)\|$. Therefore, results on the classical order of a method must be viewed with caution in the case of the Schrödinger equation, which involves discretizations of unbounded operators. Nevertheless, Magnus integrators work extremely well even with step sizes for which $\|hA(t)\|$ is large. The aim of the present paper is to explain this unexpectedly good behavior.

3. Statement of results.

In this section we state our assumptions and main results. Throughout the paper, $\|\cdot\|$ is the Euclidean norm or its induced matrix norm, or occasionally the L^2 norm of functions. We write

$$A(t) = -iH(t) = -i(U + V(t)). \tag{3.1}$$

We assume, once and for all, that the Hermitian matrix-valued function $V(t)$ and its time derivatives are bounded by

$$\left\| \frac{d^m}{dt^m} V(t) \right\| \le M_m, \qquad m = 0, 1, 2, \dots . \tag{3.2}$$

The matrix U is assumed to be symmetric positive definite, with $\|v\| \le \|Uv\|$ for all v, but no bound is assumed for the operator norm $\|U\|$. We set

$$D = U^{1/2}. \tag{3.3}$$

The typical situation is given by a discretization of the spatially continuous case where $U = -\Delta + I$, e.g., with periodic boundary conditions on a cube Q, and $V(t)$ is a bounded multiplication operator, i.e., $(V(t)v)(x) = V(x,t)v(x)$ for a real-valued smooth potential $V(x,t)$. In this continuous case we have

$$\|Dv\|^2 = \int_Q |\nabla v|^2 dx + \int_Q v^2 dx,$$

so that $\|Dv\|$ is the familiar H^1 Sobolev norm of v. In the spatially discretized case, $\|Dv\|$ can be viewed as a discrete Sobolev norm. For a space discretization with minimal grid spacing Δx, we note $\|U\| \sim \Delta x^{-2}$ and $\|D\| \sim \Delta x^{-1}$.

Our main assumptions are commutator bounds such as

$$\|[U, V(t)]v\| \le K_0 \|Dv\| \quad \text{and} \quad \|[U, \dot V(t)]v\| \le K_1 \|Dv\| \tag{3.4}$$

for all t and all vectors v. Condition (3.4) is easily verified in the spatially continuous case, with $U = -\Delta + I$ and a smooth potential $V(x,t)$ acting as a multiplication operator. The bound is obtained by noting that in one space dimension, with $' = d/dx$,

$$[U, V]v = -((Vv)'' - Vv'') = -(2V'v' + V''v),$$

with the obvious generalization to higher space dimensions. Hence, $[U, V]$ is a *first-order* differential operator, which yields (3.4). For a spectral discretization the bound (3.4) is shown, uniformly in the discretization parameter, in [8, Lemma 3.1].

Since $[A(\tau), A(\sigma)] = [U, V(\sigma) - V(\tau)] = \int_\tau^\sigma [U, \dot V(t)]\, dt$ (when $V(\sigma)$ and $V(\tau)$ commute), the second bound of (3.4) implies, for all vectors v,

$$\|[A(\tau), A(\sigma)]v\| \le K_1 h \, \|Dv\| \qquad \text{for } |\tau - \sigma| \le h. \tag{3.5}$$

THEOREM 3.1. *If $A(t)$ satisfies the commutator bound* (3.5), *then the error of the exponential midpoint rule* (2.7) *with* (2.8) *is bounded by*

$$\|y_n - y(t_n)\| \le Ch^2 \, t_n \max_{0 \le t \le t_n} \|Dy(t)\|.$$

The constant C depends only on M_m for $m \le 2$ and on K_1. In particular, C is independent of n, h, and $\|D\|$.

This error bound is to be contrasted with the error bound of the classical implicit midpoint rule $y_{n+1} = y_n + hA(t_{n+1/2})(y_n + y_{n+1})/2$, for which

$$\|y_n - y(t_n)\| \le Ch^2\, t_n \max_{0\le t\le t_n} \left\| \frac{d^3}{dt^3} y(t) \right\|.$$

Since solutions of Schrödinger equations are in general highly oscillatory, the appearance of higher time derivatives is unfavorable. On the other hand, $\|Dy(t)\|^2$ represents essentially the quantum kinetic energy, which is bounded a priori. We remark that a similar but weaker estimate for the exponential midpoint rule, with $\|D^2 y(t)\|$ instead of $\|Dy(t)\|$, was previously obtained in our paper [5] with a different proof. (Unfortunately, that paper also states an error bound for a third-order Magnus method, without detailed proof and involving the operator norm $\|U\|$, which is superseded by the results of the present paper.)

For methods of order p which contain commutator products of $A(t_n + c_j h)$ with r factors (in Example 2 we have $r = 2$ for order $p = 4$, and $r \le p - 1$ holds for all Magnus methods proposed in the literature), we assume that A satisfies, for all τ_j,

$$\left\| \left[A(\tau_k), \left[\dots, \left[A(\tau_1), \frac{d^m}{dt^m} V(\tau_0) \right] \right] \dots \right] v \right\| \le K\, \|D^k v\| \qquad \begin{cases} 0 \le m \le p, \\ k+1 \le rp. \end{cases} \tag{3.6}$$

Like (3.5), condition (3.6) is easily verified in the spatially continuous case. For a spectral space discretization of a time-dependent Schrödinger equation, we show in the appendix that (3.6) is indeed satisfied uniformly in the discretization parameter. Since $[A(\tau_1), A(\tau_0)] = [A(\tau_1), A(\tau_0) - A(\tau_1)] = [A(\tau_1), i\int_{\tau_0}^{\tau_1} \dot V(\tau) d\tau]$, condition (3.6) implies, whenever $|\tau_1 - \tau_0| \le h$,

$$\|[A(\tau_k), [\dots, [A(\tau_1), A(\tau_0)]] \dots] v\| \le Kh\, \|D^k v\|, \qquad k+1 \le rp. \tag{3.7}$$

Unlike the case of the exponential midpoint rule in Theorem 3.1, convergence of higher-order methods is shown only in the spatially discrete case under a step size restriction

$$h\, \|D\| \le c. \tag{3.8}$$

Note that this restriction is milder than the step size restriction for explicit integrators, such as Runge–Kutta methods, for which a more stringent condition $h\|D\|^2 \le c$ (i.e., $h\|A(t)\| \le \gamma$ for some constant γ) is required for stability. The classical error bounds for implicit integrators require smallness of $h\|D\|^2$ unless high temporal smoothness is supposed.

The following error bound holds for pth-order interpolatory Magnus integrators, i.e., those based on a pth-order truncation of the Magnus series and polynomial interpolation of $A(t)$ at the nodes of a pth-order quadrature formula (see section 2).

THEOREM 3.2. *If the commutator bounds* (3.6) *hold, then pth-order interpolatory Magnus integrators satisfy the error bound*

$$\|y_n - y(t_n)\| \le Ch^p\, t_n \max_{0\le t\le t_n} \|D^{p-1} y(t)\|$$

for time steps h restricted by (3.8). *The constant C depends only on M_m for $m \le p$, on K, c, and on p. In particular, C is independent of n, h, and $\|D\|$ as long as $h\|D\| \le c$.*

The error bound of Theorem 3.2 is valid also for noninterpolatory Magnus methods if the quadrature error satisfies, for all v,

$$\|\Omega_n v - \widetilde{\Omega}_n v\| \le C h^{p+1}\, \|D^{p-1} v\|, \tag{3.9}$$

where $\widetilde{\Omega}_n$ denotes the pth-order truncation of the Magnus series at t_n. This generalization of Theorem 3.2 follows directly from the proof below.

Condition (3.8) is not required for stability. If such a condition on the step size is not imposed, there is still pth-order convergence, though only for much smoother solutions: the error is then bounded by $C_p h^p \max \|D^{p-1} y(t)\| + C_{p+1} h^{p+1} \max \|D^p y(t)\| + \cdots + C_{pr} h^{pr} \max \|D^{pr-1} y(t)\|$.

Though Theorem 3.2 is formulated for arbitrary order p, we note that high-order error bounds are of limited value in the approximation of highly oscillatory solutions, for which (discretized) high-order derivatives $D^k y(t)$ have progressively much larger norms.

Theorems 3.1 and 3.2 are proved in the remainder of this article. In the following section we describe a general procedure for deriving error bounds. We will follow this procedure in detail for the exponential midpoint rule in section 5 and for fourth-order methods in section 6. This gives all the tools for the extension to the general case, which is treated in section 7.

4. General procedure for deriving error bounds.

The convergence analysis is done in two steps. In the first step we study the error which results from truncating the Magnus expansion; in the second step, we discuss the error resulting from approximating the integrals by quadrature. (In the estimates of this and the following sections, C is a generic constant, which assumes different values on different occurrences.)

Truncation of the Magnus expansion amounts to using a modified $\widetilde{\Omega}$ instead of Ω in (2.2), i.e.,

$$\widetilde{y}(t) = \exp(\widetilde{\Omega}(t)) y_0.$$

By differentiating, we obtain the approximate solution $\widetilde{y}(t)$ as the solution of the modified differential equation

$$\dot{\widetilde{y}}(t) = \widetilde{A}(t)\widetilde{y}(t) \qquad \text{with} \qquad \widetilde{A}(t) = \operatorname{dexp}_{\widetilde{\Omega}(t)}(\dot{\widetilde{\Omega}}(t)), \tag{4.1}$$

with initial value $\widetilde{y}(0) = y_0$. Note that the truncated Magnus series $\widetilde{\Omega}(t)$ and the modified operator $\widetilde{A}(t)$ are skew Hermitian if $A(t)$ is skew Hermitian. As the following lemma shows, a bound on $\widetilde{A} - A$ then immediately gives a local error bound.

LEMMA 4.1. *Let y be a solution of* (2.1) *with skew Hermitian A, $\widetilde{y}$ a solution of* (4.1). *With $E = \widetilde{A} - A$, the error satisfies*

$$\|\, \widetilde{y}(t) - y(t)\, \| \le \int_0^t \|E(\tau) y(\tau)\| d\tau.$$

Proof. We write (2.1) as $\dot{y} = A(t)y = \widetilde{A}(t)y - E(t)y$ and subtract (4.1). This shows that the error $\widetilde{\varepsilon} = \widetilde{y} - y$ satisfies

$$\dot{\widetilde{\varepsilon}} = \widetilde{A}(t)\widetilde{\varepsilon} + E(t)y, \qquad \widetilde{\varepsilon}(0) = 0.$$

Since $\widetilde{A}$ is skew Hermitian, taking the inner product with $\widetilde{\varepsilon}$ on both sides leads to

$$\langle \dot{\widetilde{\varepsilon}}, \widetilde{\varepsilon} \rangle = \langle E\, y, \widetilde{\varepsilon} \rangle \leq \|E\, y\| \, \|\widetilde{\varepsilon}\|.$$

On the other hand, $\langle \dot{\widetilde{\varepsilon}}, \widetilde{\varepsilon} \rangle = \frac{1}{2} \frac{d}{dt} \|\widetilde{\varepsilon}\|^2 = \frac{d}{dt} \|\widetilde{\varepsilon}\| \cdot \|\widetilde{\varepsilon}\|$. Integrating the inequality proves the lemma. □

A crucial step in obtaining a bound on $E = \widetilde{A} - A$ is truncating the dexp series (2.3) and providing a bound for the remainder. We define the remainder function r_p, for $p \geq 1$, via

$$\frac{e^z - 1}{z} = 1 + \frac{1}{2} z + \cdots + \frac{1}{(p-1)!}\, z^{p-2} + \frac{1}{p!}\, z^{p-1} r_p(z), \tag{4.2}$$

so that

$$\operatorname{dexp}_\Omega(B) = B + \frac{1}{2}[\Omega, B] + \cdots + \frac{1}{(p-1)!}\, \mathrm{ad}_\Omega^{p-2}(B) + \frac{1}{p!}\, r_p(\mathrm{ad}_\Omega)\big(\mathrm{ad}_\Omega^{p-1}(B)\big). \tag{4.3}$$

For $A(t)$ of the form (3.1) satisfying the conditions of section 3, we will bound the remainder term by

$$\Big\| r_p(\mathrm{ad}_{\widetilde{\Omega}(t)})\big(\mathrm{ad}_{\widetilde{\Omega}(t)}^{p-1}\big(\dot{\widetilde{\Omega}}(t)\big)\big) v \Big\| \leq C h^p \, \|D^{p-1} v\|, \qquad 0 \leq t \leq h. \tag{4.4}$$

In the case of $p > 2$, it turns out that the bound requires time steps h with (3.8), while for $p = 2$, no restriction on h is necessary.

Next we incorporate the error resulting from approximating the integrals. In the nth time step, we take $\widetilde{\Omega}(h)$ corresponding to the truncated Magnus series for $A(t_n + t)$ instead of $A(t)$, which we denote by $\widetilde{\Omega}_n$. By the quadrature approximation, $\widetilde{\Omega}_n$ is replaced by Ω_n, with which the actual computations are done. This approximation typically satisfies

$$\|\big(\widetilde{\Omega}_n - \Omega_n\big) v\| \leq C h^{p+1} \|D^{r-1} v\|, \tag{4.5}$$

where p is the order of the quadrature rule and commutator products with r factors appear in the method. For the exponential midpoint rule ($p = 2$, $r = 1$) this bound is independent of D. For pth-order interpolatory Magnus schemes (where $r \leq p - 1$) we will show that (4.5) holds and that this leads to the local error bound

$$\|\exp(\widetilde{\Omega}_n) v - \exp(\Omega_n) v\| \leq C h^{p+1} \|D^{r-1} v\|. \tag{4.6}$$

Putting both steps together, the exact solution y of (2.1) satisfies

$$y(t_{n+1}) = \exp(\Omega_n) y(t_n) + \varepsilon_n, \tag{4.7}$$

with $\varepsilon_n = y(t_{n+1}) - \exp(\widetilde{\Omega}_n) y(t_n) + \exp(\widetilde{\Omega}_n) y(t_n) - \exp(\Omega_n) y(t_n)$. By Lemma 4.1 and (4.6), this gives

$$\|\varepsilon_n\| \leq \int_{t_n}^{t_{n+1}} \|E(\tau) y(\tau)\| d\tau + C h^{p+1} \|D^{r-1} y(t_n)\|.$$

Subtracting (2.7) from (4.7) leads to the error recursion for $e_n = y_n - y(t_n)$:

$$e_{n+1} = \exp(\Omega_n) e_n + \varepsilon_n,$$

and thus

$$\|e_n\| \le \sum_{j=0}^{n-1} \|\varepsilon_j\|. \tag{4.8}$$

In summary, error bounds for general Magnus methods are obtained as follows: we have to provide a bound on $E(t)y(t)$, which basically means proving (4.4), and we have to show that the approximation Ω_n satisfies (4.6). This program is carried out in the following sections.

5. Error bounds for the exponential midpoint rule. In this section we prove Theorem 3.1. The second-order truncation of the Magnus expansion is simply

$$\widetilde{\Omega}(t) = \int_0^t A(\tau)\, d\tau, \qquad 0 \le t \le h.$$

Following the approach described in section 4, we know that $\widetilde{y}(t) = \exp(\widetilde{\Omega}(t))y_0$ solves (4.1) with

$$\widetilde{A}(t) = \operatorname{dexp}_{\widetilde{\Omega}(t)}(\dot{\widetilde{\Omega}}(t)) = A(t) + \frac{1}{2} r_2(\mathrm{ad}_{\widetilde{\Omega}(t)})\big(\mathrm{ad}_{\widetilde{\Omega}(t)}(\dot{\widetilde{\Omega}}(t))\big) =: A(t) + E_2(t), \tag{5.1}$$

where the representation (4.3) for the dexp operator was used. The remainder r_2 was defined in (4.2).

LEMMA 5.1. *r_2 satisfies* (4.4) *with $p = 2$, where the constant C depends only on M_0 of* (3.2) *and K_1 of* (3.5).

Proof. (a) We fix an arbitrary t with $0 \le t \le h$. After an orthogonal similarity transform, we may assume that $\Omega := \widetilde{\Omega}(t)$ is diagonal, $\Omega = \mathrm{diag}(\omega_k)$ with purely imaginary eigenvalues ω_k, and we define $B = \dot{\widetilde{\Omega}}(t)$. Denoting by $\bullet$ the entrywise product of matrices, we can write

$$\mathrm{ad}_\Omega(B) = \Omega B - B\Omega = Z \bullet B,$$

where $Z = (\omega_k - \omega_\ell)_{k,\ell}$. This yields

$$r_p(\mathrm{ad}_\Omega)\big(\mathrm{ad}_\Omega^{p-1}(B)\big)v = \big(R \bullet \mathrm{ad}_\Omega^{p-1}(B)\big)v,$$

where $R = (r_p(\omega_k - \omega_\ell))_{k,\ell}$. We now follow the proof of Lemma 2.2 of [5]. Note that for real x, $r_p(ix) = 1 + O(x)$, $x \to 0$, and $r_p(ix) = O(x^{-1})$, $|x| \to \infty$, and hence $r_p, r_p' \in L^2(i\mathbb{R})$. As can be seen, e.g., from formula (2.13) in [5], r_p has a Fourier transform $\widehat{r}_p \in L^1(\mathbb{R})$,

$$r_p(ix) = \int_{\mathbb{R}} e^{i\xi x} \widehat{r}_p(\xi)\, d\xi,$$

with $\|\widehat{r}_p\|_{L^1(\mathbb{R})} \le 2\pi \|r_p\|_{L^2(i\mathbb{R})}^{1/2} \|r_p'\|_{L^2(i\mathbb{R})}^{1/2}$. Consequently, the above expression can be written as

$$r_p(\mathrm{ad}_\Omega)\big(\mathrm{ad}_\Omega^{p-1}(B)\big)v = \int_{\mathbb{R}} \widehat{r}_p(\xi) \exp(\xi\Omega)\, \mathrm{ad}_\Omega^{p-1}(B) \exp(-\xi\Omega)\, v\, d\xi,$$

so that

$$\Big\| r_p(\mathrm{ad}_\Omega)\big(\mathrm{ad}_\Omega^{p-1}(B)\big)v \Big\| \le \|\widehat{r}_p\|_{L^1(\mathbb{R})} \sup_{\xi\in\mathbb{R}} \|\mathrm{ad}_\Omega^{p-1}(B) \exp(-\xi\Omega) v\|. \tag{5.2}$$

So far, this holds for general p.

(b) From now on, we set $p = 2$. Using

$$\mathrm{ad}_\Omega(B) = \mathrm{ad}_{\widetilde{\Omega}(t)}(\dot{\widetilde{\Omega}}(t)) = [\widetilde{\Omega}(t), \dot{\widetilde{\Omega}}(t)] = \int_0^t [A(\tau), A(t)]\, d\tau$$

and (3.5), we obtain for all vectors w

$$\|\mathrm{ad}_\Omega(B)w\| \le K_1 h^2 \|Dw\|.$$

Hence we have

$$\left\| r_p(\mathrm{ad}_\Omega)\big(\mathrm{ad}_\Omega^{p-1}(B)\big)v \right\| \le Ch^2 \sup_{\xi\in\mathbb{R}} \|D\exp(-\xi\Omega)v\|. \tag{5.3}$$

We now use the splitting (3.1) and write

$$\frac{i}{t}\Omega = U + \frac{1}{t}\int_0^t V(\tau)d\tau =: U + \widetilde{V}.$$

We choose $\alpha \ge 0$ such that $U+\widetilde{V}+\alpha I$ is symmetric and positive definite. To keep the notation simple, we omit the constants and denote by $\sim$ equivalent norms. Because of the boundedness of $\widetilde{V}$ and (3.3), we have for all vectors w

$$\|Dw\| = \sqrt{w^* U w} \sim \sqrt{w^*(U+\widetilde{V}+\alpha I)w} = \left\| \left(\frac{i}{t}\Omega + \alpha I\right)^{1/2} w \right\|.$$

We use this norm equivalence to bound the last factor in (5.3):

$$\begin{aligned} \|D\exp(-\xi\Omega)v\| &\sim \left\| \left(\frac{i}{t}\Omega + \alpha I\right)^{1/2} \exp(-\xi\Omega)v \right\| \\ &= \left\| \exp(-\xi\Omega) \left(\frac{i}{t}\Omega + \alpha I\right)^{1/2} v \right\| \\ &= \left\| \left(\frac{i}{t}\Omega + \alpha I\right)^{1/2} v \right\| \\ &\sim \|Dv\|. \end{aligned}$$

Inserting this into (5.3) proves the lemma. □

By definition (5.1) of E_2, this immediately yields the bound

$$\|E_2(t)y(t)\| \le Ch^2\|Dy(t)\|, \qquad 0 \le t \le h.$$

Applying Lemma 4.1 gives

$$\|\widetilde{\varepsilon}(t)\| \le Ch^3 \max_{0\le\tau\le h} \|Dy(\tau)\|. \tag{5.4}$$

The midpoint rule uses the approximation

$$\widetilde{\Omega}_n = \int_0^h A(t_n+\tau)d\tau \approx hA(t_{n+1/2}) =: \Omega_n$$

in the scheme (2.7). The midpoint rule is of order 2, and since $\|\ddot{A}(t)\| \le M_2$, the quadrature error is bounded by

$$\|\widetilde{\Omega}_n - \Omega_n\| \le \frac{1}{24} M_2 h^3.$$

The identity

$$\exp(\widetilde{\Omega}_n) - \exp(\Omega_n) = \int_0^1 \exp((1-s)\Omega_n)(\widetilde{\Omega}_n - \Omega_n)\exp(s\widetilde{\Omega}_n)\, ds$$

then yields

$$\|\exp(\widetilde{\Omega}_n) - \exp(\Omega_n)\| \le \frac{1}{24} M_2 h^3. \tag{5.5}$$

Combining (5.4) and (5.5) yields for the defects ε_j of (4.7)

$$\|\varepsilon_j\| \le Ch^3 \max_{t_j \le \tau \le t_{j+1}} \|Dy(\tau)\|.$$

By (4.8), this gives

$$\|e_n\| \le Ch^2\, t_n \max_{0 \le t \le t_n} \|Dy(t)\|,$$

which is just the statement of Theorem 3.1.

6. Error bounds for fourth-order Magnus methods. This section gives the proof of Theorem 3.2 for $p = 4$. It provides all the machinery needed for treating general-order p, but still gives an explicit presentation of the appearing terms.

A Magnus method of classical order 4 is constructed by setting

$$\dot{\widetilde{\Omega}}(t) = A(t_n + t) - \frac{1}{2}\int_0^t [A(t_n+\tau), A(t_n+t)]\, d\tau, \qquad \widetilde{\Omega}(t_n) = 0, \tag{6.1}$$

for $0 \le t \le h$. To study the local error we simplify the notation and consider the case $n = 0$. Then integration yields

$$\widetilde{\Omega}(t) = \int_0^t A(\tau)\, d\tau - \frac{1}{2}\int_0^t \int_0^\tau [A(\sigma), A(\tau)]\, d\sigma d\tau, \qquad 0 \le t \le h. \tag{6.2}$$

In this case, $\widetilde{y}(t) = \exp(\widetilde{\Omega}(t))y_0$ solves (4.1) with (partly omitting the argument t)

$$\widetilde{A}(t) = \dot{\widetilde{\Omega}}(t) + \frac{1}{2}[\widetilde{\Omega}, \dot{\widetilde{\Omega}}] + \frac{1}{6}[\widetilde{\Omega}, [\widetilde{\Omega}, \dot{\widetilde{\Omega}}]] + \frac{1}{24} r_4(\mathrm{ad}_{\widetilde{\Omega}})\big(\mathrm{ad}^3_{\widetilde{\Omega}}(\dot{\widetilde{\Omega}})\big). \tag{6.3}$$

Lemma 6.1. *If $h\|D\| \le c$, then r_4 defined in (4.2) satisfies (4.4) with $p = 4$, where the constant C depends only on K, M_0, and c.*

Proof. (a) The first part of the proof is identical to part (a) of the proof of Lemma 5.1. We write again, for fixed t with $0 \le t \le h$, $\Omega = \widetilde{\Omega}(t)$ and $B = \dot{\widetilde{\Omega}}(t)$, for $\widetilde{\Omega}(t)$ of (6.2). We start with the bound (5.2) and turn to estimating $\mathrm{ad}^3_\Omega(B)w$. Using the commutator bound (3.7) (and previously the Jacobi identity, if necessary) for terms such as, e.g.,

$$\left\| \left[\int_0^t A(\tau) d\tau, \left[\int_0^t A(\tau) d\tau, \left[\int_0^t \int_0^\tau [A(\sigma), A(\tau)]\, d\sigma d\tau, A(t) \right]\right]\right] w \right\| \le Kh^5 \|D^4 w\|,$$

it is shown under the restriction $h\|D\| \le c$ that, for all w,

$$\|\mathrm{ad}_\Omega^3(B)w\| \le Ch^4\|D^3w\|, \qquad 0 \le t \le h.$$

Inserted into (5.2), this bound yields

$$r_4(\mathrm{ad}_\Omega)\big(\mathrm{ad}_\Omega^3(B)\big)v \le Ch^4 \sup_{\xi\in\mathbb{R}} \|D^3 \exp(-\xi\Omega)v\|. \tag{6.4}$$

(b) It remains to show that the supremum can be bounded by $C\|D^3v\|$. We use the splitting (3.1) and write, still for fixed t with $0 < t \le h$,

$$\frac{i}{t}\Omega = U + \frac{1}{t}\int_0^t V(\tau)d\tau - \frac{i}{2t}\int_0^t\int_0^\tau [A(\sigma), A(\tau)]\,d\sigma d\tau =: U + \widetilde{V}.$$

By the assumptions, $\widetilde{V} = V(0) + O(h)$ is a Hermitian bounded operator, and thus there exists $\alpha \ge 0$ such that $U + \widetilde{V} + \alpha I$ is positive definite. Our next aim is to show that, for all w,

$$\|D^4w\| = \|U^2w\| \sim \left\|\left(\frac{i}{t}\Omega + \alpha I\right)^2 w\right\|. \tag{6.5}$$

We have

$$(U + \widetilde{V} + \alpha I)^2 - U^2 = 2(\widetilde{V} + \alpha I)U + [U, \widetilde{V} + \alpha I] + (\widetilde{V} + \alpha I)^2.$$

The first and the last term on the right-hand side yield bounds

$$\|2(\widetilde{V} + \alpha I)Uw\| + \|(\widetilde{V} + \alpha I)^2w\| \le C\|D^2w\| + C\|w\|. \tag{6.6}$$

Bounds for the second term are obtained from assumption (3.7). By definition of $\widetilde{V}$ and writing $U = iA(\tau) - V(\tau)$, we have

$$\begin{aligned}[U, \widetilde{V}] &= \left[U, \frac{1}{t}\int_0^t V(\tau)d\tau - \frac{i}{2t}\int_0^t\int_0^\tau [A(\sigma), A(\tau)]\,d\sigma d\tau\right] \\ &= \frac{i}{t}\int_0^t [A(\tau), V(\tau)]\,d\tau \\ &\quad + \frac{1}{2t}\int_0^t\int_0^\tau [A(0), [A(\sigma), A(\tau)]]\,d\sigma d\tau \\ &\quad + \frac{i}{2t}\int_0^t\int_0^\tau [V(0), [A(\sigma), A(\tau)]]\,d\sigma d\tau.\end{aligned}$$

By the commutator bounds (3.6) and (3.7) and the Jacobi identity, we obtain for $h\|D\| \le c$

$$\|[U, \widetilde{V}]w\| \le K\|Dw\| + \frac{1}{2}Kh^2\|D^2w\| + Kh\|D^2w\| \le C\|Dw\|. \tag{6.7}$$

Together with (6.6), this proves (6.5). Moreover, the estimates (6.6) and (6.7) show that

$$\|\big((U + \widetilde{V} + \alpha I)^2 - U^2\big)U^{-1}w\| \le C\|w\|.$$

Thus we can apply Lemma 6.2 below with $\mu = 1/2$ and $\theta = 3/4$ to show that (6.5) implies

$$\|D^3 w\| = \|U^{3/2} w\| \sim \left\| \left(\frac{i}{t}\Omega + \alpha I \right)^{3/2} w \right\|.$$

As at the end of the proof of Lemma 5.1, we then obtain

$$\|D^3 \exp(-\xi\Omega) v\| \le C \|D^3 v\|,$$

with a constant independent of $\xi \in \mathbb{R}$ and t with $0 < t \le h$. Inserting this bound into (6.4) completes the proof. □

LEMMA 6.2. *Suppose S, T are Hermitian positive definite operators such that $\|(S-T)S^{-\mu}\| \le M$ holds with $0 \le \mu < 1$. If*

$$\|Sv\| \le \|Tv\| \qquad \text{for all } v,$$

then, for $0 < \theta < 1$,

$$\|S^\theta v\| \le C \|T^\theta v\| \qquad \text{for all } v,$$

where C depends only on M and μ.

Proof. This is a reformulation of Theorem 1.4.6 in [4]. □

We are now in the position to prove a bound of $\widetilde{A}(t) - A(t)$.

LEMMA 6.3. *For $\widetilde{A}(t)$ defined in (6.3) and time steps h with $h\|D\| \le c$, the error $E_4(t) := \widetilde{A}(t) - A(t)$ is bounded, for all vectors v, by*

$$\|E_4(t) v\| \le C h^4 \|D^3 v\|, \qquad 0 \le t \le h. \tag{6.8}$$

The constant C depends only on K, M_0, M_1, and c.

Proof. We insert (6.1) and (6.2) into (6.3):

$$\begin{aligned} E_4(t) = & -\frac{1}{12} \int_0^t \int_0^t \int_0^t [A(\mu), [A(\tau), [A(\sigma), A(t)]]] \, d\sigma d\tau d\mu \\ & - \frac{1}{12} \int_0^t \int_0^t \int_0^\tau [A(\mu), [[A(\sigma), A(\tau)], A(t)]] \, d\sigma d\tau d\mu \\ & + \frac{1}{24} \int_0^t \int_0^\mu \int_0^t [[A(\sigma), A(\mu)], [A(\tau), A(t)]] d\tau d\sigma d\mu + R(t) \\ & + \frac{1}{24} r_4(\mathrm{ad}_{\widetilde{\Omega}}) \big(\mathrm{ad}^3_{\widetilde{\Omega}}(\dot{\widetilde{\Omega}})\big). \end{aligned}$$

Here, $R(t)v$ contains integrals of commutators which, by (3.7), are bounded by

$$C\big(h^5 \|D^4 v\| + h^6 \|D^5 v\|\big) \le C' h^4 \|D^3 v\|$$

for $h\|D\| \le c$. The constant C depends only on K. Then, by (3.7),

$$\|E_4(t) v\| \le C h^4 \|D^3 v\| + \frac{1}{24} \left\| r_4(\mathrm{ad}_{\widetilde{\Omega}}) \big(\mathrm{ad}^3_{\widetilde{\Omega}}(\dot{\widetilde{\Omega}})\big) v \right\|.$$

The bound (6.8) now follows from Lemma 6.1. □

Lemma 4.1 shows that $\widetilde{\varepsilon} = \widetilde{y} - y$ is bounded by

$$\|\widetilde{\varepsilon}(t)\| \le Ch^5 \max_{0\le\tau\le h} \|D^3 y(\tau)\|, \qquad 0 \le t \le h.$$

Since we want to have a fourth-order scheme, we use a quadrature formula $(b_i, c_i)_{i=1}^s$ of order $p \ge 4$. In (6.2) we replace A by its interpolation polynomial $\widehat{A}$ in the nodes $t_n + c_j h$. The integrals can then be evaluated exactly. The quadrature error for $n = 0$ is given by

$$\begin{aligned}(6.9)\qquad \widetilde{\Omega}_0 - \Omega_0 &= \int_0^h A(\tau)d\tau - \int_0^h \widehat{A}(\tau)d\tau \\ &\quad - \left(\frac{1}{2}\int_0^h\int_0^\tau [A(\sigma), A(\tau)]d\sigma d\tau - \frac{1}{2}\int_0^h\int_0^\tau [\widehat{A}(\sigma), \widehat{A}(\tau)]d\sigma d\tau\right),\end{aligned}$$

and similarly for the general nth step, with $A(t_n + \tau)$ instead of $A(\tau)$.

Lemma 6.4. *The quadrature error in the nth step satisfies*

$$(6.10)\qquad \|\big(\widetilde{\Omega}_n - \Omega_n\big)v\| \le Ch^{p+1}\|Dv\|.$$

The constant C depends only on M_m for $m \le p$ and K_1.

Proof. For ease of notation, we let $n = 0$. The error of the single integral in the representation of $\widetilde{\Omega}_n - \Omega_n$ is $O(h^{p+1})$. Assume that we use a quadrature rule with s nodes. For estimating the error of the double integral we define the interpolation error

$$J(t) := A(t) - \widehat{A}(t) = h^s \int_0^1 \widehat{\kappa}_s(\theta, \vartheta) A^{(s)}(\theta h) d\theta, \qquad 0 \le t = \vartheta h \le h,$$

where $\widehat{\kappa}_s$ denotes the Peano kernel. The difficulty in the remaining proof comes from the fact that we have only $J(t) = O(h^s)$, but we need an $O(h^p)$ estimate. We use, in addition, $J(c_i h) = 0$ and $\int_0^h J(t)\,dt = O(h^{p+1})$. For the second term in (6.9) we write

$$\begin{aligned}(6.11)\qquad &\int_0^h\int_0^\tau [A(\sigma), A(\tau)]\,d\sigma d\tau - \int_0^h\int_0^\tau [\widehat{A}(\sigma), \widehat{A}(\tau)]\,d\sigma d\tau \\ &\quad = \int_0^h\int_0^\tau \Big([\widehat{A}(\sigma), J(\tau)] + [J(\sigma), \widehat{A}(\tau)] + [J(\sigma), J(\tau)]\Big)\,d\sigma d\tau.\end{aligned}$$

Approximating the outer integral with the quadrature formula, the first term becomes

$$\int_0^h\int_0^\tau [\widehat{A}(\sigma), J(\tau)]\,d\sigma d\tau = h^{p+1}\int_0^1 \kappa_p(\theta) G^{(p)}(\theta h)\,d\theta,$$

where κ_p is the Peano kernel, and

$$G(\tau) = \int_0^\tau [\widehat{A}(\sigma), J(\tau)]\,d\sigma.$$

Using Leibniz' rule, it is seen that the dominant term of $G^{(p)}(\tau)$ is $p\,[\widehat{A}(\tau), J^{(p-1)}(\tau)]$, so that by (3.6), for any vector v,

$$\|G^{(p)}(\theta h)v\| \le C\|Dv\|.$$

This yields

$$\left\| \int_0^h \int_0^\tau [\widehat{A}(\sigma), J(\tau)]\, d\sigma d\tau\ v \right\| \le Ch^{p+1}\|Dv\|. \tag{6.12}$$

For the second term, we use partial integration:

$$\int_0^h \int_0^\tau [J(\sigma), \widehat{A}(\tau)]\, d\sigma d\tau = \left[\int_0^h J(\sigma) d\sigma, \int_0^h \widehat{A}(\mu) d\mu \right] - \int_0^h \int_0^\tau [J(\tau), \widehat{A}(\sigma)]\, d\sigma d\tau.$$

Here, for the last term, the bound was already given in (6.12). Using the quadrature formula for the integral over J, we have for the first term

$$\left[\int_0^h J(\sigma) d\sigma, \int_0^h \widehat{A}(\mu) d\mu \right] = h^{p+1} \left[\int_0^1 \kappa_p(\theta) J^{(p)}(\theta h) d\theta, \int_0^h \widehat{A}(\mu) d\mu \right].$$

Noting $J^{(p)}(t) = A^{(p)}(t)$ and using (3.6), this gives the bound

$$\left\| \int_0^h \int_0^\tau [J(\sigma), \widehat{A}(\tau)]\, d\sigma d\tau\ v \right\| \le Ch^{p+1}\|Dv\|. \tag{6.13}$$

Finally, since $\|J(t)\| = O(h^s)$,

$$\left\| \int_0^h \int_0^\tau [J(\sigma), J(\tau)]\, d\sigma\, d\tau\ v \right\| \le Ch^{2s+2}\|v\| \le Ch^{p+2}\|v\|. \tag{6.14}$$

Inserting the bounds (6.12)–(6.14) into (6.11) completes the proof. □

LEMMA 6.5. *In the situation of Lemma* 6.4,

$$\| \exp(\widetilde{\Omega}_n)v - \exp(\Omega_n)v \| \le Ch^{p+1}\|Dv\|.$$

The constant C depends only on M_m for $m \le p$ and K_1.

Proof. The variation-of-constants formula yields

$$\exp\big(\widetilde{\Omega}_n\big)v - \exp(\Omega_n)v = \int_0^1 \exp\big((1-s)\Omega_n\big)\big(\widetilde{\Omega}_n - \Omega_n\big)\exp\big(s\widetilde{\Omega}_n\big)v\, ds.$$

By (6.10) we have

$$\|(\widetilde{\Omega}_n - \Omega_n)\exp(s\widetilde{\Omega}_n)v\| \le Ch^{p+1}\|D\exp(s\widetilde{\Omega}_n)v\| \le C'h^{p+1}\|Dv\|,$$

where the last inequality is obtained as in the proof of Lemma 5.1. This gives the stated bound. □

For $p \ge 4$, the local error ε_n of the scheme (2.7) thus satisfies (4.7) with

$$\|\varepsilon_n\| \le Ch^5 \max_{t_n \le t \le t_{n+1}} \|D^3 y(t)\|.$$

Hence, with (4.8), the global error is bounded by

$$\|e_n\| \le Ct_n h^4 \max_{0 \le t \le t_n} \|D^3 y(t)\|,$$

and Theorem 3.2 is proved for $p = 4$.

7. Error bounds for higher-order Magnus integrators. The arguments of the previous section can be extended rather directly to methods of arbitrary order. In the following we describe this extension, putting the emphasis on the general structure and on a few additional considerations that become necessary. Though it would have been possible to present the general proof without first discussing the second- and fourth-order cases, we believe that it is useful to have seen and understood the explicit expressions arising in the proofs for the lower-order methods before embarking on the general case.

LEMMA 7.1. *If $\|hD\| \leq c$, then r_p defined in* (4.2) *satisfies, for a pth-order truncated Magnus expansion $\widetilde{\Omega}(t)$, the bound* (4.4), *where the constant C depends only on K, M_0, c, and p.*

Proof. For the truncated Magnus series $\widetilde{\Omega}(t)$, the expression $\mathrm{ad}^{p-1}_{\widetilde{\Omega}(t)}(\dot{\widetilde{\Omega}}(t))$ consists (after repeated use of the Jacobi identity) of a linear combination of iterated commutators of $A(\cdot)$ integrated over all but one of the independent variables over intervals bounded by h. The appearing iterated commutators and integrals are at least $(p-1)$-fold. Together with the commutator bound (3.7) and $\|hD\| \leq c$, this yields the bound

$$\left\| \mathrm{ad}^{p-1}_{\widetilde{\Omega}(t)}(\dot{\widetilde{\Omega}}(t))w \right\| \leq Ch^p \, \|D^{p-1}w\|. \tag{7.1}$$

By (5.2), this implies

$$\left\| r_p(\mathrm{ad}_{\widetilde{\Omega}(t)})\mathrm{ad}^{p-1}_{\widetilde{\Omega}(t)}(\dot{\widetilde{\Omega}}(t))v \right\| \leq Ch^p \sup_{\xi \in \mathbb{R}} \|D^{p-1} \exp(\xi \widetilde{\Omega}(t))v\| \tag{7.2}$$

for all v. By a straightforward but tedious generalization of the argument in the proof of Lemma 6.1, the supremum is bounded by

$$\sup_{\xi \in \mathbb{R}} \|D^{p-1} \exp(\xi \widetilde{\Omega}(t))v\| \leq C \, \|D^{p-1}v\|, \tag{7.3}$$

which yields the desired bound (4.4). □

LEMMA 7.2. *For $\widetilde{A}(t)$ defined in* (4.1) *and time steps h with $h\|D\| \leq c$, the error $E_p(t) := \widetilde{A}(t) - A(t)$ is bounded, for all vectors v, by*

$$\|E_p(t)v\| \leq Ch^p \, \|D^{p-1}v\|, \qquad 0 \leq t \leq h. \tag{7.4}$$

The constant C depends only on K, M_0, M_1, c, and p.

Proof. By construction of the Magnus series, $E_p(t)$ is a linear combination of at least $(p-1)$-fold integrals of iterated commutators of $A(\cdot)$ and $r_p(\mathrm{ad}_{\widetilde{\Omega}(t)})\mathrm{ad}^{p-1}_{\widetilde{\Omega}(t)}(\dot{\widetilde{\Omega}}(t))$. The stated estimate thus follows from the commutator bound (3.7), the step size bound (3.8), and Lemma 7.1. □

Consider a quadrature formula with nodes c_i $(i = 1, \ldots, s)$ and weights b_i of order p. Let $\widehat{A}(\tau)$ be the interpolation polynomial to $A(\tau)$ in the points $c_i h$, and denote by Ω_0 the expression obtained by replacing $A(\tau)$ by $\widehat{A}(\tau)$ in the expression for $\widetilde{\Omega}_0 = \widetilde{\Omega}(h)$. Similarly, let $\widetilde{\Omega}_n$ and Ω_n denote the corresponding expressions for the nth step, with $A(t_n + \tau)$ instead of $A(\tau)$.

LEMMA 7.3. *The quadrature error in the nth step satisfies*

$$\|(\widetilde{\Omega}_n - \Omega_n)v\| \le Ch^{p+1}\|D^{p-2}v\|. \tag{7.5}$$

The constant C depends only on M_m for $m \le p$ and K_1.

Proof. The proof follows the lines of the proof of Lemma 6.4. The generalization concerns the appearance of m-fold integrals of m-fold iterated commutators, for $m \le p-2$, instead of the simple commutators studied in the proof of Lemma 6.4. These terms are treated by the same techniques; they just involve more formidable expressions. By the commutator bound (3.6), this leads to an estimate involving $\|D^{p-2}v\|$ in the general situation. □

As in Lemma 6.5, this implies

$$\|\exp(\widetilde{\Omega}_n)v - \exp(\Omega_n)v\| \le Ch^{p+1}\|D^{p-2}v\|. \tag{7.6}$$

Inserting the estimates (7.4) and (7.6) into the framework of section 4 finally yields the error bound of Theorem 3.2.

8. Numerical experiments.

To illustrate the theoretical results presented in this paper, we consider the Schrödinger equation

$$i\,\frac{\partial\psi}{\partial t} = -\frac{1}{2}\Delta\psi + b(x,t)\psi, \qquad x = (x_1,\dots,x_d) \in \mathbb{R}^d,\ t > 0, \tag{8.1}$$

with a smooth (C^∞) potential $b(x,t)$ that is 2π-periodic in every coordinate direction x_j. We impose periodic initial conditions $\psi(x,0) = \psi_0(x)$. For ease of notation only, the following discussion is for the one-dimensional case $d = 1$.

A standard space discretization is given by the pseudospectral method. Here, a trigonometric polynomial

$$\psi^N(x,t) = \sum_{k=-N/2}^{N/2-1} c_k^N(t)\, e^{ikx}$$

is determined such that the equations

$$\begin{aligned} i\dot{\psi}^N(x_\ell,t) &= -\frac{1}{2}\Delta\psi^N(x_\ell,t) + b(x_\ell,t)\psi^N(x_\ell,t),\\ \psi^N(x_\ell,0) &= \psi_0(x_\ell) \end{aligned}$$

are satisfied at the mesh-points $x_\ell = 2\pi\ell/N$, with $\ell = -N/2,\dots,N/2-1$. Setting $c^N(t) = (c_k^N(t))$ the vector of Fourier coefficients, this amounts to solving

$$i\dot{c}^N = -\frac{1}{2}\widehat{\Delta}^N c^N + B^N(t)c^N, \tag{8.2}$$

where, in the case of one space dimension,

$$\widehat{\Delta}^N = (\widehat{D}^N)^2 \qquad \text{with} \quad \widehat{D}^N = \operatorname{diag}(ik) \quad (k = -N/2,\dots,N/2-1),$$

and, with F_N denoting the discrete Fourier transform of length N,

$$B^N(t) = F_N \operatorname{diag}(b(x_\ell,t))F_N^{-1}.$$

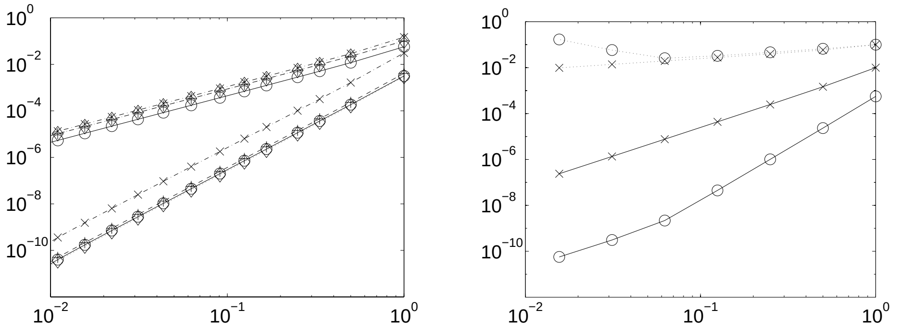


FIG. 8.1. *Error versus step sizes for the laser example: smooth and nonsmooth initial data on the left, $h\|D\| = $ const. on the right.*

We consider a one-dimensional example with data from [11], slightly modified to make the potential periodic with respect to the space interval $x \in [-\ell, \ell]$ for $\ell = 10$:

$$b(x,t) = \frac{1}{2}\frac{\pi^2}{\ell^2}\left(1 - \cos\frac{\pi x}{\ell}\right) + \sin^2(t)\frac{\pi}{\ell}\sin\frac{\pi x}{\ell}.$$

In the left-hand panel of Figure 8.1 we give precision step size diagrams at $t = 1$ for four different initial values, where we have used $N = 128$ Fourier modes for the spatial discretization. As a smooth initial value, we used the eigenstate of the unforced harmonic oscillator to the lowest energy level, $\Psi(x,0) = e^{-x^2/2}$. The convergence curves of the exponential midpoint and the fourth-order Gauss method corresponding to the smooth initial data are the solid lines marked with circles. For the other three curves, initial data of finite energy is chosen as $c^N(0) = (I - i(\widehat{D}^N)^j)^{-1}v/\rho$, $j = 1, 2, 3$, where v is a vector of normally distributed random numbers, and ρ is chosen such that $\|c^N(0)\| = 1$. For $j = 1$, the results are plotted in the dash-dotted curve marked with × symbols; for $j = 2$, we have the dashed curved marked with + symbols; and for $j = 3$, the curve is dotted marked with diamonds.

For the right-hand panel of Figure 8.1, we took the smooth initial state $\Psi(x,0) = e^{-x^2/2}$ for all curves, but varied the number of Fourier modes from $N = 32$ to $N = 2048$ and the time steps such that $Nh = 32$. This corresponds to the situation in which $\|hD\| \approx 3.5$, where $D = (-\frac{1}{2}\widehat{\Delta}^N + I)^{1/2}$. The solid line marked with the × symbols indicates the error of the midpoint rule, and the solid line marked with circles is the error for the fourth-order Gauss method. The dotted lines in the top of the picture represent the errors of the exponential midpoint and the Gauss method divided by h^2 and h^4, respectively, up to a constant factor.

9. Appendix. Commutator bounds for a spectral discretization.

We consider the pseudospectral space discretization (8.2) of the Schrödinger equation (8.1). Equation (8.2) is of the type studied in this paper, with $U = -\frac{1}{2}(\widehat{D}^N)^2 + I$ and $V(t) = B^N(t) - I$. The matrix $B^N(t)$ is circulant, with (k,l) entry equal to

$$\widehat{b}^N_{k-l}(t) = \sum_{q=-\infty}^{\infty} \widehat{b}_{k-l+qN}(t)$$

by the aliasing formula, where $\widehat{b}_j(t)$ is the jth Fourier coefficient of the 2π-periodic (in x) function $b(x,t)$. If (and only if) $b(x,t)$ is a C^∞ function of x, the Fourier coefficients $\widehat{b}_j(t)$ decay faster than any negative power of $|j|$. It then follows that the entries of the matrix $B^N(t) = (b^N_{k,l})$ are bounded by

$$|b^N_{kl}| \leq \begin{cases} \gamma_m(|k-l|+1)^{-m}, & |k-l| \leq N/2, \\ \gamma_m(N-|k-l|)^{-m}, & |k-l| > N/2, \end{cases} \tag{9.1}$$

for $k, l = -N/2, \dots, N/2-1$, with γ_m $(m = 1, 2, 3, \dots)$ independent of N.

The commutator bound (3.6) is obtained as a direct consequence of the three lemmas below, for which we need to give a further definition. We say that a sequence of matrices $\mathcal{B} = (B^N)$, with B^N of dimension $N \times N$, belongs to the class Γ^∞ if the entries satisfy estimates (9.1) with all γ_m independent of N. We denote by $\gamma(\mathcal{B}) = (\gamma_1, \gamma_2, \gamma_3, \dots)$ the sequence of smallest possible such numbers.

Lemma 9.1. *If $\mathcal{A} = (A^N)$ and $\mathcal{B} = (B^N)$ are in Γ^∞, then also $\mathcal{AB} = (A^N B^N)$ is in Γ^∞, and $\gamma(\mathcal{AB})$ is bounded in terms of $\gamma(\mathcal{A})$ and $\gamma(\mathcal{B})$.*

The proof is by direct estimation and is not given here. The following result is shown in the proof of Lemma 3.1 in [8].

Lemma 9.2. *If $\mathcal{B} = (B^N)$ is in Γ^∞, then $[\widehat{\mathcal{D}}^2, \mathcal{B}] = ([(\widehat{D}^N)^2, B^N])$ is of the form*

$$[\widehat{\mathcal{D}}^2, \mathcal{B}] = \mathcal{M}_0 + \mathcal{M}_1 \widehat{\mathcal{D}},$$

where $\mathcal{M}_0$ and $\mathcal{M}_1$ are in Γ^∞, with $\gamma(\mathcal{M}_0)$ and $\gamma(\mathcal{M}_1)$ bounded in terms of $\gamma(\mathcal{B})$.

The next lemma is proved in the same way.

Lemma 9.3. *If $\mathcal{B} = (B^N)$ is in Γ^∞, then $\widehat{\mathcal{D}}\mathcal{B} = (\widehat{D}^N B^N)$ is of the form*

$$\widehat{\mathcal{D}}\mathcal{B} = \mathcal{K}_0 + \mathcal{K}_1 \widehat{\mathcal{D}},$$

where $\mathcal{K}_0$ and $\mathcal{K}_1$ are in Γ^∞, with $\gamma(\mathcal{K}_0)$ and $\gamma(\mathcal{K}_1)$ bounded in terms of $\gamma(\mathcal{B})$.

Repeated application of these lemmas shows that

$$\left[-(\widehat{D}^N)^2 + B^N(\tau_k), \left[\dots, \left[-(\widehat{D}^N)^2 + B^N(\tau_1), \frac{d^m}{dt^m} B^N(\tau_0)\right]\right] \dots\right] = \sum_{j=0}^{k} M_j^N (\widehat{D}^N)^j,$$

with matrices M_j^N bounded independently of N and $\tau_0, \dots, \tau_k$. This gives the desired commutator bound (3.6).

Acknowledgment. We thank Katina Lorenz for stimulating discussions on commutator bounds.

REFERENCES

[1] S. Blanes, F. Casas, J.A. Oteo, and J. Ros, *Magnus and Fer expansions for matrix differential equations: The convergence problem*, J. Phys. A., 22 (1998), pp. 259–268.

[2] S. Blanes, F. Casas, and J. Ros, *Improved high order integrators based on the Magnus expansion*, BIT, 40 (2000), pp. 434–450.

[3] S. Blanes and P.C. Moan, *Splitting methods for the time-dependent Schrödinger equation*, Phys. Lett. A, 265 (2000), pp. 35–42.

[4] D. Henry, *Geometric Theory of Semilinear Parabolic Equations*, Lecture Notes in Math. 840, Springer, Berlin, Heidelberg, 1981.

[5] M. Hochbruck and C. Lubich, *Exponential integrators for quantum-classical molecular dynamics*, BIT, 39 (1999), pp. 620–634.

[6] A. Iserles, H.Z. Munthe-Kaas, S.P. Nørsett, and A. Zanna, *Lie group methods*, Acta Numer., 9 (2000), pp. 215–365.

[7] A. Iserles and S.P. Nørsett, *On the solution of linear differential equations in Lie groups*, Philos. Trans. Royal Soc. A, 357 (1999), pp. 983–1019.

[8] T. Jahnke and C. Lubich, *Error bounds for exponential operator splittings*, BIT, 40 (2000), pp. 735–744.

[9] W. Magnus, *On the exponential solution of differential equations for a linear operator*, Comm. Pure Appl. Math., 7 (1954), pp. 649–673.

[10] P.C. Moan, *On Backward Error Analysis and Nekhoroshev Stability in the Numerical Analysis of Conservative Systems of ODEs*, Ph.D. thesis, University of Cambridge, Cambridge, UK, 2002.

[11] U. Peskin, R. Kosloff, and N. Moiseyev, *The solution of the time dependent Schrödinger equation by the (t, t') method: The use of global polynomial propagators for time dependent Hamiltonians*, J. Chem. Phys., 100 (1994), pp. 8849–8855.

[12] I.V. Puzynin, A.V. Selin, and S.I. Vinitsky, *Magnus-factorized method for numerical solving the time-dependent Schrödinger equation*, Comput. Phys. Comm., 126 (2000), pp. 158–161.

SIAM J. NUMER. ANAL.
Vol. 41, No. 3, pp. 964–982

AN OPTIMAL PARALLEL NONOVERLAPPING DOMAIN DECOMPOSITION ITERATIVE PROCEDURE[*]

QINGPING DENG[†]

Abstract. A nonoverlapping domain decomposition iterative procedure is developed and analyzed for second order elliptic problems in $\mathbb{R}^N$. Its convergence is proved. The method is based on a Robin-type consistency condition with two parameters, called a transmission coefficient and a penalty coefficient, as a transmission condition together with a derivative-free transmission data updating technique on the artificial interfaces. Then the method is applied to the nonconforming finite element problems. A nonoverlapping domain decomposition iterative procedure for solving the nonconforming finite element problems of second order partial differential equations is developed and analyzed, which is directly presented to the nonconforming finite element problems without introducing any Lagrange multipliers. Its convergence is demonstrated, and the convergence rate is derived. The convergence analyses imply that the convergence rate is independent of the finite element meshes size while choosing the right parameters. Furthermore, the conclusions are extended to the unstructured finite element meshes. For both continuous problems and discrete problems, the method of this paper can be applied to general multisubdomain decompositions and implemented on parallel machines with local communications naturally. The method also allows choosing subdomains very flexibly, even as small as an individual element for finite element problems.

Key words. nonoverlapping, domain decomposition, iterative, parallel, transmission data, update, transmission coefficient, penalty coefficient, Robin boundary condition, nonconforming, finite element

AMS subject classifications. 65F10, 65N30

DOI. 10.1137/S0036142902401281

Introduction. Nonoverlapping domain decomposition methods have been studied extensively and have become very attractive for their parallelism and flexibility (cf. [1, 2, 3, 7, 8, 10, 11, 13, 14, 17, 18, 19, 20, 21, 22, 23] and the references therein). In nonoverlapping domain decomposition methods, the original domain is decomposed into subdomains; then the original problems are split into a number of subproblems over the subdomains. The subproblems could be solved in parallel or with greater independence. The main issues to develop nonoverlapping domain decomposition iterative procedures are what information should be transferred between subdomains (subproblems) and how the information is transferred. In other words, one issue is what the transmission data are and the other is how the transmission data are exchanged, that is, the strategy for updating the transmission data. The transmission data should guarantee that the solutions of subproblems could be pieced together into a reasonable approximation of the true solution of the original given problem. The updating strategy determines the costs and methods of the communications between subproblems.

The objective of this paper is to develop a nonoverlapping domain decomposition iterative procedure for second order elliptic partial differential problems and their nonconforming finite element problems. First, a nonoverlapping domain decomposition iterative procedure is developed and analyzed for second order elliptic problems in $\mathbb{R}^N$ $(N = 2, 3)$. Its convergence is proved by a "pseudoenergy" technique. The

[*] Received by the editors January 22, 2002; accepted for publication (in revised form) November 6, 2002; published electronically June 18, 2003.
http://www.siam.org/journals/sinum/41-3/40128.html

[†] Department of Mathematics, University of Tennessee, Knoxville, TN 37996 (deng@math.utk.edu).

method is based on a Robin-type consistency condition with two parameters, called a transmission coefficient and a penalty coefficient, as a transmission condition and a derivative-free transmission data updating technique on the artificial interfaces between two subdomains. Then the method is applied to the nonconforming finite element problems. A nonoverlapping domain decomposition iterative procedure for solving the nonconforming finite element problems of second order partial differential equations is also developed and analyzed. The method is directly presented to the nonconforming finite element problems without introducing any Lagrange multipliers. The convergence is proved, and the convergence rate of the method for the nonconforming finite element problems is derived. The convergence analyses show that the convergence rate is independent of the finite element mesh size, that is, optimal, while choosing the right parameters. Furthermore, the conclusions are extended to the unstructured finite element meshes.

For both continuous problems and discrete problems, the method can be applied to general multisubdomain decompositions and implemented on parallel machines with local simple communications naturally. The method allows choosing subdomains very flexibly, even as small as individual elements for finite element problems. For continuous problems, the transmission data and their updating techniques guarantee that all subproblems are always well-posed at every iteration if the initial subproblems are well-posed. On different iterative steps, the subproblems on the same subdomain are the same problems with the same kind of boundary condition but different boundary values, that is, which have the same differential operator. Hence, for finite element problems, the subproblems on the same subdomain, which are linear systems, have the same system matrix but different right-hand side terms at different iterations. Thus, we need to do only some simple operations of "matrix times vector" to solve the subproblems on every iterative step if decomposing the system matrix of the subproblems on every subdomain or finding its inverse matrix first. Last but not least, we would like to mention that if we assume each subproblem is solved by a direct method at every iterative step, then the method becomes a direct method in the single-subdomain case, that is, without decomposing the original domain; on the other hand, the method becomes a classic iterative method while choosing every individual element as a subdomain. Therefore, the method can be regarded as a bridge connecting direct methods and classic iterative methods under domain decomposition techniques. This implies that we might develop an optimal classic iterative method for solving the nonconforming finite element problems by using the method of this paper if every finite element is chosen as a subdomain.

The closely related works are [7, 8, 18]. In [7], a very similar nonoverlapping domain decomposition method but with a slightly different consistency condition on the artificial interfaces for partial differential problems is developed and analyzed. The method in [7] can also be regarded as a variant of the famous Lions method of [18], which uses the same kind of Robin-type transmission condition as [7] but different techniques for updating the transmission data on the interfaces. In [8], the method of [7] is applied into the nonconforming finite element problems, and the convergence analyses are considered. In particular, convergence rate estimates and numerical experiments are provided, which show that the convergence speed is dependent on the finite element mesh size and highly dependent on the transmission coefficient values. Other closely related work is [10, 11]. Després [10] applies the Lions method to the mixed hybrid finite element problems of the Helmholtz problems and Douglas et al. [12] apply the Lions method to mixed finite element problems by introducing a

Lagrange multiplier on the interfaces. We also refer to [1, 2, 3, 13, 14, 17, 18, 19, 20, 21, 22, 23] and the references therein for other nonoverlapping domain decomposition methods.

The outline of the paper is as follows. In section 1, some preliminaries are given. Transmission data and updating techniques on the artificial interfaces are discussed in section 2. In section 3, a nonoverlapping domain decomposition method for partial differential problems is developed and analyzed. Section 4 applies the method developed in section 3 for the nonconforming finite element problems. A nonoverlapping domain decomposition iterative procedure for solving the finite element problems is developed and analyzed. Then the convergence rate estimates are derived in section 5. The conclusions are extended to the unstructured finite element meshes in section 6. Finally, a short conclusion is presented.

1. Preliminaries. For the sake of simplicity in exposition, this paper considers the following second order elliptic problem:

$$\begin{cases} -\Delta u + \alpha(x)u = f & \text{in} \quad \Omega, \\ u = 0 & \text{on} \quad \partial\Omega, \end{cases} \tag{1.1}$$

where Ω is a bounded domain in $\mathbb{R}^N (N = 2, 3)$, $f \in L^2(\Omega)$, and $\alpha(x) \in L^\infty(\Omega)$, and is nonnegative. The problem (1.1) is a typical second order elliptic equation and also can model heat equations and wave equations by implicit difference discrete for time.

The weak formulation of (1.1) is to find $u \in H_0^1(\Omega)$ such that

$$a_\Omega(u, v) = (f, v)_\Omega \qquad \forall v \in H_0^1(\Omega), \tag{1.2}$$

where, and in the paper, for a domain $D \subset \mathbb{R}^N$, $(\cdot, \cdot)_D$ is an inner product of $L^2(D)$ and

$$a_D(u, v) = \int_D (\nabla u \cdot \nabla v + \alpha(x)uv)\, dx. \tag{1.3}$$

It is well known that problem (1.2) has a unique solution $u \in H_0^1(\Omega)$ (cf. [15]).

To describe finite element approximations for (1.2), we begin with the triangulation of Ω. Assume that $\mathcal{T}_h$ is a quasi-uniform and regular finite element triangulation of Ω and, for simplicity, $\overline{\Omega} = \cup_{\tau \in \mathcal{T}_h} \overline{\tau}$ (cf. [5]). Also, for the sake of simplicity in exposition, we here consider only the famous nonconforming Crouzeix–Raviart element (cf. [6]) on the n-simplex (triangle if $n = 2$, tetrahedron if $n = 3$) triangulation $\mathcal{T}_h$. However, it is not difficult to see that the analyses and conclusions of this paper can be easily extended to other nonconforming finite elements (for instance, the nonconforming finite elements for n-quadrilateral partition or n-simplex portion (cf. [12]) and the nonconforming finite elements for n-rectangle partition (cf. [16])). Let $S_h \subset L^2(\Omega)$ be the nonconforming Crouzeix–Raviart finite element space over the finite element mesh $\mathcal{T}_h$. Denoting N_h as the set of all face barycenters of $\mathcal{T}_h$'s element, i.e., n-simplex, in the interior of Ω and Γ_h as the set of all face barycenters of $\mathcal{T}_h$'s element on the boundary of $\partial\Omega$, we then define the finite element space S_h as follows (cf. [6]):

$$\begin{aligned} S_h = \{v :\ v|_\tau \in P_1(\tau),\ \tau \in \mathcal{T}_h,\ &v \textit{ continues at } p \in N_h \\ &\textit{and vanishes at } p \in \Gamma_h\}\,. \end{aligned} \tag{1.4}$$

Then the finite element approximate problem of (1.2) is to find $u \in S_h$ such that

$$a^h_\Omega(u,v) = (f,v)_\Omega \qquad \forall v \in S_h, \tag{1.5}$$

where, and in this paper, for a domain $D \subset \mathbb{R}^N$, a block (union) of elements of $\mathcal{T}_h$,

$$a^h_D(u,v) = \sum_{\tau\in\mathcal{T}_h, \tau\subset D} a_\tau(u,v). \tag{1.6}$$

It has been shown that the nonconforming finite element problem (1.5) has a unique solution, which has the optimal H^1 and L^2 errors and asymptotically optimal L^∞ error.

To develop a nonoverlapping domain decomposition method to solve the problem (1.2) and the nonconforming finite element problem (1.5), we decompose Ω into an arbitrary $m(\geq 2)$ of disjoint subdomains (open sets) $\Omega_1, \Omega_2, \dots, \Omega_m$; i.e., we assume that

$$\overline{\Omega} = \overline{\Omega}_1 \cup \overline{\Omega}_2 \cup \cdots \cup \overline{\Omega}_m = \Omega_1 \cup \Omega_2 \cup \cdots \cup \Omega_m \cup \Sigma, \tag{1.7}$$

$$\Sigma = \bigcup_{1\leq i\neq j\leq m} \gamma_{ij}, \qquad \gamma_{ij} = \partial\Omega_i \cap \partial\Omega_j, \qquad \Gamma_i = \partial\Omega_i \cap \partial\Omega. \tag{1.8}$$

Moreover, while considering the finite element problem (1.8), we also need to assume that the above nonoverlapping domain decomposition is aligned with $\mathcal{T}_h$; that is, every Ω_i is a block (union) of some elements or even an individual element of $\mathcal{T}_h$. Finally, we conclude this section by introducing the following notations G_r $(r = 1, 2, \dots)$, which are used in the convergence analyses:

$$\begin{cases} G_1 = \{\cup\Omega_k \mid \ \partial\Omega_k \cap \partial\Omega \ \text{ has positive measure}\}, \\ G_{r+1} = \{\cup\Omega_k \mid \ \partial\Omega_k \cap \overline{G_r} \ \text{ has positive measure}, \ \partial\Omega_k \cap G_l = \emptyset \ \forall l \leq r\}. \end{cases} \tag{1.9}$$

2. Transmission data and updating strategies. Let $u \in H^1_0(\Omega)$ be the solution of (1.1), and let $u_i = u\mid_{\Omega_i} \in H^1_{\Gamma_i}(\Omega_i)$, where, and in this paper, $H^1_{\Gamma_i}(\Omega_i)$ is a subspace of Sobolev space $H^1(\Omega_i)$ whose members vanish on Γ_i. Then u_i $(i = 1, 2, \dots, m)$ satisfies the following overdetermined subproblem:

$$\begin{cases} -\Delta u_i + \alpha(x)u_i = f & \text{in} \quad \Omega_i, \\ u_i = 0 & \text{on} \quad \Gamma_i, \\ u_i = u_j & \text{on} \quad \gamma_{ij}, \quad 1 \leq j \neq i \leq m, \\ \dfrac{\partial u_i}{\partial n_i} = -\dfrac{\partial u_j}{\partial n_j} & \text{on} \quad \gamma_{ij}, \quad 1 \leq j \neq i \leq m. \end{cases} \tag{2.1}$$

Conversely, let $u_i \in H^1_{\Gamma_i}(\Omega_i)$ $(i = 1, 2, \dots, m)$ be the solution of (2.1), and let u be a function defined over Ω satisfying $u\mid_{\Omega_i} = u_i$ $(i = 1, 2, \dots, m)$; then $u \in H^1_0(\Omega)$ is the solution of (1.1). This implies that a consistency condition on the artificial interface γ_{ij} of the problem (1.1) is

$$\begin{cases} u_i = u_j & \text{on} \quad \gamma_{ij}, \quad 1 \leq j \neq i \leq m, \\ \dfrac{\partial u_i}{\partial n_i} = -\dfrac{\partial u_j}{\partial n_j} & \text{on} \quad \gamma_{ij}, \quad 1 \leq j \neq i \leq m. \end{cases} \tag{2.2}$$

It is more convenient (cf. [7, 8, 10, 11, 18]) to replace (2.2) by the following Robin-type boundary condition on the artificial interface γ_{ij}:

$$\begin{cases} \dfrac{\partial u_i}{\partial n_i} + \lambda_{ij} u_i = -\dfrac{\partial u_j}{\partial n_j} + \lambda_{ij} u_j & \text{on} \quad \gamma_{ij}, \quad 1 \le j \ne i \le m, \\ \dfrac{\partial u_j}{\partial n_j} + \lambda_{ji} u_j = -\dfrac{\partial u_i}{\partial n_i} + \lambda_{ji} u_i & \text{on} \quad \gamma_{ij}, \quad 1 \le j \ne i \le m, \end{cases} \tag{2.3}$$

where $\lambda_{ij} = \lambda_{ji} > 0 \ \ (1 \le i \ne j \le m)$. Therefore, by using (2.3), a nonoverlapping domain decomposition iterative procedure is developed and analyzed (cf. [18]), which is based on the following subproblems:

$$\begin{cases} -\Delta u_i^{n+1} + \alpha(x) u_i^{n+1} = f & \text{in} \quad \Omega_i, \\ \dfrac{\partial u_i^{n+1}}{\partial n_i} + \lambda_{ij} u_i^{n+1} = -\dfrac{\partial u_j^n}{\partial n_j} + \lambda_{ij} u_j^n & \text{on} \ \ \gamma_{ij}, \ \ 1 \le j \le m, j \ne i, \\ u_i^{n+1} = 0 & \text{on} \quad \Gamma_i. \end{cases} \tag{2.4}$$

This procedure can be applied to general multisubdomain decompositions and implemented on parallel machines with local communications naturally. In the procedure, the transmission data on the interfaces are u_i^n and $\partial u_i^n / \partial n_i$, and they are updated by direct substitutions. In every iterate step, we have to find the extra data $\partial u_i^n / \partial n_i$ in order to update the transmission data, which is not easy, in particular, for discrete problems. Furthermore, this might cause ill-posed troubles of subproblems because of the regularities of the subproblems. This also makes the method difficult to apply to discrete problems (for instance, finite element problems (cf. [8, 10, 12])).

In order to avoid these disadvantages, a nonoverlapping domain decomposition iterative procedure is developed and analyzed by [7, 8], which is based on the following subproblems:

$$\begin{cases} -\Delta u_i^n + \alpha(x) u_i^n = f & \text{in} \quad \Omega_i, \\ \dfrac{\partial u_i^n}{\partial n_i} + \lambda_{ij} u_i^n = g_{ij}^n & \text{on} \ \ \gamma_{ij} \ \forall \ 1 \le j \le m, j \ne i, \\ u_i^n = 0 & \text{on} \quad \Gamma_i, \end{cases} \tag{2.5}$$

$$g_{ij}^{n+1} = 2\lambda_{ij} u_j^n - g_{ji}^n, \qquad 1 \le j \ne i \le m, \ \ \text{if } meas(\gamma_{ij}) > 0. \tag{2.6}$$

The precise meaning of (2.5) is nothing but the usual weak formulation, which is

(2.7)
$$a_{\Omega_i}(u_i^n, v) + \sum_{\substack{1 \le j \le m \\ j \ne i}} \lambda_{ij} \int_{\gamma_{ij}} u_i^n v ds = (f, v)_{\Omega_i} + \sum_{\substack{1 \le j \le m \\ j \ne i}} \int_{\gamma_{ij}} g_{ij}^n v ds \ \ \forall v \in H^1_{\Gamma_i}(\Omega_i),$$

and (2.6) is understood in $L^2(\Omega_i)$. This method also can be applied for general multisubdomain decompositions and implemented on parallel machines with local communications naturally. Like the above method, this method is also based on the consistency condition (2.3). However, the transmission data and the updating strategy of transmission data are different. In this procedure, the transmission data are g_{ij}^n, and the transmission data are updated by a derivative-free technique using the transmission data of the previous iterative step and the data obtained directly from the previous iterative step. Moreover, since there are no derivatives in (2.6) and (2.7) explicitly, we do not need to find any extra information (for instance, the first normal derivatives) in the iterative process. This guarantees that all subproblems are always well-posed in the iterative process if the initial subproblems are well-posed. This also makes the method easy to apply to the discrete problems (cf. [7, 8]). However, it follows from (2.6)–(2.7) that

$$\begin{aligned}\frac{\partial u_i^{n+1}}{\partial n_i} + \lambda_{ij} u_i^{n+1} &= g_{ij}^{n+1} \equiv 2\lambda_{ij} u_j^n - g_{ji}^n \\ &= 2\lambda_{ij} u_j^n - \left(\frac{\partial u_j^n}{\partial n_j} + \lambda_{ji} u_j^n\right) \\ &= -\frac{\partial u_j^n}{\partial n_j} + \lambda_{ij} u_j^n \qquad \text{on} \quad \gamma_{ij}.\end{aligned} \tag{2.8}$$

This means that this procedure is essentially equivalent to the above one (cf. [7, 8]). This procedure is applied to the nonconforming finite element problems and the convergence analyses, including convergence rate estimates, and numerical experiments are provided in [8]. The analyses have shown that the parameters λ_{ij}, called transmission coefficients, are highly effected on the convergence speed, but, whatever λ_{ij} are taken, it seems that the convergence speed is always dependent on the finite element mesh size in any case.

To improve this method and develop a method with the better convergence, we rewrite the consistency condition (2.2) by introducing extra parameter β_{ij}. It is not hard to check that (2.2) also can be replaced by the following condition on the artificial interfaces:

$$\begin{cases} \beta_{ij}\dfrac{\partial u_i}{\partial n_i} + \lambda_{ij} u_i = -\beta_{ij}\dfrac{\partial u_j}{\partial n_j} + \lambda_{ij} u_j \qquad \text{on} \quad \gamma_{ij}, \quad 1 \le i \ne j \le m, \\ \beta_{ji}\dfrac{\partial u_j}{\partial n_j} + \lambda_{ji} u_j = -\beta_{ji}\dfrac{\partial u_i}{\partial n_i} + \lambda_{ji} u_i \qquad \text{on} \quad \gamma_{ij} \quad 1 \le i \ne j \le m, \end{cases} \tag{2.9}$$

where

$$\begin{cases} \beta_{ij} = \beta_{ji} > 0, \quad 1 \le i \ne j \le m, \\ \lambda_{ij} = \lambda_{ji} > 0, \quad 1 \le i \ne j \le m. \end{cases} \tag{2.10}$$

By using the consistency condition (2.9), we can develop a nonoverlapping domain decomposition iterative procedure based on the following subproblems:

$$\begin{cases} -\Delta u_i^n + \alpha(x) u_i^n = f \qquad \text{in} \quad \Omega_i, \\ \beta_{ij}\dfrac{\partial u_i^n}{\partial n_i} + \lambda_{ij} u_i^n = g_{ij}^n \qquad \text{on} \quad \gamma_{ij} \ \forall\ 1 \le j \le m, j \ne i, \\ u_i^n = 0 \qquad \text{on} \quad \Gamma_i, \end{cases} \tag{2.11}$$

$$g_{ij}^{n+1} = 2\lambda_{ij} u_j^n - g_{ji}^n, \qquad 1 \le j \ne i \le m, \quad \text{if } meas(\gamma_{ij}) > 0. \tag{2.12}$$

Like (2.5)–(2.6), the subproblem (2.11) is understood as a usual weak formulation:

$$a_{\Omega_i}(u_i^n, v) + \sum_{\substack{1 \le j \le m \\ j \ne i}} \frac{\lambda_{ij}}{\beta_{ij}} \int_{\gamma_{ij}} u_i^n v ds = (f, v)_{\Omega_i} + \sum_{\substack{1 \le j \le m \\ j \ne i}} \frac{1}{\beta_{ij}} \int_{\gamma_{ij}} g_{ij}^n v ds \ \forall v \in H^1_{\Gamma_i}(\Omega_i), \tag{2.13}$$

and (2.12) is understood in $L^2(\Omega_i)$. It follows from (2.10)–(2.11) that

$$\begin{aligned}\beta_{ij}\frac{\partial u_i^{n+1}}{\partial n_i}+\lambda_{ij}u_i^{n+1} &= g_{ij}^{n+1}\equiv 2\lambda_{ij}u_j^n-g_{ji}^n\\ &= 2\lambda_{ij}u_j^n-\left(\beta_{ij}\frac{\partial u_j^n}{\partial n_j}+\lambda_{ji}u_j^n\right)\\ &= -\beta_{ij}\frac{\partial u_j^n}{\partial n_j}+\lambda_{ij}u_j^n \qquad \text{on } \gamma_{ij}.\end{aligned}\tag{2.14}$$

This means that the above method indeed implements the consistency condition (2.9) by an iterative process. Then, comparing (2.5)–(2.6) and (2.11)–(2.12), we have found that the only difference is at the second equalities of (2.5) and (2.11). That is, the second equality of (2.11) has an extra parameter β_{ij}. We call β_{ij} the penalty coefficient since it seems to introduce a penalty for artificial terms of the artificial interfaces of the weak formulation of (2.11) while comparing the weak formulation (2.7) and the weak formulation (2.13). We are going to show that this penalty coefficient can improve the convergence speed even for unstructured finite element meshes in the next sections.

3. A nonoverlapping domain decomposition method for (1.2).

Based on the discussions of the last section, we will develop and analyze a nonoverlapping domain decomposition method based on (2.11)–(2.13), that is, based on the Robin-type consistency condition (2.9) actually. The nonoverlapping domain decomposition iterative procedure is defined as follows.

Algorithm I.

(i) given $g_{ij}^0\in L^2(\gamma_{ij}),\ 1\le i\ne j\le m,\ meas(\gamma_{ij})>0$ arbitrarily;

(ii) then recursively find $u_i^n,\ \ i=1,2,\dots,m,$ by solving the subproblems

$$\begin{aligned}a_{\Omega_i}(u_i^n,v)&+\sum_{\substack{1\le j\le m\\ j\ne i}}\frac{\lambda_{ij}}{\beta_{ij}}\int_{\gamma_{ij}}u_i^n v ds\\ &= (f,v)_{\Omega_i}+\sum_{\substack{1\le j\le m\\ j\ne i}}\frac{1}{\beta_{ij}}\int_{\gamma_{ij}}g_{ij}^n v ds \quad \forall v\in H^1_{\Gamma_i}(\Omega_i);\end{aligned}\tag{3.1}$$

(iii) update the transmission condition data for $i=1,2,\dots,m,$

$$g_{ij}^{n+1}=2\lambda_{ij}u_j^n-g_{ji}^n,\qquad 1\le j\ne i\le m,\ \text{ if } meas(\gamma_{ij})>0,\tag{3.2}$$

where the penalty coefficient β_{ij} and the transmission coefficient λ_{ij} satisfy (2.10).

By similar arguments as those in (3.1)–(3.3) of [7], we now define the error at the iterative step n:

$$\varepsilon^n=(\varepsilon_i^n)_{1\le i\le m}=(u_i^n-u\,|_{\Omega_i})_{1\le i\le m}\in\prod_{i=1}^m H^1_{\Gamma_i}(\Omega_i).\tag{3.3}$$

Then we have

$$a_{\Omega_i}(\varepsilon_i^n,v)+\sum_{\substack{1\le j\le m\\ j\ne i}}\frac{\lambda_{ij}}{\beta_{ij}}\int_{\gamma_{ij}}\varepsilon_i^n v ds=\sum_{\substack{1\le j\le m\\ j\ne i}}\frac{1}{\beta_{ij}}\int_{\gamma_{ij}}g_{ij}^n v ds\quad \forall v\in H^1_{\Gamma_i}(\Omega_i),\tag{3.4}$$

$$g_{ij}^{n+1}=2\lambda_{ij}\varepsilon_j^n-g_{ji}^n\qquad \text{on }\ \gamma_{ij}\quad \forall\ \ 1\le j\le m,\ j\ne i,\tag{3.5}$$

where we have used g_{ij}^n to replace $g_{ij}^n-(\beta_{ij}\frac{\partial u}{\partial n_i}+\lambda_{ij}u)$.

LEMMA 3.1. *There then holds the following identity:*

$$|||g^{n+1}|||^2 = |||g^n|||^2 - 4\sum_{i=1}^{m} a_{\Omega_i}(\varepsilon_i^n, \varepsilon_i^n), \tag{3.6}$$

where

$$|||g^k|||^2 = \sum_{1\le i\ne j\le m} \frac{1}{\beta_{ij}\lambda_{ij}} \int_{\gamma_{ij}} |g_{ij}^k|^2 ds, \quad g^k = (g_{ij}^k)_{1\le i\ne j\le m}, \ k = n, n+1. \tag{3.7}$$

Proof. It follows from (3.4)–(3.5) that

$$\begin{aligned} |||g^{n+1}|||^2 &= \sum_{1\le i\ne j\le m} \frac{1}{\beta_{ij}\lambda_{ij}} \int_{\gamma_{ij}} |g_{ij}^{n+1}|^2 ds \\ &= \sum_{1\le i\ne j\le m} \frac{1}{\beta_{ij}\lambda_{ij}} \int_{\gamma_{ij}} |2\lambda_{ij}\varepsilon_j^n - g_{ji}^n|^2 ds \\ &= \sum_{1\le i\ne j\le m} \frac{1}{\beta_{ij}\lambda_{ij}} \int_{\gamma_{ij}} |g_{ji}^n|^2 ds - 4\sum_{j=1}^{m} \sum_{\substack{1\le i\le m \\ i\ne j}} \frac{1}{\beta_{ij}} \int_{\gamma_{ij}} (g_{ji}^n - \lambda_{ji}\varepsilon_j^n)\varepsilon_j^n ds \\ &= |||g^n|||^2 - 4\sum_{j=1}^{m} a_{\Omega_j}(\varepsilon_j^n, \varepsilon_j^n), \end{aligned}$$

where (2.10) has been used. Thus, (3.6) has been proved.

Moreover, from Lemma 3.1 and by an almost exact same argument as the proof of Theorem 3.2 of [7], we have the following convergence theorem for Algorithm I.

THEOREM 3.2. *Let* $u \in H_0^1(\Omega)$ *be the weak solution of* (1.2) *with reasonable regularities, and let* $u_i^n \in H_{\Gamma_i}^1(\Omega)$ $(i = 1, 2, \ldots, m)$ *be the weak solutions of* (3.1)–(3.2). *Then we have that, for any initial* $g_{ij}^0 \in L^2(\gamma_{ij})$,

$$\|u^n - u\|_{H^1} = \left(\sum_{i=1}^{m} \|u_i^n - u^n\|_{H^1(\Omega_i)}^2\right)^{1/2} \longrightarrow 0, \quad as \quad n \to \infty. \tag{3.8}$$

4. A nonoverlapping domain decomposition method for (1.5).

This section applies the method developed in the last section to the nonconforming finite element problem (1.5). A parallel nonoverlapping domain decomposition method for solving nonconforming finite element problem (1.5) is developed and analyzed. Following the approach to construct Algorithm I, we define the nonoverlapping domain decomposition iterative procedure over the aligned nonoverlapping domain decomposition (1.7)–(1.8) as follows.

ALGORITHM II.

(i) given $g_{ij}^0 \in S_h(\gamma_{ij}), \ 1 \le i \ne j \le m,$ if $meas(\gamma_{ij}) > 0$ arbitrarily;

(ii) then recursively find $u_i^n \in S_i \ (i = 1, 2, \ldots, m)$ by solving the subproblems

$$a_{\Omega_i}^h(u_i^n, v) + \sum_{\substack{1\le j\le m \\ j\ne i}} \frac{\lambda_{ij}}{\beta_{ij}} \int_{\gamma_{ij}}^{*} u_i^n v ds = (f, v)_{\Omega_i} + \sum_{\substack{1\le j\le m \\ j\ne i}} \frac{1}{\beta_{ij}} \int_{\gamma_{ij}}^{*} g_{ij}^n v ds \ \forall v \in S_i; \tag{4.1}$$

(iii) update data of the transmission condition on the interfaces

$$g_{ij}^{n+1}(p) = \ 2\lambda_{ij}u_j^n(p) - g_{ji}^n(p), \quad \text{on} \ \ p \in \gamma_{ij} \cap N_h, \ 1 \le j \ne i \le m, \tag{4.2}$$

where the penalty and transmission coefficients, β_{ij} and λ_{ij}, satisfy (2.10), $S_i = S_h \mid_{\Omega_i}$, and

$$S_h(\gamma_{ij}) = \left\{ v \in S_h \mid \; v \; = \sum_{p \in \gamma_{ij} \cap N_h} a_p \, \varphi_p \,, \quad a_p \in \mathbb{R} \right\}, \tag{4.3}$$

$$\int_{\gamma_{ij}}^{*} uvds = \sum_{p \in \gamma_{ij} \cap N_h} u(p) \, v(p) \, meas(s_p), \tag{4.4}$$

where, and in this paper, $\{\varphi_p\}_{p \in N_h}$ is the node basis of the finite element space S_h and s_p is the element face with p as its barycenter. (4.4) implies that we have used the numerical integration in (4.1) to compute the integrations on the interfaces.

We now consider the convergence of Algorithm II. Unlike the case for partial differential boundary value problems (in detail, cf. [7, 8]), we first have to give an equivalent splitting subproblem form with respect to the nonoverlapping domain decomposition for finite element problem (1.5) in order to prove the convergence of Algorithm II.

Theorem 4.1. *Let $u \in S_h$ be the solution of the finite element problem* (1.5) *and $u_i = u \mid_{\Omega_i}$. The problem* (1.5) *can be split into an equivalent splitting subproblem form. That is, there exist $g_{ij}^{*} \in S_h(\gamma_{ij})$, $1 \le i \ne j \le m$, $meas(\gamma_{ij}) > 0$, such that $u_i \in S_i$ $(i = 1, 2, \ldots, m)$ satisfies*

$$a_{\Omega_i}^h(u_i, v) + \sum_{\substack{1 \le j \le m \\ j \ne i}} \frac{\lambda_{ij}}{\beta_{ij}} \int_{\gamma_{ij}}^{*} u_i v ds = (f, v)_{\Omega_i} + \sum_{\substack{1 \le j \le m \\ j \ne i}} \frac{1}{\beta_{ij}} \int_{\gamma_{ij}}^{*} g_{ij}^{*} v ds \quad \forall v \in S_i. \tag{4.5}$$

Proof. Notice that $\{\varphi_p\}_{p \in N_h}$ is the node basis of the finite element space S_h. Then the problem (1.5) is equivalent to the following system:

$$a_{\Omega}^h(u, \varphi_p) = (f, \varphi_p)_{\Omega} \qquad \forall p \in N_h. \tag{4.6}$$

In particular, it follows from the small support property of $\varphi_p(x)$ that

$$a_{\Omega_i}^h(u, \varphi_p) - (f, \varphi_p)_{\Omega_i} = - \, [a_{\Omega_j}(u, \varphi_p) - (f, \varphi_p)_{\Omega_j}] \quad \forall p \in \gamma_{ij} \cap N_h. \tag{4.7}$$

For any γ_{ij}, any element face $s_p \subset \gamma_{ij}$ with p as its barycenter, we define G_{ij}^p as follows:

$$G_{ij}^p = - \frac{1}{meas(s_p)} [a_{\Omega_j}^h(u, \varphi_p) - (f, \varphi_p)_{\Omega_j}].$$

Therefore, we can construct $g_{ij}^{*} \in S_h(\gamma_{ij})$, $meas(\gamma_{ij}) > 0$ $(1 \le i \ne j \le m)$,

$$g_{ij}^{*}(x) = \sum_{p \in \gamma_{ij} \cap N_h} \left(\lambda_{ij} u(p) + \beta_{ij} G_{ij}^p \right) \, \varphi_p(x). \tag{4.8}$$

Thus, it follows from (4.7)–(4.8) that, for all $p \in \overline{\Omega}_i \cap N_h$, $u_i = u \mid_{\Omega_i}$ satisfies

$$a_{\Omega_i}^h(u_i, \varphi_p) + \sum_{\substack{1 \le j \le m \\ j \ne i}} \frac{\lambda_{ij}}{\beta_{ij}} \int_{\gamma_{ij}}^{*} u_i \varphi_p ds = (f, \varphi_p)_{\Omega_i} + \sum_{\substack{1 \le j \le m \\ j \ne i}} \frac{1}{\beta_{ij}} \int_{\gamma_{ij}}^{*} g_{ij}^{*} \varphi_p ds.$$

Clearly, this implies (4.5). The proof is completed.

Before discussing the convergence, we need to introduce similar notations to those in Algorithm I. Let u be the solution of the finite element problem (1.5), and let u_i^n $(1 \leq i \leq m)$ be the solutions of the subproblem (4.1) on the subdomain Ω_i at iterative step n. We then denote

$$u_i = u|_{\Omega_i}, \qquad u \ := \ (u_i)_{(1\leq i\leq m)} \in \prod_{i=1}^{m} S_i, \tag{4.9}$$

$$u^n = (u_i^n)_{(1\leq i\leq m)} \in \prod_{i=1}^{m} S_i, \qquad u^n|_{\Omega_i} \ := \ u_i^n, \tag{4.10}$$

$$e^n = (e_i^n)_{(1\leq i\leq m)} \ := \ (u_i^n - u_i) \in \prod_{i=1}^{m} S_i. \tag{4.11}$$

Clearly, e^n is indeed the error at iterative step n. Then we have

$$a_{\Omega_i}^h(e_i^n, v) + \sum_{\substack{1\leq j\leq m\\ j\neq i}} \frac{\lambda_{ij}}{\beta_{ij}} \int_{\gamma_{ij}}^{*} e_i^n v ds = \sum_{\substack{1\leq j\leq m\\ j\neq i}} \frac{1}{\beta_{ij}} \int_{\gamma_{ij}}^{*} g_{ij}^n v ds \ \forall v \in S_i, \tag{4.12}$$

$$g_{ij}^{n+1}(p) = \ 2\lambda_{ij} e_j^n(p) - g_{ji}^n(p) \quad \text{on} \ \ p \in \gamma_{ij} \cap N_h, \ \ 1 \leq j \neq i \leq m, \tag{4.13}$$

where we have used g_{ij}^n to replace $g_{ij}^n - g_{ij}^*$.

LEMMA 4.2. *We then have the following identity:*

$$a_{\Omega_i}^h(e_i^n, e_i^n) = \sum_{\substack{1\leq j\leq m\\ j\neq i}} \frac{1}{\beta_{ij}} \int_{\gamma_{ij}}^{*} (g_{ij}^n - \lambda_{ij} e_i^n) e_i^n ds. \tag{4.14}$$

LEMMA 4.3. *There then holds the following identity:*

$$|||g^{n+1}|||_*^2 = |||g^n|||_*^2 - 4 \sum_{i=1}^{m} a_{\Omega_i}^h(e_i^n, e_i^n), \tag{4.15}$$

where

$$|||g^k|||_*^2 = \sum_{1\leq i\neq j\leq m} \frac{1}{\beta_{ij}\lambda_{ij}} \int_{\gamma_{ij}}^{*} |g_{ij}^k|^2 ds, \quad g^k = (g_{ij}^k)_{1\leq i\neq j\leq m}, \ \ k = n, n+1. \tag{4.16}$$

Proof. It follows from (4.13) and (4.14) that

$$\begin{aligned} |||g^{n+1}|||_*^2 &= \sum_{1\leq i\neq j\leq m} \frac{1}{\beta_{ij}\lambda_{ij}} \int_{\gamma_{ij}}^{*} |g_{ij}^{n+1}|^2 ds \\ &= \sum_{1\leq i\neq j\leq m} \frac{1}{\beta_{ij}\lambda_{ij}} \int_{\gamma_{ij}}^{*} |2\lambda_{ij} e_j^n - g_{ji}^n|^2 ds \\ &= \sum_{1\leq i\neq j\leq m} \frac{1}{\beta_{ij}\lambda_{ij}} \int_{\gamma_{ij}}^{*} |g_{ji}^n|^2 ds - 4 \sum_{j=1}^{m} \sum_{\substack{1\leq i\leq m\\ i\neq j}} \frac{1}{\beta_{ij}} \int_{\gamma_{ij}}^{*} (g_{ji}^n - \lambda_{ji} e_j^n) e_j^n ds \\ &= |||g^n|||_*^2 - 4 \sum_{j=1}^{m} a_{\Omega_j}^h(e_j^n, e_j^n), \end{aligned}$$

where (2.10) has been used. Thus, (4.15) has been proved.

THEOREM 4.4. *Let $u \in S_h$ be the solution of the nonconforming finite element problem* (1.5), *and let $u_i^n \in S_i$ $(i = 1, 2, \ldots, m)$ be the solutions of subproblem* (4.1)–(4.2) *at iterative step n. Then we have*

$$\|u^n - u\|_h \equiv \left(\sum_{i=1}^{m} \sum_{\tau \in \mathcal{T}_h, \tau \subset \Omega_i} \|u_i^n - u\|_{H^1(\tau)}^2 \right)^{1/2} \longrightarrow 0, \quad \textit{as } n \to \infty. \tag{4.17}$$

Proof. Clearly, we indeed need to show that

$$\left(\sum_{i=1}^{m} \|e_i^n\|_{h(\Omega_i)}^2 \right)^{1/2} \equiv \left(\sum_{i=1}^{m} \sum_{\tau \in \mathcal{T}_h, \tau \subset \Omega_i} \|e_i^n\|_{H^1(\tau)}^2 \right)^{1/2} \longrightarrow 0, \quad \textit{as } n \to \infty. \tag{4.18}$$

By using Lemma 4.2, we have that for any positive integer M

$$\sum_{n=0}^{M} \left(\sum_{i=1}^{m} a_{\Omega_i}^h(e_i^n, e_i^n) \right) = \frac{1}{4}(|||g^0|||_*^2 - |||g^{M+1}|||_*^2) \geq 0. \tag{4.19}$$

This implies that

$$a_{\Omega_i}^h(e_i^n, e_i^n) \longrightarrow 0, \quad \textit{as} \quad n \to \infty, \qquad i = 1, 2, \ldots, m. \tag{4.20}$$

Therefore, if $\alpha(x) \geq \alpha_0 > 0$, one then obtains (4.17) from (4.20) immediately . We now consider the general case. Since e_i^n vanishes at the node points of $\partial\Gamma_i$, it is then easy to check that $(a_{\Omega_i}^h(\cdot, \cdot))^{1/2}$ is a norm on S_i for $\Omega_i \subset G_1$. It follows from equivalence of norms in S_i that

$$\|e_i^n\|_{h(\Omega_i)} \longrightarrow 0, \qquad \textit{as} \ \ n \to \infty \qquad \forall\, \Omega_i \subset G_1. \tag{4.21}$$

Clearly, (4.21) implies that

$$\int_{\gamma_{ij}}^{*} |e_i^n|^2 ds \longrightarrow 0, \quad \textit{as} \ \ n \to \infty \ \forall\, \Omega_i \subset G_1, \quad j \neq i. \tag{4.22}$$

Moreover, by using (4.12), (4.20), and (4.22), we can have that

$$\int_{\gamma_{ij}}^{*} |g_{ij}^n|^2 ds \longrightarrow 0, \quad \textit{as} \ \ n \to \infty \ \forall\, \Omega_i \subset G_1, \quad j \neq i. \tag{4.23}$$

Therefore, it follows from (4.13), (4.22), and (4.23) that

$$\int_{\gamma_{ij}}^{*} |g_{ij}^n|^2 ds \longrightarrow 0, \quad \textit{as} \ \ n \to \infty \ \forall\, \Omega_i \subset G_2, \ \Omega_j \subset G_1, \tag{4.24}$$

$$\int_{\gamma_{ij}}^{*} |e_i^n|^2 ds \longrightarrow 0, \quad \textit{as } n \to \infty \ \ \forall\, \Omega_i \subset G_2, \ \Omega_j \subset G_1. \tag{4.25}$$

Notice that, for $\Omega_i \subset G_2$,

$$\left(a_{\Omega_i}(\cdot, \cdot) + \sum_{\Omega_j \subset G_1} \int_{\gamma_{ij}}^{*} |\cdot|^2 ds \right)^{1/2} \tag{4.26}$$

is a norm on S_i. Hence, (4.20), (4.25), and (4.26) imply that

$$\|e_i^n\|_{h(\Omega_i)} \longrightarrow 0, \qquad as \;\; n \to \infty \quad \forall\, \Omega_i \subset G_2. \tag{4.27}$$

Similarly, we can show that, for $r \geq 3$,

$$\|e_i^n\|_{h(\Omega_i)} \longrightarrow 0, \quad as \;\; n \to \infty \quad \forall\, \Omega_i \subset G_r. \tag{4.28}$$

Since the number of G_r is less than m, we have proved (4.18). This finishes the proof.

5. Convergence rate estimates.

This section discusses the convergence rate of Algorithm II for quasi-uniform partition $\mathcal{T}_h$. Let

$$W = \prod_{i=1}^{m} S_i, \qquad \Lambda = \prod_{\substack{1 \leq i \neq j \leq m \\ meas(\gamma_{ij})>0}} S_h(\gamma_{ij}). \tag{5.1}$$

Also, let $A_f : W \times \Lambda \longmapsto W \times \Lambda$ be an affined mapping defined by (4.1)–(4.2) of Algorithm II. That is,

$$[u^{n+1}, g^{n+1}] = A_f[u^n, g^n], \tag{5.2}$$

where $[u^{n+1}, g^{n+1}]$ satisfies (3.1)–(3.2). We then have the following lemma.

LEMMA 5.1. *Let $u \in S_h$ be the solution of the problem* (1.5), *and let $u \equiv (u \mid_{\Omega_i}) \in W$. Then there exists $g \in \Lambda$ such that $[u, g]$ is a fixed point of A_f. Conversely, let $[u, g] \in W$ be a fixed point of A_f and $u \in L^2(\Omega)$ satisfying $u \mid_{\Omega_i} = u_i$. Then $u \in S_h$ is the solution of the problem* (1.5).

Proof. If $u \in S_h$ is the solution of (1.5), we have that, by taking $g = (g_{ij}^*) \in \Lambda$ in (4.5),

$$a_{\Omega_i}^h(u_i, v) + \sum_{\substack{1 \leq j \leq m \\ j \neq i}} \frac{\lambda_{ij}}{\beta_{ij}} \int_{\gamma_{ij}}^* u_i v ds = (f, v)_{\Omega_i} + \sum_{1 \leq j \leq m \top j \neq i} \frac{1}{\beta_{ij}} \int_{\gamma_{ij}}^* g_{ij} v ds \;\; \forall v \in S_i. \tag{5.3}$$

From the definition of g_{ij}^* of Lemma 5.1, it is easy to check that

$$g_{ij}(p) = \; 2\lambda_{ij} u(p) - g_{ji}(p) \;\; \forall p \in \gamma_{ij} \cap N_h, \;\; 1 \leq j \neq i \leq m, \; meas(\gamma_{ij}) > 0. \tag{5.4}$$

Therefore, (5.3) and (5.4) immediately imply that $[u, g]$ is a fixed point of A_f.

Conversely, if $[u, g]$ is a fixed point of A_f, that is, $[u, g]$ satisfies (5.3)–(5.4), it follows from (5.3)–(5.4) that by some direct computations

$$u_i(p) = u_j(p) \quad \forall\, p \in \gamma_{ij} \cap N_h, \;\; 1 \leq j \neq i \leq m, \;\; meas(\gamma_{ij}) > 0, \tag{5.5}$$

$$\sum_{i=1}^{m} a_{\Omega_i}^h(u, v) = \sum_{i=1}^{m} (f, v)_{\Omega_i} \;\; \forall\, v \in S_h. \tag{5.6}$$

It is not difficult to see from (5.5) and (5.6) that $u \in S_h$, and u is the solution of (1.5). The proof is completed.

Furthermore, if we let $A = A_f \mid_{f=0}$, $F = A_f(0, 0)$, then A is a linear mapping, indeed, which is the iterator (iterative matrix) of Algorithm II, and satisfies

$$A_f[u, g] = A[u, g] + F, \tag{5.7}$$

$$[e_i^{n+1}, g^{n+1}] = A[e_i^n, g^n]. \tag{5.8}$$

THEOREM 5.2. *Let $\rho(A)$ be the spectral radius of A. Then we have*

$$\rho(A) < 1. \tag{5.9}$$

Proof. Let μ be an eigenvalue of A, and let $[e, g] \neq [0, 0]$ be its corresponding eigenvector. Then we have that $A[e, g] = \mu[e, g]$, that is,

$$a^h_{\Omega_i}(e_i, v) + \sum_{\substack{1\le j\le m\\ j\ne i}} \frac{\lambda_{ij}}{\beta_{ij}} \int^*_{\gamma_{ij}} e_i v ds = \sum_{\substack{1\le j\le m\\ j\ne i}} \frac{1}{\beta_{ij}} \int^*_{\gamma_{ij}} g_{ij} v ds \quad \forall v \in S_i, \tag{5.10}$$

$$\mu g_{ij}(p) = \ 2\lambda_{ij} e_j(p) - g_{ji}(p) \quad \forall\ p \in \gamma_{ij} \cap N_h, \quad 1 \le j \neq i \le m. \tag{5.11}$$

Therefore, these, together with Theorem 4.1, imply that

$$\mu^2 |||g|||^2_* = |||g|||^2_* - 4 \sum_{i=1}^m a^h_{\Omega_i}(e_i, e_i). \tag{5.12}$$

This means that $|\mu| \le 1$ and $|\mu| = 1$ if and only if

$$a^h_{\Omega_i}(e_i, e_i) = 0 \qquad \forall\ i = 1, 2, \ldots, m. \tag{5.13}$$

We next show that $|\mu| < 1$, that is, $|\mu| \neq 1$. If $|\mu| = 1$, then (4.13) implies that each e_i is a constant over Ω_i. We then have that since e_i vanishes at nodal points on $\partial\Omega_i \cap \partial\Omega$

$$e_i = 0 \qquad \text{in } \ \Omega_i \qquad \forall\ \Omega_i \subset G_1. \tag{5.14}$$

Moreover, from (5.10) and (5.11), we can obtain

$$g_{ij} = 0 \qquad \text{on} \quad \gamma_{ij} \qquad \forall\ \Omega_i \subset G_1,\ j \neq i, \tag{5.15}$$

$$g_{ij} = 0 \qquad \text{on} \quad \gamma_{ij} \qquad \forall\ \Omega_i \subset G_1,\ \Omega_j \subset G_2, \tag{5.16}$$

$$e_i = 0 \qquad \text{on} \quad \gamma_{ij} \qquad \forall\ \Omega_i \subset G_1,\ \Omega_j \subset G_2. \tag{5.17}$$

Then (5.13)–(5.16) imply that

$$e_i = 0 \qquad \text{in } \ \Omega_i \qquad \forall\ \Omega_i \subset G_2. \tag{5.18}$$

Similarly, we can consider other $\Omega_i \subset G_r (r \ge 3)$. Therefore, we have shown that

$$e_i = 0 \qquad \text{in } \ \Omega_i \quad \forall\ i = 1, 2, \ldots, m, \tag{5.19}$$

$$g_{ij} = 0 \qquad \text{on } \ \gamma_{ij} \quad \forall\ 1 \le i \neq j \le m,\ meas(\gamma_{ij}) > 0, \tag{5.20}$$

that is, $[e, g] = [0, 0]$. This is a contradiction. Hence, $|\mu| < 1$. Therefore, (5.9) is proved.

THEOREM 5.3. *Assume that $\alpha(x) \ge \alpha_0 > 0$, and*

$$\begin{cases} \beta_{ij} = \beta_{ji} = \beta > 0 & \forall\ 1 \le i \neq j \le m,\ meas(\gamma_{ij}) > 0, \\ \lambda_{ij} = \lambda_{ji} = \lambda > 0 & \forall\ 1 \le i \neq j \le m,\ meas(\gamma_{ij}) > 0. \end{cases} \tag{5.21}$$

We then have the following convergence rate estimates for Algorithm II:

$$|||g^{n+1}|||_* \le \left(1 - \frac{4}{Ch^{-1}(\lambda^{-1} + \beta^{-1}\lambda) + 2}\right)^{1/2} |||g^n|||_*, \tag{5.22}$$

$$\|u^n - u\|_h \le C(\alpha) \left(1 - \frac{4}{Ch^{-1}(\lambda^{-1} + \beta^{-1}\lambda) + 2}\right)^{n/2} |||g^0|||_*, \tag{5.23}$$

where C is a constant independent of finite element mesh size h.

Proof. It follows from (4.9)–(4.14) and (5.21) that

$$\begin{aligned}
&\sum_{j\neq i} \frac{1}{\beta_{ij}\lambda_{ij}} \int_{\gamma_{ij}}^* |g_{ij}^n|^2 ds \\
&= \sum_{j\neq i} \frac{1}{\beta_{ij}\lambda_{ij}} \int_{\gamma_{ij}}^* |(g_{ij}^n - \lambda_{ij}e_i^n) + \lambda_{ij}e_i^n|^2 ds \\
&= \sum_{j\neq i} \left(\frac{1}{\beta_{ij}\lambda_{ij}} \int_{\gamma_{ij}}^* |g_{ij}^n - \lambda_{ij}e_i^n|^2 ds + \frac{\lambda_{ij}}{\beta_{ij}} \int_{\gamma_{ij}}^* |e_i^n|^2 ds \right. \\
&\qquad \left. + \frac{2}{\beta_{ij}} \int_{\gamma_{ij}}^* (g_{ij}^n - \lambda_{ij}e_i^n) e_i^n ds \right) \\
&= \sum_{j\neq i} \left(\frac{1}{\beta_{ij}\lambda_{ij}} \int_{\gamma_{ij}}^* |g_{ij}^n - \lambda_{ij}e_i^n|^2 ds + \frac{\lambda_{ij}}{\beta_{ij}} \int_{\gamma_{ij}}^* |e_i^n|^2 ds \right) \\
&\qquad + 2\, a_{\Omega_i}^h(e_i^n, e_i^n).
\end{aligned} \tag{5.24}$$

Taking $v \in S_i$ such that

$$v(p) = \begin{cases} g_{ij}^n(p) - \lambda_{ij}e_i^n(p) & \text{at } \; p \in \gamma_{ij} \cap N_h, \; j \neq i, \\ 0 & \text{at } \; p \in N_h \backslash \gamma_{ij}, \; j \neq i, \end{cases}$$

and plugging v into (4.9), we have

$$\begin{aligned}
&\sum_{j\neq i} \frac{1}{\beta_{ij}} \int_{\gamma_{ij}}^* |g_{ij}^n - \lambda_{ij}e_i^n|^2 ds = \sum_{j\neq i} \frac{1}{\beta_{ij}} \int_{\gamma_{ij}}^* (g_{ij}^n - \lambda_{ij}e_i^n) v ds \\
&\quad = a_{\Omega_i}^h(e_i^n, v) \le \left(a_{\Omega_i}^h(e_i^n, e_i^n)\right)^{1/2} \left(a_{\Omega_i}^h(v, v)\right)^{1/2}.
\end{aligned}$$

On the other hand, it follows from some direct calculations that

$$a_{\Omega_i}^h(v, v) \le Ch^{-1} \sum_{j\neq i} \int_{\gamma_{ij}}^* |v|^2 ds. \tag{5.25}$$

Then we have that

$$\sum_{j\neq i} \frac{1}{\beta_{ij}} \int_{\gamma_{ij}}^* |g_{ij}^n - \lambda_{ij}e_i^n|^2 ds \le Ch^{-1} a_{\Omega_i}^h(e_i^n, e_i^n). \tag{5.26}$$

It also follows from $\alpha(x) > \alpha_0 > 0$ and some direct calculations that

$$\sum_{j\neq i}\int_{\gamma_{ij}}^{*} |e_i^n|^2 ds \leq Ch^{-1}(e_i^n, e_i^n)_{\Omega_i} \leq Ch^{-1}a_{\Omega_i}^h(e_i^n, e_i^n). \tag{5.27}$$

Therefore, using (5.21) and (5.24)–(5.27) and summing for $i = 1, 2, \ldots, m,$ we have

$$|||g_{ij}^n|||_*^2 \leq (Ch^{-1}(\lambda^{-1} + \beta^{-1}\lambda) + 2)\sum_{i=1}^{m} a_{\Omega_i}^h(e_i^n, e_i^n). \tag{5.28}$$

Thus, from (5.28) and Lemma 4.2, we have

$$|||g_{ij}^{n+1}|||_*^2 \leq \left(1 - \frac{4}{Ch^{-1}(\lambda^{-1} + \beta^{-1}\lambda) + 2}\right)|||g_{ij}^n|||_*^2, \tag{5.29}$$

$$\begin{aligned} \|u^n - u\|_h^2 &\leq (1 + \|\alpha\|_{L^\infty(\Omega)})\sum_{i=1}^{m} a_{\Omega_i}^h(e_i^n, e_i^n) \\ &\leq \frac{1}{4}(1 + \|\alpha\|_{L^\infty(\Omega)})|||g_{ij}^n|||_*^2. \end{aligned} \tag{5.30}$$

Finally, (5.29) and (5.30) imply (5.22) and (5.23) immediately. The proof is completed.

COROLLARY 5.4. *With the assumption of Theorems* 5.2 *and* 5.3*, we then have*

$$\rho(A) \leq \left(1 - \frac{4}{Ch^{-1}(\lambda^{-1} + \beta^{-1}\lambda) + 2}\right)^{1/2}. \tag{5.31}$$

From (5.22)–(5.23) of Theorem 5.3 and (5.31) of Corollary 5.4, it is not hard to see that if we take

$$\lambda = O(h^{-1}) \qquad \text{and} \qquad \beta = O(h^{-2}), \tag{5.32}$$

then there exists a positive constant δ independent of h such that

$$\left(1 - \frac{4}{Ch^{-1}(\lambda^{-1} + \beta^{-1}\lambda) + 2}\right)^{1/2} \leq 1 - \delta. \tag{5.33}$$

This means that the method has an optimal convergence rate if (5.32) holds.

COROLLARY 5.5. *With the assumption of Theorems* 5.2 *and* 5.3*, if* (5.32) *holds, there then exists a positive constant* δ *independent of finite element mesh size* h *such that*

$$|||g^{n+1}|||_* \leq (1 - \delta)\,|||g^n|||_*, \tag{5.34}$$

$$\|u^n - u\|_h \leq C(\alpha)\,(1 - \delta)^n\,|||g^0|||_*, \tag{5.35}$$

$$\rho(A) \leq (1 - \delta)\,. \tag{5.36}$$

Notice that Algorithm II can become a classic iterative method of linear systems if every individual element of $\mathcal{T}_h$ is chosen as a subdomain. Therefore, Corollary 5.5 implies that Algorithm II indeed constructs a classic iterative method with a contracted number independent of finite element mesh size h for solving the nonconforming finite element problem (1.5). That is, an optimal iterative method can be constructed by Algorithm II.

6. Notes on unstructured finite element meshes. This section extends the conclusions of Algorithm II developed in the last two sections to the unstructured finite element meshes (cf. [4, 9]). This means that the finite element triangulation $\mathcal{T}_h$ is not a quasi-uniform finite element partition. First, we consider Theorem 4.4 and Theorem 5.2, which are about the convergence of Algorithm II. It is not hard to check that the quasi-uniform and regular properties of the finite element partition $\mathcal{T}_h$ have never been used in the proof of Theorem 4.4 and Theorem 5.2. This means that Theorem 4.4 and Theorem 5.2 still hold even for $\mathcal{T}_h$ without either quasi-uniform or regular properties.

THEOREM 6.1. *Assume that the finite element partition $\mathcal{T}_h$ is not required to be either a quasi-uniform or a regular property. Then the conclusions of Theorem* 4.4 *and Theorem* 5.2 *still hold. That is,*

$$\|u^n - u\|_h \equiv \left(\sum_{i=1}^{m} \sum_{\tau \in \mathcal{T}_h, \tau \subset \Omega_i} \|u_i^n - u\|_{H^1(\tau)}^2 \right)^{1/2} \longrightarrow 0, \quad \textit{as } n \to \infty, \tag{6.1}$$

$$\rho(A) < 1, \tag{6.2}$$

where all of the undefined notations are the same as those in Theorem 4.4 *and Theorem* 5.2.

We then consider the conclusions of the convergence rates, which include Theorem 5.3, Corollary 5.4, and Corollary 5.5. After carefully checking their proofs, we have found that the quasi-uniform and regular properties of the finite element partition $\mathcal{T}_h$ have been used in three formulas (5.25)–(5.27). In fact, not the global quasi-uniform and regular properties of $\mathcal{T}_h$ but only the local quasi-uniform and regular properties on the subdomain Ω_i have been used in (5.25)–(5.27).

Let $\mathcal{T}_{h_i}$ be the subpartition of $\mathcal{T}_h$ on Ω_i, where h_i is the mesh size of $\mathcal{T}_{h_i}$. In other words, $\mathcal{T}_{h_i}$ is the restriction of $\mathcal{T}_h$ on Ω_i. Furthermore, we assume that every subpartition $\mathcal{T}_{h_i}$ is quasi-uniform. Therefore, we have that (5.25)–(5.27) still hold. Hence, we have that

$$\sum_{j \neq i} \int_{\gamma_{ij}}^{*} |g_{ij}^n - \lambda_{ij} e_i^n|^2 ds \le C h_i^{-1} a_{\Omega_i}^h(e_i^n, e_i^n), \tag{6.3}$$

$$\sum_{j \neq i} \int_{\gamma_{ij}}^{*} |e_i^n|^2 ds \le C h^{-1} (e_i^n, e_i^n)_{\Omega_i} \le C h_i^{-1} a_{\Omega_i}^h(e_i^n, e_i^n). \tag{6.4}$$

Therefore,

$$|||g_{ij}^n|||_*^2 \le \max_{1 \le i \neq j \le m} \left[C h_i^{-1} (\lambda_{ij}^{-1} + \beta_{ij}^{-1} \lambda_{ij}) + 2 \right] \sum_{i=1}^{m} a_{\Omega_i}^h(e_i^n, e_i^n). \tag{6.5}$$

Assume that f_{ij} is any fixed element face on the interface γ_{ij} with $meas(\gamma_{ij}) > 0$. Let

$$\begin{cases} \lambda_{ij} = \lambda_{ji} = O((m(fi_{ij}))^{-1}), & 1 \le i \neq j \le m, \\ \beta_{ij} = \beta_{ji} = O((m(f_{ij}))^{-2}), & 1 \le i \neq j \le m, \end{cases} \tag{6.6}$$

where

$$m(f_{ij}) = \begin{cases} meas(f_{ij}) & \text{if } N = 2, \\ \sqrt{meas(f_{ij})} & \text{if } N = 3. \end{cases} \tag{6.7}$$

Notice that the sub-partition $\mathcal{T}_{h_i}$ is quasi-uniform. Then, $\lambda_{ij} = O(h_i^{-1})$ and $\beta_{ij} = O(h_i^{-2})$. Thus, if (6.6) holds, there then is a positive constant δ_0 independent of h_i such that

$$4 < Ch_i^{-1}(\lambda_{ij}^{-1} + \beta_{ij}^{-1}\lambda_{ij}) + 2 < \delta_0, \quad 1 \le i \ne j \le m.$$

This implies that there is a positive constant δ independent of h_i such that

$$\left(1 - \frac{4}{\max_{1\le i\ne j\le m}[(Ch_i^{-1}(\lambda_{ij}^{-1} + \beta_{ij}^{-1}\lambda_{ij}) + 2]}\right)^{1/2} \le 1 - \delta. \tag{6.8}$$

Therefore, in view of the proof of Theorem 4.4 and (6.3)–(6.8), we have the following convergence rate estimates for the unstructured finite element mesh $\mathcal{T}_h$.

THEOREM 6.2. *Assume that the finite element partition $\mathcal{T}_h$ is piecewise quasi-uniform. This means that $\mathcal{T}_h$ is an unstructured mesh, but each of its subpartitions $\mathcal{T}_{h_i}$ is a quasi-uniform mesh. Let β_{ij} and λ_{ij} satisfy* (6.6). *Then there is a positive constant δ independent of h_i $(i = 1, 2, \dots, m)$ such that*

$$|||g^{n+1}|||_* \le (1-\delta)\, |||g^n|||_*, \tag{6.9}$$

$$\|u^n - u\|_h \le C(\alpha)\,(1-\delta)^n\, |||g^0|||_*, \tag{6.10}$$

$$\rho(A) \le (1-\delta)\,, \tag{6.11}$$

where all of the undefined notations are the same as those in Corollary 5.5.

Finally, we consider a very special case, in which every individual element of $\mathcal{T}_h$ is chosen as a subdomain. For this special case, in the last section we mentioned that Algorithm II indeed produces a classic iterative method with optimal convergence if $\mathcal{T}_h$ is quasi-uniform. We now show that this conclusion is still true even for the unstructured finite element mesh $\mathcal{T}_h$, which is not quasi-uniform but regular. It is not very difficult to see that (5.25)–(5.27) hold with the same constant C for all of the subdomain Ω_i that is an individual element of $\mathcal{T}_h$ if $\mathcal{T}_h$ is regular. Also, notice that, in this special case, the element face f_{ij} of (6.6) is just the interface γ_{ij}. Thus, (6.7) can be rewritten as

$$\begin{cases} \lambda_{ij} = \lambda_{ji} = O((m(\gamma_{ij}))^{-1}), & 1 \le i \ne j \le m, \\ \beta_{ij} = \beta_{ji} = O((m(\gamma_{ij}))^{-2}), & 1 \le i \ne j \le m, \end{cases} \tag{6.12}$$

where $m(\gamma_{ij})$ is defined as $m(f_{ij})$ in (6.6). Therefore, (6.8) still holds if λ_{ij} and β_{ij} satisfy (6.12). Thus, we have the following theorem.

THEOREM 6.3. *Let $\mathcal{T}_h$ be not a quasi-uniform but a regular finite element triangulation. Assume that every individual element of $\mathcal{T}_h$ is chosen as a subdomain. If β_{ij} and λ_{ij} satisfy* (6.12), *then Algorithm* II *produces a classic iterative method with the optimal convergence rate for the linear system* (1.5). *That is, there is a positive constant δ independent of the mesh size of $\mathcal{T}_h$ such that*

$$\rho(A) \le (1-\delta)\,, \tag{6.13}$$

where, as in Corollary 5.5, *A is the iterative matrix and $\rho(A)$ is the spectral radius of A.*

7. Conclusion. In this paper, a parallel nonoverlapping domain decomposition iterative procedure based on a Robin-type consistency condition on the artificial interfaces is developed and analyzed. The key difference between the method of this paper and other similar methods is that an extra parameter β_{ij}, called penalty coefficient, is introduced. For partial differential problems, it can be regarded as a variant of the famous Lions method of [18] as well as the method of [7, 8]. However, unlike the Lions method, it does not have the regularity trouble of subproblems. For finite element problems, the weak formulation of this method can be formally derived from the weak formulation of the method in [7, 8] by introducing a penalty coefficient β_{ij} into the terms of artificial interfaces. Hence, as explained in the introduction, this method, like the method of [7, 8], can be regarded as a bridge connecting direct methods and iterative methods of linear systems in the sense of parallel algorithms. More importantly, because of introducing the penalty coefficient β_{ij}, this method could reach the optimal convergence rate even for unstructured meshes while choosing the right parameters, but the method of [7, 8] could not. Furthermore, the idea of this paper can be extended and applied to other problems and other discrete methods (for instance, mixed finite element methods, nonsymmetric and noncoercive scalar wave problems, etc.). We will consider some of these topics as well as the numerical experiments and analyses including comparisons to the other similar methods in future work. Finally, we conclude this paper with the following table of the results of two simple examples with $\Omega = [0, 1]^2$ and exact solution $u = \sin(\pi\, x)\, \sin(\pi\, y)$. The numerical experiments are implemented on a group of Sun Workstations (440 MHz CPU) by using MPI to implement the communications. In the numerical experiments, $\mathcal{T}_h$ is a three-fixed direction uniform triangulation with mesh size (leg's length) h; the domain decomposition is a uniform decomposition, in which all subdomains are congruent small squares; each subdomain (subproblem) is assigned to its own processor (machine), and no processor (machine) takes care of more than one subdomain (subproblem); the stop criterion is 10^{-5}; $\lambda_{ij} = h^{-1}$ and $\beta_{ij} = h^{-2}$.

h	$\alpha(x)$	Subdomains	Iterations	$\|u^n - u^{n-1}\|_\infty$	$\|u - u^n\|_\infty$	CPU time
.0125	1	16	27	9.9930E-6	6.5241E-4	4.93
.0100	1	25	26	6.4735E-6	5.4508E-4	4.81
.0125	0	16	28	4.8359E-6	7.1063E-4	5.05
.0100	0	25	26	8.5771E-6	6.5489E-4	4.76

Acknowledgements. The author would like to thank Dr. X. Feng and Dr. J. Xu for their helpful discussions, and two referees for their valuable comments.

REFERENCES

[1] V. I. Agoshkov, *Poincaré–Steklov's operators and domain decomposition method in finite dimensional spaces*, in First International Symposium of Domain Decomposition Methods for Partial Differential Equations, R. Glowinski, G. H. Golub, G. A. Meurant, and J. Periaux, eds., SIAM, Philadelphia, 1988, pp. 73–112.

[2] P. E. Bjørstadt and O. B. Widlund, *Iterative methods for the solution of elliptic problems on regions partitioned into substructures*, SIAM J. Numer. Anal., 23 (1986), pp. 1097–1120.

[3] J. Bramble, J. Pasciak, and G. Schatz, *An iterative method for elliptic problems on regions partitioned into substructures*, Math. Comp., 46 (1986), pp. 361–369.

[4] T. F. Chan, B. Smith, and J. Zou, *Overlapping Schwarz methods on unstructured meshes using nonmatching coarse grids*, Numer. Math., 72 (1996), pp. 149–167.

[5] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North–Holland, Amsterdam, 1978.

[6] M. Crouzeix and P.A. Raviart, *Conforming and nonconforming finite element methods for solving the stationary Stokes equations* I, Rev. Française Automat. Informat. Recherche Opérationnelle Sér. Rouge, 7 (1973), pp. 33–75.

[7] Q. Deng, *Timely communication: An analysis for a nonoverlapping domain decompositon iterative procedure*, SIAM J. Sci. Comput., 18 (1997), pp. 1517–1525.

[8] Q. Deng, *A nonoverlapping domain decomposition method for nonconforming finite element problems*, IAMJNA, submitted.

[9] Q. Deng and X. Feng, *Two-Level Overlapping Schwarz Methods for Plate Elements on Unstructured Meshes Using Non-Matching Grids*, CRM Proc. Lecture Notes 21, AMS, Providence, RI, 1999, pp. 160–170.

[10] B. Després, *Domain decomposition method and the Helmholtz problem*, in Mathematical and Numerical Aspects of Wave Propagation Phenomena, G. Cohen, L. Halpern, and P. Joly, eds., SIAM, Philadelphia, 1991, pp. 44–52.

[11] J. Douglas, Jr., P. Paes Leme, J. E. Roberts, and J. Wang, *A parallel iterative procedure applicable to the approximate solution of second order partial differential equations by mixed finite element methods*, Numer. Math., 65 (1993), pp. 95–108.

[12] J. Douglas, Jr., J. E. Santos, D. Sheen, and X. Ye, *Nonconforming Galerkin methods based on quadrilateral elements for second order elliptic problems*, M2AN Math. Model. Numer. Anal., 33 (1999), pp. 747–770.

[13] D. Funaro, A. Quarteroni, and P. Zanolli, *An iterative procedure with interface relaxation for domain decomposition methods*, SIAM J. Numer. Anal., 25 (1988), pp. 1213–1236.

[14] R. Glowinski and P. Le Tallec, *Argument Lagrangian interpretation of the nonoverlapping Schwarz alternating method*, in Third International Symposium on Domain Decomposition Methods for Partial Differential Equations, T. F. Chan, R. Glowinski, J. Periaux, and O. B. Widlund, eds., SIAM, Philadelphia, 1990, pp. 224–231.

[15] P. Grisvard, *Singularities in Boundary Value Problems*, Masson and Springer–Verlag, Paris, New York, 1992.

[16] H. Han, *Nonconforming element in the mixed finite element method*, J. Comput. Math., 3 (1984), pp. 223–233.

[17] W. Heinrichs, *Domain decomposition for fourth-order problems*, SIAM J. Numer. Anal., 30 (1993), pp. 435–453.

[18] P. L. Lions, *On the Schwarz alternating method* III: *A variant for nonoverlapping subdomains*, in Third International Symposium on Domain Decomposition Methods for Partial Differential Equations, T. F. Chan, R. Glowinski, J. Periaux, and O. B. Widlund, eds., SIAM, Philadelphia, 1990, pp. 202–223.

[19] T. Lü, T. M. Shih, and C. B. Liem, *Domain Decomposition Methods—New Numerical Techniques for Solving PDE*, Science Press, Beijing, 1992.

[20] L. D. Marini and A. Quarteroni, *A relaxation procedure for domain decomposition methods using finite elements*, Numer. Math., 55 (1989), pp. 575–598.

[21] A. M. Matsokin and S. V. Nepomnyashchikh, *On the convergence of the alternating subdomain Schwarz method without overlap*, in Approximation and Interpolation Methods, Yu. A. Kuznetsov, ed., Akad. Nauk. SSSR Sib. Otdel., Vychisl. Tsentr, Novosibirsk, Siberia, 1981, pp. 85–97.

[22] W. P. Tang, *Generalized Schwarz splitting*, SIAM J. Sci. Statist. Comput., 13 (1992), pp. 573–595.

[23] J. Xu and J. Zou, *Some nonoverlapping domain decomposition methods*, SIAM Rev., 40 (1998), pp. 857–914.

SIAM J. NUMER. ANAL.
Vol. 41, No. 3, pp. 983–1007

A LOCAL CONVERGENCE PROOF FOR THE MINVAR ALGORITHM FOR COMPUTING CONTINUOUS PIECEWISE LINEAR APPROXIMATIONS*

RICHARD E. GROFF†, PRAMOD P. KHARGONEKAR†‡, AND DANIEL E. KODITSCHEK†

Abstract. The class of continuous piecewise linear (PL) functions represents a useful family of approximants because invertibility can be readily imposed, and if a PL function is invertible, then it can be inverted in closed form. Many applications, arising, for example, in control systems and robotics, involve the simultaneous construction of a forward and inverse system model from data. Most approximation techniques require that separate forward and inverse models be trained, whereas an invertible continuous PL affords, simultaneously, the forward and inverse system model in a single representation. The `minvar` algorithm computes a continuous PL approximation to data. Local convergence of `minvar` is proven for the case when the data generating function is itself a PL function and available directly rather than through data.

Key words. piecewise linear, invertible approximation, moving mesh, triangulation

AMS subject classifications. 41-04, 41A30, 65D15

DOI. 10.1137/S0036142902402213

1. Introduction.

In this paper, we present `minvar`, a novel algorithm for computing continuous multidimensional piecewise linear (PL) approximations to data. The algorithm takes advantage of the structure of PL functions to provide a computationally effective approximation technique. This paper provides a local convergence proof for the special case when the data generating function is itself PL and is available directly rather than through discrete data.

Our interest in the PL family is driven by applications that require approximation of both forward and inverse functions from data. In xerography, for example, the print engine's color space transformation is required to stabilize color reproduction, while its inverse is required to generate printer specific color mixture commands in response to inputs expressed in device independent color coordinates [21, 23]. The field of robotics is rife with examples where changes of coordinates play a key role: in mobile robot navigation [27, 32, 33]; in the representation of gaits [31, 34]; in sensor based manipulation [8]; as well as in calibration [42]. Since a change of coordinates is a continuous and continuously invertible function, building a custom change of coordinates amounts to a search for the appropriate forward and inverse function. Representations of scalar invertible functions are required for certain machine tool calibration problems [24], for certain automobile fuel control settings [17], as well as for probability density estimation [15]. In all such settings, most approximation techniques require the construction of distinct forward and inverse representations, because the approximations are not invertible in closed form. In addition to doubling the effective training effort, accuracy suffers since the approximation of the inverse is not exactly the inverse of the forward approximation. In contrast, invertibility of PL

*Received by the editors February 6, 2002; accepted for publication (in revised form) November 27, 2002; published electronically June 18, 2003. This work was supported in part by NSF Award ECS-96322801.

http://www.siam.org/journals/sinum/41-3/40221.html

†EECS Department, 1101 Beal Ave., University of Michigan, Ann Arbor, MI 48109 (regroff@umich.edu, kod@umich.edu).

‡Current address: ECE Department, University of Florida, 300 Weil Hall, P.O. Box 116550, Gainesville, FL 32611 (ppk@ufl.edu).

functions can be verified and even imposed geometrically, that is, by well characterized and computationally effective techniques arising from geometric insights. Moreover, if a PL function is invertible, it can be inverted in closed form. Thus a single PL approximation is ideal for applications requiring the approximation of a function and its inverse.

A substantial amount of mathematical literature on real function approximation (see, for example, [6, 9, 29]), largely concerned with linear-in-parameters techniques, deals extensively with algorithms, fundamental limits, convergence rates, and families of bases in approximating functions. Recent activity has been spurred by evidence that nonlinear-in-parameters function families offer improved approximation rates in higher dimensions as compared to linear-in-parameters representations [3]. Recently, approximation methods that employ collections of local approximations have received increasing attention [1, 16, 38]. However, very little of this linear- or nonlinear-in-parameters literature addresses the problem of function approximation under the constraint of invertibility.

PL functions have been addressed in a number of different settings. Algebraic topologists used PL homeomorphisms to classify topological spaces [36] but did not address computational considerations. The study of splines, piecewise polynomials with continuity and smoothness constraints, includes PL functions [7, 10, 35]. Splines are typically extended to multiple dimensions by means of tensor products. The domain partition is then a tensor product of partitions of the individual dimensions and the approximant is the sum of tensor products of scalar spline functions. A multidimensional linear spline is then multilinear, that is, linear in each variable separately, rather than truly linear. General splines enjoy no invertibility properties. Moreover, most of the spline literature assumes the domain partition to be fixed, in which case approximation of the best L_2 spline is a linear-in-parameters problem. Allowing the partition to change introduces a nonlinear-in-parameters problem. The multivariate adaptive regression spline (MARS) literature admits a limited nonlinear parameterization by allowing the basis to adapt but does not allow general motion of the domain partition [16, 38]. The piecewise polynomial literature addresses the problem of finding (possibly discontinuous) piecewise polynomial approximations to an explicitly known scalar function. In this setting, the domain partition is considered as part of the approximation's parameterization. For scalar functions, there are results for the existence of a best approximation by possibly discontinuous piecewise polynomials under certain generalized convexity conditions [4, 18]. Algorithms similar in flavor to the scalar specialization of `minvar` were introduced in [2, 25, 26]. A treatment of discontinuous piecewise polynomial approximations on two-dimensional triangulations is provided in [39]. Also, [40] provides an algorithm for a moving mesh finite element solution to variational problems. A specialization of this moving mesh algorithm is finding the best L^p, p finite and even, continuous piecewise polynomial approximation to a function. Both of these algorithms [39, 40], as well as the piecewise polynomial literature in general [2, 25, 26], assume that the function to be approximated is available directly, and the algorithms entail steps, such as root finding, that incorporate the function intrinsically. In contrast, the `minvar` algorithm is defined for arbitrary (finite) dimension and can either use a finite set of data or directly use the function to be approximated.

Motivated by applications that require the approximation of invertible functions, we have developed the `minvar` algorithm for computing PL approximations to a set of discrete data. In the context of these applications, PL approximations offer the substantial benefit of closed form invertibility. When the domain partition is fixed, computing the best PL approximation is a linear-in-parameters problem that can be solved using classical techniques. Treating the partition as a component of the approximation's parameterization gives a much more powerful approximant, at the cost of entering the nonlinear-in-parameters problem domain. In nonlinear-in-parameters problems, one can generally expect only local, as opposed to global, convergence properties. Moving the domain partition of a PL function, or triangulation, as formally defined in the next section, has an added difficulty. A triangulation has both continuous and combinatorial parameters that interact in complex ways. Not all combinations of continuous and combinatorial parameters yield a proper triangulation. Triangulations in two and three dimensions have been studied extensively in the computational geometry literature [13, 30], but results for general dimension are more scarce, notwithstanding significant recent progress [5, 14, 28]. The price of using a family of finitely parameterized homeomorphisms, the PL approximations, is the cost of managing the combinatorial complexities of PL functions.

This paper is divided into five main sections. Section 2 provides a careful definition of the concept of a triangulation, relating it to the parameterization of PL functions. Section 3 introduces and defines the `minvar` algorithm. Section 4 provides a local convergence proof for the `minvar` algorithm when the data generating function is piecewise linear. Section 5 presents a numerical example.

2. Triangulations and PL functions.

The ability to check invertibility of a PL function, and to invert it in closed form, derives from the interplay between the PL function's combinatorial and continuous parameters. This interplay provides much power but also creates potential pitfalls. For example, changing the continuous parameters inappropriately with respect to the combinatorial structure can cause "tangles" in the domain partition. Triangulations in general dimension, the key concept in understanding PL functions, are still an area of active research in computational geometry. While the `minvar` algorithm can be stated using only an intuitive notion of triangulation, further analytical insight, such as the local convergence proof provided in section 4, is limited without a much more careful definition. This section provides definitions of triangulations and PL functions to facilitate the exposition. For further background, see [41] for an introduction to concepts in convexity and [13] for an introduction to the geometric concept of triangulations.

2.1. Simplices.

An *affine subspace* $V \subseteq \mathbb{R}^d$ is a linear subspace $L \subseteq \mathbb{R}^d$ translated by some $x_o \in \mathbb{R}^d$, i.e., $V = L + x_o$. The dimension of V is $\dim(V) := \dim(L)$. The *affine hull* of a set $\mathcal{U} \subseteq \mathbb{R}^d$, $\mathrm{aff}(\mathcal{U})$, is the smallest affine subspace containing $\mathcal{U}$. A finite set of points $\mathcal{U} \subseteq \mathbb{R}^d$ is *affinely independent* if for $i = 1, \dots, d$, no affine subspace of dimension i contains more than $i+1$ points from $\mathcal{U}$. The *convex hull* of a set $\mathcal{U} \subseteq \mathbb{R}^d$, $\mathrm{conv}(\mathcal{U})$, is the smallest convex set containing $\mathcal{U}$. A *simplex*, s, is the convex hull of a (finite) set $\mathcal{V} \subseteq \mathbb{R}^d$ of affinely independent points, $s = \mathrm{conv}(\mathcal{V})$. The set of the extreme points, or *vertices*, of s, $\mathcal{V} = \mathrm{vert}(s)$, uniquely defines s.[1] The dimension of s is $l = \dim(s) = \dim(\mathrm{aff}(s)) = \mathrm{card}\,(\mathcal{V}) - 1$, where $\mathrm{card}\,(\mathcal{V})$ is the cardinality of $\mathcal{V}$, and s is called an l-simplex. There can be at most $d+1$ affinely independent points in $\mathbb{R}^d$, and thus there are simplices of dimension $-1, 0, 1, \dots, d$, where by convention $\emptyset$ is

[1] vert is a pseudoinverse of conv, not a true inverse, since if $\mathcal{U} \subseteq s$, then $s = \mathrm{conv}(\mathcal{U} \cup \mathrm{vert}(s))$. The concept of vertices of a simplex is a special case of the more general notion of extreme points of a convex body. The Krein–Milman theorem [41] states that any convex compact set in $\mathbb{R}^d$ is the convex hull of its extreme points.

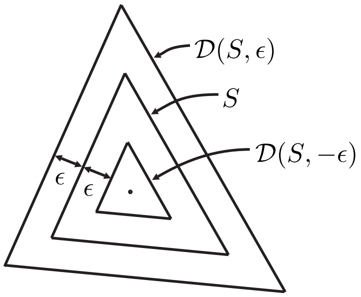


FIG. 2.1. *The* 2*-simplex* S *with its* ϵ *and* $-\epsilon$ *dilations. The point at the center is where dilation degenerates to a single point.*

considered a simplex with $\dim(\emptyset) := -1$. We may apply a partial order to simplices. Given two simplices s_1, s_2, we say that $s_1 \leq s_2$ if and only if $\operatorname{vert}(s_1) \subseteq \operatorname{vert}(s_2)$, in which case we call s_1 a face of s_2. If $\dim(s_2) = d$ and $\dim(s_1) = d-1$, then we also call s_1 a *facet* of s_2. In this paper we will predominantly be interested in d-simplices, so we adopt the convention that a capital S indicates a d-simplex, while a lowercase s denotes a simplex of any dimension.

Let S be a d-simplex with $\operatorname{vert}(S) = \{p_1, \ldots, p_{d+1}\}$. The ϵ dilation of S, written $\mathcal{D}(S,\epsilon)$, is defined as

$$\mathcal{D}(S,\epsilon) := \left\{ x = \sum_{j=1}^{d+1} \alpha_j p_j \,\middle|\, \sum_{j=1}^{d+1} \alpha_j = 1 \ \forall j, \alpha_j \geq \frac{-\epsilon}{\delta(p_j, \operatorname{aff}(\operatorname{vert}(S) - \{p_j\}))} \right\}, \tag{2.1}$$

where $\delta(p,\mathcal{U})$ is the distance from point p to the nonempty set $\mathcal{U}$. Figure 2.1 illustrates the dilation of a 2-simplex. $\mathcal{D}(S,\epsilon)$ is well defined for

$$\epsilon \geq -\left(\sum_{j=1}^{d+1} 1/\delta(p_j, \operatorname{aff}(\operatorname{vert}(S) - \{p_j\})) \right)^{-1}.$$

When equality holds, $\mathcal{D}(S,\epsilon)$ is a single point; otherwise it is a d-simplex with facets parallel to S_i, but distance $|\epsilon|$ away, with $S \subseteq \mathcal{D}(S,\epsilon)$ for $\epsilon > 0$, and $\mathcal{D}(S,\epsilon) \subseteq S$ for $\epsilon < 0$. (See Claim 3 in Appendix A for the dilation's properties.)

2.2. Triangulation. An *abstract simplicial complex* is a collection of finite sets $\mathcal{S}$ satisfying the following: if $\alpha \in \mathcal{S}$ and $\beta \subseteq \alpha$, then $\beta \in \mathcal{S}$. The *vertex set* of an abstract simplicial complex is the set $\{x \mid x \in \alpha, \alpha \in \mathcal{S}\}$.

A *geometric simplicial complex* is a collection $\mathcal{K}$ of simplices in $\mathbb{R}^d$ satisfying

1. $s_1 \in \mathcal{K}$ and $s_2 \leq s_1 \implies s_2 \in \mathcal{K}$,
2. $s_1, s_2 \in \mathcal{K} \implies s_1 \cap s_2 \leq s_1, s_2$.

The *vertex set* of a geometric simplicial complex is $\operatorname{vert}(\mathcal{K}) := \bigcup_{s \in \mathcal{K}} \operatorname{vert}(s)$. The *underlying space* of a geometric simplicial complex is $|\mathcal{K}| := \bigcup_{s \in \mathcal{K}} s$.

A *subcomplex* is a subset of a simplicial complex that is itself a simplicial complex. The *closure* of a subset $\mathcal{L} \subseteq \mathcal{K}$ is the smallest subcomplex that contains $\mathcal{L}$,

$$\operatorname{Cl} \mathcal{L} := \{\alpha \in \mathcal{K} \mid \alpha \leq \beta, \beta \in \mathcal{L}\}.$$

The *star* of a simplex s is the set of all simplices that contain s,

$$\operatorname{St} s := \{s' \in \mathcal{K} \mid s \leq s'\}.$$

The star is not in general a subcomplex.

We can parameterize[2] a geometric simplicial complex $\mathcal{K}$ in $\mathbb{R}^d$ by the pair $(P, \mathcal{S})$, where P is an indexed set of n unique points in $\mathbb{R}^d$,

$$P = \left\{ p_1, \quad p_2, \quad \dots, \quad p_n \right\},$$

and $\mathcal{S}$ is an abstract simplicial complex with vertex set $\{1, 2, \dots, n\}$. Let[3]

$$\mathcal{K}(P, \mathcal{S}) = \left\{ \operatorname{conv}(P(\alpha)) \middle| \alpha \in \mathcal{S} \right\}.$$

$\mathcal{K}(P, \mathcal{S})$ is a geometric simplicial complex if

1. for all $\alpha \in \mathcal{S}$, the points in $P(\alpha)$ are affinely independent,
2. $s_1, s_2 \in \mathcal{K}(P, \mathcal{S}) \implies s_1 \cap s_2 \leq s_1, s_2$,

and, moreover, if these properties hold, then $\operatorname{vert}(\mathcal{K}(P, \mathcal{S})) = P$. Proofs of these properties are provided in [20].

A *triangulation*[4] $\mathcal{T}$ is a geometric simplicial complex in $\mathbb{R}^d$ for which the underlying space is a k-manifold with boundary. Since a triangulation is a type of geometric simplicial complex, it can be parameterized in the same manner. We write $\mathcal{T}(P, \mathcal{S}) := \mathcal{K}(P, \mathcal{S})$ to indicate that the resulting geometric simplicial complex generated by the pair $(P, \mathcal{S})$ is a triangulation.

In this paper, we will deal only with triangulations which are d-manifolds with boundary that have a simply connected underlying space.

For notational convenience, we assume that the triangulation $\mathcal{T} = \mathcal{K}(P, \mathcal{S})$ has N d-simplices that have been indexed and named S_i, $i = 1, \dots, N$. Let $S_i, S_j \in \mathcal{T}$. We then define

$$d_{i,j} := \dim(S_i \cap S_j) = \operatorname{card}\left(\operatorname{vert}(S_i \cap S_j)\right) - 1,$$

the dimension of the face shared by S_i and S_j. Let N_i be the number of d-simplices in $\operatorname{St}\{p_i\}$,

$$N_i = \sum_{S_j \in \operatorname{St} p_i} 1.$$

2.3. PL functions.

A continuous PL function $f_{\mathcal{P}} : D \subseteq \mathbb{R}^d \to \mathbb{R}^d$ is parameterized by a triplet $\mathcal{P} = (P, Q, \mathcal{S})$. P is an indexed set of n points in the domain and Q is an indexed set of n points in the codomain,

$$P = \left\{ p_1, \quad p_2, \quad \dots, \quad p_n \right\}, \qquad Q = \left\{ q_1, \quad q_2, \quad \dots, \quad q_n \right\}.$$

[2] This is not formally a parameterization, because there are some pairs $(P, \mathcal{S})$ for which $\mathcal{K}(P, \mathcal{S})$ is not a geometric simplicial complex. However, for any geometric simplicial complex $\mathcal{K}$, we can write down a pair $(P, \mathcal{S})$ such that $\mathcal{K} = \mathcal{K}(P, \mathcal{S})$.

[3] Formally, an indexed set P of n points in $\mathbb{R}^d$ is a map $P : \{1, 2, \dots, n\} \to \mathbb{R}^d$. The ith member of P is $P(i)$, which we generally write as p_i for notational convenience. Here we extend the notion of P sets. Let $\alpha \subseteq \{1, 2, \dots, n\}$; then $P(\alpha) := \left\{ P(i) \middle| i \in \alpha \right\}$.

[4] There is no formal definition of triangulation in geometry [13]. The definition of triangulation used here is slightly more general than the one used in [14], which requires that the underlying space be the convex hull of the vertex set. A triangulation as defined here which has a simply connected underlying space may be transformed to a triangulation as defined in [14] by a PL homeomorphism. The key concept in our definition is that a triangulation has good local volume properties everywhere. The underlying space has no "thin" spots. In topology, a triangulation of a topological space $\mathcal{X}$ is formally defined as a geometric simplicial complex $\mathcal{K}$ coupled with a homeomorphism between $|\mathcal{K}|$ and $\mathcal{X}$. The definition of triangulation used in this paper is more narrow than the topological notion.

$\mathcal{S}$ is an abstract simplicial complex of indices with $\operatorname{vert}(\mathcal{S}) = \{1, \ldots, n\}$ such that $\mathcal{T}(P, \mathcal{S})$ is a triangulation and $|\mathcal{T}(P, \mathcal{S})| = D$. This defines a continuous PL function $f_{\mathcal{P}}$ such that $f_{\mathcal{P}}(p_i) = q_i$, and for any $S \in \mathcal{T}(P, \mathcal{S})$, $f_{\mathcal{P}}(x)$ is affine on S. For a d-simplex $S_i \in \mathcal{T}(P, \mathcal{S})$ with $\operatorname{vert}(S_i) = \{p_{i_1}, p_{i_2}, \ldots, p_{i_{d+1}}\}$, the PL function $f_{\mathcal{P}}(x)$ for $x \in S_i$ is given by

$$f_{\mathcal{P}}\big|_{S_i}(x) = \begin{bmatrix} q_{i_1} & q_{i_2} & \cdots & q_{i_{d+1}} \end{bmatrix} \begin{bmatrix} p_{i_1} & p_{i_2} & \cdots & p_{i_{d+1}} \\ 1 & 1 & & 1 \end{bmatrix}^{-1} \begin{bmatrix} x \\ 1 \end{bmatrix}. \tag{2.2}$$

Equation (2.2) uses a homogeneous representation for the rightmost two factors, though $f_{\mathcal{P}}|_{S_i}$ can be equivalently expressed in the more typical form as $A_i x + b_i$. The d-simplices of $\mathcal{T}(P, \mathcal{S})$ are a cover for D. If $S_i \cap S_j \neq \emptyset$, and $f_{\mathcal{P}}\big|_{S_i}(x) = A_i x + b_i$ and $f_{\mathcal{P}}\big|_{S_j}(x) = A_j x + b_j$, then $A_i x + b_i = A_j x + b_j$ for $x \in S_i \cap S_j$. This follows from Claim 5 in Appendix A, which states that $(A_i - A_j)$ has a null space of dimension $d_{i,j}$ parallel to $\operatorname{aff}(S_i \cap S_j)$.

One of the most compelling properties of PL functions is the ability to check invertibility and invert in closed form. Let $f_{\mathcal{P}}$ be a PL function parameterized by $\mathcal{P} = (P, Q, \mathcal{S})$. If $\mathcal{T}(Q, \mathcal{S})$ is a triangulation, then the PL function is invertible, and the inverse $f_{\mathcal{P}}{}^{-1}$ is a PL function parameterized by $\mathcal{P}^{-1} = (Q, P, \mathcal{S})$. This is proven in Claim 7 in Appendix A.

Another important fact applied in proving the main result of section 4 is the continuity of a PL function in its continuous parameters. This claim, stated formally in Claim 6 in Appendix A, establishes that two PL functions with the same combinatorial structure are close in the L_∞ sense if their vertices are close in the Euclidean sense.

We call a PL function $f_{\mathcal{P}}$ parameterized by $\mathcal{P} = (P, Q, \mathcal{S})$ *nondegenerate* if for all $p_i \in P$ such that $p_i \notin \partial\, |\mathcal{T}(P, \mathcal{S})|$ the matrix H_i is full rank, where

$$H_i = \left(\frac{1}{N_i} \sum_{S_j \in \operatorname{St}\{p_i\}} A_j{}^{\mathrm{T}} A_j \right) - \overline{A^i}^{\mathrm{T}} \overline{A^i}, \qquad \text{where } \overline{A^i} = \frac{1}{N_i} \sum_{S_j \in \operatorname{St}\{p_i\}} A_j.$$

Intuitively, nondegeneracy of $f_{\mathcal{P}}$ requires that for any $p_i \notin \partial\, |\mathcal{T}(P, \mathcal{S})|$ not all of the affine functions that $f_{\mathcal{P}}$ takes in the surrounding d-simplices are parallel.

3. The minvar algorithm.

The `minvar` algorithm is an iterative scheme to generate a locally good PL approximation to data. Similar to algorithms proposed in the possibly discontinuous piecewise polynomial approximation literature [2, 25, 26, 39], `minvar` takes advantage of the structure of PL functions. Let $\mathcal{Z} = \{(x_i, y_i)\}_{i=1}^{N_s}$, where $x_i \in D \subseteq \mathbb{R}^d$ and $y_i \in \mathbb{R}^d$, be the set of input-output data to be approximated. The `minvar` algorithm iteratively improves a PL approximation to the data, $f_{\mathcal{P}}^{(k)}$, parameterized by $\mathcal{P}^{(k)} = (P^{(k)}, Q^{(k)}, \mathcal{S})$, such that $\left|\mathcal{T}(P^{(k)}, \mathcal{S})\right| = D$. (The superscript in parentheses indicates the iteration number.) The algorithm breaks down into two stages. The first stage partitions the data according to the d-simplices of $\mathcal{T}(P^{(k)}, \mathcal{S})$ and computes the least squares linear approximations for each subset of the data. This set of linear approximations is the optimal possibly discontinuous PL approximation on the partition $\mathcal{T}(P^{(k)}, \mathcal{S})$. The second stage chooses $(P^{(k+1)}, Q^{(k+1)})$ to make $f_{\mathcal{P}}^{(k+1)}$, a continuous PL function, be "close" to the discontinuous approximation from the first stage. The stages are then iterated.

Recall from the previous section that the domain of a PL function is $|\mathcal{T}(P, \mathcal{S})|$, so moving a vertex $p_i \in \partial\, |\mathcal{T}(P, \mathcal{S})|$ will change the domain of definition of the PL function. Since we desire a fixed domain for the PL function, the present exposition considers vertices on the domain boundary to be fixed. This can be relaxed to allow boundary vertices that are not extreme points to move in appropriately chosen affine subspaces using a constrained version of the cost function from step 3 of the `minvar` algorithm. Computational details of constrained motion will be presented in a subsequent paper on engineering applications of `minvar`.

From an initial parameterization $\mathcal{P}^{(0)} = (P^{(0)}, Q^{(0)}, \mathcal{S})$, the `minvar` algorithm generates a sequence of parameterizations, $\mathcal{P}^{(k)} = (P^{(k)}, Q^{(k)}, \mathcal{S})$, as follows:

1. Partition the data set $\mathcal{Z}$ into subsets $\mathcal{Z}_j$ corresponding to the d-simplices, $S_1, \ldots, S_N$ of $\mathcal{T}(P^{(k)}, \mathcal{S})$, breaking multiple memberships (data points that lie on the boundary between d-simplices) systematically.
2. Compute the least squares affine approximation, $L_j(x)$, for each subset $\mathcal{Z}_j$.
3. Update the vertex locations,

$$(3.1)\quad p_i^{(k+1)} = \arg\min_{x\in\mathbb{R}^d} \operatorname{var} L^i(x) + \lambda \left\| x - p_i^{(k)} \right\|^2 \qquad \forall\, p_i^{(k)} \notin \partial \left| \mathcal{T}(P^{(k)}, \mathcal{S}) \right|,$$

$$(3.2)\quad p_i^{(k+1)} = p_i^{(k)} \qquad \text{otherwise,}$$

$$(3.3)\quad q_i^{(k+1)} = \frac{1}{N_i} \sum_{S_j \in \mathrm{St}\{p_i\}} L^i\left(p_i^{(k+1)}\right) \qquad \forall\, i,$$

where $L_j(x) = A_j x + b_j$ and

$$\operatorname{var} L^i(x) = \sum_{S_j \in \mathrm{St}\{p_i\}} \left(L_j(x) - \frac{1}{N_i} \sum_{S_k \in \mathrm{St}\{p_i\}} L_k(x) \right)^2 .$$

4. If the vertices are not converged, then $k \leftarrow k+1$, go to 1.

Notice that (3.1) is a positive definite quadratic function of x, and thus can be minimized in closed form by "completing the square," with the solution given by

$$p_i^{(k+1)} = -{H_i}^{-1} h_i, \tag{3.4}$$

where

$$H_i = \left[\frac{1}{N_i} \sum_{S_j \in \mathrm{St}\{p_i\}} {A_j}^{\mathrm{T}} A_j \right] - \overline{A^i}^{\mathrm{T}} \overline{A^i} + \lambda I,$$

$$h_i = \left[\frac{1}{N_i} \sum_{S_j \in \mathrm{St}\{p_i\}} {A_j}^{\mathrm{T}} b_j \right] - \overline{A^i}^{\mathrm{T}} \overline{b^i} - \lambda p_i^{(k)},$$

$$\overline{A^i} = \frac{1}{N_i} \sum_{S_j \in \mathrm{St}\{p_i\}} A_j \qquad \overline{b^i} = \frac{1}{N_i} \sum_{S_j \in \mathrm{St}\{p_i\}} b_j .$$

The nonnegative quantity $\operatorname{var} L^i(x)$ measures, as a function of location in the domain x, how tightly clustered the range values generated from the least squares approximations on d-simplices in $\mathrm{St}\{p_i\}$ are. In the case of a scalar domain, an interior vertex p_i is in at most two 1-simplices. Thus, if the least squares approximations are not parallel, then $\operatorname{var} L^i(x_c) = 0$ at and only at x_c, the domain value of the point at which the least squares approximations intersect. In the scalar case, the `minvar` algorithm with $\lambda = 0$ moves the domain and codomain vertices to the intersection point of the least squares approximations. We called our initial scalar algorithm the "Graph Intersection" algorithm [22] due to this fact. For dimensions higher than 1, there is generically no unique intersection point for the least squares approximations surrounding p_i, due to the geometry of triangulations. Rather than the intersection point, `minvar` with $\lambda = 0$ picks the point where the range values are most tightly clustered.

The λ term in (3.1) is a regularization. It guarantees that (3.1) will have a unique minimum, even if all the least squared approximations are parallel. More importantly, in the implementation of `minvar` the λ parameter can be tuned to prevent a vertex from jumping long distances and creating a "tangle" in the domain triangulation of the approximation. A tangle is when movement of the vertices causes $\mathcal{T}(P, \mathcal{S})$ to no longer be a geometric simplicial complex. That is, either a simplex has been "flattened" or there are simplices whose intersection is not another simplex from the complex. This generally occurs when a domain vertex moves through one of its opposing faces. Methods for detecting and correcting triangulation tangles will be covered in a subsequent paper on using `minvar` in engineering applications.

In this exposition, `minvar` does not modify the combinatorial structure, $\mathcal{S}$, of the PL approximation. Heuristics for adapting the domain triangulation of a two-dimensional PL function are presented in [12, 11] for interpolation and [39] for approximation. These heuristics flip edges in the domain triangulation to improve a local goodness criterion, similar to a method for computing the planar Delaunay triangulation. Generalizing these heuristics to higher dimensions is difficult because local topological changes of the triangulation in dimensions greater than two are more complex than edge flipping [28, 14]. Nonetheless, we find that adaptation of the combinatorial parameters of the PL function via topological flipping provides significant benefit in practice. Techniques for adapting the combinatorial structure will be presented in a subsequent paper on engineering applications of `minvar`.

4. A local convergence proof for the `minvar` algorithm.

We turn now to the central result: a local convergence proof for the `minvar` algorithm. The result is for the "approximation," as opposed to "estimation," version of the `minvar` algorithm. That is, the data generating function is considered to be directly available in closed form, rather than through a set of discrete data. In this case, the least squares approximations from step 2 become L_2 orthogonal projections of $f^*_{\mathcal{P}}|_{S_i}$, the data generating function restricted to S_i, to the space of affine functions. Since the data or data generating function only appear in step 2, the approximation version may be viewed as the limit behavior of the estimation version when provided with an unbounded quantity of uniformly distributed data.

Theorem 4.1 shows that, if the data generating function is a nondegenerate PL function and the approximation is initialized "close enough" to the data generating function, then the `minvar` algorithm with $\lambda = 0$ will cause the approximation to converge to the data generating function in the L_∞ sense. In this case, "close enough" means that the initial approximation shares the same combinatorial structure as the data generating function, and the vertices of the approximation start close to the corresponding vertices of the data generating function. Examining `minvar` when $\lambda = 0$ admits a simpler proof while capturing the essence of the algorithm. Similar results could be obtained for $\lambda > 0$, though the convergence rate would be slower. An additional technical condition, that the data generating function be nondegenerate, is required when $\lambda = 0$ in order to guarantee existence of a unique solution to (3.1), whereas for $\lambda > 0$ the regularized variance minimization in (3.1) is guaranteed to have a unique solution.

In this paper, unless otherwise noted, vector norms are the standard Euclidean norm and matrix norms are the induced two norm.

THEOREM 4.1. *Let $f^*_{\mathcal{P}}$ be a nondegenerate PL data generating function parameterized by $\mathcal{P}^* = (P^*, Q^*, \mathcal{S}^*)$. Let $\epsilon_0 = \epsilon_0(\mathcal{P}^*)$ be given by* (4.7)*. Let the initial approximation $f^{(0)}_{\mathcal{P}}$ be parameterized by $(P^{(0)}, Q^{(0)}, \mathcal{S}^*)$, satisfying for some $\epsilon < \epsilon_0$, $\|p^{(0)}_i - p^*_i\| < \epsilon$, for all i. Then application of the* `minvar` *algorithm with $\lambda = 0$ yields a sequence of approximations satisfying*

$$\lim_{j\to\infty} \left\| f^{(j)}_{\mathcal{P}} - f^*_{\mathcal{P}} \right\|_\infty = 0.$$

Proof. Proposition 4.5 shows that iteration of the `minvar` algorithm causes the vertices of the approximation to converge to the vertices of the data generating function. By Claim 6 in Appendix A, a PL function is continuous in its vertices. The theorem follows directly. □

The theorem follows readily from Proposition 4.5, which likewise follows readily from Proposition 4.4. The statements and proofs of the propositions and lemmas follow in the next subsections, but first we offer a short sketch of the structure of the proof. The essence of Proposition 4.4 is that when the distances between the vertices of the approximation and the corresponding vertices of the data generating function are bounded by ϵ, then after one iteration of the `minvar` algorithm the distances will be bounded by a constant times ϵ^2. This result is established by applying two lemmas corresponding to the two stages of the algorithm. Lemma 4.2 proves that if the distances between corresponding vertices are bounded by ϵ, then the perturbation of the least squares affine map over a given simplex of the approximation from the affine map in the corresponding simplex of the data generating function is bounded by a constant times ϵ^2. Lemma 4.3 proves that if the perturbation of the least squares affine map over a simplex of the approximation from the affine map in the corresponding simplex of the data generating function is bounded by Δ, then the variance minimization will place the new vertices of the approximation such that the distance between them and the corresponding vertices of the data generating function are bounded by a constant times Δ. The combination of Lemmas 4.2 and 4.3 provides Proposition 4.4.

The quadratic rate of convergence in Proposition 4.4 arises from the hypothesis that the data generating function is piecewise linear and close to the initial approximation. Without this assumption, Lemma 4.2 would fail to provide an ϵ^2 perturbation in the least squares affine approximations. In this case, we suspect the convergence rate of the algorithm to be linear. Convergence may be slower on fine triangulations, but since this algorithm is intended primarily for use with a discrete set of data, the fineness of the triangulation is inherently limited by the amount of data provided. In applications, `minvar` can run triangulations of practical size in a few minutes.

4.1. Lemmas and propositions. This section states the lemmas and propositions, while the proofs are provided in the following section. First, we introduce several reoccurring constants. These constants may be interpreted geometrically as minima or maxima of different measures of the "radii" of d-simplices in the triangulation $\mathcal{T}(P^*, \mathcal{S}^*)$ of the data generating function. The first measures the maximum inter-vertex distance between "connected" vertices,

$$r_1 := \max_{\substack{S^* \in \mathcal{T}(P^*, \mathcal{S}^*) \\ p_i^*, p_j^* \in S^*}} \left\| p_i^* - p_j^* \right\| . \tag{4.1}$$

The second measures the minimum distance of a vertex to its opposing hyperplanes,

$$r_2 := \min_{\substack{S^* \in \mathcal{T}(P^*, \mathcal{S}^*) \\ p^* \in \mathrm{vert}(S^*)}} \delta(p^*, \mathrm{aff}(\mathrm{vert}(S^*) - \{p^*\})) . \tag{4.2}$$

The third measures the d-simplex which can be dilated the least before it intersects simplices outside its immediate neighborhood.

$$r_3 := \min_{S^* \in \mathcal{T}(P^*, \mathcal{S}^*)} \sup \left\{ \epsilon \middle| \mathcal{D}(S^*, \epsilon) \subseteq |\mathrm{Cl}\,\mathrm{St} S^*| \right\} . \tag{4.3}$$

The first lemma shows that if the domain vertices of the approximation are close to the domain vertices of the data generating function, then least squares affine fit in a simplex S_i is a perturbation away from the affine function that the data generating function takes in S_i^*. Moreover, the perturbation is quadratic in the bound on the distance between the approximation and data generating function's domain vertices. We write $\Pi(f)$ to denote the L_2 orthogonal projection of the function f onto the space of affine functions.

LEMMA 4.2. *Let $f_{\mathcal{P}}^*$ be a PL data generating function parameterized by $\mathcal{P}^* = (P^*, Q^*, \mathcal{S}^*)$. Consider a PL approximation $f_{\mathcal{P}}$ parameterized by $\mathcal{P} = (P, Q, \mathcal{S}^*)$. Let $\epsilon < \epsilon_c$, where*

$$\epsilon_c := \min \left\{ \tfrac{1}{2(d+1)} r_2, \quad r_3, \quad 1 \right\} . \tag{4.4}$$

Consider the simplices S_i^* and S_i. Let $x_c \in S_i^*$. Let $f_{\mathcal{P}}^*|_{S_i^*}(x) = A_i^*(x - x_c) + b_i^*$. If $\left\| p_j - p_j^* \right\| < \epsilon$ for all $p_j^* \in S_i^*$, then the least squares approximation to $f_{\mathcal{P}}^*$ on S_i, $\Pi(f_{\mathcal{P}}^*|_{S_i})(x) = \hat{A}_i(x - x_c) + \hat{b}_i$, satisfies the property

$$\left\| \begin{bmatrix} \hat{A}_i{}^{\mathrm{T}} - A_i^{*\mathrm{T}} \\ \hat{b}_i{}^{\mathrm{T}} - b_i^{*\mathrm{T}} \end{bmatrix} \right\|_2 < c_{1,i} \epsilon^2, \tag{4.5}$$

where $c_{1,i} = c_{1,i}(\mathcal{P}^*)$ is given by (4.18).

The second lemma considers one set of affine functions that all intersect at a common point and another set of affine functions which are perturbations of the first set of functions. It is shown that performing the variance minimization, equivalent to (3.1) with $\lambda = 0$, on the second set of functions generates a point whose distance from the intersection point is linear in the norm of the perturbations.

LEMMA 4.3. *Let L^* be a set of N affine maps, $L_1^*, \ldots L_N^*$, such that all intersect at (p^*, q^*) and are written as $L_i^*(x) = A_i^*(x - p^*) + q^*$, and such that H^*, given by* (4.20)*, is full rank. Let L be a set of perturbed affine maps, $L_1, \ldots L_N$, expressed as $L_i(x) = \hat{A}_i(x - p^*) + \hat{q}_i$, which satisfy the property*

$$\left\| \begin{bmatrix} \hat{A}_i{}^{\mathrm{T}} - A_i^{*\mathrm{T}} \\ \hat{q}_i{}^{\mathrm{T}} - q^{*\mathrm{T}} \end{bmatrix} \right\| < \Delta \tag{4.6}$$

for $\Delta < \Delta_0$, where $\Delta_0 = \Delta_0(A_i^*, p^*, q^*)$ is given by (4.19)*. Let p' and q' be given by*

$$p' = \arg\min_x \quad \operatorname{var} L(x),$$
$$q' = \frac{1}{N}\sum_{i=1}^{N} L(p').$$

Then p' and q' satisfy

$$\|p' - p^*\| < c_2\Delta,$$
$$\|q' - q^*\| < c_3\Delta,$$

where $c_2 = c_2(A_i^*, p^*, q^*)$ and $c_3 = c_3(A_i^*, p^*, q^*)$ are given by (4.23) *and* (4.24)*.*

The first proposition brings the two lemmas together to show that a single step of the `minvar` algorithm induces a quadratic change in the distance of the approximation vertices to the data generating function vertices.

Proposition 4.4. *Let $f^*_{\mathcal{P}}$ be a nondegenerate PL data generating function parameterized by $\mathcal{P}^* = (P^*, Q^*, \mathcal{S}^*)$. Let $\epsilon < \epsilon_d$,*

$$\epsilon_d := \min\left\{\epsilon_c, \quad \sqrt{\frac{\Delta_0^m}{c_1}}\right\},$$

where $\epsilon_c = \epsilon_c(\mathcal{P}^*)$ is given by (4.4)*, $\Delta_0^m = \Delta_0^m(\mathcal{P}^*)$ by* (4.25)*, and $c_1 = c_1(\mathcal{P}^*)$ by* (4.27)*.*

If the PL approximation $f_{\mathcal{P}}$ parameterized by $(P, Q, \mathcal{S}^*)$ satisfies $\|p_i - p_i^*\| < \epsilon$ for all i, then one iteration of the `minvar` *algorithm with $\lambda = 0$ gives the new approximation f'_P parameterized by $(P', Q', \mathcal{S}^*)$, which satisfies*

$$\|p'_i - p_i^*\| < c_4\epsilon^2,$$
$$\|q'_i - q_i^*\| < c_5\epsilon^2,$$

for all i, where $c_4 = c_4(\mathcal{P}^*)$ and $c_5 = c_5(\mathcal{P}^*)$ are given by (4.31) *and* (4.32)*.*

The second proposition applies the first proposition to show that iteration of the `minvar` algorithm causes convergence of the vertices of the approximation to the vertices of the data generating function.

Proposition 4.5. *Let $f^*_{\mathcal{P}}$ be a nondegenerate PL data generating function parameterized by $(P^*, Q^*, \mathcal{S}^*)$. Let*

$$\epsilon_0 = \min\left\{\epsilon_d, \ \frac{1}{c_4}\right\}, \tag{4.7}$$

where ϵ_d and c_4 are given in Proposition 4.4*. If for some $0 < \epsilon < \epsilon_0$ the initial PL approximation $f_P^{(0)}$ with parameterization $(P^{(0)}, Q^{(0)}, \mathcal{S}^*)$ satisfies $\|p_i^{(0)} - p_i^*\| < \epsilon$ for all i, then iteration of the* `minvar` *algorithm with $\lambda = 0$ gives a sequence of approximations $f_P^{(k)}$ satisfying*

$$\lim_{k\to\infty} \left\|p_i^{(k)} - p_i^*\right\| = 0,$$
$$\lim_{k\to\infty} \left\|q_i^{(k)} - q_i^*\right\| = 0$$

for all i.

4.2. Proofs of lemmas and propositions. This section presents proofs of the lemmas and propositions stated in the previous section.

Proof of Lemma 4.2. Let $\varphi_i(x) := A_i^*(x - x_c) + b_i^*$, the extension of $f_{\mathcal{P}}^*|_{S_i^*}$ to the entire domain. Let $\psi_i(x) := f_{\mathcal{P}}^*(x) - \varphi_i(x)$.

The orthogonal projection Π is a linear operator, and, moreover, for g affine, $\Pi(g) = g$. It follows that

$$\begin{aligned} \Pi(f_{\mathcal{P}}^*|_{S_i}) &= \Pi(\varphi_i|_{S_i}) + \Pi(\psi_i|_{S_i}) \\ &= \varphi_i + \Pi(\psi_i|_{S_i}). \end{aligned} \tag{4.8}$$

Let $\hat{A}_i$ and $\hat{b}_i$ be such that $\Pi(f_{\mathcal{P}}^*|_{S_i})(x) = \hat{A}_i(x - x_c) + \hat{b}_i$. Then from (4.8) it follows that $\Pi(\psi_i|_{S_i}) = (\hat{A}_i - A_i^*)(x - x_c) + (\hat{b}_i - b_i^*)$. Moreover, since $\Pi(\psi_i|_{S_i}) = \Pi(f_{\mathcal{P}}^*|_{S_i} - \varphi_i|_{S_i})$, we can compute $(\hat{A}_i - A_i^*)$ and $(\hat{b}_i - b_i^*)$ using the formula for the L_2 orthogonal projection of $f_{\mathcal{P}}^*|_{S_i} - \varphi_i|_{S_i}$,

$$\begin{bmatrix} (\hat{A}_i - A_i^*)^{\mathrm{T}} \\ (\hat{b}_i - b_i^*)^{\mathrm{T}} \end{bmatrix} = {S_{xx,i}}^{-1} S_{xy,i},$$

$$S_{xx,i} = \int_{S_i} \begin{bmatrix} x - x_c \\ 1 \end{bmatrix} \begin{bmatrix} x^{\mathrm{T}} - {x_c}^{\mathrm{T}} & 1 \end{bmatrix} dx, \qquad S_{xy,i} = \int_{S_i} \begin{bmatrix} x - x_c \\ 1 \end{bmatrix} (f_{\mathcal{P}}^*(x) - \varphi_i(x))^{\mathrm{T}} dx.$$

The submultiplicative property holds for the induced two norm,

$$\left\| \begin{bmatrix} (\hat{A}_i - A_i^*)^{\mathrm{T}} \\ (\hat{b}_i - b_i^*)^{\mathrm{T}} \end{bmatrix} \right\| \leq \left\| {S_{xx,i}}^{-1} \right\| \left\| S_{xy,i} \right\|,$$

so we can independently establish bounds on $\left\| {S_{xx,i}}^{-1} \right\|$ and $\| S_{xy,i} \|$. We will proceed to bound $\left\| {S_{xx,i}}^{-1} \right\|$. Since S_i is a d-simplex, it follows from calculus and the definition of $S_{xx,i}$ that $S_{xx,i}$ is a positive definite matrix. Let M_i^* be given by

$$M_i^* := \int_{\mathcal{D}(S_i^*, -\epsilon_c)} \begin{bmatrix} x - x_c \\ 1 \end{bmatrix} \begin{bmatrix} x^{\mathrm{T}} - {x_c}^{\mathrm{T}} & 1 \end{bmatrix} dx.$$

Since $0 < \epsilon < \epsilon_c$ by hypothesis and $\epsilon_c \leq \frac{1}{2(d+1)} r_2$ by definition, it follows from Claim 3 in Appendix A that $\mathcal{D}(S_i^*, -\epsilon_c)$ is a d-simplex. Thus M_i^* is also positive definite, and hence invertible. Since $\epsilon < \epsilon_c$ and the vertices or S_i are all less than ϵ away from the vertices of S_i^*, it follows that $\mathcal{D}(S_i^*, -\epsilon_c) \subseteq S_i$. Thus, $x^{\mathrm{T}} M_i^* x < x^{\mathrm{T}} S_{xx,i} x$ for all x, which implies that $\lambda_{\min}(M) < \lambda_{\min}(S_{xx,i})$. This provides the bound on $\left\| {S_{xx,i}}^{-1} \right\|$,

$$\left\| {S_{xx,i}}^{-1} \right\| < \left\| {M_i^*}^{-1} \right\|.$$

Now we proceed to $\| S_{xy,i} \|$. By the properties of norms,

$$\| S_{xy,i} \| \leq \int_{S_i} \left\| \begin{bmatrix} x - x_c \\ 1 \end{bmatrix} \right\| \| \psi_i(x) \| \, dx. \tag{4.9}$$

Let $\mathcal{C}(S_i, \epsilon) = \{ x \in \mathbb{R}^d | \delta(x, S_i) \leq \epsilon \}$. By hypothesis, the vertices of S_i are less than ϵ away from the vertices of S_i^*, so $\operatorname{vert} S_i \subseteq \mathcal{C}(S_i^*, \epsilon)$. Moreover, since both S_i and

$\mathcal{C}(S_i^*, \epsilon)$ are convex, $S_i \subseteq \mathcal{D}(S_i^*, \epsilon)$. The integrand in (4.9) is nonnegative definite, so (4.9) is bounded by

$$\leq \int_{\mathcal{C}(S_i^*,\epsilon)} \left\| \begin{bmatrix} x - x_c \\ 1 \end{bmatrix} \right\| \|(\psi_i(x))\| \, dx.$$

By hypothesis $x_c \in S_i^*$. By the definition of r_1 and since $\epsilon < \epsilon_c$, it follows that for all $x \in \mathcal{C}(S_i^*, \epsilon)$, $\|x - x_c\| \leq \bar{r} := r_1 + 2\epsilon_c$. Thus, the integral above is further bounded by

$$\leq \sqrt{1 + \bar{r}^2} \int_{\mathcal{C}(S_i^*,\epsilon)} \|(\psi_i(x))\| \, dx. \tag{4.10}$$

Once again the integrand is nonnegative definite, so (4.10) can be bounded by integrating over $\mathcal{D}(S_i^*, \epsilon)$, since $\mathcal{C}(S_i^*, \epsilon) \subseteq \mathcal{D}(S_i^*, \epsilon)$,

$$\leq \sqrt{1 + \bar{r}^2} \int_{\mathcal{D}(S_i^*,\epsilon)} \|(\psi_i(x))\| \, dx \tag{4.11}$$

$$= \sqrt{1 + \bar{r}^2} \sum_{j=1}^{N} \int_{U_j} \|(\psi_i(x))\| \, dx, \tag{4.12}$$

where $U_j = \mathcal{D}(S_i^*, \epsilon) \cap S_j^*$ and N is the total number of d-simplices in the domain triangulation. Since $\psi_i(x) = 0$ on S_i^*, the term corresponding to $j = i$ in (4.12) is 0. By hypothesis $\epsilon < \epsilon_c \leq r_3$, so then following from the definition of r_3, $S_j^* \cap \mathcal{D}(S_i^*, \epsilon) \neq \emptyset$ if and only if $S_j^* \in \operatorname{St} S_i^*$. Thus the terms in (4.12) are only nonzero for j such that S_j^* is incident to S_i^*. Consider such a term,

$$\int_{U_j} \|\psi_i(x)\| \, dx = \int_{U_j} \left\| \left(A_j^* - A_i^*\right)(x - x_c) + b_j^* - b_i^* \right\| dx,$$

where $f_{\mathcal{P}}^*|_{S_j^*}(x) = A_j^*(x - x_c) + b_j^*$. By Claim 5 in Appendix A, there exists $x_O \in S_j^* \cap S_i^* \subseteq U_j$ such that $\left(A_j^* - A_i^*\right)(x_O - x_c) + b_j^* - b_i^* = 0$. Applying the change of coordinates $y = x - x_O$ gives

$$\int_{U_j} \|\psi_i(x)\| \, dx = \int_{U_j - x_O} \left\| \left(A_j^* - A_i^*\right) y \right\| \, dy. \tag{4.13}$$

Let L be the linear subspace parallel to $\operatorname{aff}(S_i^* \cap S_j^*)$. Recall that $\dim L = \dim S_i^* \cap S_j^* := d_{i,j}$. By Claim 5 in Appendix A, $L \subseteq \mathcal{N}(\left(A_j^* - A_i^*\right))$. Let $v_1, \ldots, v_{d_{i,j}}$ be an orthonormal basis for L. Let $v_{d_{i,j}+1}, \ldots, v_d$ be an orthonormal basis for $L^{\perp}$. Then $P = \begin{bmatrix} v_1 & v_2 & \cdots & v_d \end{bmatrix}$ is an orthogonal matrix. Rewrite (4.13) under the change of coordinates $z = P^{\mathrm{T}} y$,

$$= \int_{P^{\mathrm{T}}(U_j - x_O)} \left\| \left(A_j^* - A_i^*\right) P z \right\| \, dz. \tag{4.14}$$

Since the integrand is a nonnegative definite function, we may bound (4.14) by increasing the volume over which the integrand is integrated. By Claim 4 in Appendix A, there exists $\kappa_{i,j}$ such that for all $x \in U_j$, $\delta\big(x, \operatorname{aff}(S_i^* \cap S_j^*)\big) < \kappa_{i,j}\epsilon$. Equivalently $\delta(y, L) < \kappa_{i,j}\epsilon$ for any $y \in U_j - x_O$, from which it follows that the projection of y onto $L^{\perp}$ must have magnitude less than $\kappa_{i,j}\epsilon$. Moreover, by the definition of $\bar{r}$, the projection of y onto L must have magnitude less than $\bar{r}$. It follows that $P^{\mathrm{T}}(U_j - x_O) \subseteq [-\bar{r}, \bar{r}]^{d_{i,j}} \times B_{\bar{d}_{i,j}}(\kappa_{i,j}\epsilon)$, where $\bar{d}_{i,j} = d - d_{i,j}$ and $B_{\bar{d}_{i,j}}(\kappa_{i,j}\epsilon)$ is the $\bar{d}_{i,j}$-dimensional ball of radius $\kappa_{i,j}\epsilon$. Then (4.14) is bounded by

$$\leq \int_{[-\bar{r},\bar{r}]^{d_{i,j}} \times B_{\bar{d}_{i,j}}(\kappa_{i,j}\epsilon)} \left\| \left(A_j^* - A_i^*\right) P z \right\| \; dz. \tag{4.15}$$

The first $d_{i,j}$ columns of $\left(A_j^* - A_i^*\right) P$ are zero, since the first $d_{i,j}$ columns of P are in the nullspace of $\left(A_j^* - A_i^*\right)$. Thus, the integrand in (4.15) has no dependence on $z_1, z_2, \ldots, z_{d_{i,j}}$, so we can integrate through for $z_1, \ldots, z_{d_{i,j}}$, giving

$$= (2\bar{r})^{d_{i,j}} \int_{B_{\bar{d}_{i,j}}(\kappa_{i,j}\epsilon)} \left\| \left(A_j^* - A_i^*\right) P \begin{bmatrix} I \\ 0 \end{bmatrix} \bar{z}_1 \right\| \; dz_{d_{i,i}+1} \ldots dz_d, \tag{4.16}$$

where ${\bar{z}_1}^{\mathrm{T}} = \begin{bmatrix} z_{d_{i,j}+1} & \cdots & z_d \end{bmatrix}^{\mathrm{T}}$. Since the first $d_{i,j}$ columns of $\left(A_j^* - A_i^*\right) P$ are zero and P is orthogonal, it follows that $\|\left(A_j^* - A_i^*\right) P \begin{bmatrix} I & 0 \end{bmatrix}^{\mathrm{T}}\| \leq \|A_j^* - A_i^*\|$. Thus, (4.16) can be further bound by

$$\leq (2\bar{r})^{d_{i,j}} \left\| A_j^* - A_i^* \right\| \int_{B_{\bar{d}_{i,j}}(\kappa_{i,j}\epsilon)} \|\bar{z}_1\| \; d\bar{z}_1. \tag{4.17}$$

From calculus (see, for example, [37]) it can be shown that[5]

$$\int_{B_k(\epsilon)} \|w\| \; dw = \frac{k}{k+1} \frac{\pi^{k/2}}{\Gamma(k/2)} \epsilon^{k+1}.$$

Applying this with (4.17) to (4.12), and then simplifying using the fact that $\epsilon^m \leq \epsilon^2$ for $m \geq 2$ since $\epsilon < \epsilon_c \leq 1$, gives

$$\|S_{xy,i}\| \leq \sqrt{1 + \bar{r}^2} \; l_i \max_{\substack{j \text{ s.t. } j \neq i, \\ S_j^* \in \mathrm{St} S_i^*}} \left((2\bar{r})^{d_{i,j}} \left\| A_j^* - A_i^* \right\| {\kappa_{i,j}}^{1+\bar{d}_{i,j}} \frac{\bar{d}_{i,j}}{\bar{d}_{i,j}+1} \frac{\pi^{(\bar{d}_{i,j})/2}}{\Gamma((\bar{d}_{i,j})/2)} \right) \epsilon^2,$$

where $l_i = \sum_{S_j^* \in \mathrm{St} S_i^*} 1$. Then

$$\begin{aligned} \left\| \begin{bmatrix} {\hat{A}_i}^{\mathrm{T}} - A_i^{*\mathrm{T}} \\ {\hat{b}_i}^{\mathrm{T}} - b_i^{*\mathrm{T}} \end{bmatrix} \right\| &\leq \left\| {S_{xx,i}}^{-1} \right\| \|S_{xy,i}\| \\ &< c_{1,i} \epsilon^2, \end{aligned}$$

where

(4.18)

$$c_{1,i} := \left\| M_i^{*-1} \right\| \sqrt{1 + \bar{r}^2} \; l_i \max_{\substack{j \text{ s.t. } j \neq i, \\ S_j^* \in \mathrm{St} S_i^*}} \left((2\bar{r})^{d_{i,j}} \left\| A_j^* - A_i^* \right\| {\kappa_{i,j}}^{1+\bar{d}_{i,j}} \frac{\bar{d}_{i,j}}{\bar{d}_{i,j}+1} \frac{\pi^{(\bar{d}_{i,j})/2}}{\Gamma((\bar{d}_{i,j})/2)} \right)$$

[5] For even k, $\Gamma(k/2) = (k/2)!$. For odd k, let $k' = \frac{1}{2}(k-1)$; then $\Gamma(k/2) = \Gamma(\frac{1}{2} + k') = \sqrt{\pi} \frac{(2k'+2)!}{(k+1)! 4^{k'+1}}$.

and $\bar{d}_{i,j} = d - d_{i,j}$, $l_i = \sum_{S_j^* \in \mathrm{St} S_i^*} 1$, and $\bar{r} = r_1 + 2\epsilon_c$. □

Proof of Lemma 4.3. Let the constant Δ_0 from the statement of the lemma be given by

$$\Delta_0 := \min \left\{ \frac{1}{N} \sum_{j=1}^{N} \|A_j^*\| \quad , \quad \left(\frac{12}{N} \|H^{*-1}\| \sum_{j=1}^{N} \|A_j^*\| \right)^{-1} \right\}, \tag{4.19}$$

where H^* is given by (4.20).

Solving $p' = \arg\min_x \operatorname{var} \; L(x)$ is equivalent to solving (3.1) with $\lambda = 0$. As with (3.1), a closed form expression for p' can be found by "completing the square." Specifically, $p' = H^{-1}h$,

$$H := \left(\frac{1}{N} \sum_{j=1}^{N} \hat{A}_j{}^{\mathrm{T}} \hat{A}_j \right) - \overline{\hat{A}}^{\mathrm{T}} \overline{\hat{A}},$$

$$h := \left(\frac{1}{N} \sum_{j=1}^{N} \hat{A}_j{}^{\mathrm{T}} (\hat{q}_j - \hat{A}_j p^*) \right) - \overline{\hat{A}}^{\mathrm{T}} \overline{\hat{b}},$$

where $\overline{\hat{A}} = \frac{1}{N} \sum_{j=1}^{N} \hat{A}_j$ and $\overline{\hat{b}} = \frac{1}{N} \sum_{j=1}^{N} (\hat{q}_j - \hat{A}_j p^*)$. Let $\tilde{A}_j = \hat{A}_j - A_j^*$ and $\tilde{q}_j = \hat{q}_j - q^*$. Then $H = H^* + \tilde{H}$ and $h = h^* + \tilde{h}$, with

$$H^* := \left(\frac{1}{N} \sum_{j=1}^{N} A_j^{*\mathrm{T}} A_j^* \right) - \overline{A^*}^{\mathrm{T}} \overline{A^*}, \tag{4.20}$$

$$h^* := \left(\frac{1}{N} \sum_{j=1}^{N} A_j^{*\mathrm{T}} (q^* - A_j^* p^*) \right) - \overline{A^*}^{\mathrm{T}},$$

$$\tilde{H} = \frac{1}{N} \left[\sum_{j=1}^{N} (A_j^* + \tilde{A}_j)^{\mathrm{T}} \tilde{A}_j + \sum_{j=1}^{N} \tilde{A}_j{}^{\mathrm{T}} A_j^* \right] - \overline{A^*}^{\mathrm{T}} \overline{\tilde{A}} - \overline{\tilde{A}}^{\mathrm{T}} \overline{A^*} - \overline{\tilde{A}}^{\mathrm{T}} \overline{\tilde{A}},$$

$$\tilde{h} = \frac{1}{N} \left[\sum_{j=1}^{N} (A_j^* + \tilde{A}_j)^{\mathrm{T}} (\tilde{q}_j - \tilde{A}_j p^*) + \sum_{j=1}^{N} \tilde{A}_j{}^{\mathrm{T}} (q^* - A_j^* p^*) \right] - \overline{A^*}^{\mathrm{T}} \overline{\tilde{b}} - \overline{\tilde{A}}^{\mathrm{T}} \overline{b^*} - \overline{\tilde{A}}^{\mathrm{T}} \overline{\tilde{b}},$$

where

$$\overline{A^*} = \frac{1}{N} \sum_{j=1}^{N} A_j^*, \qquad \overline{b^*} = \frac{1}{N} \sum_{j=1}^{N} (q^* - A_j^* p^*),$$

$$\overline{\tilde{A}} = \frac{1}{N} \sum_{j=1}^{N} \tilde{A}_j, \qquad \overline{\tilde{b}} = \frac{1}{N} \sum_{j=1}^{N} (\tilde{q}_j - \tilde{A}_j p^*).$$

Notice that H^* and h^* depend only on A_j^*, p^*, and q^*. Moreover, since all functions in L^* go through (p^*, q^*), it must be that $p^* = \arg\min_x \operatorname{var} \; L^*(x)$, and thus $p^* =$

${H^*}^{-1}h^*$. Rewriting $p' = H^{-1}h$, gives $(H^* + \tilde{H})\,(p^* + (p' - p^*)) = h^* + \tilde{h}$. Applying $H^*p^* = h^*$ and solving for $p' - p^*$ yields

$$p' - p^* = \left(H^* + \tilde{H}\right)^{-1}\left(\tilde{h} - \tilde{H}p^*\right).$$

From the hypothesis it follows that $\|\tilde{A}_j\| < \Delta$ and $\|\tilde{q}_j\| < \Delta$. Applying these bounds and the properties of norms, it follows after some computation that

$$\begin{aligned}\left\|\tilde{H}\right\| &\le 2\Delta^2 + \frac{4\Delta}{N}\sum_{j=1}^{N}\left\|A_j^*\right\|,\\ \left\|\tilde{h}\right\| &\le 2\left(1 + \|p^*\|\right)\Delta^2 + \frac{2\Delta}{N}\sum_{j=1}^{N}\left[\left\|A_j^*\right\|\left(1 + \|p^*\|\right) + \|q^*\| + \left\|A_j^*\right\|\|p^*\|\right].\end{aligned}$$

Since $\Delta < \Delta_0$ and by definition $\Delta_0 \le \frac{1}{N}\sum_{j=1}^{N}\left\|A_j^*\right\|$, the above bounds may be further simplified to

$$\left\|\tilde{H}\right\| < \left(\frac{6}{N}\sum_{j=1}^{N}\left\|A_j^*\right\|\right)\Delta, \tag{4.21}$$

$$\left\|\tilde{h}\right\| < \left(\frac{2}{N}\sum_{j=1}^{N}\left[2\left\|A_j^*\right\|\left(1 + \|p^*\|\right) + \|q^*\| + \left\|A_j^*\right\|\|p^*\|\right]\right)\Delta. \tag{4.22}$$

Also by definition $\Delta_0 \le (\frac{12}{N}\|{H^*}^{-1}\|\sum_{j=1}^{N}\|A_j^*\|)^{-1}$, so the bound in (4.21) can be simplified to $\|\tilde{H}\| < \frac{1}{2\|{H^*}^{-1}\|}$, and thus $\|{H^*}^{-1}\tilde{H}\| < \frac{1}{2}$. From [19], if $M \in \mathbb{R}^{n\times n}$ and $\|M\| < 1$, then $I - M$ is nonsingular and $\|(I - M)^{-1}\| \le \frac{1}{1-\|M\|}$. Some computation using this fact and the bound on $\|{H^*}^{-1}\tilde{H}\|$ provides

$$\left\|(H^* + \tilde{H})^{-1}\right\| < 2\left\|{H^*}^{-1}\right\|.$$

So then

$$\begin{aligned}\|p' - p^*\| &\le \left\|\left(H^* + \tilde{H}\right)^{-1}\right\|\left\|\tilde{h} - \tilde{H}p^*\right\|\\ &\le 2\left\|{H^*}^{-1}\right\|\left(\left\|\tilde{h}\right\| + \left\|\tilde{H}\right\|\|p^*\|\right)\\ &< c_2\Delta,\end{aligned}$$

$$\text{where} \qquad c_2 := \frac{4\left\|{H^*}^{-1}\right\|}{N}\sum_{j=1}^{N}\left(6\left\|A_j^*\right\|\|p^*\| + 2\left\|A_j^*\right\| + \|q^*\|\right), \tag{4.23}$$

which is the first part of the desired result. Applying this bound and the definition of q', we find after some computation that

$$\|q' - q^*\| < c_3\Delta,$$

$$\text{where} \quad c_3 := 1 + \frac{2c_2}{N} \sum_{j=1}^{N} \left\| A_j^* \right\| \tag{4.24}$$

which completes the desired result. □

Proof of Proposition 4.4. Let

$$\Delta_0^m = \min_{\substack{i \text{ s.t.} \\ p_i^* \in P^*}} \left\{ \frac{1}{N_i} \sum_{j=1}^{N_i} \left\| A_{i_j}^* \right\| \quad , \quad \left(\frac{12}{N_i} \left\| H_i^{*-1} \right\| \sum_{j=1}^{N_i} \left\| A_{i_j}^* \right\| \right)^{-1} \right\}, \tag{4.25}$$

where $S_{i_1}^*, \dots, S_{i_{N_i}}^*$ are the N_i d-simplices in $\mathrm{St}\{p_i^*\}$, and

$$H_i^* = \left(\frac{1}{N_i} \sum_{j=1}^{N_i} A_{i_j}^{*\,\mathrm{T}} A_{i_j}^* \right) - \left(\frac{1}{N_i} \sum_{j=1}^{N_i} A_{i_j}^* \right)^{\mathrm{T}} \left(\frac{1}{N_i} \sum_{j=1}^{N_i} A_{i_j}^* \right).$$

We will examine the effect of a single iteration of the `minvar` algorithm on $p_i \in P$, $q_i \in Q$. The results will hold independently of i, giving the desired result.

The first stage of the `minvar` algorithm calculates the least squares projection $\Pi(f_{\mathcal{P}}^*|_{S_i})$ in each d-simplex S_i of the approximation (step 2 of the algorithm; since this proposition addresses the approximation version of the problem, there is no partitioning of data to be performed in step 1). Let $S_{i_1}, \dots, S_{i_{N_i}}$ be the d-simplices in $\mathrm{St}\{p_i\}$. Since $f_{\mathcal{P}}^*$ and $f_{\mathcal{P}}$ are parameterized with the same abstract simplicial complex $\mathcal{S}^*$, $S_{i_1}^*, \dots, S_{i_{N_i}}^*$ are the d-simplices in $\mathrm{St}\{p_i^*\}$. Let $f_{\mathcal{P}}^*|_{S_{i_j}^*}(x) = A_{i_j}^*(x - p_i^*) + q_i^*$ and $\Pi(f_{\mathcal{P}}^*|_{S_i}) = \hat{A}_{i_j}(x - p_i^*) + \hat{q}_{i_j}$. Since $\epsilon < \epsilon_d \leq \epsilon_c$, it follows from Lemma 4.2 that for each $j = 1, \dots, N_i$,

$$\left\| \begin{bmatrix} \hat{A}_{i_j}{}^{\mathrm{T}} - A_{i_j}^{*\,\mathrm{T}} \\ \hat{q}_{i_j}{}^{\mathrm{T}} - q_i^{*\,\mathrm{T}} \end{bmatrix} \right\| < c_{1,i_j} \epsilon^2, \tag{4.26}$$

where c_{1,i_j} is given by (4.18). Let

$$c_1 = \max_{i=1,..,N} c_{1,i}, \tag{4.27}$$

where N is the total number of d-simplices in $\mathcal{T}(P, \mathcal{S}^*)$. Then for $j = 1 \dots, N_i$,

$$\left\| \begin{bmatrix} \hat{A}_{i_j}{}^{\mathrm{T}} - A_{i_j}^{*\,\mathrm{T}} \\ \hat{b}_{i_j}{}^{\mathrm{T}} - b_{i_j}^{*\,\mathrm{T}} \end{bmatrix} \right\| < c_1 \epsilon^2. \tag{4.28}$$

Step 3 of `minvar` moves the knots taking $(p_i, q_i) \to (p_i', q_i')$. Since $\epsilon < \epsilon_d \leq \sqrt{\Delta_0^m / c_1}$ by hypothesis, it follows that $c_1 \epsilon^2 < \Delta_0^m$. Thus, (4.28) implies that the bound in (4.6) is satisfied, permitting application of Lemma 4.3, which gives

$$\| p_i' - p_i^* \| < c_{2,i} c_1 \epsilon^2,$$
$$\| q_i' - q_i^* \| < c_{3,i} c_1 \epsilon^2,$$

where

$$c_{2,i} := \frac{4 \left\| H_i^{*-1} \right\|}{N_i} \sum_{j=1}^{N_i} \left(6 \left\| A_{i_j}^* \right\| \| p_i^* \| + 2 \left\| A_{i_j}^* \right\| + \| q_i^* \| \right), \tag{4.29}$$

$$c_{3,i} := 1 + \frac{2c_{2,i}}{N_i} \sum_{j=1}^{N_i} \left\| A_{i_j}^* \right\|. \tag{4.30}$$

For each p_i, q_i we can compute such bounds. Let

$$c_4 := c_1 \max_i c_{2,i}, \tag{4.31}$$

$$c_5 := c_1 \max_i c_{3,i}. \tag{4.32}$$

Then, for all i, p_i' and q_i' satisfy

$$\|p_i' - p_i^*\| < c_4\epsilon^2, \tag{4.33}$$

$$\|q_i' - q_i^*\| < c_5\epsilon^2, \tag{4.34}$$

which is the desired result. □

Proof of Proposition 4.5. First we establish by induction that for $k \geq 1$,

$$\frac{1}{c_4}\left(c_4\epsilon\right)^{2^k} < \epsilon_d, \tag{4.35}$$

$$\left\|p_i^{(k)} - p_i^*\right\| < \frac{1}{c_4}\left(c_4\epsilon\right)^{2^k}, \tag{4.36}$$

$$\left\|q_i^{(k)} - q_i^*\right\| < \frac{c_5}{c_4}\left(c_4\epsilon\right)^{2^k}. \tag{4.37}$$

For $k = 1$, $c_4\epsilon^2 < \epsilon_d$ since $\epsilon < \epsilon_0 \leq \sqrt{\epsilon_d/c_4}$. For $k > 1$, $\frac{1}{c_4}\left(c_4\epsilon\right)^{2^k} < \epsilon_d$ by the induction hypothesis. Moreover, $c_4\epsilon < 1$ since $\epsilon < \epsilon_0 \leq \frac{1}{c_4}$. Then $\frac{1}{c_4}(c_4\epsilon)^{2^{(k+1)}} = (\frac{1}{c_4}\left(c_4\epsilon\right)^{2^k})\left(c_4\epsilon\right)^{2^k} < \epsilon_d$. This establishes (4.35). Since $\epsilon < \epsilon_0 \leq \epsilon_d$, it follows that for $k = 1$, after a single iteration of the `minvar` algorithm, (4.36) and (4.37) will hold by Proposition 4.4. For $k > 1$, for all i, $\|p_i^{(k)} - p_i^*\| < \frac{1}{c_4}\left(c_4\epsilon\right)^{2^k}$ by the induction hypothesis. From (4.35), proven above, $\frac{1}{c_4}\left(c_4\epsilon\right)^{2^k} < \epsilon_d$. Since for all i, $\|p_i^{(k)} - p_i^*\| < \epsilon_d$, it follows from Proposition 4.4 that

$$\left\|p_i^{(k+1)} - p_i^*\right\| < c_4\left(\frac{1}{c_4}\left(c_4\epsilon\right)^{2^k}\right)^2 = \frac{1}{c_4}\left(c_4\epsilon\right)^{2^{(k+1)}},$$

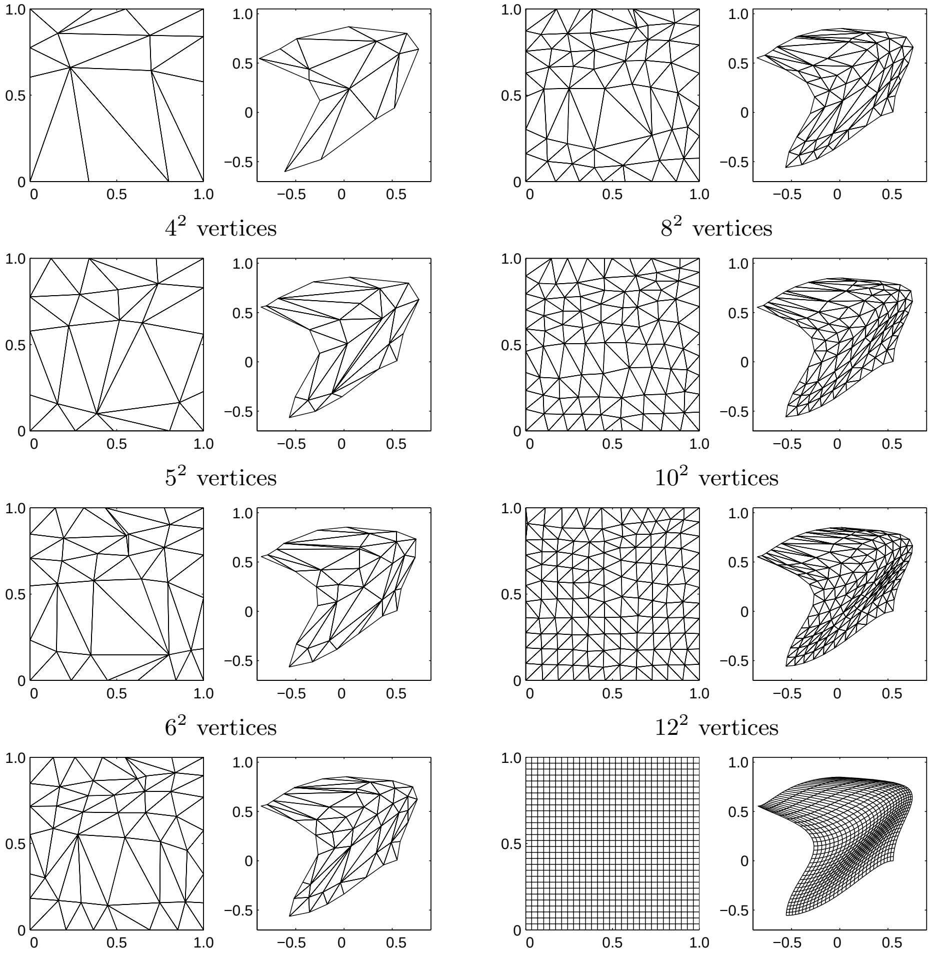
$$\left\|q_i^{(k+1)} - q_i^*\right\| < c_5\left(\frac{1}{c_4}\left(c_4\epsilon\right)^{2^k}\right)^2 = \frac{c_5}{c_4}\left(c_4\epsilon\right)^{2^{(k+1)}},$$

which establishes (4.36) and (4.37). Since $c_4\epsilon < 1$, as argued above, it follows that (4.36) and (4.37) go to 0 as k goes to infinity. □

5. Numerical example. This section presents an example of the `minvar` algorithm's performance on a "test function," $f : [0,1]^2 \to \mathbb{R}^2$, given by

$$f(x) = \begin{bmatrix} \tanh\frac{5}{8}\left(2x_1 - 4{x_2}^4 + 3{x_2}^2 - 1\right) \\ \tanh\frac{5}{8}\left(2{x_1}^2 - \ \ {x_1}^4 + 2x_2 - 1\right) \end{bmatrix}, \tag{5.1}$$

which is invertible over the domain $D = [0,1]^2$. The implementation of `minvar` constructs an approximation to a discrete set of data and includes constrained motion of boundary vertices as well as data dependent retriangulation. Since the test function is neither piecewise linear nor directly available, Theorem 4.1 provides no performance guarantees, but good performance under these circumstances suggests `minvar`'s broader applicability. Two sets of numerical studies are presented. The first



4^2 vertices

8^2 vertices

5^2 vertices

10^2 vertices

6^2 vertices

12^2 vertices

7^2 vertices

test function

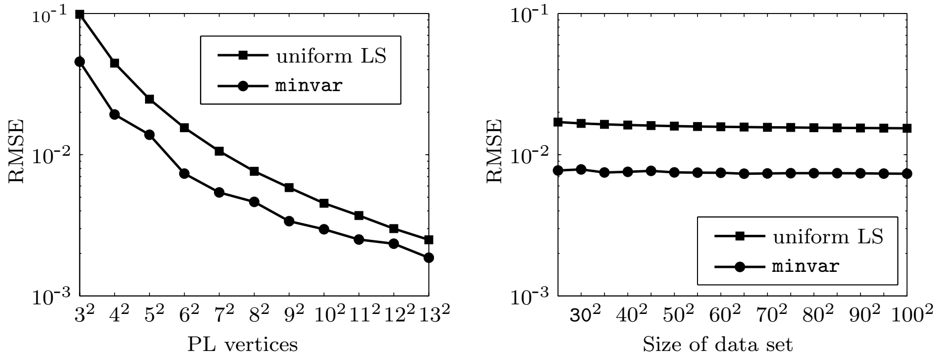
Fig. 5.1. *Visualizations of the test function (lower right) and a series of PL approximations to the test function. In each subfigure, the domain is displayed on the left and the range on the right. Note that the approximations are invertible, since there are no tangles in the range triangulations.*

examines the effects of varying the number of vertices in the PL approximation, and the second examines how data set size affects approximations with a fixed number of vertices.

The first set of experiments fits PL approximations of differing sizes to a single data set. The data set was generated by sampling the test function on an 80×80 uniform grid over the domain. For each $n = 2, \dots, 13$, three different PL approximations with n^2 vertices were computed: (i) the least squares continuous PL approximation on a fixed uniform triangulation of the domain (referred to as "uniform LS"), (ii) `minvar`, initialized on a uniform triangulation of the domain (referred to as "`minvar`"), and (iii) the least squares continuous PL approximation on the final triangulation from `minvar` (referred to as "`minvar` LS"). Recall that when the domain triangulation is fixed, the least squares continuous PL approximation problem becomes linear-in-parameters and the solution can be computed directly [7, 35]. Figure 5.1 shows the test function and several exemplars of the `minvar` approximations. Table 5.1 and

Table 5.1
RMSE of approximations by uniform LS, `minvar`*, and* `minvar` *LS.*

Vertices	Uniform LS	`minvar`	`minvar` LS
2^2	1.79801e-01	1.79818e-01	1.79801e-01
3^2	9.85617e-02	4.56123e-02	4.48297e-02
4^2	4.45294e-02	1.92465e-02	1.89605e-02
5^2	2.47517e-02	1.38427e-02	1.36523e-02
6^2	1.55282e-02	7.35190e-03	7.25304e-03
7^2	1.05933e-02	5.40420e-03	5.34596e-03
8^2	7.63439e-03	4.63040e-03	4.56706e-03
9^2	5.84815e-03	3.38636e-03	3.34218e-03
10^2	4.53419e-03	2.96688e-03	2.92793e-03
11^2	3.71421e-03	2.50778e-03	2.47792e-03
12^2	2.99647e-03	2.34302e-03	2.32279e-03
13^2	2.49846e-03	1.86386e-03	1.84316e-03



(a) (b)

Fig. 5.2. *RMSE performance of* `minvar` *compared to a least squares continuous PL approximation on a uniform triangulation.* (a) *Vertices in approximation vs. RMSE for approximations to the* 80×80 *data set.* (b) *Density of training data set vs. RMSE on validation data for approximations with* 6^2 *vertices.*

Figure 5.2(a) show the root mean square error (RMSE) of the approximations as a function of the number of vertices. For 2^2 domain vertices, all of them are on the corners of the domain and must remain fixed, so `minvar` can change only the range vertices. Since `minvar` is not guaranteed to give the least squares continuous PL approximation for a given triangulation, it is not surprising that the uniform LS approximation's RMSE is slightly lower than `minvar`'s in this case. The RMSE difference between `minvar` and `minvar` LS approximations is less than 2%. Least squares could be applied as a post processing step to `minvar`, but since the differences are relatively small, this might not be necessary in application settings. From the triangulations of the `minvar` approximations in Figure 5.1, the domain triangulations move farther for lower numbers of vertices. As the number of vertices increases, the triangulations visually seem to deviate less from the initial uniform triangulation. The RMSE performance of `minvar` reflects this, giving the biggest reductions in RMSE as compared to uniform LS for triangulations with 3^2 to 6^2 vertices. Since this study uses initial conditions in which the vertices are on a uniform grid, approximations with large numbers of vertices may be getting caught in local minima near their initial conditions. In this case, performance could be improved by refining converged less dense approximations to create initial conditions for the dense approximations [39].

In the second set of experiments, PL approximations with 6^2 vertices were trained using data sets of varying size. The PL approximations were chosen to have 6^2 vertices because, as mentioned above, this is in the region of sizes where the performance gains from `minvar` are greatest. The data sets were generated by sampling the test function on a uniform $n \times n$ grid, $n = 25, 30, 35, \dots, 100$. The approximations were compared using a validation data set generated by evaluating the test function at 1000 points sampled from a uniform probability distribution over the domain. Figure 5.2(b) shows the validation set RMSE for `minvar` and uniform LS. With a 20×20 data set, `minvar` fails to run with a 6^2 vertex approximation, because as the vertices move, several simplices shrink to the point that they do not contain enough data to make the linear least squares approximation unique. Data sparsity is a serious issue in this type of local approximation.

From this example, `minvar` shows marked benefit when the approximation has relatively few vertices compared to the complexity of the test function. We expect that `minvar`'s performance on higher order (more vertices) approximations could be improved by seeding initial conditions based on lower order approximations. The algorithm produces a consistent approximation to variously sized data sets, so long as there is enough data for it to run.

6. Conclusion.

Numerous applications require the simultaneous approximation of a function and its inverse from a set of discrete data. While there is a substantial literature on function approximation, very little of it addresses the constraint of invertibility. The inverse of a continuous PL function can be computed in closed form, which is ideal for applications requiring the approximation of a function and its inverse. In the PL literature, the partition is often fixed, in which case the minimum squared error approximation problem is linear-in-parameters. The problem becomes nonlinear-in-parameters when the domain partition is allowed to move.

The `minvar` algorithm is a novel method for computing continuous PL approximations to data. Rather than using gradient descent on the parameters, `minvar` takes advantage of the structure of PL functions, iteratively moving the vertices of the approximation based on local least squares fits. The `minvar` algorithm is proven to converge locally in the special case when the data generating function is itself PL and available directly rather than through discrete data. While this result seems very natural, complexity in the proof arises from the interaction of the domain triangulations of the data generating function and its approximation. Indeed, many difficulties in constructing PL approximations, such as triangulation tangles, arise primarily from the combinatorial properties of PL functions. For general approximation problems, this added complexity may cause PL approximation to appear less attractive than other nonlinear-in-parameters approximation techniques, but for an important subset of applications the PL function's closed form invertibility makes the combinatorial complexity cost effective.

The present work can be extended in several directions. The analysis here addresses the approximation rather than the estimation version of the problem. A formal connection between the approximation and estimation versions could be constructed in the appropriate statistical framework. Similarly, there is no study of the effects of noise on the convergence properties of `minvar`. Since data from the data generating function are used only for computing the least squares approximations, and least squares has good noise properties, the authors expect that the `minvar` algorithm will also have good noise properties. The present analytical work considers only PL data generating functions. A desirable extension would be to show that the algorithm converges to the locally best PL approximation, given that one exists. For scalar functions there are generalized convexity conditions that characterize existence and uniqueness of (possibly discontinuous) PL approximations [4, 18, 26], but the authors are unaware of similar results for dimensions greater than one or with continuous piecewise approximations. Numerical experience suggests that `minvar` does have good convergence properties on other types of functions. Practical application of `minvar` raises a number of challenging issues such as constrained motion of vertices on the boundary of the domain, methods to avoid and correct triangulation tangling, and retriangulation or adaptation of the combinatorial parameters of the PL approximation. Due to space limitations, these topics will be addressed in detail in a subsequent paper on the use of `minvar` in engineering applications.

Appendix A. Geometric properties of simplices and triangulations. Proving properties of the `minvar` algorithms requires some insight into the underlying geometric structures upon which PL functions are constructed. This appendix presents a number of geometry facts used in the proof of convergence. Proofs of these facts are available in [20].

A.1. Barycentric coordinates as distances. It is often convenient to represent points in $\mathbb{R}^d$ using barycentric coordinates with respect to the vertices of some d-simplex. Let $p_1, \ldots, p_{d+1}$ be affinely independent points in $\mathbb{R}^d$. Let $x \in \mathbb{R}^d$ and let $\bar{\alpha} = \begin{bmatrix}\alpha_1 & \cdots & \alpha_{d+1}\end{bmatrix}^{\mathrm{T}}$ be such that $x = \sum_{i=1}^{d+1} \alpha_i p_i$ and $\sum_{i=1}^{d+1} \alpha_i = 1$. $\bar{\alpha}$ are called the *barycentric coordinates* of x with respect to $p_1, \ldots, p_{d+1}$. If $\bar{\alpha} \in \Delta_{d+1} := \{\bar{\alpha} \in \mathbb{R}^{d+1} | \alpha_i \geq 0, \sum_{i=1}^{d+1} \alpha_i = 1\}$, then $x \in \operatorname{conv}(p_1, \ldots, p_{d+1})$, and Δ_d is called the standard d-simplex.

The distance between a point $x \in \mathbb{R}^d$ and a nonempty set $A \subseteq \mathbb{R}^d$ is well defined [41] and written as $\delta(x, A) := \inf_{z \in A} \|x - z\|$. Let $H_i = \operatorname{aff}(\{p_1, \ldots, p_{d+1}\} - \{p_i\})$, the hyperplane opposing p_i. Let (a_i, c_i) be an implicit representation for H_i, that is, $H_i = \left\{z \in \mathbb{R}^d | {a_i}^{\mathrm{T}} z + c_i = 0\right\}$. The distance of a point $x \in \mathbb{R}^d$ to H_i is given by

$$\delta(x, H_i) = \frac{\left|{a_i}^{\mathrm{T}} x + c_i\right|}{\|a_i\|}.$$

We define $\delta_s(x, H_i)$ as the *signed distance* of x from H_i. That is, $|\delta_s(x, H_i)| = \delta(x, H_i)$, and $\delta_s(x, H_i) > 0$ for x on the same side of H_i as p_i, and $\delta_s(x, H_i) < 0$ for x on the opposite side of H_i as p_i.

By the following claim, the barycentric coordinates of x can be interpreted as the scaled distances of x from hyperplanes $H_1, \ldots, H_{d+1}$.

Claim 1. *Let $x \in \mathbb{R}^d$. Let $\bar{\alpha} = \begin{bmatrix}\alpha_1 & \cdots & \alpha_{d+1}\end{bmatrix}^{\mathrm{T}}$ be the barycentric coordinates of x with respect to the affinely independent points $p_1, \ldots, p_{d+1} \in \mathbb{R}^d$. Then $\delta_s(x, H_i) = \alpha_i \delta(p_i, H_i)$.*

Similarly, the barycentric coordinates of x can be used to measure the distance to an affine subspace that is the intersection of two or more of $H_1, \ldots, H_{d+1}$.

Claim 2. *Let $x \in \mathbb{R}^d$. Let $\bar{\alpha} = \begin{bmatrix}\alpha_1 & \cdots & \alpha_{d+1}\end{bmatrix}^{\mathrm{T}}$ be barycentric coordinates of x with respect to the affinely independent points $p_1, \ldots, p_{d+1}$. The distance from x to the affine subspace $A = \operatorname{aff}(\{p_1, \ldots p_k\})$ is given by $\delta(x, A) = {\bar{\alpha}_s}^{\mathrm{T}} G \bar{\alpha}_s$, where $\bar{\alpha}_s = \begin{bmatrix}\alpha_{k+1} & \alpha_{k+2} & \cdots & \alpha_{d+1}\end{bmatrix}^{\mathrm{T}}$ and $G \in \mathbb{R}^{(d-k+1) \times (d-k+1)}$ is a positive definite matrix whose entries depend only on $p_1, \ldots, p_k$.*

The explicit form for G, which derives from the Gram determinants used in the proof, is provided in [20].

A.2. Properties of the dilation. The dilation arises in Lemma 4.2 in measuring how the approximation's mismatched triangulation affects the least squares affine approximations. The first claim establishes the general properties of the dilation, and the second claim establishes a relationship between incident d-simplices that is required for the proof of Lemma 4.2.

CLAIM 3. *Let $S \subseteq \mathbb{R}^d$ be a d-simplex with vertices $p_1, \ldots, p_{d+1}$. Let*

$$\epsilon_{\min} = -\left(\sum_{i=1}^{d+1} 1/\delta(p_i, H_i)\right)^{-1}, \tag{A.1}$$

where H_i is the opposing hyperplane to p_i, as defined above. $\mathcal{D}(S, \epsilon_{\min})$ is a single point. For $\epsilon > \epsilon_{\min}$, $\mathcal{D}(S, \epsilon)$ is a d-simplex, with faces parallel to and translated distance $|\epsilon|$ away from the faces of S. For $\epsilon_{\min} \leq \epsilon \leq 0$, $\mathcal{D}(S, \epsilon) \subseteq S$, while for $\epsilon \geq 0$, $S \subseteq \mathcal{D}(S, \epsilon)$.

CLAIM 4. *Let $S_a, S_b \subseteq \mathbb{R}^d$ be incident d-simplices, that is, $S_a \cap S_b = s_{ab}$, where $s_{ab} \leq S_a, S_b$ is a $(k-1)$-simplex, $1 \leq k < d-1$. Let $\operatorname{vert} S_a = \{p_1, \ldots, p_{d+1}\}$, $\operatorname{vert} S_b = \{p_1, \ldots, p_k, q_{k+1}, \ldots, q_{d+1}\}$, and $\operatorname{vert} s_{ab} = \{p_1, \ldots, p_k\}$. Let $A = \operatorname{aff} s_{ab}$. Then $\exists \kappa_{a,b} > 0$ such that for all $\epsilon > 0$, if $x \in \mathcal{D}(S_a, \epsilon) \cap S_b$, then $\delta(x, A) < \kappa_{a,b}\epsilon$.*

A.3. Properties of PL functions. Claims pertaining to the parameterization, continuity, and invertibility of PL functions are provided below.

CLAIM 5. *Let $f_{\mathcal{P}}$ be a continuous PL function parameterized by $(P, Q, \mathcal{S})$. Let $S_i, S_j \in \mathcal{T}(P, \mathcal{S})$ be such that $S_i \cap S_j \neq \emptyset$. Let $S_i \cap S_j$ be a $(k-1)$-simplex, so then S_i and S_j share k vertices in common. Let $\operatorname{vert} S_i = \{p_{i_1}, \ldots, p_{i_{d+1}}\}$, $\operatorname{vert} S_j = \{p_{i_1}, \ldots, p_{i_k}, p_{j_{k+1}}, \ldots, p_{j_{d+1}}\}$, and $\operatorname{vert} S_i \cap S_j = \{p_{i_1}, \ldots, p_{i_k}\}$. Then,*

1. *$f_{\mathcal{P}}$ in S_i and S_j is given by*

$$f_{\mathcal{P}}\big|_{S_i}(x) = U_i V_i^{-1} \begin{bmatrix} x^{\mathrm{T}} & 1 \end{bmatrix}^{\mathrm{T}}, \qquad f_{\mathcal{P}}\big|_{S_j}(x) = U_j V_j^{-1} \begin{bmatrix} x^{\mathrm{T}} & 1 \end{bmatrix}^{\mathrm{T}},$$
$$U_i = \begin{bmatrix} q_{i_1} & \cdots & q_{i_{d+1}} \end{bmatrix}, \qquad U_j = \begin{bmatrix} q_{i_1} & \cdots & q_{i_k} & q_{j_{k+1}} & \cdots & q_{j_{d+1}} \end{bmatrix},$$
$$V_i = \begin{bmatrix} p_{i_1} & \cdots & p_{i_{d+1}} \\ 1 & & 1 \end{bmatrix}, \qquad V_j = \begin{bmatrix} p_{i_1} & \cdots & p_{i_k} & p_{j_{k+1}} & \cdots & p_{j_{d+1}} \\ 1 & & 1 & 1 & & 1 \end{bmatrix}.$$

Then $U_i V_i^{-1} \begin{bmatrix} x^{\mathrm{T}} & 1 \end{bmatrix}^{\mathrm{T}} = U_j V_j^{-1} \begin{bmatrix} x^{\mathrm{T}} & 1 \end{bmatrix}^{\mathrm{T}}$ for $x \in \operatorname{aff}(S_i \cap S_j)$.

2. *In nonhomogeneous form, let $f_{\mathcal{P}}\big|_{S_i}(x) = A_i x + b_i$ and $f_{\mathcal{P}}\big|_{S_j}(x) = A_j x + b_j$. Let L be the linear subspace parallel to $\operatorname{aff}(S_i \cap S_j)$. Then $L \subseteq \mathcal{N}(A_i - A_j)$.*

CLAIM 6 (continuity in vertices). *Consider two continuous PL functions, $f_{\mathcal{P}}^*$ parameterized by $\mathcal{P}^* = (P^*, Q^*, \mathcal{S}^*)$ and $f_{\mathcal{P}}$ parameterized by $(P, Q, \mathcal{S}^*)$, such that $|\mathcal{T}(P, \mathcal{S}^*)| = |\mathcal{T}(P^*, \mathcal{S}^*)|$. Let $c > 0$. There exists $c' = c'(\mathcal{P}^*, c)$ such that, for $0 < \epsilon < r_3$, where*

$$r_3 = \min_{S^* \in \mathcal{T}(P^*, \mathcal{S}^*)} \sup \left\{ \epsilon \,\middle|\, \mathcal{D}(S^*, \epsilon) \subseteq |\operatorname{Cl} \operatorname{St} S^*| \right\}. \tag{A.2}$$

If $\|p_i - p_i^*\| < \epsilon$ and $\|q_i - q_i^*\| < c\epsilon$ for all i, then

$$\|f_{\mathcal{P}} - f_{\mathcal{P}}^*\|_\infty < c'\epsilon.$$

That is, a continuous PL function is continuous in its vertices.

The explicit form for c' is provided in [20].

CLAIM 7. *Let $f_{\mathcal{P}}$ be a PL function parameterized by $\mathcal{P} = (P, Q, \mathcal{S})$. If $\mathcal{T}(Q, \mathcal{S})$ is also a triangulation, then the PL function is invertible on its range, and the inverse, ${f_{\mathcal{P}}}^{-1}$, is parameterized by $\mathcal{P}^{-1} = (Q, P, \mathcal{S})$.*

REFERENCES

[1] C. G. ATKESON, A. W. MOORE, AND S. SCHAAL, *Locally weighted learning*, Artificial Intelligence Rev., 11 (1997), pp. 11–73.

[2] M. J. BAINES, *Algorithms for optimal discontinuous piecewise linear and constant L_2 fits to continuous functions with adjustable nodes in one and two dimensions*, Math. Comp., 62 (1994), pp. 645–669.

[3] A. R. BARRON, *Universal approximation bounds for superpositions of a sigmoidal function*, IEEE Trans. Inform. Theory, 39 (1993), pp. 930–945.

[4] D. L. BARROW, C. K. CHUI, P. W. SMITH, AND J. D. WARD, *Unicity of best mean approximation by second order splines with variable knots*, Math. Comp., 32 (1978), pp. 1131–1143.

[5] K. Q. BROWN, *Voronoi diagrams from convex hulls*, Inform. Process. Lett., 9 (1979), pp. 223–228.

[6] E. W. CHENEY, *Introduction to Approximation Theory*, McGraw-Hill, New York, 1966.

[7] C. K. CHUI, *Multivariate Splines*, CBMS-NSF Regional Conf. Ser. in Appl. Math 54, SIAM, Philadelphia, 1988.

[8] N. J. COWAN, J. D. WEINGARTEN, AND D. E. KODITSCHEK, *Visual servoing via navigation functions*, IEEE Trans. Rob. Autom., 18 (2002), pp. 521–533.

[9] P. J. DAVIS, *Interpolation and Approximation*, Dover, New York, 1975.

[10] C. DE BOOR, K. HÖLLIG, AND S. RIEMENSCHNEIDER, *Box Splines*, Springer-Verlag, New York, 1993.

[11] N. DYN, D. LEVIN, AND S. RIPPA, *Algorithms for the construction of data dependent triangulations*, in Algorithms for Approximation, II, Chapman and Hall, London, 1990, pp. 185–192.

[12] N. DYN, D. LEVIN, AND S. RIPPA, *Data dependent triangulations for piecewise linear interpolation*, IMA J. Numer. Anal., 10 (1990), pp. 137–154.

[13] H. EDELSBRUNNER, *Geometry and Topology for Mesh Generation*, Cambridge University Press, Cambridge, UK, 2001.

[14] H. EDELSBRUNNER AND N. SHAH, *Incremental topological flipping works for regular triangulations*, Algorithmica, 15 (1996), pp. 223–241.

[15] S. FIORI, *Probability density function learning by unsupervised neurons*, Internat. J. Neural Systems, (2001), pp. 399–417.

[16] J. H. FRIEDMAN, *Multivariate adaptive regression splines*, Ann. Statist., 19 (1991), pp. 1–141.

[17] S. FURRY AND J. KAINZ, *Rapid algorithm development applied to engine management systems*, in Proceedings of the 1998 SAE International Congress & Exposition, Detroit, MI, 1998.

[18] R. C. GAYLE AND J. M. WOLFE, *Unicity in piecewise polynomial L_1-approximation via an algorithm*, Math. Comp., 65 (1996), pp. 647–660.

[19] G. E. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, MD, 1996.

[20] R. E. GROFF, P. P. KHARGONEKAR, AND D. E. KODITSCHEK, *A Local Convergence Proof for the Minvar Algorithm for Computing Continuous Piecewise Linear Approximations*, Technical report CSE-TR-464-02, University of Michigan, Ann Arbor, MI, 2002.

[21] R. E. GROFF, P. P. KHARGONEKAR, D. E. KODITSCHEK, T. T. THIERET, AND L. MESTHA, *Modeling and control of color xerographic processes*, in Proceedings of the 38th IEEE Conference on Decision and Control, Vol. 2, Phoenix, AZ, 1999, pp. 1697–1702.

[22] R. E. GROFF, D. E. KODITSCHEK, AND P. P. KHARGONEKAR, *Piecewise linear homeomorphisms: The scalar case*, in Proceedings of the International Joint Conference on Neural Networks, Como, Italy, 2000.

[23] R. E. GROFF, D. E. KODITSCHEK, P. P. KHARGONEKAR, AND T. T. THIERET, *Representation of color space transformations for effective calibration and control*, in IS&T's NIP16: International Conference on Digital Printing Technology, Society for Imaging Science and Technology, Vancouver, BC, Canada, 2000, pp. 255–260.

[24] N. HAI, J. NI, AND J. YUAN, *Generalized model formulation technique for error synthesis and error compensation on machine tools*, in Proceedings of the NAMRX XXVI Conference, Society of Manufacturing Engineers, Atlanta, GA, 1998.

[25] J. B. KIOUSTELIDIS, *Optimal segmented approximations*, Computing, 24 (1980), pp. 1–8.

[26] J. B. Kioustelidis, *Optimal segmented polynomial L_s-approximations*, Computing, 26 (1981), pp. 239–246.

[27] D. E. Koditschek and E. Rimon, *Robot navigation functions on manifolds with boundary*, Adv. in Appl. Math., 11 (1990), pp. 412–442.

[28] C. L. Lawson, *Properties of n-dimensional triangulations*, Comput. Aided Geom. Design, 3 (1986), pp. 231–246.

[29] G. G. Lorentz, *Approximation of Functions*, Holt, Rhinehart and Winston, New York, 1966.

[30] F. P. Preparata and M. I. Shamos, *Computational Geometry: An Introduction*, Springer-Verlag, New York, 1985.

[31] M. H. Raibert, *Legged Robots That Balance*, MIT Press, Cambridge, MA, 1986.

[32] E. Rimon and D. E. Koditschek, *The construction of analytic diffeomorphisms for exact robot navigation on star worlds*, Trans. Amer. Math. Soc., 327 (1991), pp. 71–115.

[33] E. Rimon and D. E. Koditschek, *Exact robot navigation using artificial potential fields*, IEEE Trans. Rob. Autom., 8 (1992), pp. 501–518.

[34] W. J. Schwind and D. E. Koditschek, *Approximating the stance map of a 2 dof monoped runner*, J. Nonlinear Sci., 10 (2000), pp. 533–568.

[35] E. V. Shikin and A. I. Plis, *Handbook on Splines for the User*, CRC Press, Boca Raton, FL, 1995.

[36] E. H. Spanier, *Algebraic Topology*, McGraw-Hill, New York, 1966.

[37] M. Spivak, *Calculus on Manifolds*, Westview Press, Boulder, CO, 1965.

[38] C. J. Stone, M. H. Hansen, C. Kooperberg, and Y. K. Truong, *Polynomial splines and their tensor products in extended linear modeling*, Ann. Statist., 25 (1997), pp. 1371–1470.

[39] Y. Tourigny and M. J. Baines, *Analysis of an algorithm for generating locally optimal meshes for L_2 approximation by discontinuous piecewise polynomials*, Math. Comp., 66 (1997), pp. 623–650.

[40] Y. Tourigny and F. Hülsemann, *A new moving mesh algorithm for the finite element solution of variational problems*, SIAM J. Numer. Anal., 35 (1998), pp. 1416–1438.

[41] R. Webster, *Convexity*, Oxford University Press, New York, 1994.

[42] J. C. Ziegert and P. Datseris, *Basic considerations for robot calibration*, International Journal of Robotics and Automation, 4 (1989), pp. 158–166.

SIAM J. Numer. Anal.
Vol. 41, No. 3, pp. 1008–1021

CONSISTENCY OF GENERALIZED FINITE DIFFERENCE SCHEMES FOR THE STOCHASTIC HJB EQUATION*

J. FRÉDÉRIC BONNANS† AND HOUSNAA ZIDANI‡

Abstract. We analyze a class of numerical schemes for solving the HJB equation for stochastic control problems, which enters the framework of Markov chain approximations and generalizes the usual finite difference method. The latter is known to be monotonic, and hence valid, only if the scaled covariance matrix is dominant diagonal. We generalize this result by, given the set of neighboring points allowed to enter the scheme, showing how to compute effectively the class of covariance matrices that is consistent with this set of points. We perform this computation for several cases in dimensions 2, 3, and 4.

Key words. stochastic control, finite differences, viscosity solutions, consistency, HJB equation

AMS subject classifications. 93E20, 49L99

DOI. 10.1137/S0036142901387336

1. Motivation.

This paper is devoted to the discussion of numerical algorithms for solving stochastic optimal control problems. In order to simplify the presentation of the main ideas, consider the following model problem (Fleming and Rishel [4] and Lions and Bensoussan [7]):

$$(P_x) \qquad \begin{cases} \operatorname{Min} W(x,u) = \mathbb{E}\int_0^\infty \ell(y_{x,u}(t), u(t))e^{-\lambda t}\mathrm{d}t; \\ \begin{cases} \mathrm{d}y_{x,u}(t) = f(y_{x,u}(t), u(t))\mathrm{d}t + \sigma(y_{x,u}(t), u(t))\mathrm{d}w(t), \\ y_{x,u}(0) = x, \end{cases} \\ u(t) \in U, \quad t \in [0, \infty[. \end{cases}$$

Here $y_{x,u}(t) \in \mathbb{R}^n$ is the state variable, $u(t) \in \mathbb{R}^m$ is the control variable that for almost all t must belong to the set $U \subset \mathbb{R}^m$, $\lambda > 0$ is the discounting factor, $\ell : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ is a distributed cost, $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ is a deterministic dynamics, $\sigma(\cdot,\cdot)$ is a mapping from $\mathbb{R}^n \times \mathbb{R}^m$ into the space of $n \times r$ matrices, and w is a standard r dimensional Brownian motion. We are assuming full observation of the state, and we are looking for a control in the class of feedback controls. In what follows we assume f, σ, and ℓ to be Lipschitz and bounded. Then the solution to the stochastic differential equation and the associated cost are well defined (see, e.g., Fleming and Soner [5]). The covariance matrix is defined as

$$a(x,u) := \sigma(x,u)\sigma(x,u)^\top \quad \forall \ (x,u) \in \mathbb{R}^n \times \mathbb{R}^m, \tag{1.1}$$

where by $\top$ we denote the transposition operator. It is known (see Lions [8] and Fleming and Soner [5]) that the value function V of problem (P_x), defined by $V(x) =$

*Received by the editors April 4, 2001; accepted for publication (in revised form) December 6, 2002; published electronically June 18, 2003.
http://www.siam.org/journals/sinum/41-3/38733.html

†Inria-Rocquencourt, Domaine de Voluceau, BP 105, 78153 Le Chesnay, France (Frederic.Bonnans@inria.fr).

‡Unité de Mathématiques Appliquées, ENSTA, 32 Boulevard Victor, 75739 Paris Cedex 15, France (zidani@ensta.fr).

$\inf_u W(x,u)$, is the unique bounded viscosity solution of the Hamilton–Jacobi–Bellman (HJB) equation

$$\lambda V(x) \quad = \quad \mathcal{H}(x, V_x(x), V_{xx}(x)) \quad \forall\, x \in \mathbb{R}^n, \tag{1.2}$$

the Hamiltonian $\mathcal{H}$ being defined as

$$\mathcal{H}(x,p,Q) := \inf_{u\in U} \left\{ \ell(x,u) + f(x,u)\cdot p + \tfrac{1}{2} \sum_{i,j=1}^{n} a_{ij}(x,u)\, Q_{ij} \right\}, \tag{1.3}$$

where $x \in \mathbb{R}^n$, $p \in \mathbb{R}^n$, and Q is an $n\times n$ symmetric matrix. A basic idea for discretizing this problem is as follows (an up-to-date synthesis of this approach is given in Kushner and Dupuis [6]). Consider a regular grid G^h of discretization of the state space $\mathbb{R}^n$, with discretization steps $h = (h_1, \ldots, h_n)$. With the coordinate $k = (k_1, \ldots, k_n)$ in $\mathbb{Z}^n$ is associated the point $x_k \in \mathbb{R}^n$ of the form

$$x_k := (k_1 h_1, \ldots, k_n h_n). \tag{1.4}$$

Of course the real computations should be performed on a finite grid. However, we will not discuss this point, and we rather analyze the result of computations on this infinite grid. Let us consider an optimal control problem for a Markov chain on the grid G^h. Let $\{X^h_q, q \geq 0\}$ be the states of the Markov chain at time q, with transition probabilities denoted $p^h(x,y\,|\,u)$, where $u \in U$ is the canonical control value. Let Δt^h be an *interpolation interval* satisfying $\Delta t^h \to 0$ as $h \to 0$, and let $\mathbb{E}^{h,u}_{k,q}$ be the conditional expectation of X^h_{q+1}, given that $\{X^h_q = x_k\}$, and the control value u. Suppose that the chain obeys the following *local consistency* conditions:

$$\mathbb{E}^{h,u}_{k,q}\left[X^h_{q+1} - x_k\right] = \Delta t^h f(x_k,u) + o(\Delta t^h), \tag{1.5a}$$

$$\mathrm{Cov}^{h,u}_{k,q}\left[X^h_{q+1} - x_k\right] = \Delta t^h a(x_k,u) + o(\Delta t^h), \tag{1.5b}$$

$$\sup_q \left|X^h_{q+1} - X^h_q\right| \to 0. \tag{1.5c}$$

A possible adaptation for the cost function to this Markov chain is the following:

$$W^h(x,u^h) = \Delta t^h \mathbb{E}\left[\sum_{q\geq 0} \ell(X^h_q, u^h_q)(1+\lambda\Delta t^h)^{-q-1}\right], \tag{1.6}$$

where $u^h = (u^h_q)$, and $u^h_q \in U$ denote the random variable which represents the control action for the chain at discrete time q. Then the dynamic programming equation for the controlled chain $\{X^h_q, q\geq 0\}$ and the cost (1.6) is

$$V^h(x_k) = (1+\lambda\Delta t^h)^{-1} \operatorname*{Min}_{u\in U} \left[\Delta t^h \ell(x_k,u) + \sum_{y\in G^h} p(x_k, y \mid u) V^h(y)\right] \tag{1.7}$$

for $x_k \in G^h$. It is known that the function V^h converges uniformly over compact sets to the value function V for the original problem, as $h \to 0$, whenever the “local consistency” conditions (1.5) are satisfied, the interpolation interval possibly depending on (x,u); see Kushner and Dupuis [6].

Now the Markov chain approximation method consists of finding a chain $\{X_q^h\}$ satisfying the "local consistency" (1.5). A standard way for the construction of such an approximating chain is to use the finite difference approximations. However, this works only if the matrix a has a dominant diagonal (see section 3 for details), whereas this matrix may be an arbitrary semidefinite positive matrix. In some cases it is possible to make a change of variables in the state space in order for this hypothesis to be satisfied; see, e.g., Kushner and Dupuis [6, section 5.4]. However, when the control enters the matrix σ, and hence also in a, this is no longer possible in general. By contrast, the Markov chain approximation method is, in principle, able to handle the case when the covariance matrix is not dominant diagonal. In fact, relation (1.5) essentially gives linear relations (to be satisfied approximately) on the transition probabilities, while the latter have to be nonnegative and of sum equal to 1. Several questions then arise. First of all, since the Markov chain represents the discretization of a partial differential equation, it is highly desirable to limit the transitions from one point of the grid to other points that are not too far away. Also, for computational complexity reasons, the number of transitions should be as small as possible.

We are led then to the following question. Given a point in the grid with coordinate k and a control, choose a set of other points in the grid to which transitions are allowed. For instance we may allow transitions to points for which the coordinates k' are such that $|k'_i - k_i| \leq 1$ for all i. More generally, we choose a set of neighbors defined by constraints on the difference of coordinates $k' - k$. Is it possible then to compute consistent transition probabilities? In other words, what is the class of covariance matrices that is compatible with such a choice of possible transitions? And then what is the cost of computing the transition probabilities themselves? Finally, on what basis should we choose the neighbors?

These are several delicate questions. The paper is essentially devoted to the first of them, i.e., how to check the consistency condition. Note that our results apply also to finite horizon problems, in which the value function depends on time and space, since the analysis of consistency for these problems leads basically to questions of the same nature. Similarly, we discuss only explicit schemes, but implicit schemes (in connection to the policy iteration methods; see Kushner and Dupuis [6, section 6.2]) also lead to the same questions, and our results apply also to this case. Note that in the case of a covariance matrix that is a smooth function of the state only, it is possible to state a consistent approximation using finite elements; see Chung, Hanson, and Xu [3]. However, it is not easy to extend this idea to the case when the covariance matrix either is not differentiable or depends also on the control.

2. Generalized finite differences.

Let us present a generalization of the usual finite difference schemes; we will see later that these generalized finite differences are in fact a particular case of Markov chain approximation. Let $\varphi = \{\varphi_k\}$ be a real valued function over $\mathbb{Z}^n$. With $\xi \in \mathbb{Z}^n$, associate the *shift* operator δ_ξ defined by $\delta_\xi \varphi_k := \varphi_{\xi+k}$. Consider the finite difference operator $\Delta_\xi = \delta_\xi + \delta_{-\xi} - 2\delta_0$; in other words,

$$\Delta_\xi \varphi_k := \varphi_{k+\xi} + \varphi_{k-\xi} - 2\varphi_k = \varphi_{k+\xi} - \varphi_k - (\varphi_k - \varphi_{k-\xi}). \tag{2.1}$$

If Φ is a C^2 (twice continuously differentiable) function over $\mathbb{R}^n$, and $\varphi_k = \Phi(x_k)$ for all k, then by a standard Taylor expansion we have that

$$\Delta_\xi \varphi_k := \sum_{i,j=1}^{n} h_i h_j \xi_i \xi_j \Phi_{x_i x_j} + o(\|h\|^2). \tag{2.2}$$

For instance, when ξ is equal to e_i (the ith element of the natural basis of $\mathbb{R}^n$) and $e_i \pm e_j$, respectively, we obtain

$$\left\{ \begin{array}{lcl} \Delta_{e_i}\varphi_k & = & (h_i)^2\Phi_{x_ix_i} + o(\|h\|^2), \\ \Delta_{e_i \pm e_j}\varphi_k & = & (h_i)^2\Phi_{x_ix_i} + (h_j)^2\Phi_{x_jx_j} \pm 2h_ih_j\Phi_{x_ix_j} + o(\|h\|^2). \end{array} \right. \tag{2.3}$$

Denote v_k the approximation of the value function V at x_k. Let $D_k^u v_k$ be a notation for the upwind spatial finite difference

$$(D_k^u v_k)_i = \frac{v_{k+e_i} - v_k}{h_i} \quad \text{if} \quad f(x_k,u)_i \geq 0, \quad \frac{v_k - v_{k-e_i}}{h_i} \quad \text{if not.} \tag{2.4}$$

Now let $\mathcal{S}$ be a finite set of $\mathbb{Z}^n \setminus \{0\}$ containing $\{e_1, \dots, e_n\}$. We consider explicit schemes based on the difference operators that we just discussed, namely

$$\lambda v_k \quad = \quad \inf_{u \in U} \left\{ \ell(x_k,u) + f(x_k,u) \cdot D_k^u v_k + \sum_{\xi \in \mathcal{S}} \alpha_{k,\xi}^u \Delta_\xi v_k \right\} \tag{2.5}$$

for all $k \in \mathbb{Z}^n$. We will see soon how to choose the coefficients $\alpha_{k,\xi}^u$ in order to have a convergent approximation. Note that, since $\Delta_\xi = \Delta_{-\xi}$, we may assume without loss of generality that either $\alpha_{k,\xi}^u$ or $\alpha_{k,-\xi}^u$ is zero for all ξ. In particular, we may assume that $\alpha_{k,-e_i}^u$ is zero for all i. Note that there are possibilities other than (2.4) for discretizing the first-order term. For instance, it may be useful to consider centered differences in order to obtain (if the solution is smooth enough) higher orders of accuracy. However, since the difficulty for obtaining consistency lies in the discretization of the second-order term in the HJB equation, we will not elaborate on this.

Let $\Delta t^h > 0$ denote a *fictitious* time step (fictitious in the sense that the discrete scheme involves space, but not time, so that this time step has no influence on the solution). Multiplying (2.5) by Δt^h and adding v_k on both sides, we get

$$\begin{array}{l} v_k := (1 + \lambda\Delta t^h)^{-1}, \\ \displaystyle\inf_{u \in U} \left\{ v_k + \Delta t^h\, \ell(x_k,u) + \Delta t^h\, f(x_k,u) \cdot D_k^u v_k + \Delta t^h \sum_{\xi \in \mathcal{S}} \alpha_{k,\xi}^u \Delta_\xi v_k \right\}. \end{array} \tag{2.6}$$

With straightforward calculations, we can remark that the approximation (2.6) can be written in the form of (1.7), with the following transition probabilities:

$$\begin{aligned} & p^h(x_k, x_k \mid u) = 1 - \Delta t^h \sum_{i=1}^n \left(\frac{|f_i(x_k,u)|}{h_i} + 2\sum_{\xi \in \mathcal{S}} \alpha_{k,\xi}^u \right), \\ & p^h(x_k, x_{k \pm e_i} \mid u) = \Delta t^h \left(\frac{f_i^{\pm}(x_k,u)}{h_i} + \alpha_{k,e_i}^u \right), \\ & p^h(x_k, x_{k \pm \xi} \mid u) = \Delta t^h\, \alpha_{k,\xi}^u \qquad \text{for } \xi \in \mathcal{S}, \xi \neq e_i, \\ & p^h(x_k, y) = 0 \qquad \text{for } y \notin x_{k+\mathcal{S}}, \end{aligned}$$

where $f_i^+(x_k,u) = \max(f_i(x_k,u),0)$ and $f_i^-(x_k,u) = -\min(f_i(x_k,u),0)$.

Note that the sum of transition probabilities is, whatever the choice of coefficients $\alpha_{k,\xi}^u$, equal to one. However, that these transition probabilities are nonnegative adds the following condition on $\alpha^u_{k,\xi}$:

$$\alpha^u_{k,\xi} \geq 0 \ \ \forall\, (\xi, k, u) \in \mathcal{S} \times \mathbb{Z}^n \times U, \tag{2.7a}$$

$$\sum_{i=1}^{n} \frac{|f_i(x_k, u)|}{h_i} + 2\sum_{\xi\in\mathcal{S}} \alpha^u_{k,\xi} \leq \left(\Delta t^h\right)^{-1} \ \ \forall\, (k, u) \in \mathbb{Z}^n \times U. \tag{2.7b}$$

The second condition (2.7b) is always satisfied, when Δt^h is small enough, if the left-hand side of (2.7b) is uniformly bounded. We obtain in (2.16) such a bound. Here again we could take the more general point of view of having a time step depending on (x, u). Again, we prefer not to be general in order to concentrate on the main difficulties.

Assume (2.7) to be satisfied (we will see that (2.7b) is satisfied as soon as the time step is small enough) so that the scheme is a Markov chain approximation (of a specific type), since transition probabilities to points of the form $x_{k\pm\xi}$ are equal if $\xi \neq e_i$ for some i. We therefore concentrate on the local consistency condition (1.5). Since the terms multiplied by each coefficient $\alpha^u_{k,\xi}$ have a mean equal to x_k, we have that

$$\mathbb{E}^{h,u}_{k,q}\left[X^h_{q+1} - x_k\right] = \Delta t^h \sum_i f^+_i(x_k, u)e_i - \Delta t^h \sum_i f^-_i(x_k, u)e_i = \Delta t^h f(x_k, u) \tag{2.8}$$

so that condition (1.5a) is always satisfied. This in turn implies, denoting $\hat{x} := x_k + \Delta t^h f(x_k, u)$,

$$\mathrm{Cov}^{h,u}_{k,q}\left[X^h_{q+1} - x_k\right] = \mathbb{E}^{h,u}_{k,q}\left[(X^h_{q+1} - \hat{x}_k)(X^h_{q+1} - \hat{x}_k)^\top\right] + o(\Delta t^h) \tag{2.9}$$

and, therefore,

$$\mathrm{Cov}^{h,u}_{k,q}\left[X^h_{q+1} - x_k\right] = \Delta t^h \sum_{\xi\in\mathcal{S}} \sum_{i,j} h_i h_j \xi_i \xi_j \alpha^u_{k,\xi} e_i e_j^\top + o(\Delta t^h). \tag{2.10}$$

In view of (2.10), and since $\Delta t^h \to 0$ as $h \to 0$, local consistency holds iff we have

$$\sum_{\xi\in\mathcal{S}} \sum_{i,j} h_i h_j \xi_i \xi_j \alpha^u_{k,\xi} e_i e_j^\top = a(x_k, u) + o(1). \tag{2.11}$$

In what follows, we discuss the *strong consistency* property

$$\sum_{i,j} h_i h_j \xi_i \xi_j \alpha^u_{k,\xi} e_i e_j^\top = a(x_k, u) \quad \forall\ k \in \mathbb{Z}^n. \tag{2.12}$$

Let a^h denote the scaled covariance matrix $\{a_{ij}/h_i h_j\}$. Then a condition equivalent to (2.12) is

$$\sum_{\xi\in\mathcal{S}} \alpha^u_{k,\xi} \xi \xi^\top = a^h(x_k, u) \quad \forall\ k \in \mathbb{Z}^n. \tag{2.13}$$

Since every $\alpha^u_{k,\xi}$ is nonnegative, strong consistency means that the symmetric matrix $a^h(x_k, u)$ belongs, for all k and u, to the cone generated by the set $\{\xi\xi^\top; \xi \in \mathcal{S}\}$ that we denote

$$\mathcal{C}(\mathcal{S}) := \left\{ \sum_{\xi\in\mathcal{S}} \alpha_\xi \xi \xi^\top; \alpha \in \mathbb{R}^{|\mathcal{S}|}_+ \right\}. \tag{2.14}$$

Note that strong consistency implies a bound on the coefficients $\alpha^u_{k,\xi}$, which in turn allows us to obtain an estimate of the fictitious time step.

LEMMA 2.1. *Assume that the strong consistency condition holds. Then*

$$\sum_{\xi\in\mathcal{S}} \alpha^u_{k,\xi} \leq \operatorname{trace} a^h(x_k,u), \tag{2.15}$$

and hence condition (2.7b) *for the fictitious time step is satisfied whenever*

$$\sum_{i=1}^{n} \frac{\|f_i\|_\infty}{h_i} + 2\|\operatorname{trace} a^h\|_\infty \leq \left(\Delta t^h\right)^{-1}. \tag{2.16}$$

Proof. Taking the trace of both sides of (2.13), and since the trace of $\xi\xi^\top$ is greater than or equal to 1, obtain (2.15). The second part of the lemma is immediate. □

It follows from this lemma that, when $h \downarrow 0$, we may take Δt^h of order $O(\min_i h_i^2)$, as expected.

Now we can summarize the results of this section in the following theorem.

THEOREM 2.2. *Let $\mathcal{S}$ be a fixed finite set of $\mathbb{Z}^n \setminus \{0\}$ containing $\{e_1,\dots,e_n\}$, and let h be a fixed step size. Assume that, for every $k \in \mathbb{Z}^n$ and every $u \in U$, the scaled covariance matrix $a^h(x_k,u)$ belongs to the cone $\mathcal{C}(\mathcal{S})$. Then the scheme* (2.6) *is a consistent Markov chain approximation whenever the coefficients $(\alpha^u_{k,\xi})$ are nonnegative and satisfy condition* (2.13)*, the time step being such that* (2.16) *is satisfied.*

As said before, condition (2.16) is not really restrictive since Δt^h is just a fictitious time step. Note that implicit schemes can be used, as already mentioned, in connection with the policy iteration algorithm, and in that case it is easily seen that one can take a time step of order $O(\min_i h_i)$. Similar results hold in the finite horizon case. The most important condition in the above theorem is that the scaled matrix might belong to the cone $\mathcal{C}(\mathcal{S})$. Before going on the characterization of $\mathcal{C}(\mathcal{S})$, we will first compare our scheme to the classical finite differences approximations.

Remark 2.1. The results of this section are close to the analysis in section 5.4.4 of [6], where consistency for an arbitrary set of transition to neighbors is discussed. The point of view of this paper is rather to fix the set $\mathcal{S}$ of the neighbors allowed to enter the scheme and then to characterize the class of covariance matrices for which consistency holds. We will see in sections 4 and 5 (and this is the main novelty of the paper) how to obtain an effective characterization.

Remark 2.2. It is possible to study consistency taking the point of view of the discretization of the HJB equation (1.2), the solution being defined in the sense of viscosity. Barles and Souganidis [2] give a systematic way of obtaining convergent approximation schemes for second-order partial differential equations whose solution satisfies some strong uniqueness property. Their approach applies to (1.2) and leads to the same conditions as those of Theorem 2.2.

3. Classical finite differences approximations.

Let us show that the generalized finite difference algorithm, given in the above section, is indeed a generalization of the classical finite differences approximations that we recall now. Let Φ be a C^2 function over $\mathbb{R}^n$, and let $\varphi_k := \Phi(x_k)$ for all k. Given any $\xi \in \mathbb{Z}^n$, we can approximate the second-order derivatives of Φ by the following finite differences:

$$\frac{\delta_{\xi+e_i+e_j} - \delta_{\xi+e_i} - \delta_{\xi+e_j} + \delta_\xi}{h_i h_j}\varphi_k = \Phi_{x_i x_j}(x_k) + o(1). \tag{3.1}$$

Denote the corresponding operators as follows:

$$d_{ij}^{\xi} := \frac{\delta_{\xi+e_i+e_j} - \delta_{\xi+e_i} - \delta_{\xi+e_j} + \delta_{\xi}}{h_i h_j}. \tag{3.2}$$

Viewing i (resp., j) as the first (resp., second) coordinate, when $\xi = 0$, we call this operator d_{ij}^{ξ} the right upper approximation of $\Phi_{x_i x_j}$. We can similarly define left upper, right lower, and left lower approximations of $\Phi_{x_i x_j}$ by taking ξ equal to $-e_i$, $-e_j$, and $-e_i - e_j$, respectively. By combining these amounts, we can define centered approximations; the corresponding operators are along the main and second diagonals:

$$D_{ij}^{+} := \tfrac{1}{2}(d_{ij}^{0} + d_{ij}^{-e_i-e_j}), \qquad D_{ij}^{-} := \tfrac{1}{2}(d_{ij}^{-e_i} + d_{ij}^{-e_j}). \tag{3.3}$$

In other words,

$$\begin{aligned} D_{ij}^{+} &= \frac{1}{2h_i h_j}(\delta_{e_i+e_j} + \delta_{-e_i-e_j} + 2\delta_0 - \delta_{e_i} - \delta_{-e_i} - \delta_{e_j} - \delta_{-e_j}), \\ D_{ij}^{-} &= \frac{1}{2h_i h_j}(\delta_{e_i} + \delta_{-e_i} + \delta_{e_j} + \delta_{-e_j} - \delta_{e_i-e_j} - \delta_{e_j-e_i} - 2\delta_0). \end{aligned} \tag{3.4}$$

In addition, for the approximation of diagonal second-order derivatives we take the standard centered formula

$$D_{ii} := \frac{\delta_{e_i} + \delta_{-e_i} - 2\delta_0}{h_i h_i}. \tag{3.5}$$

The classical finite differences approximation of (1.2) is

$$\begin{aligned} \lambda v_k &= \inf_{u \in U} \left(\ell(x_k,u) + f(x_k,u) \cdot D_k^u v_k + \tfrac{1}{2} \sum_{i,j=1}^{n} a_{ij}(x_k,u) D_{ij}^{\pm} v_k \right) \\ &\quad \forall\ k \in \mathbb{Z}^n,\ \ q \in \mathbb{N}, \\ y_k^0 &= 0 \quad \forall\ k \in \mathbb{Z}^n, \end{aligned} \tag{3.6}$$

where if $i \neq j$, $D_{ij}^{\pm}$ is equal either to D_{ij}^{+} or D_{ij}^{-} and D_k^u is the upwind spatial finite difference defined in (2.4). The above scheme is equivalent to the following one:

$$\begin{aligned} &v_k := (1+\lambda \Delta t^h)^{-1}, \\ &\qquad \inf_{u \in U} \left\{ v_k + \Delta t^h \ell(x_k,u) + \Delta t^h f(x_k,u) \cdot D_k^u v_k + \tfrac{1}{2}\Delta t^h \sum_{i,j=1}^{n} a_{ij}(x_k,u) D_{ij}^{\pm} v_k. \right\}. \end{aligned} \tag{3.7}$$

It is known that this scheme is a consistent Markov chain approximation under restrictive assumptions that we make explicit now (this is a reformulation of known results; see, e.g., [6] or [9]).

Lemma 3.1. *The classical finite differences approximation scheme can be interpreted as a consistent Markov chain approximation iff the following three conditions hold:*

(i) *If $i \neq j$ is such that $a_{ij}(x_k,u) \neq 0$, then $D_{ij}^{\pm} = D_{ij}^{+}$ if $a_{ij}(x_k,u) > 0$ and $D_{ij}^{\pm} = D_{ij}^{-}$ if $a_{ij}(x_k,u) < 0$;*

(ii) *The matrix* $a^h(x_k,u)$ *is dominant diagonal or, equivalently,*

$$\frac{a_{ii}(x_k,u)}{h_i} \geq \sum_{j\neq i} \frac{|a_{ij}(x_k,u)|}{h_j} \quad \forall \ i=1,\dots,n; \tag{3.8}$$

(iii) *The time step* Δt^h *satisfies the following condition:*

$$\sum_{i=1}^{n} \frac{|f(x_k,u)_i|}{h_i} + \sum_{i=1}^{n} \left(2\frac{a_{ii}(x_k,u)}{h_i^2} - \sum_{j\neq i} \frac{|a_{ij}(x_k,u)|}{h_i h_j} \right) \leq \left(\Delta t^h\right)^{-1}. \tag{3.9}$$

We now make explicit the link between the two approaches by expressing the classical finite differences approximation scheme as a Markov chain approximation scheme. If the conditions of the above lemma are satisfied, then we can write the approximation of second-order terms as

$$\begin{aligned} \sum_{i,j=1}^{n} a_{ij}(x_k,u) D_{ij}^{\pm} &= \sum_{\substack{i\neq j \\ a_{ij}>0}} \frac{a_{ij}}{h_i h_j} \Delta_{e_i+e_j} - \sum_{\substack{i\neq j \\ a_{ij}<0}} \frac{a_{ij}}{h_i h_j} \Delta_{e_i-e_j} \\ &\quad + \sum_{i} \left(\frac{a_{ii}}{(h_i)^2} - \sum_{j\neq i} \frac{|a_{ij}|}{h_i h_j} \right) \Delta_{e_i}. \end{aligned} \tag{3.10}$$

The weights of the transitions are nonnegative iff condition (ii) of Lemma 3.1 is satisfied. It follows that the classical finite difference scheme is equivalent to the generalized finite difference scheme where the set $\mathcal{S}$ is equal to

$$\hat{\mathcal{S}} := \{e_1,\dots,e_n\} \cup \{e_i \pm e_j, 1 \leq i \neq j \leq n\}.$$

We have that $\mathcal{C}(\hat{\mathcal{S}})$ is precisely the cone of dominant diagonal matrices.

4. Characterization of finitely generated cones. Let us come back to the analysis of the generalized finite difference method. In what follows we will concentrate on characterizations of the strong consistency condition, with special attention to the case when $\mathcal{S}$ is the set $\mathcal{S}^q$ of neighboring points of order q, defined by

$$\mathcal{S}^q := \{\xi \in \mathbb{Z}^n; |\xi_i| \leq q, \ i=1,\dots,n\}. \tag{4.1}$$

Characterizing a finitely generated cone happens to be a classical problem of convex analysis and polyhedral combinatorics, and it can be solved using the notion of polar cone. Let us recall these classical results; an excellent reference on this subject is Pulleyblank [10].

Let $\mathcal{C}$ be a nonempty closed convex cone in $\mathbb{R}^p$. The associated (positively) polar cone is

$$\mathcal{C}^* = \{x^* \in X^* \ ; \ \langle x^*, x\rangle \geq 0 \quad \forall x \in \mathcal{C}\}.$$

It is known that $(\mathcal{C}^*)^* = \mathcal{C}$. Let $\mathcal{C}$ be finitely generated, say, by $g_1,\dots,g_q$. Then $\mathcal{C}^* = \bigcap_i \{x^* \in X^* \ ; \ \langle x^*, g_i\rangle \geq 0\}$. It happens that the set $\mathcal{C}^*$ is also finitely generated, say, by $g_1^*,\dots,g_r^*$; this dual generator can be computed by a certain recursion. Since $\mathcal{C} = (\mathcal{C}^*)^*$, it follows that

$$\mathcal{C} = \{x; \ \langle g_i^*, x\rangle \geq 0, \ i=1,\dots,r\}.$$

This means that the cone $\mathcal{C}$ is characterized by a finite number of linear inequalities whose coefficients can be computed.

Let us specialize this result to the case of the cone $\mathcal{C}(\mathcal{S})$. Let $\mathcal{M}$ be the set of symmetric matrices, and let $\mathcal{M}_+$ be the set of symmetric definite positive matrices. Using the Frobenius scalar product $A \cdot B = \sum_{i,j} A_{ij} B_{ij}$, for which $B \cdot \xi\xi^\top = \xi^\top B \xi$, for all square $n \times n$ symmetric matrices B and n dimensional vectors ξ, we have that the polar cone is

$$\mathcal{C}(\mathcal{S}^q)^* = \left\{ B \in \mathcal{M}\ ;\ \xi^\top B \xi \geq 0 \quad \forall \xi \in \mathcal{S} \right\}. \tag{4.2}$$

Consider the example when $\mathcal{S} = \mathcal{S}^q$ is defined in (4.1). Using the fact that $\mathcal{C}(\mathcal{S}^q)$ is strictly increasing with q and is a subset of the cone $\mathcal{M}_+$, and $(\mathcal{M}_+)^* = \mathcal{M}_+$, we have the infinite chain of strict inclusions

$$\mathcal{C}(\mathcal{S}^1) \subset \mathcal{C}(\mathcal{S}^2) \cdots \subset \mathcal{M}_+ \subset \cdots \mathcal{C}(\mathcal{S}^2)^* \subset \mathcal{C}(\mathcal{S}^1)^*. \tag{4.3}$$

It can be noticed that, since the cone $\mathcal{C}(\mathcal{S}^q)$ contains every nonnegative diagonal matrix, each element of its dual has a nonnegative diagonal.

An important observation is that $\mathcal{S}^q$, and therefore also $\mathcal{C}(\mathcal{S}^q)$, are invariant through the linear transformations in $\mathbb{R}^n$ that correspond to a permutation of coordinates and also correspond to the change of sign of coordinates. The permutation of coordinates i and j of $\xi \in \mathbb{R}^n$ result in the permutation of elements of $\xi\xi^\top$ of coordinates (i,k) and (j,k), and (k,i) and (k,j), for all k, while changing the sign of ξ_i results in changing the sign of elements of $\xi\xi^\top$ of coordinates (i,j) for $j \neq i$. Since these transformations are self-adjoint, for each $B \in \mathcal{C}(\mathcal{S}^q)^*$, the matrices obtained by the same (adjoint) transformations (so that the scalar product with B remains invariant) also belong to $\mathcal{C}(\mathcal{S}^q)^*$. In particular, a generator of $\mathcal{C}(\mathcal{S}^q)^*$ can be partitioned into classes of equivalence corresponding to the above mentioned transformations. This allows us to give a compact description of the set of generators.

5. Specific examples.

We have performed the computation of generators of dual cones using the Qhull algorithm by Barber, Dobkin, and Huhdanpaa [1]. The latter computes, given a finite set in $\mathbb{R}^m$, a minimal set of linear inequalities characterizing its convex hull. This computation is made using the floating-point arithmetic of the C language. However, the risk of numerical errors due to the floating-point arithmetic is limited, since we were able to compute a scaling of the data for which all coefficients are small integers, up to an absolute precision of 10^{-10}.

The link between the convex hull of a finite set and the generator of a dual cone is as follows. Consider a generator $g_1, \dots, g_n$, and set $g_0 := 0$. Then compute a minimal characterization of the convex hull of $g_0, \dots, g_n$ of the form $\langle g_i^*, \cdot \rangle \geq b_i$, $i = 1, \dots, r$. A minimal generator of the dual cone is given by the homogeneous inequalities; i.e., the dual cone is

$$\{ g \in \mathbb{R}^m;\ \ \langle g_i^*, g \rangle \geq b_i, \ \ i \in I \}$$

with $I := \{1 \leq i \leq r;\ \ b_i = 0\}$.

Our actual computations deal with spaces of symmetric matrices of size n. Each of them can be represented by its upper triangular part, and thus is viewed as an element of $\mathbb{R}^m$, $m = \frac{1}{2}n(n+1)$; in particular, $m = 3$, 6, and 10 for $n = 2$, 3, and 4, respectively.

Once a generator of the dual cone has been obtained, it remains to identify the classes of equivalence (defined in the previous section) in order to obtain compact expression. This was done by sorting the elements following the (ordered) weights of diagonal elements (the latter being, as we already know, nonnegative). It appears that this suffices for identifying the equivalence classes, as can be checked by generating them using the formulas given below and comparing both sets.

Dimension 2. In the case $n = 2$, we computed characterizations of the sets $\mathcal{C}(\mathcal{S}^q)$, $q = 1$ to 10. We display detailed results for $q = 1$ to 7. The set $\mathcal{C}(\mathcal{S}^1)$ is characterized by 4 constraints and 1 equivalence class:

$$a_{ii} \geq |a_{ij}|, \quad 1 \leq i \neq j \leq 2.$$

The set $\mathcal{C}(\mathcal{S}^2)$ is characterized by 8 constraints and 2 equivalence classes:

$$\begin{cases} 2a_{ii} \geq |a_{ij}|, \\ 2a_{ii} + a_{jj} \geq 3|a_{ij}| \end{cases}$$

for $1 \leq i \neq j \leq 2$. The set $\mathcal{C}(\mathcal{S}^3)$ is characterized by 16 constraints and 4 equivalence classes:

$$\begin{cases} 3a_{ii} \geq |a_{ij}|, \\ 3a_{ii} + 2a_{jj} \geq 5|a_{ij}|, \\ 6a_{ii} + a_{jj} \geq 5|a_{ij}|, \\ 6a_{ii} + 2a_{jj} \geq 7|a_{ij}| \end{cases}$$

for $1 \leq i \neq j \leq 2$. The set $\mathcal{C}(\mathcal{S}^4)$ is characterized by 24 constraints and 6 equivalence classes:

$$\begin{cases} 4a_{ii} \geq |a_{ij}|, \\ 4a_{ii} + 3a_{jj} \geq 7|a_{ij}|, \\ 6a_{ii} + a_{jj} \geq 5|a_{ij}|, \\ 6a_{ii} + 2a_{jj} \geq 7|a_{ij}|, \\ 12a_{ii} + a_{jj} \geq 7|a_{ij}|, \\ 12a_{ii} + 6a_{jj} \geq 17|a_{ij}| \end{cases}$$

for $1 \leq i \neq j \leq 2$. The set $\mathcal{C}(\mathcal{S}^5)$ is characterized by 40 constraints and 10 equivalence classes:

$$\begin{cases} 5a_{ii} \geq |a_{ij}|, \\ 5a_{ii} + 4a_{jj} \geq 9|a_{ij}|, \\ 10a_{ii} + 2a_{jj} \geq 9|a_{ij}|, \\ 10a_{ii} + 3a_{jj} \geq 11|a_{ij}|, \\ 12a_{ii} + a_{jj} \geq 7|a_{ij}|, \\ 12a_{ii} + 6a_{jj} \geq 17|a_{ij}|, \\ 15a_{ii} + 2a_{jj} \geq 11|a_{ij}|, \\ 15a_{ii} + 6a_{jj} \geq 19|a_{ij}|, \\ 20a_{ii} + a_{jj} \geq 9|a_{ij}|, \\ 20a_{ii} + 12a_{jj} \geq 31|a_{ij}| \end{cases}$$

for $1 \leq i \neq j \leq 2$. The set $\mathcal{C}(\mathcal{S}^6)$ is characterized by 48 constraints and 12 equivalence classes:

$$\begin{cases} 6_{ii} \geq |a_{ij}|, \\ 6a_{ii} + 5a_{jj} \geq 11|a_{ij}|, \\ 10a_{ii} + 2a_{jj} \geq 9|a_{ij}|, \\ 10a_{ii} + 3a_{jj} \geq 11|a_{ij}|, \\ 12a_{ii} + a_{jj} \geq 7|a_{ij}|, \\ 12a_{ii} + 6a_{jj} \geq 17|a_{ij}|, \\ 15a_{ii} + 2a_{jj} \geq 11|a_{ij}|, \\ 15a_{ii} + 6a_{jj} \geq 19|a_{ij}|, \\ 20a_{ii} + a_{jj} \geq 9|a_{ij}|, \\ 20a_{ii} + 12a_{jj} \geq 31|a_{ij}|, \\ 30a_{ii} + a_{jj} \geq 11|a_{ij}|, \\ 30a_{ii} + 20a_{jj} \geq 49|a_{ij}| \end{cases}$$

for $1 \leq i \neq j \leq 2$. The set $\mathcal{C}(\mathcal{S}^7)$ is characterized by 72 constraints and 18 equivalence classes:

$$\begin{cases} 7a_{ii} \geq |a_{ij}|, \\ 7a_{ii} + 6a_{jj} \geq 13|a_{ij}|, \\ 14a_{ii} + 3a_{jj} \geq 13|a_{ij}|, \\ 14a_{ii} + 4a_{jj} \geq 15|a_{ij}|, \\ 15a_{ii} + 2a_{jj} \geq 11|a_{ij}|, \\ 15a_{ii} + 6a_{jj} \geq 19|a_{ij}|, \\ 20a_{ii} + a_{jj} \geq 9|a_{ij}|, \\ 20a_{ii} + 12a_{jj} \geq 31|a_{ij}|, \\ 21a_{ii} + 2a_{jj} \geq 13|a_{ij}|, \\ 21a_{ii} + 10a_{jj} \geq 29|a_{ij}|, \\ 28a_{ii} + 2a_{jj} \geq 15|a_{ij}|, \\ 28a_{ii} + 15a_{jj} \geq 41|a_{ij}|, \\ 30a_{ii} + a_{jj} \geq 11|a_{ij}|, \\ 30a_{ii} + 20a_{jj} \geq 49|a_{ij}|, \\ 35a_{ii} + 6a_{jj} \geq 29|a_{ij}|, \\ 35a_{ii} + 12a_{jj} \geq 41|a_{ij}|, \\ 42a_{ii} + a_{jj} \geq 13|a_{ij}|, \\ 42a_{ii} + 30a_{jj} \geq 71|a_{ij}|. \end{cases}$$

Dimension 3. When $n = 3$, we computed characterizations of the sets $\mathcal{C}(\mathcal{S}^q)$, $q = 1$ to 2. The set $\mathcal{C}(\mathcal{S}^1)$ is characterized by

$$\begin{cases} a_{ii} \geq |a_{ij}|, \\ a_{ii} + a_{jj} \geq (-1)^p a_{ik} + (-1)^q a_{jk} + 2(-1)^{p+q+1} a_{ij} \end{cases}$$

for $i \neq j \neq k$ and $p, q \in \{1, 2\}$. As was expected, this cone is larger than the cone of dominant diagonal matrices. The set $\mathcal{C}(\mathcal{S}^2)$ is characterized by

$$\begin{cases}
2a_{ii} \geq |a_{ij}|,\\
2a_{ii} + a_{jj} \geq 3|a_{ij}|,\\
2a_{ii} + 2a_{jj} \geq 4(-1)^p a_{ij} + (-1)^q a_{jk} - (-1)^{p+q} a_{ik},\\
2a_{ii} + 2a_{jj} + a_{kk} \geq 4(-1)^p a_{ij} + 3(-1)^q a_{jk} - 3(-1)^{p+q} a_{ik},\\
3a_{ii} + 2a_{jj} + 2a_{kk} \geq 5(-1)^p a_{ij} + 4(-1)^q a_{jk} - 5(-1)^{p+q} a_{ik},\\
6a_{ii} + a_{jj} + a_{kk} \geq 5(-1)^p a_{ij} + 2(-1)^q a_{jk} - 5(-1)^{p+q} a_{ik},\\
6a_{ii} + 2a_{jj} + a_{kk} \geq 7(-1)^p a_{ij} + 3(-1)^q a_{jk} - 5(-1)^{p+q} a_{ik},\\
6a_{ii} + 2a_{jj} + 2a_{kk} \geq 7(-1)^p a_{ij} + 4(-1)^q a_{jk} - 7(-1)^{p+q} a_{ik},\\
8a_{ii} + 2a_{jj} \geq 8(-1)^p a_{ij} + (-1)^q a_{jk} - 2(-1)^{p+q} a_{ik},\\
8a_{ii} + 2a_{jj} + a_{kk} \geq 8(-1)^p a_{ij} + 3(-1)^q a_{jk} - 6(-1)^{p+q} a_{ik},\\
8a_{ii} + 3a_{jj} + 2a_{kk} \geq 10(-1)^p a_{ij} + 5(-1)^q a_{jk} - 8(-1)^{p+q} a_{ik},\\
8a_{ii} + 6a_{jj} + 2a_{kk} \geq 14(-1)^p a_{ij} + 7(-1)^q a_{jk} - 8(-1)^{p+q} a_{ik},\\
12a_{ii} + 2a_{jj} + a_{kk} \geq 10(-1)^p a_{ij} + 3(-1)^q a_{jk} - 7(-1)^{p+q} a_{ik},\\
12a_{ii} + 4a_{jj} + a_{kk} \geq 14(-1)^p a_{ij} + 4(-1)^q a_{jk} - 7(-1)^{p+q} a_{ik},\\
12a_{ii} + 6a_{jj} + 2a_{kk} \geq 17(-1)^p a_{ij} + 7(-1)^q a_{jk} - 10(-1)^{p+q} a_{ik},\\
12a_{ii} + 6a_{jj} + 4a_{kk} \geq 17(-1)^p a_{ij} + 10(-1)^q a_{jk} - 14(-1)^{p+q} a_{ik},\\
18a_{ii} + 2a_{jj} + a_{kk} \geq 12(-1)^p a_{ij} + 3(-1)^q a_{jk} - 9(-1)^{p+q} a_{ik},\\
18a_{ii} + 8a_{jj} + a_{kk} \geq 24(-1)^p a_{ij} + 6(-1)^q a_{jk} - 9(-1)^{p+q} a_{ik},\\
18a_{ii} + 10a_{jj} + 2a_{kk} \geq 27(-1)^p a_{ij} + 9(-1)^q a_{jk} - 12(-1)^{p+q} a_{ik}.
\end{cases}$$

Dimension 4. When $n = 4$, the set $\mathcal{C}(\mathcal{S}^1)$ is characterized by

$$\left\{\begin{array}{ll}
a_{ii} & \geq |a_{ij}|,\\
a_{ii} + a_{jj} & \geq (-1)^p a_{ik} + (-1)^q a_{jk} - 2(-1)^{p+q} a_{ij},\\
a_{ii} + a_{jj} + a_{kk} & \geq (-1)^p a_{il} + (-1)^q a_{jl} + (-1)^r a_{kl}\\
 & -2(-1)^{p+q} a_{ij} - 2(-1)^{p+r} a_{ik} - 2(-1)^{q+r} a_{jk},\\
2a_{ii} + a_{jj} + a_{kk} + a_{ll} & \geq 3(-1)^p a_{ij} + 3(-1)^q a_{ik} + 3(-1)^r a_{il}\\
 & -2(-1)^{p+q} a_{jk} - 2(-1)^{p+r} a_{jl} - 2(-1)^{q+r} a_{kl},\\
4a_{ii} + a_{jj} + a_{kk} & \geq 2(-1)^p a_{il} + (-1)^q a_{jl} + (-1)^r a_{kl}\\
 & -4(-1)^{p+q} a_{ij} - 4(-1)^{p+r} a_{ik} - 2(-1)^{q+r} a_{jk},\\
4a_{ii} + 2a_{jj} + a_{kk} + a_{ll} & \geq 6(-1)^p a_{ij} + 4(-1)^q a_{ik} + 4(-1)^r a_{il}\\
 & -3(-1)^{p+q} a_{jk} - 3(-1)^{p+r} a_{jl} - 2(-1)^{q+r} a_{kl}
\end{array}\right.$$

for $i \neq j \neq k \neq l$ and $p, q, r \in \{1, 2\}$.

Summary of results. The following table summarizes the various steps of our calculation and highlights the importance of reduction of constraints using the classes of equivalence.

Here $\mathcal{S}^*$ is the set of matrices in $\mathcal{S}$ of trace not greater than 1.

n	q	Size of generator of primal cone	# of constraints defining $\mathcal{S}^*$	# of constraints defining $\mathcal{C}$	# of classes of equivalence
2	1	4	6	4	1
2	2	8	13	8	2
2	3	16	27	16	4
2	4	24	39	24	6
2	5	40	67	40	10
2	6	48	87	48	12
2	7	72	123	72	18
2	8	88	159	88	22
2	9	112	203	112	28
2	10	128	239	128	32
3	1	13	31	24	2
3	2	49	563	372	19
4	1	40	476	328	6

6. Discussion of results. In this paper we have worked in the framework of "Markov chain approximations" discussed in Kushner and Dupuis [6]. Our main result is a method for computing a characterization of the class of covariance matrices that are strongly consistent with an a priori choice of neighboring points to which transitions are allowed. In the computation and display of the results, we use in an essential way the property of invariance of these cones with respect to some transformations. Although we are looking for linear inequalities with integer coefficients, the computations were made in floating-point arithmetic. However, once properly scaled, the results are up to a precision of 10^{-10} equal to very small integers, as may be seen in the table above, and hence it seems that these results are exact, despite the fact that our method is not a mathematical proof (we tried an exact approach based on computer algebra, but without success, since many singularities were encountered). So, we have performed the computations, giving explicit results, for dimensions of the state space between 2 and 4, and when only a limited number of neighboring points are allowed (which is a highly desirable feature). This said, it seems that we have performed the computations for essentially all cases for which the numerical resolution of the stochastic HJB equation is of reasonable complexity. Indeed, when the number of linear inequalities characterizing strongly consistent matrices is large, we may expect that computing the coefficient of the algorithm will be expensive.

On the other hand, our results are only a preliminary step towards an efficient numerical algorithm. There are two main difficulties. The first is designing fast algorithms for computing the coefficients $\alpha_{k,\xi}^u$. The latter are, by definition, solutions of a linear programming problem, but using a linear programming solver for each control at each point of the grid would be inefficient. The second difficulty is dealing with the case when consistency does not hold, e.g., by approximating the matrix $a(x_k, u)$ by a consistent matrix and then performing an error analysis. We are now pursuing some research in these directions.

Acknowledgment. We thank the three anonymous referees for their useful remarks.

REFERENCES

[1] C. Barber, D. Dobkin, and H. Huhdanpaa, *The quickhull algorithm for convex hulls*, ACM Trans. Math. Software, 22 (1996), pp. 469–483.

[2] G. Barles and P. E. Souganidis, *Convergence of approximation schemes for fully nonlinear second order equations*, Asymptotic Anal., 4 (1991), pp. 271–283.

[3] S. Chung, F. Hanson, and H. Xu, *Parallel stochastic dynamic programming: Finite element methods*, Linear Algebra Appl., 172 (1992), pp. 197–218.

[4] W. H. Fleming and R. Rishel, *Deterministic and Stochastic Optimal Control*, Appl. Math. 1, Springer-Verlag, New York, 1975.

[5] W. H. Fleming and H. Soner, *Controlled Markov Processes and Viscosity Solutions*, Springer-Verlag, New York, 1993.

[6] H. J. Kushner and P. G. Dupuis, *Numerical Methods for Stochastic Control Problems in Continuous Time*, 2nd ed., Appl. Math. 24, Springer-Verlag, New York, 2001.

[7] J. L. Lions and A. Bensoussan, *Application des inéquations variationnelles en contrôle stochastique*, Méthodes Mathématiques de l'Informatique 6, Dunod, Paris, 1978.

[8] P. Lions, *Optimal control of diffusion processes and Hamilton-Jacobi-Bellman equations. Part 2: Viscosity solutions and uniqueness*, Comm. Partial Differential Equations, 8 (1983), pp. 1220–1276.

[9] P. Lions and B. Mercier, *Approximation numérique des équations de Hamilton-Jacobi-Bellman*, RAIRO Anal. Numér., 14 (1980), pp. 369–393.

[10] W. Pulleyblank, *Polyhedral combinatorics*, in Optimization, G. L. Nemhauser, A. H. G. Rinnooy Kan, and M. J. Todd, eds., Elsevier, Amsterdam, 1989.

SIAM J. NUMER. ANAL.
Vol. 41, No. 3, pp. 1022–1031

SCALED PIVOTS AND SCALED PARTIAL PIVOTING STRATEGIES*

J. M. PEÑA[†]

Abstract. Scaled pivots for Gaussian elimination of an $n \times n$ matrix are introduced. They are used to obtain bounds for the Skeel condition number of the resulting upper triangular matrix and for a growth factor which has been introduced by Amodio and Mazzia [*BIT*, 39 (1999), pp. 385–402]. A bound of this growth factor for row scaled partial pivoting strategies is also included. It is proved that the Skeel condition number of an $n \times n$ upper triangular matrix which is strictly diagonally dominant by rows is bounded above by a number which is independent of n. It is also shown that the calculation of the n scaled pivots associated with a pivoting strategy presenting nice properties when applied to nonsingular $n \times n$ M-matrices adds $\mathcal{O}(n)$ elementary operations to the complete cost of this pivoting strategy.

Key words. growth factor, Gaussian elimination, scaled partial pivoting, Skeel condition number

AMS subject classifications. 65F05, 65G05, 65G99

DOI. 10.1137/S0036142901395163

1. Introduction and basic notations.

The growth factor is an indicator of the stability of Gaussian elimination. The classical growth factor of an $n \times n$ matrix $A = (a_{ij})_{1 \le i,j \le n}$ is the number

$$\rho^W(A) := \frac{\max_{i,j,k} |a_{ij}^{(k)}|}{\max_{i,j} |a_{ij}|}. \tag{1.1}$$

On p. 398 of [1], Amodio and Mazzia have introduced the growth factor

$$\rho(A) := \frac{\max_k \|A^{(k)}\|_\infty}{\|A\|_\infty} \tag{1.2}$$

and have shown its nice behavior for the error analysis of Gaussian elimination. Conditioning is another important concept to be considered in such error analysis. The traditional condition number of a matrix A with respect to the norm $\| \cdot \|_\infty$ is given by $\kappa(A) := \|A\|_\infty \, \|A^{-1}\|_\infty$. Given a matrix $B = (b_{ij})_{1 \le i,j \le n}$, we shall denote $|B|$ the matrix of absolute values of the entries of B. If we write $A \le B$, it means $a_{ij} \le b_{ij}$ for all i, j. The Skeel condition number (cf. [9]) of a matrix A is defined as

$$\mathrm{Cond}(A) = \| \, |A^{-1}| \, |A| \, \|_\infty.$$

Let us mention two nice properties of $\mathrm{Cond}(A)$. The Skeel condition number of a matrix A is less than or equal to $\kappa(A)$, and it can be much smaller. In contrast with $\kappa(A)$, $\mathrm{Cond}(A)$ is invariant under row scaling.

The first topic considered in this paper is the introduction of the scaled pivots associated with the performance of Gaussian elimination of a nonsingular $n \times n$ matrix

*Received by the editors September 14, 2001; accepted for publication (in revised form) December 23, 2002; published electronically June 18, 2003. This research was partially supported by Spanish Research grant BFM2000-1253.

http://www.siam.org/journals/sinum/41-3/39516.html

[†]Departamento de Matemática Aplicada, Universidad de Zaragoza, 50009 Zaragoza, Spain (jmpena@posta.unizar.es).

with any given pivoting strategy. A scaled pivot is the quotient between the absolute value of the pivot and the norm of the corresponding row. The computational cost of calculating all possible scaled pivots is $\mathcal{O}(n^2)$ elementary operations per step of Gaussian elimination. If we have obtained n pivots after the complete performance of a given pivoting strategy, the additional cost of calculating the corresponding n scaled pivots is $\mathcal{O}(n^2)$ elementary operations. In section 2 the scaled pivots are used to obtain bounds for the growth factor (1.2). In section 3 we apply them to derive bounds for the Skeel condition number of the upper triangular matrix U resulting after Gaussian elimination. As a consequence of these bounds we prove in Corollary 3.3 that the Skeel condition number of an $n \times n$ upper triangular matrix which is strictly diagonally dominant by rows is bounded above by a number which is independent of n.

The second main topic of this paper is related to scaled partial pivoting strategies, that is, strategies which incorporate row scaling *implicitly.* In [5] some properties of these strategies with respect to the Skeel condition number of the upper triangular matrix U resulting after Gaussian elimination were analyzed. We also proved that if there exists a permutation matrix P such that the LU-factorization of the matrix $B = PA$ satisfies $|LU| = |L|\,|U|$ (the property which can be used to derive small componentwise backward errors), then P is associated with the row scaled partial pivoting for any strictly monotone vector norm. In contrast with the usual growth factor (1.1), in section 4 we get specific bounds for the growth factor (1.2) in the case of row scaled partial pivoting. A disadvantage of scaled partial pivoting strategies is their high computational cost: $\mathcal{O}(n^3)$ elementary operations for the complete factorization. However, we shall see in section 5 that, for important classes of matrices, these strategies can be implemented without computational cost or with less computational cost than partial pivoting, and they present better stability properties than partial pivoting. On the other hand, we also show in section 5 that the calculation of the n scaled pivots associated with a pivoting strategy presenting nice properties when applied to nonsingular $n \times n$ M-matrices adds $\mathcal{O}(n)$ elementary operations to the complete cost of this strategy.

2. Scaled pivots and the growth factor.

Given $k \in \{1, 2, \ldots, n\}$, let α, β be two increasing sequence of k positive integers less than or equal to n. Then we denote $A[\alpha|\beta]$ the $k \times k$ submatrix of A containing rows numbered by α and columns numbered by β. Gaussian elimination with a given pivoting strategy, for nonsingular matrices $A = (a_{ij})_{1 \le i,j \le n}$, consists of a succession of at most $n-1$ major steps resulting in a sequence of matrices as follows:

$$A = A^{(1)} \longrightarrow \tilde{A}^{(1)} \longrightarrow A^{(2)} \longrightarrow \tilde{A}^{(2)} \longrightarrow \cdots \longrightarrow A^{(n)} = \tilde{A}^{(n)} = U, \tag{2.1}$$

where $A^{(t)} = (a_{ij}^{(t)})_{1 \le i,j \le n}$ has zeros below its main diagonal in the first $t-1$ columns. The matrix $\tilde{A}^{(t)} = (\tilde{a}_{ij}^{(t)})_{1 \le i,j \le n}$ is obtained from the matrix $A^{(t)}$ by reordering the rows and/or columns $t, t+1, \ldots, n$ of $A^{(t)}$ according to the given pivoting strategy and satisfying $\tilde{a}_{tt}^{(t)} \neq 0$. To obtain $A^{(t+1)}$ from $\tilde{A}^{(t)}$ we produce zeros in column t below the *pivot element* $\tilde{a}_{tt}^{(t)}$ by subtracting multiples of row t from the rows beneath it. If P and/or Q are permutation matrices such that the Gaussian elimination of $B = PAQ$ can be performed without row exchanges, then the first row of $\tilde{A}^{(t)}[t, \ldots, n]$ coincides with the first row of $B^{(t)}[t, \ldots, n]$, and the other rows coincide up to the order. If $B = P^T AP$, we say that we have performed *symmetric pivoting.* Let $r_i^{(t)}$ (resp., $\tilde{r}_i^{(t)}$) denote the ith row ($t \le i \le n$) of the submatrix $A^{(t)}[1, \ldots, n|t, t+1, \ldots, n]$ (resp., $\tilde{A}^{(t)}[1, \ldots, n|t, t+1, \ldots, n]$).

Given a pivoting strategy which has produced the pivots $\tilde{a}_{11}^{(1)}, \dots, \tilde{a}_{nn}^{(n)}$, let us define the kth *scaled pivot* as

$$p_k := \frac{|\tilde{a}_{kk}^{(k)}|}{\|\tilde{r}_k^{(k)}\|_1} \tag{2.2}$$

for each $k = 1, \dots, n$. Calculating the n scaled pivots $p_1, \dots, p_n$ associated with the pivoting strategy adds $\mathcal{O}(n^2)$ elementary operations to the complete cost of the strategy. Obviously, $p_k \le 1$ for all k and, if A is an $n \times n$ matrix, $p_n = 1$. We shall see that these numbers provide information on the growth factor ρ and on the Skeel condition number of the upper triangular matrix U.

A *row* (resp., *symmetric*) *scaled partial pivoting* strategy for the norm $\|\cdot\|_1$ consists of an implicit scaling by the norm $\|\cdot\|_1$ followed by partial (resp., symmetric and partial) pivoting. For each t $(1 \le t \le n-1)$, these strategies look for an integer $\hat{i}_t$ $(t \le \hat{i}_t \le n)$ such that

$$p_t = \frac{|a_{\hat{i}_t t}^{(t)}|}{\|r_{\hat{i}_t}^{(t)}\|_1} = \max_{t \le i \le n} \frac{|a_{it}^{(t)}|}{\|r_i^{(t)}\|_1} \tag{2.3}$$

(resp., $p_t = \frac{|a_{\hat{i}_t \hat{i}_t}^{(t)}|}{\|r_{\hat{i}_t}^{(t)}\|_1} = \max_{t \le i \le n} \frac{|a_{ii}^{(t)}|}{\|r_i^{(t)}\|_1}$). Observe that the scaled pivots p_k of a row (resp., symmetric) scaled partial pivoting strategy satisfy for each $i = k, k+1, \dots, n$

$$p_k \ge \frac{|\tilde{a}_{ik}^{(k)}|}{\|\tilde{r}_i^{(k)}\|_1} \tag{2.4}$$

and, respectively,

$$p_k \ge \frac{|\tilde{a}_{ii}^{(k)}|}{\|\tilde{r}_i^{(k)}\|_1}. \tag{2.5}$$

The scaled pivots can be calculated after performing *any* pivoting strategy. As we have commented above, the calculation of the n scaled pivots associated with a pivoting strategy adds $\mathcal{O}(n^2)$ elementary operations to the complete cost of the strategy.

The following result will be applied to show the usefulness of the scaled pivots in order to estimate the growth factor.

PROPOSITION 2.1. *If $A^{(k+1)}$ is obtained from $\tilde{A}^{(k)}$ in the kth step of Gaussian elimination with a given pivoting strategy, then one has for all $i = k+1, \dots, n$,*

$$\|r_i^{(k+1)}\|_1 \le \|\tilde{r}_i^{(k)}\|_1 + \left(\frac{1}{p_k} - 2\right) |\tilde{a}_{ik}^{(k)}|, \tag{2.6}$$

where p_k is the scaled pivot given by (2.2), *and the equality in* (2.6) *can be achieved.*

Proof. For $i = k+1, \dots, n$, $j = k+1, \dots, n$, we have

$$|a_{ij}^{(k+1)}| = \left| \tilde{a}_{ij}^{(k)} - \frac{\tilde{a}_{ik}^{(k)}}{\tilde{a}_{kk}^{(k)}} \tilde{a}_{kj}^{(k)} \right| \le |\tilde{a}_{ij}^{(k)}| + \left| \frac{\tilde{a}_{kj}^{(k)}}{\tilde{a}_{kk}^{(k)}} \right| |\tilde{a}_{ik}^{(k)}|. \tag{2.7}$$

Therefore

$$\|r_i^{(k+1)}\|_1 \le \sum_{j=k+1}^{n} \left(|\tilde{a}_{ij}^{(k)}| + \left| \frac{\tilde{a}_{kj}^{(k)}}{\tilde{a}_{kk}^{(k)}} \right| |\tilde{a}_{ik}^{(k)}| \right)$$

and so

$$\|r_i^{(k+1)}\|_1 \le (\|\tilde{r}_i^{(k)}\|_1 - |\tilde{a}_{ik}^{(k)}|) + |\tilde{a}_{ik}^{(k)}| \sum_{j=k+1}^{n} \frac{|\tilde{a}_{kj}^{(k)}|}{|\tilde{a}_{kk}^{(k)}|},$$

which can be written as

$$\|r_i^{(k+1)}\|_1 \le (\|\tilde{r}_i^{(k)}\|_1 - |\tilde{a}_{ik}^{(k)}|) + |\tilde{a}_{ik}^{(k)}| \frac{\|\tilde{r}_k^{(k)}\|_1 - |\tilde{a}_{kk}^{(k)}|}{|\tilde{a}_{kk}^{(k)}|}.$$

Hence, (2.6) follows.

We can observe that inequalities (2.7) can become equalities, depending on the sign of the elements of $\tilde{A}^{(k)}$. If (2.7) holds as an equality for each $i = k+1, \dots, n$, $j = k+1, \dots, n$, then the following inequalities become equalities and (2.6) also holds as an equality. □

Observe that the row (resp., symmetric) scaled partial pivoting chooses the maximal scaled pivots among all pivoting strategies interchanging rows (resp., the same rows and columns). From Proposition 2.1 we see that maximizing the scaled pivots is related to minimizing the quotients

$$\|r_i^{(k+1)}\|_1 / \|\tilde{r}_i^{(k)}\|_1.$$

Hence, we can say that the scaled partial pivoting strategies satisfy a "local optimality" property in the sense that the growth of row norms at a single step of Gaussian elimination is minimized.

We now can derive from Proposition 2.1 the following consequences on the growth factor ρ in terms of the scaled pivots.

COROLLARY 2.2. *Let $p_{i_1}, \dots, p_{i_r}$ be the scaled pivots less than $1/2$, and let ρ be the growth factor* (1.2) *of Gaussian elimination with a given pivoting strategy. Then $\rho = 1$ if $r = 0$ and $1 \le \rho \le \prod_{1 \le k \le r}((1 - p_{i_k})/p_{i_k})$ if $r > 0$.*

Proof. If $i \in \{1, \dots, k\}$, then $\|r_i^{(k+1)}\|_1 = \|\tilde{r}_i^{(k)}\|_1$. Let us assume that $i \in \{k+1, \dots, n\}$. If $p_k \ge 1/2$, then $\frac{1}{p_k} - 2 \le 0$ and we deduce from (2.6) that $\|r_i^{(k+1)}\|_1 \le \|\tilde{r}_i^{(k)}\|_1$. If $p_k < 1/2$, it is sufficient to apply (2.6) and observe that $(\frac{1}{p_k} - 2)|\tilde{a}_{ik}^{(k)}| \le (\frac{1}{p_k} - 2)\|\tilde{r}_i^{(k)}\|_1$ in order to derive $\|r_i^{(k+1)}\|_1 \le \frac{1-p_k}{p_k}\|\tilde{r}_i^{(k)}\|_1$. Iterating the previous arguments, we can deduce the result. □

Although the bound of Corollary 2.2 is not as sharp as the bound of Proposition 2.1, it can be useful in some cases. The case $r = 0$ in the previous corollary corresponds to $p_k \ge 1/2$ for all $k = 1, \dots, n-1$, which holds if and only if the upper triangular matrix obtained after Gaussian elimination is diagonally dominant. In this case, $\rho = 1$ (let us recall that, in this case, we proved in [7] that the usual growth factor satisfies $\rho^W \le n - 1$). On the other hand, Corollary 2.2 can be useful if we compare it with the estimate $\|A\|/\|U\|$ used in LAPACK [2, p. 241] for the inverse of the growth factor. Although LAPACK uses only this estimate for Gaussian elimination with row exchanges, if it is used without row exchanges it can be unreliable. For instance, if we consider the matrix

$$A = A^{(1)} = \begin{pmatrix} \varepsilon & 1 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix}$$

and apply Gaussian elimination without row exchanges, we obtain the matrices

$$A^{(2)} = \begin{pmatrix} \varepsilon & 1 & 0 \\ 0 & 1 & 0 \\ 0 & -\frac{1}{\varepsilon} & 1 \end{pmatrix}, \quad A^{(3)} = U = \begin{pmatrix} \varepsilon & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Then $\|U\|/\|A\| = 1$ and it does not show the growth corresponding to the matrix $A^{(2)}$: in fact, $\rho(A) = 1/\varepsilon$. The bound provided by Corollary 2.2 is also $1/\varepsilon$.

3. Scaled pivots and the condition number. Let us recall that Cond(A) represents the Skeel condition number of a matrix A. Given the scaled pivots p_k (2.6) of a nonsingular matrix, let us consider the *minimal scaled pivot* p

$$p := \min_{1 \le k \le n} p_k. \tag{3.1}$$

Let us observe that $0 < p \le 1$.

The following result provides a bound for the Skeel condition number of the triangular matrix resulting after Gaussian elimination in terms of the minimal scaled pivot.

THEOREM 3.1. *Let $U = (u_{ij})_{1\le i,j\le n}$ be the upper triangular matrix obtained after performing Gaussian elimination on a nonsingular matrix A with a pivoting strategy with minimal scaled pivot p (see (3.1)). Then Cond$(U) \le \frac{2-2p}{1-2p}\left(\frac{1-p}{p}\right)^{n-1} - \frac{1}{1-2p}$ if $p \ne \frac{1}{2}$ and Cond$(U) \le 2n-1$ if $p = \frac{1}{2}$.*

Proof. Let $V := D^{-1}U$, where D is the diagonal matrix whose (i,i)-entry is u_{ii} for all i. Then Cond$(U) =$ Cond(V), and $V = (V_{ij})_{1\le i,j\le n}$ is upper triangular with $V_{ii} = 1$ and

$$\sum_{j=i+1}^{n} |V_{ij}| = \sum_{j=i+1}^{n} \frac{|u_{ij}|}{|u_{ii}|} = \frac{\sum_{j=i}^{n} |u_{ij}| - |u_{ii}|}{|u_{ii}|} = \frac{1}{p_i} - 1 = \frac{1-p_i}{p_i} \le \frac{1-p}{p} \tag{3.2}$$

for all i.

If we compute V^{-1} by Gauss–Jordan, starting from the last column, we can easily obtain the following bound for the absolute value of $(V^{-1})_{ij}$ for any $i \in \{1,\dots,n\}$ and $j \ge i$:

(3.3)
$$|(V^{-1})_{ij}| \le |V_{ij}| + |V_{i,j-1}|\,|(V^{-1})_{j-1,j}| + |V_{i,j-2}|\,|(V^{-1})_{j-2,j}| + \cdots + |V_{i,i+1}|\,|(V^{-1})_{i+1,j}|.$$

Let us see by induction on $j-i$ that $|(V^{-1})_{ij}| \le \left(\frac{1-p}{p}\right)^{j-i}$. It holds when $j-i=0$ because $|(V^{-1})_{ii}| = 1$ and when $j-i=1$ because $|(V^{-1})_{i,i+1}| = |V_{i,i+1}| \le \frac{1-p}{p}$ by (3.2). Let us assume that it holds when $j-i \le k$, and let us prove it when $j-i = k+1$. In this case, if we apply (3.2) and the induction hypothesis to (3.3) we derive

$$|(V^{-1})_{ij}| \le \left(\frac{1-p}{p}\right)^{k} \sum_{j\ne i} |V_{ij}| \le \left(\frac{1-p}{p}\right)^{k+1} = \left(\frac{1-p}{p}\right)^{j-i}.$$

Let $W := |V^{-1}|\,|V|$. Then, taking into account (3.2) and that, by (3.2) $\sum_{j=k}^{n} |V_{kj}| \leq \frac{1}{p}$, we have for each $i \in \{1, \dots, n\}$ that

(3.4)

$$\sum_{j=i}^{n} |w_{ij}| = \sum_{j=i}^{n} \sum_{k=i}^{j} |(V^{-1})_{ik}||V_{kj}| = \sum_{k=i}^{n} \sum_{j=k}^{n} |(V^{-1})_{ik}||V_{kj}|$$
$$= \left(\sum_{k=i}^{n-1} \left(\sum_{j=k}^{n} |(V^{-1})_{ik}||V_{kj}| \right) \right) + |(V^{-1})_{in}||V_{nn}| \leq \sum_{k=i}^{n-1} \frac{1}{p} \left(\frac{1-p}{p} \right)^{k-i} + \left(\frac{1-p}{p} \right)^{n-i}.$$

If $p = 1/2$, then we obtain from (3.4)

$$\sum_{j=i}^{n} |w_{ij}| \leq 2n - 2i + 1. \tag{3.5}$$

If $p \neq 1/2$, then we can deduce from (3.4)

(3.6)

$$\sum_{j=i}^{n} |w_{ij}| \leq \frac{1}{p} \frac{((1-p)/p)^{n-i} - 1}{((1-p)/p) - 1} + \left(\frac{1-p}{p} \right)^{n-i} = \left(\frac{1}{1-2p} + 1 \right) \left(\frac{1-p}{p} \right)^{n-i} - \frac{1}{1-2p}$$
$$= \left(\frac{2-2p}{1-2p} \right) \left(\frac{1-p}{p} \right)^{n-i} - \frac{1}{1-2p},$$

and the result follows. □

Remark 3.2. The proof of the previous result can be applied to any nonsingular upper triangular matrix $U = (u_{ij})_{1 \leq i,j \leq n}$ such that $(|u_{ii}| / \sum_{j=i}^{n} |u_{ij}|) \geq p$ for all i. An analogous result to Theorem 3.1 also can be deduced for any nonsingular lower triangular matrix $L = (l_{ij})_{1 \leq i,j \leq n}$ such that $(|l_{ii}| / \sum_{j=1}^{i} |l_{ij}|) \geq p$ for all i. When $p = 1/2$ the matrices are diagonally dominant by rows and the bound of Theorem 3.1 for their Skeel condition number coincides with that obtained in Proposition 2.1 of [7]. The case $p > 1/2$ corresponds to the case of $n \times n$ matrices which are *strictly* diagonally dominant by rows. In this case $1 - 2p < 0$ and so we can derive the following result from Theorem 3.1, which provides a bound for the Skeel condition number which does not depend on n.

COROLLARY 3.3. *Let $U = (u_{ij})_{1 \leq i,j \leq n}$ be an upper triangular matrix which is strictly diagonally dominant by rows, and let $p := \min_{1 \leq i \leq n} \{|u_{ii}| / \sum_{j=i}^{n} |u_{ij}|\}$. Then Cond$(U) \leq \frac{1}{2p-1}$.*

In order to illustrate the use of the bound provided by Theorem 3.1, let us consider the matrices

$$A = \begin{pmatrix} 2 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 2 \end{pmatrix}, \quad U = \begin{pmatrix} 2 & -1 & 0 & 0 \\ 0 & \frac{3}{2} & -1 & 0 \\ 0 & 0 & \frac{4}{3} & -1 \\ 0 & 0 & 0 & \frac{5}{4} \end{pmatrix}.$$

Matrices with the structure of A often appear when applying the finite difference method for a boundary value problem. If we perform Gaussian elimination without row exchanges to A, we obtain the matrix U. Since U is strictly diagonally dominant by rows, we can apply Corollary 3.3 and obtain $\mathrm{Cond}(U) \leq 7$ ($p = 4/7$). However, if we apply the bound of Theorem 3.1, we get $\mathrm{Cond}(U) \leq 4.46875$. In this case, $\mathrm{Cond}(U) = 10/3$.

In section 5 we shall include examples where we also deal with $\kappa(U)$. On the other hand, the relationship of the row scaled partial pivoting and the Skeel condition number of U was analyzed in [5]. In the next section we provide some properties of the row scaled partial pivoting with respect to the growth factor ρ (1.2).

4. Row scaled partial pivoting strategies.

In this section, we obtain bounds for the growth factor ρ of row scaled partial pivoting strategies.

The usual growth factor ρ^W (1.1) of row scaled partial pivoting cannot be bounded above by the bound 2^{n-1} of partial pivoting (see [4, p. 192]). In Theorem 5.1 of [1] it was proved that the growth factor ρ (1.2) of partial pivoting is also bounded above by 2^{n-1}. The following result provides an upper bound for ρ when using row scaled partial pivoting, which is lower than that obtained from applying Corollary 2.2.

THEOREM 4.1. *Let $p_{i_1}, \ldots, p_{i_r}$ be the scaled pivots less than $1/2$ of the row scaled partial pivoting strategy for $\|\cdot\|_1$ for a nonsingular $n \times n$ matrix A. Then the growth factor ρ (1.2) for this strategy satisfies*

$$1 \leq \rho \leq 2^r \prod_{1 \leq k \leq r} (1 - p_{i_k}).$$

Proof. If $i \in \{1, \ldots, k\}$, then $\|r_i^{(k+1)}\|_1 = \|\tilde{r}_i^{(k)}\|_1$. Let us assume that $i \in \{k+1, \ldots, n\}$. If $p_k \geq 1/2$, then $\frac{1}{p_k} - 2 \leq 0$ and we deduce from (2.6) that $\|r_i^{(k+1)}\|_1 \leq \|\tilde{r}_i^{(k)}\|_1$. Let us assume that $p_k < 1/2$. Then $\frac{1}{p_k} - 2 > 0$ and so $1 - 2p_k > 0$.

Now, from (2.6) and (2.2) we deduce that for $i = k+1, \ldots, n$

$$\|r_i^{(k+1)}\|_1 \leq \|\tilde{r}_i^{(k)}\|_1 + (1 - 2p_k)|\tilde{a}_{ik}^{(k)}| \frac{\|\tilde{r}_k^{(k)}\|_1}{|\tilde{a}_{kk}^{(k)}|}$$

or, equivalently,

$$\|r_i^{(k+1)}\|_1 \leq \|\tilde{r}_i^{(k)}\|_1 + (1 - 2p_k)\|\tilde{r}_i^{(k)}\|_1 \frac{|\tilde{a}_{ik}^{(k)}|}{\|\tilde{r}_i^{(k)}\|_1} \frac{\|\tilde{r}_k^{(k)}\|_1}{|\tilde{a}_{kk}^{(k)}|}.$$

Taking into account (2.4) and $1 - 2p_k > 0$, we can derive

$$\|r_i^{(k+1)}\|_1 \leq \|\tilde{r}_i^{(k)}\|_1 + (1 - 2p_k)\|\tilde{r}_i^{(k)}\|_1 = 2(1 - p_k)\|\tilde{r}_i^{(k)}\|_1.$$

Iterating the previous arguments, we can deduce the result. □

Let us observe that, in contrast with section 2, in the following result we do not have to assume a bound for the scaled pivots in order to obtain a bound for the growth factor ρ. The proof in the case that there are scaled pivots less than $1/2$ is a consequence of Theorem 4.1 and the fact that $p_n = 1$. Otherwise, $\rho = 1$ by Corollary 2.2.

COROLLARY 4.2. *The growth factor ρ (1.2) of the row scaled partial pivoting strategy for the norm $\|\cdot\|_1$ for a nonsingular $n \times n$ matrix A satisfies*

$$1 \leq \rho \leq 2^{n-1}.$$

5. Examples. This section includes some examples where the results of this paper can be applied.

The stability of Gaussian elimination without pivoting when the coefficient matrix is diagonally dominant is well known (see pp. 288–289 of [11] and [10, pp. 122–123]). In Theorem 5.2 of [2] it was proved that for such matrices the growth factor $\rho = 1$. In previous sections, we have obtained nice bounds in the case that all scaled pivots are greater than or equal to $1/2$ or, equivalently, when the upper triangular matrix obtained after Gaussian elimination with a pivoting strategy is diagonally dominant by rows. In this case, we have proved that, for an $n \times n$ matrix, the growth factor $\rho = 1$ and the Skeel condition number $\mathrm{Cond}(U) \leq 2n-1$. Let us now present some classes of matrices and pivoting strategies satisfying the previous properties. A nonsingular matrix A is an *M-matrix* if it has positive diagonal entries and nonpositive off-diagonal entries, and A^{-1} is nonnegative. M-matrices have very important applications, for instance, in iterative methods in numerical analysis, in the analysis of dynamical systems, in economics, and in mathematical programming. Inverses of M-matrices arise, for instance, in the solution of certain integral equations and in certain physical problems. Given a matrix $A = (a_{ij})_{1\leq i,j\leq n}$, the *comparison matrix* $\mathcal{M}(A) = (m_{ij})_{1\leq i,j\leq n}$ is defined by $m_{ii} := |a_{ii}|$ and $m_{ij} := -|a_{ij}|$ if $i \neq j$. Finally, A is an *H-matrix* if its comparison matrix is an M-matrix. The following pivoting strategies were introduced in [7]. A row (resp., symmetric) *maximal relative diagonal dominance* (m.r.d.d.) pivoting is a pivoting (resp., symmetric pivoting) which chooses as pivot at the tth step $(1 \leq t \leq n-1)$ a row i_t satisfying $\frac{|a_{i_t t}^{(t)}|}{\sum_{j>t} |a_{i_t j}^{(t)}|} = \max_{t\leq i\leq n}\{\frac{|a_{it}^{(t)}|}{\sum_{j>t} |a_{ij}^{(t)}|}\}$ (resp., $\frac{|a_{i_t i_t}^{(t)}|}{\sum_{j\geq t, j\neq i_t} |a_{i_t j}^{(t)}|} = \max_{t\leq i\leq n}\{\frac{|a_{ii}^{(t)}|}{\sum_{j\geq t, j\neq i} |a_{ij}^{(t)}|}\}$). By Proposition 4.5 of [7], the symmetric m.r.d.d. pivoting strategy coincides with the symmetric scaled partial pivoting strategy for $\|\cdot\|_1$. A row (resp., symmetric) *maximal absolute diagonal dominance* (m.a.d.d.) pivoting is a row (resp., symmetric) pivoting which chooses as pivot at the tth step $(1 \leq t \leq n-1)$ a row i_t satisfying $|a_{i_t t}^{(t)}| - \sum_{j>t} |a_{i_t j}^{(t)}| = \max_{t\leq i\leq n}\{|a_{it}^{(t)}| - \sum_{j>t} |a_{ij}^{(t)}|\}$ (resp., $|a_{i_t i_t}^{(t)}| - \sum_{j\geq t, j\neq i_t} |a_{i_t j}^{(t)}| = \max_{t\leq i\leq n}\{|a_{ii}^{(t)}| - \sum_{j\geq t, j\neq i} |a_{ij}^{(t)}|\}$).

In Proposition 4.3 of [7] it was proved that if A is a nonsingular H-matrix, then any symmetric diagonally dominant pivoting strategy leads to an upper triangular matrix U which is diagonally dominant by rows. If A is a nonsingular M-matrix, then it is known (see, for instance, the proof of Theorem 4.4 of [7]) that the matrix U is in fact strictly diagonally dominant by rows. In this case, all scaled pivots are greater that $1/2$. If p is the minimal scaled pivot (3.1), then $\mathrm{Cond}(U) \leq (1/(2p-1))$, which is a bound independent of n (see Remark 3.2 and Corollary 3.3).

Given a nonsingular H-matrix A and any symmetric diagonally dominant pivoting strategy, since $\rho = 1$, we also have $\|U\|_\infty \leq \|A\|_\infty$. If A is an M-matrix, then, taking into account that the triangular matrices L and U associated with the strategies are also M-matrices, we can extend the arguments provided in [8] to any such strategy in order to prove that $\kappa(U) \leq \kappa(A)$ (see also [1]).

As shown in Remark 4.6 and Proposition 4.7 of [7], if A is a nonsingular $n \times n$ M-matrix, the symmetric m.a.d.d. pivoting strategy consists of choosing as pivot row in each step the row whose elements give a maximal sum, and it adds $\mathcal{O}(n^2)$ elementary operations to the computational cost of complete Gaussian elimination. Following the notations of Proposition 4.7 of [7], let $e := (1,\dots,1)^T$ and $b_1 := Ae$. The symmetric m.a.d.d. pivoting strategy produces the sequence of matrices (2.1) and the corresponding sequence of vectors:

$$b_1 = b_1^{(1)} \longrightarrow \tilde{b_1}^{(1)} \longrightarrow b_1^{(2)} \longrightarrow \tilde{b_1}^{(2)} \longrightarrow \cdots \longrightarrow b_1^{(n)} = \tilde{b_1}^{(n)} = c.$$

Taking into account that each component c_k of c gives the sum of the elements of the corresponding kth row of $U = (u_{ij})_{1 \le i,j \le n}$, that U is also a nonsingular M-matrix and so it has positive diagonal entries and nonpositive off-diagonal entries, and that $\tilde{a}_{kk}^{(k)} = u_k$, we can deduce from (2.2) that each scaled pivot p_k satisfies

$$p_k = \frac{u_{kk}}{2u_{kk} - c_k}. \tag{5.1}$$

Therefore, the calculation of each scaled pivot p_k requires one multiplication, one subtraction, and one division and so the calculation of the n scaled pivots associated with the symmetric m.a.d.d. pivoting strategy adds $\mathcal{O}(n)$ elementary operations to the complete computational cost of this strategy.

In addition to the good properties of row scaled partial pivoting strategies presented in section 4, let us recall that in [5] it was proved that, given a nonsingular matrix A, if there exists a permutation matrix P such that the LU-factorization of the matrix $B = PA$ satisfies $|LU| = |L|\,|U|$, then P is associated with the row scaled partial pivoting for any strictly monotone vector norm. This can be used to derive nice backward error bounds (see also [3] and section 9.2 of [4]). These strategies satisfy the following property.

PROPOSITION 5.1. *Let $B = (b_{ij})_{1 \le i,j \le n}$ be a nonsingular matrix. If the LU-factorization of B satisfies $|LU| = |L|\,|U|$, then $|b_{ij}^{(k)}| \le |b_{ij}|$ for all i, j, k.*

Proof. It is easy to deduce the equation

$$B^{(k)}[k, \dots, n] = L[k, \dots, n]U[k, \dots, n]. \tag{5.2}$$

Clearly,

$$|L[k, \dots, n]U[k, \dots, n]| \le |L|\,|U|. \tag{5.3}$$

The result follows from (5.2), (5.3), and the hypothesis. ☐

As a consequence of the previous result, the corresponding strategies satisfy $\rho = 1$ and $\rho^W = 1$. Let us mention some classes of matrices for which the associated pivoting strategies satisfy Proposition 5.1. *Totally positive* matrices are matrices with all their minors nonnegative. Totally positive matrices arise naturally in many areas of mathematics, statistics, and economics. Inverses of totally positive matrices appear, for instance, in difference approximations of boundary value problems of fourth order ordinary differential equations. For the following classes of matrices the row scaled partial pivoting for any strictly monotone vector norm produces no row exchanges and satisfies the hypothesis of Proposition 5.1 (see [5]): nonsingular totally positive matrices, inverses of totally positive matrices, inverses of M-matrices, tridiagonal symmetric positive definite matrices, and tridiagonal M-matrices.

Let us finish with an example in which $B = PA$ satisfies Proposition 5.1 for a permutation matrix P different from the identity. A class of matrices containing totally positive matrices is given by the sign-regular matrices. An $n \times m$ matrix A is *sign-regular* if, for each k ($1 \le k \le \min\{n, m\}$), all $k \times k$ submatrices of A have a determinant with the same nonstrict sign. In [6] there was introduced a pivoting strategy for sign-regular matrices which was called *first-last pivoting* due to the fact that we choose as pivot row at each step of Gaussian elimination either the first or the last row among all possible rows. If A is a nonsingular $n \times n$ sign-regular matrix, the first-last pivoting strategy adds at most $2n-2$ subtractions and $4n-4$ multiplications to the computational cost of complete Gaussian elimination. By Corollary 3.7 of [6], the matrix $B = PA$ (where P is the permutation matrix associated with the first-last pivoting strategy) satisfies the hypothesis of Proposition 5.1. Therefore, the first-last pivoting produces for nonsingular sign-regular matrices the same row exchanges as any row scaled partial pivoting for a strictly monotone vector norm.

REFERENCES

[1] P. Amodio and F. Mazzia, *A new approach to backward error analysis of LU factorization*, BIT, 39 (1999), pp. 385–402.

[2] E. Anderson, Z. Bai, C. Bischof, S. Blackford, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, and D. Sorensen, *LAPACK Users' Guide, Third Edition*, Software Environ. Tools 9, SIAM, Philadelphia, 1999.

[3] C. de Boor and A. Pinkus, *Backward error analysis for totally positive linear systems*, Numer. Math., 27 (1977), pp. 485–490.

[4] N. J. Higham, *Accuracy and Stability of Numerical Algorithms*, SIAM, Philadelphia, 1996.

[5] J. M. Peña, *Pivoting strategies leading to small bounds of the errors for certain linear systems*, IMA J. Numer. Anal., 16 (1996), pp. 141–153.

[6] J. M. Peña, *Backward stability of a pivoting strategy for sign-regular linear systems*, BIT, 37 (1997), pp. 910–924.

[7] J. M. Peña, *Pivoting strategies leading to diagonal dominance by rows*, Numer. Math., 81 (1998), pp. 293–304.

[8] J. M. Peña, *A note on a paper by P. Amodio and F. Mazzia*, BIT, 41 (2001), pp. 640–643.

[9] R. D. Skeel, *Scaling for numerical stability in Gaussian elimination*, J. Assoc. Comput. Mach., 26 (1979), pp. 494–526.

[10] B. Wendroff, *Theoretical Numerical Analysis*, Academic Press, New York, 1966.

[11] J. H. Wilkinson, *Error analysis of direct methods of matrix inversion*, J. Assoc. Comput. Mach., 8 (1961), pp. 281–330.

SIAM J. Numer. Anal.
Vol. 41, No. 3, pp. 1032–1051

SHARP CFL, DISCRETE KINETIC FORMULATION, AND ENTROPIC SCHEMES FOR SCALAR CONSERVATION LAWS*

CHARALAMBOS MAKRIDAKIS† and BENOÎT PERTHAME‡

Abstract. We consider semidiscrete and fully discrete conservative finite volume schemes approximating the solution to one-dimensional scalar conservation law. We show that all E-schemes are associated with a discrete kinetic formulation with a nonnegative kinetic defect measure. This construction provides an alternative proof of the discrete local entropy inequalities with simple expressions of the discrete entropy fluxes. In contrast to the known results, which are restricted to CFL of the form $\lambda Q \leq 1/2$, our proof holds under "sharp" CFL conditions.

Key words. scalar conservation laws, kinetic formulations, finite volume methods, CFL condition

AMS subject classifications. 35L65, 65M12, 76L05

DOI. 10.1137/S0036142902402997

1. Introduction.

We consider conservative schemes approximating the scalar conservation law

$$\frac{\partial}{\partial t}u + \frac{\partial}{\partial x}A(u) = 0, \tag{1.1}$$

$$u(t=0,x) = u^0(x) \in L^1 \cap L^\infty(\mathbb{R}). \tag{1.2}$$

We assume and denote

$$A(0) = 0, \qquad a(\cdot) = A'(\cdot).$$

As usual [6], [17], [18], (1.1) is completed by the family of entropy inequalities; for any convex function S, there holds

$$\frac{\partial}{\partial t}S(u) + \frac{\partial}{\partial x}\eta^S(u) \leq 0, \tag{1.3}$$

with $\eta^S(u) = \int_0^u S'(\xi)\, a(\xi)\, d\xi$.

The purpose of this paper is to give new proofs under improved CFL conditions for discrete local entropy inequalities for a wide class of conservative entropic schemes for (1.1), the E-schemes (see [14]), and to investigate the connection between these schemes and the discretization of the kinetic formulation of the conservation law as introduced in [13], [12]. In fact, the proofs of the local entropy inequalities follow (as in the continuous case) from the positiveness of the defect measure appearing in the discrete kinetic formulation of the schemes.

*Received by the editors February 22, 2002; accepted for publication (in revised form) September 6, 2002; published electronically June 18, 2003. This work was partially supported by HYKE European Programme HPRN-CT-2002-00282.

http://www.siam.org/journals/sinum/41-3/40299.html

†Department of Applied Mathematics, University of Crete, 71409 Heraklion-Crete, Greece and Institute of Applied and Computational Mathematics, FORTH, 71110 Heraklion-Crete, Greece (makr@math.uoc.gr).

‡Département de Mathématiques et Applications, UMR 8553, Ecole Normale Supérieure, 45, rue d'Ulm, 75230 Paris Cedex 05, France (Benoit.Perthame@ens.fr).

In order to describe our results we introduce the following notation for the discretization. For simplicity we take a uniform mesh, but the analysis covers the nonuniform case with appropriate modifications; cf. Remark 3.6.

- $h > 0$ is the uniform mesh size,
- $x_{i+1/2} = (i + 1/2)h$, $i \in \mathbb{Z}$, are the cell interfaces,
- C_i denotes the cell $(x_{i-1/2}, x_{i+1/2})$,
- Δt is the time step, $t^n = n\Delta t$,
- $v_i(t)$ (resp., v_i^n) denotes the solution to the numerical scheme,
- $\lambda = \frac{\Delta t}{h}$.

The principle of the finite volume method consists of conservation approximations of the solution cell-averages

$$u_i(t) = \frac{1}{h_i}\int_{C_i} u(t,x)\,dx \qquad \text{(semidiscrete case)},$$

$$u_i^n = \frac{1}{h_i}\int_{C_i} u(t^n,x)\,dx \qquad \text{(fully discrete case)}.$$

We study both semidiscrete and fully discrete conservative schemes based on a two point numerical flux $\mathcal{A} = \mathcal{A}(u,v)$. It is chosen so that the numerical fluxes, called below $A_{i+1/2}(t)$, approximate the exact fluxes $A(u(t, x_{i+1/2}))$. We thus require the numerical fluxes to be *consistent*, i.e., $\mathcal{A}(u,u) = A(u)$ [8], [9], [10], [11]. The semidiscrete scheme is defined by

$$\begin{aligned} h\frac{d}{dt}v_i(t) + A_{i+1/2}(t) - A_{i-1/2}(t) &= 0, \qquad i \in \mathbb{Z}, \\ v_i(t=0) = v_i^0 &\in l^1(\mathbb{Z}) \quad \text{given}, \\ A_{i+1/2}(t) &= \mathcal{A}(v_i(t), v_{i+1}(t)). \end{aligned} \tag{1.4}$$

The corresponding fully discrete scheme that we consider is

$$\begin{aligned} v_i^{n+1} - v_i^n + \lambda\big(A_{i+1/2}^n - A_{i-1/2}^n\big) &= 0, \qquad i \in \mathbb{Z}, \\ v_i(t=0) = v_i^0 &\in l^1(\mathbb{Z}) \quad \text{given}, \\ A_{i+1/2}^n &= \mathcal{A}(v_i^n, v_{i+1}^n), \end{aligned} \tag{1.5}$$

where $\mathcal{A}$ is the same numerical flux.

Local entropy inequalities. It is well known that a key property that guarantees the convergence of the schemes to the unique entropy solution of the conservation law is to satisfy a discrete version of the entropy inequalities associated with (1.1). Namely, we are interested in schemes (1.4) for which in-cell entropy inequalities hold; i.e., for any convex function S, there holds

$$\begin{aligned} h\frac{d}{dt}S(v_i(t)) + \eta_{i+1/2} - \eta_{i-1/2} &\le 0, \\ \eta_{i+1/2} &= \eta(S; v_i(t), v_{i+1}(t)) \end{aligned} \tag{1.6}$$

for some appropriate entropy discrete flux function $\eta(S\,; u, v)$. For fully discrete schemes we require

$$S(v_i^{n+1}) - S(v_i^n) + \lambda\big(\eta_{i+1/2}^n - \eta_{i-1/2}^n\big) \le 0, \qquad i \in \mathbb{Z}. \tag{1.7}$$

A related class of schemes are the *E-schemes* introduced by Osher [14] in the semidiscrete case. These are the schemes for which the Lipschitz continuous function $\mathcal{A}(u,v)$ satisfies

$$\begin{aligned} &\mathcal{A}(u,v) \le A(\xi) \qquad \text{for } u \le \xi \le v, \\ &\mathcal{A}(u,v) \ge A(\xi) \qquad \text{for } v \le \xi \le u. \end{aligned} \tag{1.8}$$

In Osher [14], it was shown that (1.6) follows from the E-property for the flux. In the fully discrete case, Tadmor [19] showed that, under certain CFL limitations, E-schemes satisfy (1.7). Tadmor's seminal approach requires writing (1.5) in its viscosity form

$$\begin{aligned} v_i^{n+1} = v_i^n &- \frac{\lambda}{2}\big(A(v_{i+1}^n) - A(v_{i-1}^n) \big) \\ &+ Q_{i+1/2}\big(v_{i+1}^n - v_i^n\big) - Q_{i-1/2}\big(v_i^n - v_{i-1}^n\big), \end{aligned} \tag{1.9}$$

where the *viscosity coefficient* $Q_{i+1/2}$ is

$$\begin{aligned} &Q_{i+1/2} = Q(v_i^n, v_{i+1}^n), \\ &Q(u,v) = \lambda \, \frac{A(u) + A(v) - 2\mathcal{A}(u,v)}{v-u}\,. \end{aligned} \tag{1.10}$$

In [19] was shown that if $\mathcal{A}(u,v)$ satisfies the E-property and the CFL conditions

$$\begin{aligned} &Q_{i+1/2} \le \frac{1}{2}, \\ &\lambda \max_\xi |a(\xi)| \le \frac{1}{2} \end{aligned} \tag{1.11}$$

are met, then the fully discrete scheme (1.5) satisfies the in-cell entropy inequalities (1.7). Later this proof was extended to the multidimensional finite volume setting in [5], [1]; see also [15] for an improved version. As it was already noticed in [19], (1.11) are stronger than one would like them to be. Indeed, consider the two limiting cases for Q that correspond to Godunov and Lax–Friedrichs schemes (denote Q^G and Q^{LF} their numerical viscosity coefficients). Then one can check that for all E-schemes [14], [19], $Q^G \le Q$, but Godunov's scheme is known to satisfy (1.7) under the following sharp CFL: $\lambda \max_\xi |a(\xi)| \le 1$. In addition, $Q^{LF} \equiv 1$, i.e., (1.11), is also restrictive. The reason behind this restricted CFL is the method of proof in [19] which splits the numerical cell into two half subcells and reduces the numerical viscosity of any E-scheme in a convex combination of Godunov and modified Lax–Friedrichs schemes. This splitting into two subcells avoids analyzing the interaction of waves but leads to the restricted CFL.

In what follows we show that indeed (1.11) can be relaxed to the "sharp" conditions

$$\begin{aligned} &Q_{i+1/2} \le 1, \\ &\lambda \max_\xi |a(\xi)| \le 1\,. \end{aligned} \tag{1.12}$$

Our proofs do not rely on the above comparison with Godunov and modified Lax–Friedrichs; rather it is based on the kinetic formulation of E-schemes that we present in what follows.

Kinetic formulation. To each one of the schemes considered we will associate a discrete kinetic scheme. To do that we first consider the kinetic formulation of the conservation law (1.1) introduced in [13] (see also [16]):

$$\frac{\partial}{\partial t}f(x,t,\xi)+a(\xi)\frac{\partial}{\partial x}f(x,t,\xi)=\frac{\partial}{\partial \xi}m(x,t,\xi)\,. \tag{1.13}$$

Then $f(x,t,\xi)=\chi(\xi,u(x,t))$, and m is a nonnegative bounded measure with compact support with respect to ξ if and only if

$$u(x,t)=\int_{\mathbb{R}}f(x,t,\xi)d\xi$$

is the unique entropy solution of the conservation law (1.1). The kinetic equation has incorporated all the entropy inequalities (1.3). We use the standard notation for the signed characteristic function, $a\in\mathbb{R}$,

$$\chi(\xi,a)=\begin{cases}1, & 0<\xi\le a,\\ -1, & a\le\xi<0,\\ 0 & \text{otherwise}\,.\end{cases} \tag{1.14}$$

For later reference note the following key property of χ that allows us to derive (1.3) integrating (1.13) against $S'(\xi)\,d\xi$; for all the continuous functions S

$$\int\chi(\xi,a)S'(\xi)d\xi=S(a)-S(0)\quad\text{for all } a\in\mathbb{R}. \tag{1.15}$$

One of the results of this paper is that, for any E-flux, an appropriate upwind discretization of the linear transport part of (1.13) provides a discrete kinetic formulation. In other words, when $v_i(t)$, $i\in\mathbb{Z}$, is given through (1.4), then there also holds

$$\begin{aligned}(1.16)\quad &h\,\frac{\partial}{\partial t}\chi(\xi,v_i(t))+\big[a_+(\xi,v_i,v_{i+1})\chi(\xi,v_i(t))-a_-(\xi,v_i,v_{i+1})\chi(\xi,v_{i+1}(t))\big]\\ &\qquad-\big[a_+(\xi,v_{i-1},v_i)\chi(\xi,v_{i-1}(t))-a_-(\xi,v_{i-1},v_i)\chi(\xi,v_i(t))\big]=\frac{\partial}{\partial\xi}m_i(t,\xi),\end{aligned}$$

where the functions in the right-hand side—called the kinetic defect measures—satisfy

$$\begin{aligned}&m_i(t,\xi)=m_-(\xi,v_{i-1},v_i)+m_+(\xi,v_i,v_{i+1}),\quad m_\pm(\cdot,u,v)\ge 0,\\ &m_\pm(\cdot,u,v)\text{ vanish outside of the nonordered interval }(u,v),\end{aligned} \tag{1.17}$$

and the numerical speeds $a_\pm$, bounded by quantities of order $\frac{\partial}{\partial u}\mathcal{A}$, $\frac{\partial}{\partial v}\mathcal{A}$, or $a(\xi)$, satisfy

$$\begin{aligned}&a_\pm(\xi,u,v)\ge 0,\\ &a_+(\xi,u,v)=\max(0,a(\xi)),\quad a_-(\xi,u,v)=\max(0,-a(\xi))\quad\text{for }\xi\notin(u,v),\\ &\mathcal{A}(u,v)=\int_{\mathbb{R}}a_+(\xi,u,v)\,\chi(\xi,u)\,d\xi-\int_{\mathbb{R}}a_-(\xi,u,v)\,\chi(\xi,v)\,d\xi\,.\end{aligned} \tag{1.18}$$

Therefore a simple ξ integration shows that from a formula (1.16) one derives a semidiscrete scheme (1.4), and the entropy flux in (1.6) follows by integrating (1.16) against $S'(\xi)$:

$$\eta(S;u,v) = \int_{\mathbb{R}} a_+(\xi,u,v)\, S'(\xi)\, \chi(\xi,u)\, d\xi - \int_{\mathbb{R}} a_-(\xi,u,v)\, S'(\xi)\, \chi(\xi,v)\, d\xi\,. \tag{1.19}$$

At this level one can observe that, for an E-scheme, the S-linear entropy flux is not unique and several choices of a_+, a_- are possible that lead to different discrete entropy fluxes.

Notice that the most natural and simple example from this point of view is the Engquist–Osher scheme [7], where (1.16) holds with

$$\begin{aligned} &a_+(\xi) = \max(0, a(\xi)), \quad a_-(\xi) = \max(0, -a(\xi)),\\ &\mathcal{A}_{EO}(u,v) = A_+(u) + A_-(v),\\ &A_+(u) = \int_0^u a_+(\xi)\, d\xi, \qquad A_-(u) = \int_0^u a_-(\xi)\, d\xi. \end{aligned} \tag{1.20}$$

This case has the remarkable property that the discrete kinetic formulation (1.16) is a linear equation on χ, a fundamental property in the continuous formulation (1.13) which allows, for instance, a convergence proof of the Engquist–Osher scheme based on merely L^∞ bounds (see [2]). This simple case is also a model for kinetic schemes for systems of conservation laws [3], [20], [16] and allows us to give another convergence proof [21]. An alternative proof based on the framework of [13] and a kinetic formulation of Godunov's finite volume scheme was given in [22].

To recover fully discrete schemes (1.5) by a kinetic formulation is more intricate and therefore we may have to introduce more general discretizations of the linear transport part of (1.13). We thus define the following.

Definition 1.1. *The function $a(\xi,u,v)$, which is integrable and has compact support with respect to ξ, is called a discrete kinetic flux corresponding to $\mathcal{A}(u,v)$ if*

$$\begin{aligned} &\int_{\mathbb{R}} a(\xi,u,v) d\xi = \mathcal{A}(u,v)\,,\\ &a(\xi,u,u) = a(\xi)\chi(\xi,u) = A'(\xi)\chi(\xi,u)\,. \end{aligned} \tag{1.21}$$

In the semidiscrete case our choice can be, e.g., $a(\xi,u,v) = a_+(\xi,u,v)\,\chi(\xi,u) + a_-(\xi,u,v)\,\chi(\xi,v)$, but in the fully discrete case we have to consider more general representation formulas.

In section 2 we investigate the semidiscrete scheme (1.4), and we prove in Theorem 2.1 that E-schemes are characterized by the existence of a semidiscrete kinetic formulation (1.16). In fact, we show first that the existence of a more general discrete kinetic formulation (cf. (2.1)) is equivalent to the fact that $\mathcal{A}(u,v)$ is an E-flux. Towards this goal a crucial step is that the integrand of the discrete kinetic flux, defined in (2.8), should satisfy the requirements provided by Lemma 2.5 and further by Proposition 2.7.

In section 3 we investigate the fully discrete scheme (1.5) and the existence of a discrete kinetic flux corresponding to $\mathcal{A}(u,v)$, $a(\xi,u,v)$ such that

$$\begin{aligned} \chi(\xi,v_i^{n+1}) - \chi(\xi,v_i^n) + \lambda[a(\xi,v_i^n,v_{i+1}^n) - a(\xi,v_{i-1}^n,v_i^n)]&\\ = \frac{\partial}{\partial \xi} m_i^n(\xi),& \end{aligned} \tag{1.22}$$

with m_i^n a nonnegative function as in (1.17). Our main result is that if $\mathcal{A}(u,v)$ is an E-flux and the CFL conditions (1.12) are met, then we can construct $a(\xi,u,v) = a_\lambda(\xi,u,v)$ such that (1.22) is a kinetic formulation of (1.5) with m_i^n nonnegative, Theorem 3.1. Then the in-cell entropy inequalities (1.7) follow with entropy flux $\eta_{i+1/2} = \eta(\lambda, S\,;\,u,v) = \int_{\mathbb{R}} S'(\xi)a_\lambda(\xi,u,v)\,d\xi$. The proof is constructive, and the conditions on $a(\xi,u,v)$ derived in section 2 for the semidiscrete problem are particularly useful in the analysis.

In section 4 we give the construction, and additional explicit formulas, for the Engquist–Osher scheme. This section can be viewed as a model for the generic construction in section 2.

2. Semidiscrete schemes.

In this section we investigate general three point semidiscrete scheme (1.4) with consistent flux $\mathcal{A}(u,v)$. We prove the equivalence between three properties; the E-property, the possibility of writing a kinetic discretization as (1.16), and the existence of discrete entropy fluxes in (1.6).

Namely, the main result of this section is the following theorem.

THEOREM 2.1. *Consider the semidiscrete scheme* (1.4) *with a consistent discrete flux $\mathcal{A}(u,v)$. The following three properties are equivalent:*

(i) *$\mathcal{A}(u,v)$ is an E-flux as defined in* (1.8);

(ii) *all the in-cell entropy inequalities* (1.6), *i.e., for any convex function S, are satisfied;*

(iii) *there exists a discrete kinetic flux $a(\xi,u,v)$ corresponding to $\mathcal{A}(u,v)$, and nonnegative functions m_i satisfying* (1.17), *such that the kinetic formulation of* (1.4) *holds:*

$$h\,\frac{\partial}{\partial t}\chi(\xi,v_i(t)) + [a(\xi,v_i,v_{i+1}) - a(\xi,v_{i-1},v_i)] = \frac{\partial}{\partial \xi}m_i(t,\xi). \tag{2.1}$$

The entropy fluxes in (1.6), *as well as $a(\xi,u,v)$, are not unique, and a possible relation is*

$$\eta(S;u,v) = \int_{\mathbb{R}} S'(\xi)\,a(\xi,u,v)\,d\xi.$$

In addition, $a(\xi,u,v)$ admits an "upwind" splitting of the form (1.18).

We first recall the equivalence between properties (i) and (ii) for the sake of completeness. The property (iii) is then derived in several steps. We conclude this section with an explicit construction of discrete kinetic fluxes like (1.18).

Proof of Theorem 2.1. (i) ⇔ (ii). We depart from (ii). Multiplying (1.4) by $S'(v_i(t))$, we obtain that the in-cell entropy inequality is equivalent to the existence of $\eta(S,\cdot,\cdot)$ such that, for all values v_i, $v_{i\pm1}$ and all convex S, we have

$$\eta(S;v_i,v_{i+1}) - \eta(S;v_{i-1},v_i) \le S'(v_i)[\mathcal{A}(v_i,v_{i+1}) - \mathcal{A}(v_{i-1},v_i)],$$

which is equivalent to

$$\begin{cases}\eta(S;v_i,v_{i+1}) - \eta(S;v_i,v_i) \le S'(v_i)[\mathcal{A}(v_i,v_{i+1}) - A(v_i)],\\ \eta(S;v_i,v_i) - \eta(S;v_{i-1},v_i) \le S'(v_i)[A(v_i) - \mathcal{A}(v_{i-1},v_i)],\end{cases}$$

which is again equivalent, for all u, v, and S convex, to the existence of a function $\eta(S,\cdot,\cdot)$ such that

$$\eta(S;v,v) - S'(v)[A(v) - \mathcal{A}(u,v)] \le \eta(S;u,v) \le S'(u)[\mathcal{A}(u,v) - A(u)] + \eta(S;u,u).$$

Denoting $\eta(S;v)=\eta(S;v,v)$, the above inequality is obviously equivalent to (and then the choice of $\eta(S;u,v)$ is anything in between)

$$\eta(S;v)-S'(v)[A(v)-\mathcal{A}(u,v)]\le S'(u)[\mathcal{A}(u,v)-A(u)]+\eta(S;u),$$

or, in other words,

$$\eta(S;v)-\eta(S;u)\le S'(v)A(v)-S'(u)A(u)-\mathcal{A}(u,v)[S'(v)-S'(u)],$$

which is equivalent to

$$\begin{cases}\dfrac{\partial}{\partial u}\eta(S;u)=S'(u)a(u),\\ \displaystyle\int_v^u S''(\zeta)A(\zeta)d\zeta\le\mathcal{A}(u,v)[S'(u)-S'(v)],\end{cases}$$

and it remains to choose, as a generating family for S convex, the family $S''(\zeta)=\delta(\zeta-\xi)$ to recover the equivalence with the E-property (1.8).

We would like to conclude with noticing that the entropy fluxes are automatically consistent; i.e., the relation $\frac{\partial}{\partial u}\eta(S;u,u)=S'(u)a(u)$ is derived from (ii). □

We now introduce some steps towards the semidiscrete kinetic formulation (iii). We start with the following lemma.

LEMMA 2.2. *Let $a(\xi,u,v)$ be a discrete kinetic flux corresponding to $\mathcal{A}(u,v)$; then we have*

$$m_i(t,\xi)=m_+(\xi;v_i,v_{i+1})+m_-(\xi;v_{i-1},v_i), \tag{2.2}$$

with

$$\begin{aligned}m_+(\xi;u,v)&=\int_{-\infty}^{\xi}\delta(\zeta-u)[A(u)-\mathcal{A}(u,v)]\,d\zeta+\int_{-\infty}^{\xi}[\,a(\zeta,u,v)-a(\zeta)\chi(\zeta,u)]\,d\zeta,\\ m_-(\xi;u,v)&=-\int_{-\infty}^{\xi}\delta(\zeta-v)[A(v)-\mathcal{A}(u,v)]\,d\zeta-\int_{-\infty}^{\xi}[\,a(\zeta,u,v)-a(\zeta)\chi(\zeta,v)\,]\,d\zeta\,.\end{aligned} \tag{2.3}$$

Moreover, m_i is nonnegative for any value of its arguments if and only if both m_+ and m_- are nonnegative for any value of their arguments.

Proof. It is a simple matter to check that

$$\frac{\partial}{\partial t}\chi(\xi,v_i(t))=\delta(\xi-v_i(t))\,\frac{d}{dt}v_i(t)\,.$$

Then using the above formula and (1.4) in (2.1), we get

$$\begin{aligned}-\delta(\xi-v_i(t))\,[A_{i+1/2}-A_{i-1/2}]+[a(\xi,v_i(t),v_{i+1}(t))-a(\xi,v_{i-1}(t),v_i(t))]\\ =\frac{\partial}{\partial\xi}m_i(\xi,t)\end{aligned}$$

or, equivalently,

$$\begin{aligned}\frac{\partial}{\partial\xi}m_i(\xi,t)=\delta(\xi-v_i(t))\,[\,(A(v_i(t))-A_{i+1/2})-(A(v_i(t))-A_{i-1/2})\,]\\ +[a(\xi,v_i(t),v_{i+1}(t))-a(\xi)\chi(\xi,v_i(t))]\\ -[a(\xi,v_{i-1}(t),v_i(t))-a(\xi)\chi(\xi,v_i(t))\,]\,.\end{aligned}$$

Since we want m_i to have bounded support, we can integrate to obtain

$$\begin{aligned} m_i(\xi,t) = &\int_{-\infty}^{\xi} \delta(\zeta - v_i(t))\,[\,(A(v_i(t)) - A_{i+1/2}) - (A(v_i(t)) - A_{i-1/2})\,]\,d\zeta \\ &+ \int_{-\infty}^{\xi} [a(\zeta, v_i(t), v_{i+1}(t)) - a(\zeta)\chi(\zeta, v_i(t))] d\zeta \\ &- \int_{-\infty}^{\xi} [a(\zeta, v_{i-1}(t), v_i(t)) - a(\zeta)\chi(\zeta, v_i(t))\,]\,d\zeta \\ = &\; m_+(\xi, v_i(t), v_{i+1}(t)) + m_-(\xi, v_{i-1}(t), v_i(t)). \end{aligned}$$

By the definition of the discrete kinetic fluxes (Definition 1.1) and the consistency of the flux $\mathcal{A}(u,v)$ we see that

$$m_+(\xi, v, v) = 0, \qquad m_-(\xi, v, v) = 0;$$

thus m is nonnegative if and only if both m_+ and m_- are nonnegative. □

Remark 2.3. Since $m_\pm(+\infty; u, v) = m_\pm(-\infty; u, v) = 0$ we can see that (2.3) takes the form

$$\begin{aligned} m_+(\xi; u, v) &= \int_{-\infty}^{\xi} [\,a(\zeta, u, v) - a(\zeta)\chi(\zeta, u)\,]\,d\zeta \quad \text{for} \quad \xi < u, \\ m_+(\xi; u, v) &= -\int_{\xi}^{+\infty} [\,a(\zeta, u, v) - a(\zeta)\chi(\zeta, u)\,]\,d\zeta \quad \text{for} \quad u < \xi, \\ m_-(\xi; u, v) &= -\int_{-\infty}^{\xi} [\,a(\zeta, u, v) - a(\zeta)\chi(\zeta, v)\,]\,d\zeta \quad \text{for} \quad \xi < v, \\ m_-(\xi; u, v) &= \int_{\xi}^{+\infty} [\,a(\zeta, u, v) - a(\zeta)\chi(\zeta, v)\,]\,d\zeta \quad \text{for} \quad v < \xi. \end{aligned} \tag{2.4}$$

We proceed by further reducing the form of m_i. We need some more notation. For $u, v \in \mathbb{R}$ we denote $I_{u,v}$ the interval that they define. Also,

$$I_{u,v} = [m, M], \quad \text{where } m = \min\{u, v\} \text{ and } M = \max\{u, v\}\,. \tag{2.5}$$

We then notice the identity

$$\chi(\xi, u) = \chi(\xi, v) \quad \text{for } \xi \in \mathbb{R} \setminus I_{u,v}\,. \tag{2.6}$$

Then one may check the following lemma.

LEMMA 2.4. *Assume that $a(\xi, u, v)$ is a discrete kinetic flux corresponding to $\mathcal{A}(u, v)$. If m_+ and m_- are both nonnegative, then $a(\xi, u, v)$ satisfies the consistency condition outside the interval $I_{u,v}$:*

$$a(\xi, u, v) = a(\xi)\chi(\xi, u) = a(\xi)\chi(\xi, v), \quad \xi \in \mathbb{R} \setminus I_{u,v}\,. \tag{2.7}$$

Conversely, if (2.7) *is satisfied, then m_+ and m_- vanish (and therefore are nonnegative) outside the interval $I_{u,v}$.*

Proof. Assume first that $\xi < m < 0$; then $\chi(\xi, u) = \chi(\xi, v) = 0$. In addition, both m_+ and m_- are nonnegative; therefore (2.4) implies that

$$\int_{-\infty}^{\xi} a(\zeta, u, v)\,d\zeta = 0.$$

Since ξ is arbitrary, $a(\xi,u,v)=0$ for $\xi<m<0$. Similarly, if $\xi>M>0$, $a(\xi,u,v)=0$. In the case where $M<0$ and $M<\xi$, again by (2.4) we have

$$\int_{\xi}^{+\infty}[\,a(\zeta,u,v)-a(\zeta)\chi(\zeta,u)\,]\,d\zeta=0.$$

Since $M<\xi$ is arbitrary, we conclude that $a(\xi,u,v)=a(\xi)\chi(\xi,u)=a(\xi)\chi(\xi,v)$. The proof is similar in the case $\xi<m$, $m>0$. □

In view of Lemma 2.4 we are able to define a function $\mathcal{A}(\xi,u,v)$ of three variables as

$$\mathcal{A}(\xi,u,v)=\int_{-\infty}^{\xi}a(\zeta,u,v)\,d\zeta\,. \tag{2.8}$$

It is to be noted that $\mathcal{A}(\xi,u,v)$ should not be confused with the discrete flux $\mathcal{A}(u,v)$, although they are of course related depending on the values of ξ since by Definition 1.1

$$\mathcal{A}(+\infty,u,v)=\mathcal{A}(u,v).$$

In the next lemma we derive conditions for the discrete kinetic flux in $I_{u,v}$ by using its integrand $\mathcal{A}(\xi,u,v)$.

Lemma 2.5. *Assume that $a(\xi,u,v)$ is a discrete kinetic flux corresponding to $\mathcal{A}(u,v)$, and that $\mathcal{A}(\xi,u,v)$ is defined by* (2.8)*. Let m_+ and m_- both be nonnegative. Then the following conditions are satisfied in the interval $I_{u,v}$:*

$$u\le\xi\le v\qquad\begin{cases}A(\xi)\ge\mathcal{A}(\xi,u,v)\ge\mathcal{A}(u,v) & \textit{when}\ \ \xi\ge 0,\\ 0\ge\mathcal{A}(\xi,u,v)\ge\mathcal{A}(u,v)-A(\xi) & \textit{when}\ \ \xi<0\end{cases} \tag{2.9}$$

and

$$v\le\xi\le u\qquad\begin{cases}\mathcal{A}(u,v)\ge\mathcal{A}(\xi,u,v)\ge A(\xi) & \textit{when}\ \ \xi\ge 0,\\ \mathcal{A}(u,v)-A(\xi)\ge\mathcal{A}(\xi,u,v)\ge 0 & \textit{when}\ \ \xi<0.\end{cases} \tag{2.10}$$

Conversely, if (2.9), (2.10) *are satisfied, then m_+ and m_- are nonnegative in the interval $I_{u,v}$.*

Proof. We treat only the case $u<\xi<v$ since the other is similar. Then equations (2.3) imply that

$$m_+(\xi,u,v)=A(u)-\mathcal{A}(u,v)+\mathcal{A}(\xi,u,v)-\int_{-\infty}^{\xi}a(\zeta)\chi(\zeta,u)\,d\zeta\,, \tag{2.11}$$

$$m_-(\xi,u,v)=-\mathcal{A}(\xi,u,v)+\int_{-\infty}^{\xi}a(\zeta)\chi(\zeta,v)\,d\zeta\,. \tag{2.12}$$

But then it is easy to check that

$$-\int_{-\infty}^{\xi}a(\zeta)\chi(\zeta,u)\,d\zeta=A(\xi)\mathbf{I}_{\{\xi<0\}}-A(u) \tag{2.13}$$

and

$$\int_{-\infty}^{\xi}a(\zeta)\chi(\zeta,v)\,d\zeta=A(\xi)\mathbf{I}_{\{\xi>0\}}. \tag{2.14}$$

Since both m_+ and m_- should be nonnegative, (2.9) follows. The converse is also immediate by using the above identities. ∎

Remark 2.6. From Lemma 2.4 we deduce that both m_+ and m_- are supported in $I_{u,v}$. Further, by the proof of the Lemma 2.5 we have the following formulas:

$$u \le \xi \le v: \quad \begin{aligned} &m_+(\xi;u,v) = \mathcal{A}(\xi,u,v) - \mathcal{A}(u,v) + A(\xi)\mathbf{1}_{\{\xi<0\}},\\ &m_-(\xi;u,v) = A(\xi)\mathbf{1}_{\{\xi>0\}} - \mathcal{A}(\xi,u,v)\end{aligned} \tag{2.15}$$

and

$$v \le \xi \le u: \quad \begin{aligned} &m_+(\xi;u,v) = \mathcal{A}(\xi,u,v) - A(\xi)\mathbf{1}_{\{\xi>0\}},\\ &m_-(\xi;u,v) = \mathcal{A}(u,v) - \mathcal{A}(\xi,u,v) - A(\xi)\mathbf{1}_{\{\xi<0\}}.\end{aligned} \tag{2.16}$$

We have now the following result.

PROPOSITION 2.7. *Assume that we have at our disposal a Lipschitz function $\mathcal{A}(\xi,u,v)$ which satisfies* (2.9) *and* (2.10) *and the endpoint values*

$$\textit{for } u \le v \quad \begin{cases} \mathcal{A}(u,u,v) = A(u)\mathbf{1}_{\{u>0\}},\\ \mathcal{A}(v,u,v) = \mathcal{A}(u,v) - A(v)\mathbf{1}_{\{v<0\}} \end{cases} \tag{2.17}$$

and

$$\textit{for } v \le u \quad \begin{cases} \mathcal{A}(u,u,v) = \mathcal{A}(u,v) - A(u)\mathbf{1}_{\{u<0\}},\\ \mathcal{A}(v,u,v) = A(v)\mathbf{1}_{\{v>0\}}. \end{cases} \tag{2.18}$$

Then $a(\xi,u,v)$ is well defined by

$$\begin{aligned} &a(\xi,u,v) = \frac{\partial}{\partial\xi}\mathcal{A}(\xi,u,v)\,, \quad \xi \in I_{u,v},\\ &a(\xi,u,v) = a(\xi)\chi(\xi,u) = a(\xi)\chi(\xi,v), \quad \xi \in \mathbb{R}\setminus I_{u,v}. \end{aligned} \tag{2.19}$$

In addition, $a(\xi,u,v)$ is a discrete kinetic flux corresponding to $\mathcal{A}(u,v)$ and (2.1) *is a kinetic formulation of* (1.4) *with m_i nonnegative.*

Proof. Having (2.8) and (2.6) in mind, we first extend $\mathcal{A}(\xi,u,v)$ outside the interval $I_{u,v}$ by letting

$$\mathcal{A}(\xi,u,v) = \int_{-\infty}^{\xi} a(\zeta)\chi(\zeta,u)d\zeta, \quad \xi \le m,$$

and

$$\mathcal{A}(\xi,u,v) = \int_{-\infty}^{m} a(\zeta)\chi(\zeta,u)d\zeta + A(M,u,v) - A(m,u,v) + \int_{M}^{\xi} a(\zeta)\chi(\zeta,u)d\zeta, \quad \xi \ge M.$$

Then the function $\mathcal{A}(\cdot,u,v)$ is a well defined, continuous function and $a(\xi,u,v)$ is the derivative of $\mathcal{A}(\xi,u,v)$, $\xi \in \mathbb{R}$. Then it is easy to see that since $\mathcal{A}(\xi,u,v)$ satisfies (2.9) and (2.10) with equalities at the endpoints of the interval $I_{u,v}$,

$$\mathcal{A}(+\infty,u,v) = \int_{\mathbb{R}} a(\xi,u,v)d\xi = A(u,v)\,;$$

i.e., $a(\xi,u,v)$ is a discrete kinetic flux corresponding to $\mathcal{A}(u,v)$. The proof is complete in view of Lemmas 2.4 and 2.5. □

We are now ready to complete the proof of the last equivalence in Theorem 2.1.

Proof of Theorem 2.1. (i) ⇔ (iii). Assume first (i), i.e., that $\mathcal{A}(u,v)$ is an E-flux. Then one can construct a discrete kinetic flux as in Lemma 2.5 and Proposition 2.7. Indeed, one choice of $\mathcal{A}(\xi,u,v)$ in $I_{u,v}$ is

$$\text{for } u \le \xi \le v: \quad \mathcal{A}(\xi,u,v) = \max\{\mathcal{A}(u,\xi)\,,\, \mathcal{A}(u,v)\} - A(\xi)\mathbf{1}_{\{\xi<0\}} \tag{2.20}$$

and

$$\text{for } v \le \xi \le u: \quad \mathcal{A}(\xi,u,v) = \min\{\mathcal{A}(\xi,v)\,,\, \mathcal{A}(u,v)\} - A(\xi)\mathbf{1}_{\{\xi<0\}}. \tag{2.21}$$

Then since $\mathcal{A}$ is an E-flux, it is straightforward to verify that $\mathcal{A}(\xi,u,v)$ satisfies (2.9) and (2.10) with equalities at the endpoints of the interval $I_{u,v}$. Therefore Proposition 2.7 implies that (iii) holds.

Conversely, if (iii) holds with m nonnegative, then $\mathcal{A}(\xi,u,v)$ defined in (2.8) should satisfy (2.9) and (2.10). But then necessarily $\mathcal{A}(u,v)$ satisfies

$$\begin{aligned} &\text{if } u \le \xi \le v: \quad A(\xi) \ge \mathcal{A}(u,v)\,,\\ &\text{if } v \le \xi \le u: \quad \mathcal{A}(u,v) \ge A(\xi)\,; \end{aligned}$$

i.e., $\mathcal{A}(u,v)$ is an E-flux and (i) is proved. □

End of the proof of Theorem 2.1. It remains to consider another choice in Lemma 2.5 and Proposition 2.7 in order to obtain the refined kinetic formulation (1.16) with signed speeds. We built an admissible (i.e., that satisfies (2.9), (2.10), (2.17), (2.18)) Lipschitz function $\mathcal{A}(\xi,u,v)$ which is nonincreasing in ξ for $u<v$ and increasing in ξ for $v<u$. We give the formula and skip the tedious but easy proof:

$$u \le \xi \le v\colon \quad \mathcal{A}(\xi,u,v) = \begin{cases} \max\left\{\mathcal{A}(u,v)\,,\, \min\limits_{\max(0,u)\le\zeta\le\xi} \mathcal{A}(u,\zeta)\right\} & \text{when } \xi \ge 0,\\ \max\limits_{\xi\le\zeta\le\min(0,v)} \{\max(\mathcal{A}(u,v)\,,\, \mathcal{A}(u,\zeta)) - A(\zeta)\} & \text{when } \xi < 0, \end{cases} \tag{2.22}$$

$$v \le \xi \le u\colon \quad \mathcal{A}(\xi,u,v) = \begin{cases} \min\left\{\mathcal{A}(u,v)\,,\, \max\limits_{\max(0,v)\le\zeta\le\xi} \mathcal{A}(\zeta,v)\right\} & \text{when } \xi \ge 0,\\ \min\limits_{\xi\le\zeta\le\min(0,u)} \{\min(\mathcal{A}(u,v), \mathcal{A}(\zeta,v)) - A(\zeta)\} & \text{when } \xi < 0. \end{cases} \tag{2.23}$$

Thanks to the monotonicity of $\mathcal{A}(\xi,u,v)$, one readily checks that indeed (1.16) holds with

$$a_\pm(\xi,u,v) = \left|\frac{\partial}{\partial\xi}\mathcal{A}(\xi,u,v)\right| \quad \text{for } \xi \in I_{u,v},$$

$$a_\pm(\xi,u,v) = a_\pm(\xi) \quad \text{for } \xi \in \mathbb{R}\setminus I_{u,v}. \qquad □$$

3. Fully discrete schemes. For a given fully discrete scheme (1.5) we will associate a discrete kinetic formulation as follows. Assume that we are given approximations at level $n : v_i^n,\ i \in \mathbb{Z}$. Define then the approximations at the next level as

$$\begin{aligned} f_i^{n+1} &= \chi(\xi, v_i^n) - \lambda[a(\xi, v_i^n, v_{i+1}^n) - a(\xi, v_{i-1}^n, v_i^n)] \quad \text{and} \\ v_i^{n+1} &= \int f_i^{n+1}(\xi) d\xi. \end{aligned} \tag{3.1}$$

We call (3.1) *a kinetic formulation* of the difference scheme (1.5) if $a(\xi, u, v)$ is a discrete kinetic flux corresponding to $\mathcal{A}(u, v)$ and there exist measures with compact support with respect to ξ, m_i^n such that

$$\chi(\xi, v_i^{n+1}) - f_i^{n+1} = \frac{\partial}{\partial \xi} m_i^n(\xi)\,. \tag{3.2}$$

Then integrating (3.2) with respect to ξ we recover the scheme (1.5). In such a case the discrete kinetic scheme can be written in a compact form as

$$\begin{aligned} \chi(\xi, v_i^{n+1}) - \chi(\xi, v_i^n) + \lambda[a(\xi, v_i^n, v_{i+1}^n) - a(\xi, v_{i-1}^n, v_i^n)] \\ = \frac{\partial}{\partial \xi} m_i^n(\xi)\,. \end{aligned} \tag{3.3}$$

In this section we will investigate under what conditions on $\mathcal{A}(u, v)$ and $a(\xi, u, v)$ the scheme (3.1) is a kinetic formulation of (1.5) with nonnegative $m_i^n(\xi)$, i.e., under what conditions (3.3) holds with $m_i^n(\xi)$ nonnegative. This will imply that the scheme satisfies all local discrete entropy inequalities.

THEOREM 3.1. *Consider a conservative scheme* (1.5) *with a consistent discrete flux* $\mathcal{A} = \mathcal{A}(u, v)$. *Assume the following:*

(i) $\mathcal{A}(u, v)$ *is an E-flux;*

(ii) *the CFL condition* (1.12) *is satisfied.*

Then there exists a discrete kinetic flux corresponding to $\mathcal{A}(u, v)$, $a = a_\lambda(\xi, u, v)$, *and a nonnegative measure* m *such that* (3.3) *is a kinetic formulation of* (1.5). *Consequently, all the in-cell entropy inequalities, i.e., for any convex function* S, *hold true:*

$$S\Big(v_i^{n+1}\Big) - S\Big(v_i^n\Big) + \lambda\Big[\,\eta_{i+1/2}^n - \eta_{i-1/2}^n\,\Big] \le 0,$$

with discrete entropy flux

$$\eta_{i+1/2}^n = \eta(S; v_i^n, v_{i+1}^n), \qquad \eta(S; u, v) = \int_{\mathbb{R}} S'(\xi)\, a(\xi, u, v)\, d\xi.$$

Remark 3.2. In our construction, the discrete kinetic flux depends on λ. We especially do not answer the open question to know whether, for E-schemes and under the CFL condition (1.12), there are in-cell entropy inequalities with $\eta(S)$ independent of λ. Because of this difference, it seems that a reverse theorem is wrong; the existence of a fully discrete kinetic formulation with a_λ, or of in-cell entropy inequalities with $\eta_\lambda(S)$, does not imply the E-property. Note that still in the construction of [19] the discrete entropy flux depends on λ. We recall that a weaker property, called “ordered schemes” (restrict the E-property to $\xi = u$ or v) is enough to have a TVD scheme; see [16].

As in the previous section, for $u, v \in \mathbb{R}$ we denote $I_{u,v}$ the interval that they define. We will need the following lemmas.

LEMMA 3.3. *Let $a(\xi, u, v)$ be a discrete kinetic flux corresponding to $\mathcal{A}(u, v)$, and assume that* (3.3) *is a kinetic formulation of* (1.5). *Setting*

$$I_{v_i^n, v_i^{n+1}} = [m, M], \qquad \text{where } m = \min\{v_i^n, v_i^{n+1}\} \text{ and } M = \max\{v_i^n, v_i^{n+1}\}\,, \tag{3.4}$$

we have for $\xi \in \mathbb{R} \setminus I_{v_i^n, v_i^{n+1}}$

(3.5)
$$m(\xi; v_{i-1}^n, v_i^n, v_{i+1}^n) = \lambda \int_{-\infty}^{\xi} [\, a(\zeta, v_i^n, v_{i+1}^n) - a(\zeta, v_{i-1}^n, v_i^n)]\, d\zeta \quad \text{for} \quad \xi < m,$$
$$m(\xi; v_{i-1}^n, v_i^n, v_{i+1}^n) = -\lambda \int_{\xi}^{+\infty} [\, a(\zeta, v_i^n, v_{i+1}^n) - a(\zeta, v_{i-1}^n, v_i^n)]\, d\zeta \quad \text{for} \quad M < \xi.$$

Proof. Assume first that $\xi < m < 0$; then $\chi(\xi, v_i^{n+1}) = \chi(\xi, v_i^n) = 0$. Also if $\xi < m$ and $m > 0$, then for $\xi < 0$, $\chi(\xi, v_i^{n+1}) = \chi(\xi, v_i^n) = 0$ and for $\xi > 0$, $\chi(\xi, v_i^{n+1}) = \chi(\xi, v_i^n) = 1$. Therefore

$$\int_{-\infty}^{\xi} \chi(\zeta, v_i^{n+1}) - \chi(\zeta, v_i^n)\, d\zeta = 0 \qquad \text{for } \xi < m\,.$$

Similarly, we show that

$$\int_{-\infty}^{\xi} \chi(\zeta, v_i^{n+1}) - \chi(\zeta, v_i^n)\, d\zeta = v_i^{n+1}) - v_i^n \qquad \text{for } \xi > M\,,$$

and therefore (3.5) follows in view of (1.5) and Definition 1.1. □

LEMMA 3.4. *Under the assumptions of Lemma* 3.3 *we have for $\xi \in I_{v_i^n, v_i^{n+1}}$*

(3.6)
$$m(\xi; v_{i-1}^n, v_i^n, v_{i+1}^n) = \xi - v_i^n + \lambda \int_{-\infty}^{\xi} [a(\zeta, v_i^n, v_{i+1}^n) - a(\zeta, v_{i-1}^n, v_i^n)] d\zeta \quad \text{for} \quad v_i^n < \xi,$$
$$m(\xi; v_{i-1}^n, v_i^n, v_{i+1}^n) = v_i^n - \xi - \lambda \int_{\xi}^{+\infty} [a(\zeta, v_i^n, v_{i+1}^n) - a(\zeta, v_{i-1}^n, v_i^n)] d\zeta \quad \text{for} \quad \xi < v_i^n.$$

Proof. Assume first that $v_i^n < \xi < v_i^{n+1}$. Then one can verify that

$$\int_{-\infty}^{\xi} \chi(\zeta, v_i^{n+1}) - \chi(\zeta, v_i^n)\, d\zeta = \xi - v_i^n\,,$$

which implies the first equality of (3.6). Similarly, there holds

$$\int_{-\infty}^{\xi} \chi(\zeta, v_i^{n+1}) - \chi(\zeta, v_i^n)\, d\zeta = v_i^{n+1} - \xi \qquad \text{for } v_i^{n+1} < \xi < v_i^n\,,$$

and in this case (3.6) follows again in view of (1.5) and Definition 1.1. □

We are ready now to prove the main result in this section.

Proof of Theorem 3.1. For the given discrete flux $\mathcal{A}(u, v)$ we first observe that if $a(\xi, u, v)$ is a function that is constructed according to Proposition 2.7, i.e., if $a(\xi, u, v)$

is a kinetic flux for the semidiscrete scheme, then $m(\xi; v_{i-1}^n, v_i^n, v_{i+1}^n)$ in Lemma 3.3 are nonnegative. Indeed, if $\xi < m$, then $\xi < v_i^n$ and

$$m(\xi; v_{i-1}^n, v_i^n, v_{i+1}^n) = m_+(\xi; v_i^n, v_{i+1}^n) + m_-(\xi; v_{i-1}^n, v_i^n)\,,$$

where m_+ and m_- are defined in (2.4) for $\xi < v_i^n$. A similar relation holds for $\xi > M$. The same reasoning implies that

$$\int_{-\infty}^{z} [\,a(\zeta, v_i^n, v_{i+1}^n) - a(\zeta, v_{i-1}^n, v_i^n)]\, d\zeta \geq 0 \quad \text{for all} \quad z < v_i^n$$

and

$$-\int_{z}^{+\infty} [\,a(\zeta, v_i^n, v_{i+1}^n) - a(\zeta, v_{i-1}^n, v_i^n)]\, d\zeta \geq 0 \quad \text{for all} \quad v_i^n < z\,.$$

Therefore, in such a case, $m(\xi; v_{i-1}^n, v_i^n, v_{i+1}^n)$ in Lemma 3.4 will be nonnegative if we are able to show that

$$\begin{aligned} &\xi - v_i^n + \lambda \int_{v_i^n}^{\xi} [\,a(\zeta, v_i^n, v_{i+1}^n) - a(\zeta, v_{i-1}^n, v_i^n)] d\zeta \geq 0 \quad \text{for} \quad v_i^n < \xi < v_i^{n+1},\\ &v_i^n - \xi - \lambda \int_{\xi}^{v_i^n} [\,a(\zeta, v_i^n, v_{i+1}^n) - a(\zeta, v_{i-1}^n, v_i^n)]\, d\zeta \geq 0 \quad \text{for} \quad v_i^{n+1} < \xi < v_i^n. \end{aligned} \tag{3.7}$$

Next, for $u, \overline{v}, \underline{v} \in \mathbb{R}$, let

$$\overline{u} = u - \lambda\big(\mathcal{A}(u, \overline{v}) - \mathcal{A}(\underline{v}, u)\,\big)\,. \tag{3.8}$$

The proof of the theorem is therefore reduced on finding a discrete kinetic flux $a(\xi, u, v)$ such that

(a) $a(\xi, u, v)$ satisfies the requirements of Proposition 2.7;

(b) for any $u, \overline{v}, \underline{v} \in \mathbb{R}$

$$\begin{aligned} M(\underline{v}, u, \overline{v}) &= \xi - u + \lambda \int_{u}^{\xi} [\,a(\zeta, u, \overline{v}) - a(\zeta, \underline{v}, u)] d\zeta \geq 0 \quad \text{for} \quad u < \xi < \overline{u},\\ M(\underline{v}, u, \overline{v}) &= u - \xi - \lambda \int_{\xi}^{u} [\,a(\zeta, u, \overline{v}) - a(\zeta, \underline{v}, u)]\, d\zeta \geq 0 \quad \text{for} \quad \overline{u} < \xi < u. \end{aligned}$$

In what follows we show that a discrete kinetic flux that satisfies (a) and (b) indeed exists. To motivate our construction we will consider first the cases

(I) $u < \xi < \overline{u},\ \ \xi < \{\overline{v}, \underline{v}\}$,

(II) $\overline{u} < \xi < u,\ \ \{\overline{v}, \underline{v}\} < \xi$.

In case (I) we have ($\mathcal{A}(\xi, u, v)$ is defined in (2.8))

$$\begin{aligned} M(\underline{v}, u, \overline{v}) &= \xi - u + \lambda\Big(\mathcal{A}(\xi, u, \overline{v}) - \mathcal{A}(u, u, \overline{v}) - \mathcal{A}(\xi, \underline{v}, u) + \mathcal{A}(u, \underline{v}, u)\Big)\\ &= \xi - u + \lambda\Big(\mathcal{A}(\xi, u, \overline{v}) - \mathcal{A}(\xi, \underline{v}, u)\Big)\\ &= \frac{1}{2}(\xi - u) + \lambda\mathcal{A}(\xi, u, \overline{v}) - \frac{\lambda}{2}\Big[A(u) + A(\xi)\Big]\\ &\quad + \frac{1}{2}(\xi - u) - \lambda\mathcal{A}(\xi, \underline{v}, u) + \frac{\lambda}{2}\Big[A(u) + A(\xi)\Big]\,, \end{aligned}$$

where we have used that (cf. Proposition 2.7)

$$\mathcal{A}(u,u,\overline{v}) = \mathcal{A}(u,\underline{v},u) = \int_{-\infty}^{u} a(\zeta)\chi(\zeta,u)\,d\zeta\,.$$

Assume for a moment that $\xi > 0$. Then $M(\underline{v},u,\overline{v})$ is nonnegative if

$$\begin{aligned} &\frac{1}{2}(u-\xi) + \frac{\lambda}{2}\Big[A(u)+A(\xi)\Big] \le \lambda\mathcal{A}(\xi,u,\overline{v}),\\ &\lambda\mathcal{A}(\xi,\underline{v},u) \le \frac{1}{2}(\xi-u) + \frac{\lambda}{2}\Big[A(u)+A(\xi)\Big]\,. \end{aligned} \tag{3.9}$$

Similarly, in case (II) we have

$$\begin{aligned} M(\underline{v},u,\overline{v}) &= u-\xi-\lambda\mathcal{A}(u,u,\overline{v})+\lambda\mathcal{A}(\xi,u,\overline{v})+\lambda\mathcal{A}(u,\underline{v},u)-\lambda\mathcal{A}(\xi,\underline{v},u)\\ &= u-\xi-\lambda\mathcal{A}(u,\overline{v})+\lambda\mathcal{A}(\xi,u,\overline{v})+\lambda\mathcal{A}(\underline{v},u)-\lambda\mathcal{A}(\xi,\underline{v},u)\\ &= \frac{1}{2}(u-\xi)-\lambda\mathcal{A}(u,\overline{v})+\lambda\mathcal{A}(\xi,u,\overline{v})+\frac{\lambda}{2}\Big[A(u)-A(\xi)\Big]\\ &\quad+\frac{1}{2}(u-\xi)+\lambda\mathcal{A}(\underline{v},u)-\lambda\mathcal{A}(\xi,\underline{v},u)-\frac{\lambda}{2}\Big[A(u)-A(\xi)\Big]\,, \end{aligned}$$

where we have used that (cf. Proposition 2.7)

$$\mathcal{A}(u,u,\overline{v}) = \mathcal{A}(u,\overline{v}) - \int_{u}^{+\infty} a(\zeta)\chi(\zeta,u)\,d\zeta\,,$$

and the similar relation for $\mathcal{A}(u,\underline{v},u)$. Still assuming $\xi > 0$, then $M(\underline{v},u,\overline{v})$ will be nonnegative if

$$\begin{aligned} &\frac{1}{2}(\xi-u)+\lambda\mathcal{A}(u,\overline{v})-\frac{\lambda}{2}\Big[A(u)-A(\xi)\Big] \le \lambda\mathcal{A}(\xi,u,\overline{v}),\\ &\lambda\mathcal{A}(\xi,\underline{v},u) \le \frac{1}{2}(u-\xi)+\lambda\mathcal{A}(\underline{v},u)-\frac{\lambda}{2}\Big[A(u)-A(\xi)\Big]\,. \end{aligned} \tag{3.10}$$

Relations (3.9) and (3.10) suggest the following:

$$\begin{aligned} \text{for } u\le\xi\le v: \quad &\frac{1}{2}(u-\xi)+\frac{\lambda}{2}\Big[A(u)+A(\xi)\Big]\\ &\le \lambda\mathcal{A}(\xi,u,v) \le \frac{1}{2}(v-\xi)+\lambda\mathcal{A}(u,v)-\frac{\lambda}{2}\Big[A(v)-A(\xi)\Big] \end{aligned} \tag{3.11}$$

and

$$\begin{aligned} \text{for } v\le\xi\le u: \quad &\frac{1}{2}(\xi-u)+\lambda\mathcal{A}(u,v)-\frac{\lambda}{2}\Big[A(u)-A(\xi)\Big]\\ &\le \lambda\mathcal{A}(\xi,u,v) \le \frac{1}{2}(\xi-v)+\frac{\lambda}{2}\Big[A(v)+A(\xi)\Big]\,. \end{aligned} \tag{3.12}$$

It will be convenient to introduce the following notation:

$$\begin{aligned} \lambda\overline{A}_{v>u}(\xi) &= \frac{1}{2}(v-\xi)+\lambda\mathcal{A}(u,v)-\frac{\lambda}{2}\Big[A(v)-A(\xi)\Big],\\ \lambda\underline{A}_{v>u}(\xi) &= \frac{1}{2}(u-\xi)+\frac{\lambda}{2}\Big[A(u)+A(\xi)\Big] \end{aligned} \tag{3.13}$$

and

$$\begin{aligned}\lambda\overline{A}_{u>v}(\xi) &= \frac{1}{2}(\xi - v) + \frac{\lambda}{2}\Big[A(v) + A(\xi)\Big],\\ \lambda\underline{A}_{u>v}(\xi) &= \frac{1}{2}(\xi - u) + \lambda\mathcal{A}(u,v) - \frac{\lambda}{2}\Big[A(u) - A(\xi)\Big]\,.\end{aligned}\tag{3.14}$$

A crucial fact is that, despite the E-property, we have indeed

$$\underline{A}_{v>u} \le \overline{A}_{v>u} \quad \text{and} \quad \underline{A}_{u>v} \le \overline{A}_{u>v}\,.$$

This is because by (1.12),

$$\lambda\overline{A}_{v>u} - \lambda\underline{A}_{v>u} = \frac{1}{2}(v-u) + \lambda\mathcal{A}(u,v) - \frac{\lambda}{2}\big[A(v)+A(u)\big] = \frac{1}{2}(v-u)\,\big[1-Q(u,v)\big] \ge 0$$

and

$$\lambda\overline{A}_{u>v} - \lambda\underline{A}_{u>v} = \frac{1}{2}(u-v) - \lambda\mathcal{A}(u,v) + \frac{\lambda}{2}\big[A(v)+A(u)\big] = \frac{1}{2}(u-v)\,\big[1-Q(u,v)\big] \ge 0.$$

Next, since we are looking for a flux that will satisfy (2.9) and (2.10), it is useful for what follows to notice that (1.12) implies that for $\xi \in I_{u,v}$

$$\overline{A}_{v>u}(\xi) \ge \mathcal{A}(u,v), \qquad \underline{A}_{v>u}(\xi) \le A(\xi)\tag{3.15}$$

and

$$\overline{A}_{u>v}(\xi) \ge A(\xi), \qquad \underline{A}_{u>v}(\xi) \le \mathcal{A}(u,v)\,.\tag{3.16}$$

We are ready now to define

$$\begin{aligned}B_{v>u}(\xi) &= \min\{\overline{A}_{v>u}(\xi), A(\xi)\} \quad \text{for } u \le \xi \le v,\\ B_{u>v}(\xi) &= \min\{\overline{A}_{u>v}(\xi), \mathcal{A}(u,v)\} \quad \text{for } v \le \xi \le u\,.\end{aligned}\tag{3.17}$$

Then, since $\mathcal{A}(u,v)$ is an E-flux the above relationships imply that

$$\mathcal{A}(u,v) \le B_{v>u}(\xi) \le A(\xi) \quad \text{and} \quad \underline{A}_{v>u}(\xi) \le B_{v>u}(\xi) \le \overline{A}_{v>u}(\xi)\,.\tag{3.18}$$

In addition,

$$B_{v>u}(u) = \min\{\overline{A}_{v>u}(u), A(u)\} = A(u) \text{ and } B_{v>u}(v) = \min\{\mathcal{A}(u,v), A(v)\} = \mathcal{A}(u,v),\tag{3.19}$$

where in the first equality in (3.19) we used $\lambda\overline{A}_{v>u}(u) - \lambda A(u) = \frac{1}{2}(v-u)\big[1 - Q(u,v)\big] \ge 0$. Similar relations hold for $B_{u>v}$. Hence, in the cases under consideration and for $\xi > 0$, it suffices to define in $I_{u,v}$

$$A(\xi,u,v) = \begin{cases} B_{v>u}(\xi) & \text{for } u \le \xi \le v,\ \xi > 0,\\ B_{u>v}(\xi) & \text{for } v \le \xi \le u,\ \xi > 0\,.\end{cases}\tag{3.20}$$

It is clear now that the right extension of $A(\xi,u,v)$ when $\xi < 0$ is

$$A(\xi,u,v) = \begin{cases} B_{v>u}(\xi) - A(\xi) & \text{for } u \le \xi \le v,\ \xi < 0,\\ B_{u>v}(\xi) - A(\xi) & \text{for } v \le \xi \le u,\ \xi < 0\,.\end{cases}\tag{3.21}$$

It is straightforward to see that this choice satisfies (a) and (b) in cases (I) and (II) and for $\xi < 0$.

In all the other cases the above choice of $A(\xi, u, v)$ satisfies (a) and (b). Property (a) is clear in any case. Property (b) is a consequence of (3.18) (and its corresponding relation for $u > v$) and of the CFL condition $\lambda \max_\xi |a(\xi)| \leq 1$. To illustrate this we consider only the case $\overline{u} \leq \xi \leq u$, $\overline{v} < \xi < 0$, $u < \underline{v} < 0$, the other cases being similar. Indeed,

$$\begin{aligned}
M(\underline{v}, u, \overline{v}) &= u - \xi - \lambda\mathcal{A}(u, u, \overline{v}) + \lambda\mathcal{A}(\xi, u, \overline{v}) + \lambda \int_\xi^u a(\zeta)\chi(\zeta, u) d\zeta \\
&= u - \xi - \lambda\mathcal{A}(u, \overline{v}) + \lambda \int_u^{+\infty} a(\zeta)\chi(\zeta, u) d\zeta + \lambda\mathcal{A}(\xi, u, \overline{v}) \\
&\geq \frac{1}{2}(u - \xi) - \frac{\lambda}{2}\Big[A(u) - A(\xi)\Big] \\
&\quad - \lambda A(\xi) + \lambda \int_u^{+\infty} a(\zeta)\chi(\zeta, u) d\zeta \\
&= \frac{1}{2}(u - \xi) - \frac{\lambda}{2}\Big[A(u) - A(\xi)\Big] + \lambda\Big[A(u) - A(\xi)\Big] \\
&= \frac{1}{2}(u - \xi) + \frac{\lambda}{2}\Big[A(u) - A(\xi)\Big] \geq 0\,.
\end{aligned}$$

The proof of the kinetic formulation is therefore complete.

The proof of the local entropy inequalities is immediate, again after integrating (1.22) against $S'(\xi)\, d\xi$. □

Remark 3.5. It should be noted that in the proof of the previous theorem the CFL condition (1.12) was used in its local form (only specific v_i^n appear)

$$\lambda Q_{i+1/2} = \lambda Q(v_i^n, v_{i+1}^n) \leq 1.$$

Indeed, tracing back in the proof we see that $\underline{v}$, u, and $\overline{v}$ represent the values v_{i-1}^n, v_i^n, and v_{i+1}^n.

Remark 3.6. Note that in the semidiscrete case the construction of the discrete kinetic flux is local and therefore the choice of uniform mesh is done only for notational simplicity. In the fully discrete case more care is needed in the construction of the discrete kinetic flux when nonuniform mesh is considered, essentially since $a(\xi, u, v)$ depends on λ. Indeed, denoting λ' the CFL number corresponding to the next interval in the construction of the previous theorem, a modification on the choice of $B_{v>u}(\xi)$, $B_{u>v}(\xi)$ is needed depending of the sign of $1 - \lambda'/\lambda$; e.g., one may choose

$$(3.22) \quad B_{v>u}(\xi) = \begin{cases} \min\{\overline{A}_{v>u}(\xi), A(\xi)\} & \text{for } u \leq \xi \leq v, \quad \text{provided } \lambda \geq \lambda', \\ \max\{\underline{A}_{v>u}(\xi), A(u, v)\} & \text{for } u \leq \xi \leq v, \quad \text{provided } \lambda' \geq \lambda. \end{cases}$$

4. Engquist–Osher scheme.

4.1. Semidiscrete Engquist–Osher scheme. We give first a direct proof for the kinetic formulation of the Engquist–Osher scheme [7]. We also give explicit formulas for the kinetic defect measures $m_\pm$.

THEOREM 4.1. *There is a unique nonnegative function $m_i(t, \xi)$ with bounded support in ξ such that the scheme* (1.4) *with Engquist–Osher flux* (1.20) *is equivalent* *to the kinetic equation* (1.16)–(1.17), *with* $a_\pm(\xi, u, v) = a_\pm(\xi)$. *Moreover, we have the bound*

$$\sum_{i\in\mathbb{Z}} h \int_0^\infty m_i(t,\xi)\,dt \le \mathbf{I}_{\{\xi\ge 0\}} \|(v_i^0 - \xi)_+\|_{l^1} + \mathbf{I}_{\{\xi\le 0\}} \|(\xi - v^0)_+\|_{l^1} \le \|v^0\|_{l^1},$$

and the functions $m_\pm$ *are given by*

$$\begin{aligned} m_+(\xi; u, v) &= A_-(u) - A_-(\xi) \quad \text{for } u \le \xi \le v,\\ m_+(\xi; u, v) &= A_-(v) - A_-(\xi) \quad \text{for } v \le \xi \le u, \end{aligned}$$

$$\begin{aligned} m_-(\xi; u, v) &= A_+(\xi) - A_+(u) \quad \text{for } u \le \xi \le v,\\ m_-(\xi; u, v) &= A_+(\xi) - A_+(v) \quad \text{for } v \le \xi \le u. \end{aligned} \tag{4.1}$$

In other words, the Engquist–Osher scheme is nothing but a linear upwind discretization of the kinetic formulation. Also Theorem 2.1(ii) implies that the Engquist–Osher scheme satisfies all local entropy inequalities.

Proof. As in Lemma 2.2 we see that

$$m_i(t,\xi) = [m_+(\xi; v_i(t), v_{i+1}(t)) + m_-(\xi; v_{i-1}(t), v_i(t))],$$

where the functions m_+ satisfy

$$\begin{aligned} &\frac{\partial}{\partial\xi} m_+(\xi; v_i(t), v_{i+1}(t))\\ &= \delta(\xi - v_i(t))[A(v_i(t)) - A_{i+1/2}] + [a_+(\xi)\chi(\xi, v_i(t)) - a_-(\xi)\chi(\xi, v_{i+1}(t))] - a(\xi)\chi(\xi, v_i(t)) \end{aligned}$$

and

$$\begin{aligned} &\frac{\partial}{\partial\xi} m_-(\xi; v_{i-1}(t), v_i(t))\\ &= \delta(\xi - v_i(t))[-A_{i-1/2} - A(v_i(t))] - [a_+(\xi)\chi(\xi, v_{i-1}(t)) - a_-(\xi)\chi(\xi, v_i(t))] + a(\xi)\chi(\xi, v_i(t)). \end{aligned}$$

Then, with a slight change of notation,

$$\begin{aligned} &\frac{\partial}{\partial\xi} m_+(\xi; u, v)\\ &= -\delta(\xi - u)\int_{\mathbb{R}} a_-(\zeta)\,[\chi(\zeta; u) - \chi(\zeta; v)]\,d\zeta + a_-(\xi)\,[\chi(\xi; u) - \chi(\xi; v)]. \end{aligned}$$

Notice that the integral in ξ of the right-hand side of this identity vanishes. Therefore since m_+ should have compact support in ξ,

$$m_+(\xi; u, v) = 0 \qquad \text{for } \xi \notin [u, v] \quad \text{(nonordered interval)}.$$

Indeed, the brackets $[\chi(\zeta; u) - \chi(\zeta; v)]$ and $[\chi(\xi; u) - \chi(\xi; v)]$ are supported in $[u, v]$. Also, they have the same sign as $u - v$. Therefore, either $v < u$ and $\frac{\partial}{\partial\xi} m_+(\xi; u, v)$ is positive between v and u and thus $m_+(\xi; u, v)$ vanishes for $\xi \in]-\infty, v]$ or it is nonnegative for $\xi \in [v, u]$ and has a jump at $\xi = u$ and thus vanishes for $\xi > u$. Either $v > u$ and a similar argument shows that $m_+(\xi; u, v)$ is again nonnegative. The same argument as before shows that $m_-(\xi; u, v)$ is nonnegative also.

Finally, the bound on the discrete kinetic defect measure is obtained as in the continuous case. We first argue for $\xi_0 \geq 0$. We use Kruzkov's entropy $S^+_{\xi_0}(\xi) = (\xi - \xi_0)_+$, $S^{+\prime\prime}(\xi)_{\xi_0} = \delta(\xi = \xi_0)$, and we multiply (1.16) by $S^{+\prime}_{\xi_0}(\xi)$, integrate in ξ and in time, and sum up on i. Taking into account the sign in the quantity

$$\int_{\mathbb{R}} S^{+\prime}_{\xi_0}(\xi)\, \chi(\xi; v_i(t))\, d\xi \geq 0,$$

we obtain the inequality, for $\xi_0 \geq 0$,

$$\begin{aligned}\sum_{i\in\mathbb{Z}} h \int_0^\infty m_i(t,\xi_0)\, dt &\leq \sum_{i\in\mathbb{Z}} h\, \|(v_i^0 - \xi)_+\|_{l^1} \\ &\leq \|v^0\|_{l^1}.\end{aligned}$$

A similar argument for $\xi_0 \leq 0$ concludes the proof of the proposition. □

4.2. Fully discrete Engquist–Osher scheme. As in section 3, we depart from the fully discrete Engquist–Osher scheme (1.5) and define $f_i^{n+1}(\xi)$ by the formula

$$\begin{aligned} f_i^{n+1}(\xi) - \chi(\xi, v_i^n) + \lambda\big[a_+(\xi)\chi(\xi, v_i^n) - a_-(\xi)\chi(\xi, v_{i+1}^n)\big] \\ - \lambda\big[a_+(\xi)\chi(\xi, v_{i-1}^n) - a_-(\xi)\chi(\xi, v_i^n)\big] = 0, \end{aligned} \tag{4.2}$$

which can also be written under the kinetic form

$$\begin{aligned} \chi(\xi, u_i^{n+1}) - \chi(\xi, v_i^n) + \lambda\big[a_+(\xi)\chi(\xi, v_i^n) - a_-(\xi)\chi(\xi, v_{i+1}^n)\big] \\ - \lambda\big[a_+(\xi)\chi(\xi, v_{i-1}^n) - a_-(\xi)\chi(\xi, v_i^n)\big] = \frac{\partial}{\partial\xi} m_i^n(\xi), \end{aligned} \tag{4.3}$$

where $m_i^n(\xi)$ vanishes at $\pm\infty$ because

$$u_i^{n+1} = \int_{\mathbb{R}} f_i^{n+1}(\xi)\, d\xi. \tag{4.4}$$

We claim that for λ small enough this is a kinetic formulation.

Theorem 4.2. *Consider the scheme* (1.5), (1.20), *and assume the CFL condition*

$$\lambda \max_\xi |a(\xi)| \leq 1; \tag{4.5}$$

then (4.3) *holds with* $m_i^n(\xi) \geq 0$ *satisfying* (1.17), *and thus it is a kinetic formulation of the Engquist–Osher scheme.*

Proof. Using (4.4) and a variant of Brenier's lemma ([4] or [16, Chap. 2.2]), the property

$$0 \leq \operatorname{sgn}(\xi)\, f_i^{n+1}(\xi) \leq 1 \tag{4.6}$$

is enough to ensure that $m_i^n(\xi) \geq 0$, using the relation

$$\frac{\partial}{\partial\xi} m_i^n(\xi) = \chi(\xi, u_i^{n+1}) - f_i^{n+1}(\xi).$$

To check the signs in (4.6), we rewrite (4.2) as

$$\begin{aligned} f_i^{n+1}(\xi) = \chi(\xi, v_i^n)\Big(1 - \lambda\, a_+(\xi) - \lambda\, a_-(\xi)\Big) \\ + \lambda\, \chi(\xi, v_{i+1}^n) a_-(\xi) + \lambda\, \chi(\xi, v_{i+1}^n) a_+(\xi). \end{aligned} \tag{4.7}$$

To prove the property (4.6), it is enough to notice that this is a convex combination of χ's, a property which follows obviously from $a_\pm \geq 0$ and (4.5). □

Acknowledgment. The authors would like to thank the referees of this paper for their useful comments.

REFERENCES

[1] S. Benharbit, A. Chalabi, and J. P. Vila, *Numerical viscosity and convergence of finite volume methods for conservation laws with boundary conditions*, SIAM J. Numer. Anal., 32 (1995), pp. 775–796.

[2] R. Botchorishvili, B. Perthame, and A. Vasseur, *Equilibrium schemes for scalar conservation laws with stiff sources*, Math. Comp., 72 (2003), pp. 131–157.

[3] F. Bouchut, *Construction of BGK models with a family of kinetic entropies for a given system of conservation laws*, J. Statist. Phys., 95 (1999), pp. 113–170.

[4] Y. Brenier, *Résolution d'équations d'évolution quasilinéaires en dimensions N d'espace à l'aide d'équations linéaires en dimensions $N+1$*, J. Differential Equations, 50 (1982), pp. 375–390.

[5] B. Cockburn, F. Coquel, and P. G. LeFloch, *An error estimate for finite volume methods for multidimensional conservation laws*, Math. Comp., 63 (1994), pp. 77–103.

[6] C. M. Dafermos, *Hyperbolic Conservation Laws in Continuum Physics*, Grundlehren Math. Wiss. 325, Springer-Verlag, Berlin, 2000.

[7] B. Enquist and S. Osher, *Stable and entropy satisfying approximations for transonic flow calculations*, Math. Comp., 31 (1980), pp. 45–75.

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

[9] E. Godlewski and P. A. Raviart, *Hyperbolic Systems of Conservation Laws*, Ellipses, Paris, 1991.

[10] D. Kröner, *Numerical Schemes for Conservation Laws*, Wiley, Chichester, Teubner, Stuttgart, 1997.

[11] R. J. LeVeque, *Numerical Methods for Conservation Laws*, Lectures Math. ETH Zurich, Birkhaüser, Basel, 1992.

[12] P.-L. Lions, B. Perthame, and E. Tadmor, *Formulation cinétique des lois de conservation scalaires multidimensionnelles*, C. R. Acad. Sci. Paris Sér. I Math., 312 (1991), pp. 97–102.

[13] P.-L. Lions, B. Perthame, and E. Tadmor, *A kinetic formulation of multidimensional scalar conservation*, J. Amer. Math. Soc., 7 (1994), pp. 169–191.

[14] S. Osher, *Riemann solvers, the entropy condition, and difference approximations*, SIAM J. Numer. Anal., 21 (1984), pp. 217–235.

[15] S. Noelle, *Convergence of higher order finite volume schemes on irregular grids*, Adv. Comput. Math., 3 (1995), pp. 197–218.

[16] B. Perthame, *Kinetic Formulations of Conservation Laws*, Oxford University Press, Oxford, UK, 2002.

[17] D. Serre, *Systèmes Hyperboliques de Lois de Conservation.* I, Diderot, Paris, 1996.

[18] D. Serre, *Systèmes Hyperbolique de Lois de Conservation.* II, Diderot, Paris, 1996.

[19] E. Tadmor, *Numerical viscosity and the entropy condition for conservative difference schemes*, Math. Comp., 43 (1984), pp. 369–381.

[20] A. Vasseur, *Convergence of a semi-discrete kinetic scheme for the system of isentropic gas dynamics with $\gamma = 3$*, Indiana Univ. Math. J., 48 (1999), pp. 347–364.

[21] A. Vasseur, *Kinetic semidiscretization of scalar conservation laws and convergence by using averaging lemmas*, SIAM J. Numer. Anal., 36 (1999), pp. 465–474.

[22] M. Westdickenberg and S. Noelle, *A new convergence proof for finite volume schemes using the kinetic formulation of conservation laws*, SIAM J. Numer. Anal., 37 (2000), pp. 742–757.

SIAM J. NUMER. ANAL.
Vol. 41, No. 3, pp. 1052–1073

CONVERGENCE RATE ANALYSIS OF A MULTIPLICATIVE SCHWARZ METHOD FOR VARIATIONAL INEQUALITIES*

LORI BADEA†, XUE-CHENG TAI‡, AND JUNPING WANG§

Abstract. This paper derives a linear convergence for the Schwarz overlapping domain decomposition method when applied to constrained minimization problems. The convergence analysis is based on a minimization approach to the corresponding functional over a convex set. A general framework of convergence is established for some multiplicative Schwarz algorithm. The abstract theory is particularly applied to some obstacle problems, which yields a linear convergence for the corresponding Schwarz overlapping domain decomposition method of one and two levels. Numerical experiments are presented to confirm the convergence estimate derived in this paper.

Key words. domain decomposition, variational inequalities, finite element methods, obstacle problems

AMS subject classifications. 65N55, 65N30, 65J15

DOI. 10.1137/S0036142901393607

1. Introduction.

The study of domain decomposition methods was motivated by the increasing need of fast numerical solutions for problems in science and engineering. Such practical problems are often of very large scale and are extremely difficult to solve by using classical approaches. The domain decomposition method has the capability of providing new numerical algorithms which are efficient and parallelizable. The Schwarz overlapping domain decomposition method represents a typical thinking of parallelization and shall be the main focus of this paper.

The Schwarz method consists of two categories which have been traditionally classified as multiplicative and additive methods. The multiplicative Schwarz replicates the well-known Gauss–Seidel iteration for linear systems in a block fashion, while the additive Schwarz method resembles the Jacobi iteration in numerical linear algebra. Both methods have been well studied for second order elliptic problems for the last two decades. Details can be found from [2, 4, 6, 13, 14, 15, 16, 18, 20, 32] and the references cited therein. However, to the authors' knowledge, there are very few existing results which are satisfactorily developed for the Schwarz method when applied to constrained minimization problems.

The main objective of this paper is to establish a convergence rate estimate for the overlapping domain decomposition method for variational inequalities. The result is inspired by the classical analysis of [4] for linear second order elliptic problems and extends some of the new techniques for nonlinear problems of [26, 27, 29, 30]. The essential idea is to decompose the global approximating space into subspaces, which

*Received by the editors August 14, 2001; accepted for publication (in revised form) January 8, 2003; published electronically June 18, 2003.
http://www.siam.org/journals/sinum/41-3/39360.html

†Institute of Mathematics, Romanian Academy of Sciences, P.O. Box 1-764, RO-70700 Bucharest, Romania (lbadea@imar.ro). The research of this author was supported by IMAR under contract ICA1-CT-2000-70022 with the European Commission and by the Norwegian Research Council under contract 128224/431.

‡Department of Mathematics, University of Bergen, Johannes Brunsgate 12, 5007, Bergen, Norway (tai@mi.uib.no). The research of this author was supported by IMAR under contract ICA1-CT-2000-70022 with the European Commission and by the Norwegian Research Council under contract 128224/431.

§Department of Mathematical and Computer Sciences, Colorado School of Mines, Golden, CO 80401 (jwang@mines.edu).

is the key idea behind the latest convergence analysis for domain decomposition and multigrid methods. We shall first establish an abstract framework for the convergence of general minimization problems and then apply it to some obstacle problems by verifying the assumptions of the abstract theory.

A brief review of the existing work on the domain decomposition methods for variational inequalities is as follows. In [1], Badea proved a convergence of a domain decomposition algorithm which is based on minimizing quadratic functionals in a Hilbert space. A convergence rate was established there by using the maximum principle for the problem. A similar method was later proposed and analyzed in [3] as a new member of the additive Schwarz methods. Various one-level overlapping domain decomposition methods have been studied in [10, 12, 17, 19, 21, 24, 25]. A linear convergence for the one-level overlapping domain decomposition method was derived recently in [29, 22, 33] under the condition that the iterative solution increases or decreases monotonically to the true solution. It is known that we can linearize the obstacle problem first and then apply domain decomposition methods for the linearized problems; see, for example, [11]. Our approach is applied directly to the obstacle problem, and no linearization is necessary in the domain decomposition scheme.

Both the one-level and two-level domain decomposition methods are considered in this paper. As it is well known, the two-level method makes use of a coarse level, and its convergence is quite challenging in theory. In fact, the convergence for two-level algorithms has not been fully understood so far in the literature. The only ones we know are from [26, 27, 28]; see also [31] for a two-level algebraic method for the Signorini problem. The method proposed in [26, 27] relies on a decomposition of the convex set, which is different from the algorithm to be studied in the present paper. For the approach to be taken here, the subproblems can be solved in parallel or sequentially. Numerical tests and convergence rate analysis for the parallel version have been done in [28] for domain decomposition and multigrid methods. In this paper, we shall give a convergence rate estimate for the sequential method and concentrate only on the one-level and two-level domain decomposition methods. To the authors' knowledge, our result is the first that gives an explicit convergence rate estimate for this two-level Schwarz method for variational inequalities. For the one-level method, our estimate does not require any monotone property of the iterative solution. Moreover, we give an explicit relation between the convergence rate and the overlapping size. The convergence rate analysis for the multigrid method with the sequential approach is much more difficult and remains open.

This paper is organized as follows. In section 2, we present some abstract domain decomposition algorithms for general convex minimization problems over convex constraint sets. In section 3, we state an abstract result of convergence based on some assumptions for the spatial decomposition. In section 4, we apply the abstract convergence result to a specific obstacle problem by verifying all the conditions required for the abstract theory. To validate our convergence theory, we present some numerical results in section 5 for a two-sided obstacle problem. Finally, in section 6, we provide a complete proof for the main convergence estimate for constrained minimization problems.

2. Algorithm description.

Given a reflexive Banach space V and a convex functional $F : V \mapsto \mathbb{R}$, we consider the following optimization problem:

$$\min_{v \in K} F(v), \quad K \subset V, \tag{2.1}$$

where K is a closed convex subset of V. We are interested in the case where the space V can be decomposed into a sum of subspaces V_i, i.e.,

$$V = V_1 + V_2 + \cdots + V_m = \sum_{i=1}^{m} V_i \; . \tag{2.2}$$

This means that for any $v \in V$, there exists $v_i \in V_i$ such that $v = \sum_{i=1}^{m} v_i$.

With the decomposition (2.2), there are two different ways to solve the nonlinear problem (2.1). The first approach is to decompose K into a sum of $K_i \subset V_i$, $i = 1, 2, \ldots, m$, i.e.,

$$K = K_1 + K_2 + \cdots + K_m = \sum_{i=1}^{m} K_i,$$

and then to solve a minimization problem over each subset K_i in parallel or sequentially. The convergence rate analysis and numerical experiments for this approach have been conducted in [26, 27]. The approach of [26] could handle one- and two-level domain decomposition methods as well as the multigrid method. The second approach does not involve any decomposition of the convex set K and is illustrated in Algorithms 1 and 2.

Algorithm 1. *For a given $u^n \in K$ and $\rho \in (0, 1/m)$, compute $e_i^{n+1} \in V_i$ in parallel for $i = 1, 2, \ldots, m$ such that*

$$e_i^{n+1} = \arg \min_{v_i + u^n \in K, \; v_i \in V_i} G(v_i) \quad \textit{with} \quad G(v_i) = F(u^n + v_i) \tag{2.3}$$

and then update

$$u^{n+1} := u^n + \rho \sum_{i=1}^{m} e_i^{n+1}.$$

Algorithm 2. *For a given $u^n \in K$, compute $e_i^{n+1} \in V_i$ sequentially for $i = 1, 2, \ldots, m$ such that*

$$e_i^{n+1} = \arg \min_{v_i + u^{n+\frac{i-1}{m}} \in K, \; v_i \in V_i} G(v_i) \quad \textit{with} \quad G(v_i) = F(u^{n+\frac{i-1}{m}} + v_i) \tag{2.4}$$

and update

$$u^{n+\frac{i}{m}} := u^{n+\frac{i-1}{m}} + e_i^{n+1}.$$

The algorithms introduced in [1] and [3] are in the same spirit as Algorithms 1 and 2. A convergence rate analysis for Algorithm 1 has been established in [28] for domain decomposition and multigrid methods and in [3] for domain decomposition methods. The objective of this paper is to study Algorithm 2 and derive a linear convergence. The conditions for the convergence of Algorithm 2 differ from those for Algorithm 1. In addition, the analysis turns out to be more complicated than Algorithm 1. The techniques used in the analysis are extensions of those presented in [26, 29, 30].

3. An abstract theory of convergence.

Assume that the minimization functional F is Gâteaux differentiable (see [8]) and that there exists a constant $\kappa > 0$ such that

$$\langle F'(w) - F'(v), w - v\rangle \geq \kappa \|w - v\|_V^2 \quad \forall w, v \in V. \tag{3.1}$$

Here $\langle \cdot, \cdot \rangle$ is the duality pairing between V and its dual space V', i.e., the value of a linear function at an element of V. Under the condition (3.1), problem (2.1) has a unique solution; see [8, p. 35]. For some nonlinear problems, the constant κ may depend on v and w and the analysis given here is still applicable; see [30, Rem. 2.1] for more information. Our abstract convergence theory is based on the following two assumptions inspired from [1].

Assumption 1. There exists a constant $C_1 > 0$ such that for any $w, v \in K$ and $s_i \in V_i$ with $w + \sum_{j=1}^i s_j \in K$, $i = 1, \dots, m$, there exist $z_i \in V_i$ satisfying

$$\begin{cases} \text{(a) } v - w = \displaystyle\sum_{i=1}^m z_i, \qquad \text{(b) } w + \displaystyle\sum_{j=1}^{i-1} s_j + z_i \in K \quad \text{for } i = 1, \dots, m, \\ \text{(c) } \left(\displaystyle\sum_{i=1}^m \|z_i\|_V^2\right)^{\frac{1}{2}} \leq C_1 \left(\|v - w\|_V^2 + \displaystyle\sum_{j=1}^m \|s_j\|_V^2\right)^{\frac{1}{2}}. \end{cases} \tag{3.2}$$

Assumption 2. There exists a constant $C_2 > 0$ which is the least constant satisfying the following inequality for any $w_{ij} \in V, u_i \in V_i,$ and $v_j \in V_j :$

$$\sum_{i,j=1}^m |\langle F'(w_{ij} + u_i) - F'(w_{ij}), v_j\rangle| \leq C_2 \left(\sum_{i=1}^m \|u_i\|_V^2\right)^{\frac{1}{2}} \left(\sum_{j=1}^m \|v_j\|_V^2\right)^{\frac{1}{2}}. \tag{3.3}$$

Let u be the unique solution of (2.1). Our main result of the convergence estimate can be stated as follows.

THEOREM 3.1. *Assume that the space decomposition satisfies* (3.2), (3.3), *and assume that the functional F satisfies* (3.1). *Then for the iterative approximation $\{u^n\}_{n=1}^\infty$ given by Algorithm* 2, *we have*

$$\frac{|F(u^{n+1}) - F(u)|}{|F(u^n) - F(u)|} \leq 1 - \frac{1}{(\sqrt{1 + C^*} + \sqrt{C^*})^2} \tag{3.4}$$

and

$$||u^n - u|\,|_V^2 \leq \frac{2}{\kappa} \left[1 - \frac{1}{(\sqrt{1 + C^*} + \sqrt{C^*})^2}\right]^n |F(u^0) - F(u)|, \tag{3.5}$$

where

$$C^* = \left((1 + C_1)C_2 + \frac{(C_1 C_2)^2}{2\kappa}\right)\frac{2}{\kappa}. \tag{3.6}$$

In order to prove the theorem, we need to combine the special assumption (3.2) with the techniques used in [26]. The proof is tedious and rather complex, and it is postponed to section 6.

4. Application to obstacle problems.

The objective of this section is to apply the abstract convergence theory to obstacle problems and to derive a linear convergence for the corresponding domain decomposition algorithm. To this end, let $\Omega \subset \mathbb{R}^d$ be an open bounded and connected domain with a polyhedral boundary. Consider the problem that seeks an unknown function $u = u(x)$ on Ω satisfying

$$a(u, v-u) \geq f(v-u) \quad \forall v \in K, \tag{4.1}$$

where

$$\begin{aligned} a(v,w) &= \int_\Omega \nabla v \cdot \nabla w \; dx, \\ K &= \{v \in H_0^1(\Omega) | \;\; \alpha(x) \leq v(x) \leq \beta(x) \text{ a.e. in } \Omega\}, \end{aligned} \tag{4.2}$$

$\alpha(x)$ and $\beta(x)$ are two obstacle functions in $L^\infty(\Omega)$, and $f(\cdot)$ is a bounded linear functional on the Sobolev space $H_0^1(\Omega)$. It is well known that the above problem is equivalent to the following minimization problem (see [9], for instance):

$$\min_{v \in K} F(v), \quad F(v) = \frac{1}{2} a(v,v) - f(v). \tag{4.3}$$

For the obstacle problem (4.1), the reflexive Banach space is given by $V = H_0^1(\Omega)$. Correspondingly, we have $\kappa = 1$ in assumption (3.1). We point out that our algorithms and the convergence estimate presented in the previous section are valid for a general class of optimization problems in which the optimization functional F is a strongly convex functional satisfying (3.1).

We use the standard notation for Sobolev spaces $H_0^k(\Omega)$ and $W_0^{k,p}(\Omega)$ and their norms and seminorms. In particular, for a given subdomain $D \subset \Omega$ and $v \in H_0^1(D)$, we shall always extend v with zero in $\Omega \backslash D$, i.e.,

$$H_0^1(D) = \{v | \quad v \in H^1(\Omega), \quad v = 0 \text{ in } \Omega \backslash D\}.$$

Throughout the paper, C will be used to denote a generic constant that does not depend on mesh parameters of the finite element partitions introduced later.

4.1. Numerical approximation and technical tools.

The domain Ω is first partitioned into a coarse mesh denoted $\mathcal{T}_H$ with a mesh size H. Next, we refine the partition $\mathcal{T}_H$ and obtain a fine mesh partition $\mathcal{T}_h$ with a mesh size $h < H$. We assume that both the coarse and fine meshes are shape-regular (see [7]).

Let $S_H \subset W_0^{1,\infty}(\Omega)$ and $S_h \subset W_0^{1,\infty}(\Omega)$ be the continuous, piecewise linear finite element spaces associated with $\mathcal{T}_H$ and $\mathcal{T}_h$, respectively. More precisely, we have

$$S_H = \left\{ v \in W_0^{1,\infty}(\Omega) \middle| \quad v|_\tau \in P_1(\tau) \, \forall \tau \in \mathcal{T}_H \right\}$$

and

$$S_h = \left\{ v \in W_0^{1,\infty}(\Omega) \middle| \quad v|_\tau \in P_1(\tau) \, \forall \tau \in \mathcal{T}_h \right\}.$$

The obstacle problem (4.1) is approximated by a finite element function $u_h(x) \in K \cap S_h$ satisfying

$$a(u_h, v - u_h) \geq f(v - u_h) \quad \forall v \in K \cap S_h. \tag{4.4}$$

Let $\{\Omega_i\}_{i=1}^M$ be a nonoverlapping domain decomposition for Ω, and each Ω_i is the union of some coarse mesh elements. For each Ω_i, we consider an enlarged subdomain Ω_i^δ consisting of elements $\tau \in \mathcal{T}_h$ with $\operatorname{dist}(\tau, \Omega_i) \le \delta \le H$. The union of Ω_i^δ covers $\bar{\Omega}$ with overlaps of size δ. Let us denote the piecewise linear finite element space with vanishing values on the boundary $\partial\Omega_i^\delta$ as $S_h(\Omega_i^\delta)$. It is not hard to show that

$$S_h = \sum_{i=1}^{M} S_h(\Omega_i^\delta) \quad \text{and} \quad S_h = S_H + \sum_{i=1}^{M} S_h(\Omega_i^\delta)\ . \tag{4.5}$$

For the overlapping subdomains, assume that there exist m colors such that each subdomain Ω_i^δ can be marked with one color, and the subdomains with the same color will not intersect with each other. For suitable decompositions, one can choose $m = 2$ if $d = 1$, $m \le 4$ if $d = 2$, and $m \le 8$ if $d = 3$. Let Ω_i^c be the union of the subdomains with the ith color, and

$$V_i = \{v \in S_h | \quad v(x) = 0, \quad x \notin \Omega_i^c\}$$

for $i = 1, 2, \ldots, m$. By denoting $V_0 = S_H$ and $V = S_h$, we see from (4.5) that

$$\text{(a)} \quad V = \sum_{i=1}^{m} V_i \qquad \text{and} \qquad \text{(b)} \quad V = V_0 + \sum_{i=1}^{m} V_i. \tag{4.6}$$

Associated with the subdomains, we consider some functions $\theta_j^i \in C^1(\overline{\Omega})$, $i = 1, 2, \ldots, m$, $j = i, \ldots, m$, such that for any $i = 1, 2, \ldots, m$ we have

$$\operatorname{supp}(\theta_j^i) \subset \bar{\Omega}_j^c,\ 0 \le \theta_j^i \le 1 \quad \forall j = i, \ldots, m, \text{ and } \sum_{j=i}^{m} \theta_j^i = 1 \text{ in } \bigcup_{j=i}^{m} \Omega_j^c. \tag{4.7}$$

More precisely, θ_j^1 is a partition of unity with respect to the subdomains Ω_j^c, $j = 1, 2, \ldots, m$; θ_j^2 is a partition of unity with the subdomains Ω_j^c, $j = 2, \ldots, m$; i.e., the subdomains with the first color are dropped. Accordingly, θ_j^i is a partition of unity with respect to the subdomains Ω_j^c, $j = i, \ldots, m$, where the subdomains Ω_j^c, $j = 1, 2 \ldots,\ i-1$, are dropped. Due to the overlapping property, the preceding functions can be constructed to satisfy

$$|\nabla \theta_j^i| \le C/\delta. \tag{4.8}$$

In the following, I_h denotes the Lagrangian interpolation operator which uses the function values at the nodes of a given mesh $\mathcal{T}_h$ with a mesh size h. The following estimate is correct due to the special structure of the functions θ_j^i:

$$\|I_h(\theta_j^i v)\|_0 \le C\|v\|_0, \quad |I_h(\theta_j^i v)|_1 \le C\|v\|_1 + \frac{1}{\delta}\|v\|_0 \quad \forall\, i, j, \quad \forall\, v \in S_h. \tag{4.9}$$

We also need a nonlinear interpolation operator $I_H^\ominus : S_h \mapsto S_H$ introduced in [26, 27]. Denote $\mathcal{N}_H = \{x_0^i\}_{i=1}^{n_0}$ all the interior nodes for $\mathcal{T}_H$. For a given x_0^i, let ω_i be the union of the mesh elements of $\mathcal{T}_H$ having x_0^i as one of its vertices, i.e.,

$$\omega_i := \bigcup \{\tau \in \mathcal{T}_H, x_0^i \in \bar{\tau}\}.$$

Let $\{\phi_0^i\}_{i=1}^{n_0}$ be the associated nodal basis functions. It is clear that ω_i is the support of ϕ_0^i. Given a nodal point $x_0^i \in \mathcal{N}_H$ and a $v \in S_h$, let $I_i v = \min_{\omega_i} v(x)$. The interpolation function is then defined as

$$I_H^{\ominus} v := \sum_{x_0^i \in \mathcal{N}_H} (I_i v)\phi_0^i(x). \tag{4.10}$$

From the definition, it is easy to see that

$$I_H^{\ominus} v \le v \quad \forall v \in S_h, \tag{4.11}$$

$$I_H^{\ominus} v \ge 0 \quad \forall v \ge 0, v \in S_h. \tag{4.12}$$

Moreover, the interpolation for a given $v \in S_h$ on a finer mesh is always no smaller than the corresponding interpolation on a coarser mesh due to the fact that each coarser mesh element contains several finer mesh elements, i.e.,

$$I_{H_1}^{\ominus} v \le I_{H_2}^{\ominus} v \quad \forall H_1 \ge H_2 \ge h, \quad \forall v \in S_h. \tag{4.13}$$

Define

$$c_d = \begin{cases} C & \text{if} \quad d = 1, \\ C(1 + |\log(H/h)|^{\frac{1}{2}}) & \text{if} \quad d = 2, \\ C(H/h)^{\frac{1}{2}} & \text{if} \quad d = 3. \end{cases}$$

Using Lemma 2.3 in [5], it was proven in [26, 27] that the following approximation properties are correct for the nonlinear interpolation operator $I_H^{\ominus}$.

THEOREM 4.1. *For any $v, w \in S_h$, it is true that*

$$\|I_H^{\ominus} v - I_H^{\ominus} w - (v - w)\|_0 \le c_d H |v - w|_1, \tag{4.14}$$

$$\|I_H^{\ominus} v \ - \ v\|_0 \le c_d H |v|_1, \tag{4.15}$$

$$|I_H^{\ominus} v - I_H^{\ominus} w|_1 \le c_d |v - w|_1. \tag{4.16}$$

4.2. Two-level domain decomposition methods. In the two-level domain decomposition method, the coarse level space S_H is used in the iterative scheme for correction. As a result, the analysis will be based on the space decomposition as given in (4.6.b). Our goal is to verify Assumptions 1 and 2. Notice that the verification for Assumption 2 is straightforward and is essentially the same as for linear problems. We are left with the verification of Assumption 1 by finding the smallest constant C_1 which satisfies (3.2). We use V_0 to denote the coarse mesh and, correspondingly, all the summation index in Assumptions 1 and 2 will start from 0 to m.

The following lemma is stated for a general convex constraint set K defined by constraints on the function values at the fine mesh nodes, and it originates from a similar one given in [1] for the Sobolev spaces. Assume that v, w, $w + \sum_{j=0}^{i} s_j \in K$, $s_i \in V_i$, $i = 0, 1, \dots, m$, holds true for a general convex subset. Choose a $v_0 \in V_0$ such that

$$v - v_0 \in K, \qquad v_0 + w + s_0 \in K. \tag{4.17}$$

We then define z_i, $i = 0, 1, 2, \dots, m$, recursively by

$$z_0 = s_0 + v_0, \quad z_i = I_h\biggl(\theta_i^i\biggl(v - w - \sum_{j=0}^{i-1} z_j\biggr) + (1 - \theta_i^i)s_i\biggr), \quad i = 1, \dots, m. \tag{4.18}$$

LEMMA 4.2. *For a general convex subset $K \subset H_0^1(\Omega)$, assume that v, w, $w + \sum_{j=0}^{i} s_j \in K$, $s_i \in V_i$, $i = 1, \ldots, m$, and assume that v_0 satisfies* (4.17). *Then the functions z_i, $i = 1, \ldots, m$, defined in* (4.18) *satisfy*

$$z_i \in V_i, \quad z_i + w + \sum_{j=0}^{i-1} s_j \in K, \tag{4.19}$$

$$v - w - \sum_{j=0}^{i} z_j \in H_0^1\left(\bigcup_{j=i+1}^{m} \Omega_j^c\right), \tag{4.20}$$

$$v - \sum_{j=0}^{i} z_j + \sum_{j=0}^{i} s_j \in K. \tag{4.21}$$

Proof. The conclusion shall be proved by induction. For $i = 1$, we get from (4.18) that

$$z_1 = I_h(\theta_1^1(v - w - z_0) + (1 - \theta_1^1)s_1). \tag{4.22}$$

Due to the fact that $\theta_1^1 = 0$, $s_1 = 0$ in $\Omega \backslash \Omega_1^c$, it is true that $z_i = 0$ in $\Omega \backslash \Omega_1^c$ and thus $z_1 \in V_1$. Using (4.17), the assumption that $w + s_0 + s_1 \in K$, and the fact that $0 \le \theta_1^1 \le 1$, it is not hard to see that

$$z_1 + w + s_0 = I_h(\theta_1^1(v - v_0) + (1 - \theta_1^1)(w + s_0 + s_1)) \in K.$$

As $I_h(v - w - z_0) = v - w - z_0$, one gets from (4.22) that

$$v - w - z_0 - z_1 = I_h((1 - \theta_1^1)(v - w - z_0 - s_1)). \tag{4.23}$$

From (4.7), one obtains that $\theta_1^1 = 1$ in $\Omega_1^c \backslash \cup_{j=2}^{m} \Omega_j$. Combining it with the above equality we get

$$v - w - z_0 - z_1 \in H_0^1\left(\bigcup_{j=2}^{m} \Omega_j^c\right). \tag{4.24}$$

Furthermore, one gets from (4.17), the assumption that $w + s_0 + s_1 \in K$, the fact that $0 \le \theta_1^1 \le 1$, and (4.23) that

$$\begin{aligned} &v - z_0 - z_1 + s_0 + s_1 \\ &= I_h((1 - \theta_1^1)(v - z_0 + s_0) + \theta_1^1(w + s_0 + s_1)) \\ &= I_h((1 - \theta_1^1)(v - v_0) + \theta_1^1(w + s_0 + s_1)) \in K. \end{aligned}$$

In what follows, we shall assume that a z_i defined by (4.18) satisfies (4.19)–(4.21); then we shall prove that z_{i+1} also satisfies (4.19)–(4.21). From (4.18), we see that

$$z_{i+1} = I_h\left(\theta_{i+1}^{i+1}\left(v - w - \sum_{j=0}^{i} z_j\right) + (1 - \theta_{i+1}^{i+1})s_{i+1}\right). \tag{4.25}$$

Using the fact that

$$\theta_{i+1}^{i+1} \in H_0^1(\Omega_{i+1}^c), \quad s_{i+1} \in H_0^1(\Omega_{i+1}^c),$$

and from (4.20), we see that $z_{i+1} \in H_0^1(\Omega_{i+1}^c)$ and thus $z_{i+1} \in V_{i+1}$. In addition, one gets by using (4.20), (4.25), the assumption $w + \sum_{j=0}^{i} s_j \in K$, and the fact that $0 \le \theta_{i+1}^{i+1} \le 1$ that

$$\begin{aligned} z_{i+1} + w + \sum_{j=0}^{i} s_j & \\ = I_h\Bigg(\theta_{i+1}^{i+1}\Bigg(v + \sum_{j=0}^{i} s_j - \sum_{j=0}^{i} z_j\Bigg) + (1-\theta_{i+1}^{i+1})\Bigg(w + \sum_{j=0}^{i} s_j\Bigg)\Bigg) \in K. \end{aligned} \tag{4.26}$$

From (4.25), it is easy to calculate that

$$\begin{aligned} v - w - \sum_{j=0}^{i+1} z_j &= v - w - \sum_{j=0}^{i} z_j - z_{i+1} \\ &= I_h\Bigg((1-\theta_{i+1}^{i+1})\Bigg(v - w - \sum_{j=0}^{i} z_j - s_{i+1}\Bigg)\Bigg). \end{aligned} \tag{4.27}$$

Using the fact that $s_{i+1} \in H_0^1(\Omega_{i+1}^c)$, $\theta_{i+1}^{i+1} = 1$ in $\Omega_{i+1}^c \setminus \cup_{k=i+2}^m \Omega_k^c$, and from (4.20), one obtains

$$v - w - \sum_{j=0}^{i+1} z_j \in H_0^1\left(\bigcup_{j=i+2}^{m} \Omega_j^c\right).$$

To verify (4.21) for $i+1$, one gets from (4.27), (4.20), the assumption $w + \sum_{j=0}^{i+1} s_j \in K$, and the fact $0 \le \theta_{i+1}^{i+1} \le 1$ that

$$\begin{aligned} v - \sum_{j=0}^{i+1} z_j + \sum_{j=0}^{i+1} s_j & \\ = I_h\Bigg(\theta_{i+1}^{i+1}\Bigg(v - \sum_{j=0}^{i} z_j + \sum_{j=0}^{i} s_j\Bigg) + (1-\theta_{i+1}^{i+1})\Bigg(w + \sum_{j=0}^{i+1} s_j\Bigg)\Bigg) \in K. \end{aligned}$$

Thus, we have proved by induction that (4.19)–(4.21) are correct for all z_i defined as in (4.18). □

Assume from now on that the convex set K is given as in (4.2). For any $v, w+s_0 \in K$, let

$$\sigma^{\oplus} = I_h \max(0, v - w - s_0), \qquad \sigma^{\ominus} = I_h \max(0, w + s_0 - v),$$

and define

$$v_0 = I_H^{\ominus}\sigma^{\oplus} - I_H^{\ominus}\sigma^{\ominus}. \tag{4.28}$$

Due to the special structure of $\sigma^{\ominus}$ and $\sigma^{\oplus}$, it is not hard to show that

$$|\sigma^{\oplus}|_1 \le C|v - w - s_0|_1, \quad |\sigma^{\ominus}|_1 \le C|v - w - s_0|_1. \tag{4.29}$$

Thus, from (4.14)–(4.16) and the fact that $v - w - s_0 = \sigma^{\oplus} - \sigma^{\ominus}$ one obtains

$$\begin{aligned}&\|v_0 - (v - w - s_0)\|_l \\ &\leq \|I_H^{\ominus}\sigma^{\oplus} - I_H^{\ominus}\sigma^{\ominus} - (\sigma^{\oplus} - \sigma^{\ominus})\|_l \\ &\leq c_d H^{1-l}|\sigma^{\oplus}|_1 + c_d H^{1-l}|\sigma^{\ominus}|_1 \\ &\leq c_d H^{1-l}|v - w - s_0|_1, \ l = 0, 1.\end{aligned} \tag{4.30}$$

As $\alpha(x) \leq v, w + s_0 \leq \beta(x)$, there follows that

$$v - w - s_0 \leq \min(\beta - w - s_0, v - \alpha), \quad w + s_0 - v \leq \min(\beta - v, w + s_0 - \alpha).$$

Note that $\min(\beta - w - s_0, v - \alpha) \geq 0$ and $\min(\beta - v, w + s_0 - \alpha) \geq 0$. It follows from properties (4.11) and (4.12) that

$$\begin{aligned}0 \leq I_H^{\ominus}\sigma^{\oplus} \leq \min(\beta - w - s_0, v - \alpha), \\ 0 \leq I_H^{\ominus}\sigma^{\ominus} \leq \min(\beta - v, w + s_0 - \alpha),\end{aligned}$$

which implies that $v_0 = I_H^{\ominus}\sigma^{\oplus} - I_h^{\ominus}\sigma^{\ominus}$ satisfies

$$\max(v - \beta, \alpha - w - s_0) \leq v_0 \leq \min(\beta - w - s_0, v - \alpha).$$

The above inequality shows that

$$\alpha(x) \leq v_0 + w + s_0 \leq \beta(x), \qquad \alpha(x) \leq v - v_0 \leq \beta(x), \tag{4.31}$$

which means that v_0, defined in (4.28), satisfies (4.17) when K is given as in (4.2).

Lemma 4.3. *Let v_0 be given as in* (4.28). *Then the functions z_i, $i = 0, 1, 2, \ldots, m$, defined in* (4.18) *satisfy*

$$\|v - w - z_0\|_0 \leq c_d H\left(|v - w|_1 + |s_0|_1\right), \tag{4.32}$$

$$|v - w - z_0|_1 \leq c_d\left(|v - w|_1 + |s_0|_1\right), \tag{4.33}$$

$$\left\|v - w - \sum_{j=0}^{i} z_j\right\|_0 \leq c_d H\biggl(|v - w|_1 + \sum_{j=0}^{i}|s_j|_1\biggr), \quad i = 1, \ldots, m, \tag{4.34}$$

$$\left|v - w - \sum_{j=0}^{i} z_j\right|_1 \leq c_d\biggl(1 + \frac{H}{\delta}\biggr)\biggl(|v - w|_1 + \sum_{j=0}^{i}|s_j|_1\biggr), \quad i = 1, \ldots, m. \tag{4.35}$$

Proof. The estimates (4.32) and (4.33) follow from (4.30). We shall establish (4.34) and (4.35) by induction. Since $s_i \in H_0^1(\Omega_i^c)$ and Ω_i^c, $i = 1, \ldots, m$, contains many disjoint subdomains with size proportional to H, then the Friedrich–Poincaré inequality can be employed to yield

$$\|s_i\|_0 \leq CH|s_i|_1, \quad i = 1, 2, \ldots, m. \tag{4.36}$$

Now applying (4.9), (4.30), and (4.36) to (4.23) gives

$$\begin{aligned}&\|v - w - z_0 - z_1\|_0 \qquad \text{(using (4.9) and (4.23))} \\ &\leq C\|v - w - z_0 - s_1\|_0 \qquad \text{(using } z_0 = v_0 + s_0 \text{ and (4.30))} \\ &\leq c_d H|v - w - s_0|_1 + \|s_1\|_0 \qquad \text{(using (4.36))} \\ &\leq c_d H(|v - w|_1 + |s_0|_1 + |s_1|_1).\end{aligned} \tag{4.37}$$

Similarly, one arrives at

$$
\begin{aligned}
(4.38)\qquad & |v-w-z_0-z_1|_1 \\
& \le C\|v-w-v_0-s_0-s_1\|_1+\frac{C}{\delta}\|v-w-v_0-s_0-s_1\|_0 \\
& \le C\|v-w-s_0-v_0\|_1+C\|s_1\|_1 \\
& \quad +\frac{C}{\delta}\|v-w-s_0-v_0\|_0+\frac{C}{\delta}\|s_1\|_0 \\
& \le c_d\left(1+\frac{H}{\delta}\right)|v-w-s_0|_1+C\left(1+\frac{H}{\delta}\right)|s_1|_1 \\
& \le c_d\left(1+\frac{H}{\delta}\right)(|v-w|_1+|s_0|_1+|s_1|_1).
\end{aligned}
$$

Now, let us assume that (4.35) and (4.34) are correct for i, and we shall show that they are also correct for $i+1$. To this end, it follows from (4.9) and (4.27) that

$$\left\|v-w-\sum_{j=0}^{i+1}z_j\right\|_0 \le C\left(\left\|v-w-\sum_{j=0}^{i}z_j\right\|_0+\|s_{i+1}\|_0\right),$$

which, with the help of (4.36), shows that (4.34) is correct for $i+1$ if it is correct for i. Finally, using again (4.9) and (4.27), we have

$$
\begin{aligned}
\left|v-w-\sum_{j=0}^{i+1}z_j\right|_1 &\le +C\left(\left|v-w-\sum_{j=0}^{i}z_j\right|_1+|s_{i+1}|_1\right) \\
&\quad+\left(C+\frac{1}{\delta}\right)\left(\left\|v-w-\sum_{j=0}^{i}z_j\right\|_0+\|s_{i+1}\|_0\right).
\end{aligned}
$$

Thus, it follows from (4.34) and (4.36) that (4.35) is correct for $i+1$ if it is correct for the index i. This completes the proof of the lemma. □

THEOREM 4.4. *The estimate* (3.2) *in Assumption* 1 *holds true for the decomposition* (4.6.b) *with*

$$C_1=c_d\left(1+\frac{H}{\delta}\right). \tag{4.39}$$

Proof. Since $\theta_m^m\equiv 1$, then from (4.27) we conclude that

$$v-w-\sum_{j=0}^{m}z_j=0 \text{ in } \Omega,$$

which shows that (3.2.a) is valid. Condition (3.2.b) has been shown to be valid for z_0 and z_i in (4.31) and (4.19). It follows from

$$
\begin{aligned}
\|z_0\|_1 &\le \|v-w-z_0\|_1+\|v-w\|_1, \\
\|z_i\|_1 &\le \left\|v-w-\sum_{j=0}^{i}z_j\right\|_1+\left\|v-w-\sum_{j=0}^{i-1}z_j\right\|_1, \quad i=1,\ldots,m,
\end{aligned}
$$

and (4.32)–(4.35) that (3.2.c) holds true with C_1 being given in (4.39). We point out that the generic constant depends on m, which is the number of colors for the subdomains. □

The estimate (3.3) in Assumption 2 has been shown to be correct for the decomposition (4.6.b) with $C_2 = \sqrt{m+1}$ and m being the number of colors; see [30] and [20, 32] for details. Thus, all the conditions of the abstract convergence Theorem 3.1 are verified for the proposed domain decomposition method for the obstacle problem. As a consequence of Theorem 3.1, we see that the convergence rate of Algorithm 2 for the obstacle problem is given by

$$\frac{F(u^{n+1}) - F(u)}{F(u^n) - F(u)} \le 1 - \frac{1}{1 + (c_d H/\delta)^2}$$

or

$$||u^n - u||_1^2 \le \frac{2}{\kappa} \left[1 - \frac{1}{1 + (c_d H/\delta)^2}\right]^n [F(u^0) - F(u)].$$

4.3. Domain decomposition methods without coarse levels. When no coarse levels are used in the domain decomposition method, the finite element space $V = S_h$ can be decomposed into subspaces as given in (4.6.a). In this case, Algorithm 2 turns out to be the classical Schwarz alternating method for the corresponding minimization problem. We want to show that the abstract convergence Theorem 3.1 can be applied to yield a linear convergence for the Schwarz method, in which the rate of convergence depends only on the overlapping size. Furthermore, our result is more useful than those presented in [22, 29, 33] since no monotonicity is assumed on the iterative approximations.

Let $v, w \in K$ and $s_i \in V_i$ satisfy $w + \sum_{j=1}^{i} s_j \in K$. We define z_i recursively by

$$z_i = I_h\biggl(\theta_i^i\biggl(v - w - \sum_{j=1}^{i-1} v_j\biggr) + (1-\theta_i^i)s_i\biggr), \quad i = 1, \dots, m. \tag{4.40}$$

By repeating the proof as for Lemma 4.2, we obtain the following result.

LEMMA 4.5. *For a general convex subset $K \subset H_0^1(\Omega)$, assume that v, w, $w + \sum_{j=1}^{i} s_j \in K$, $s_i \in V_i$ for $i = 1, \dots, m$. Let z_i, $i = 1, \dots, m$, be defined as in* (4.40). *Then we have*

$$z_i \in V_i, \quad z_i + w + \sum_{j=1}^{i-1} s_j \in K, \tag{4.41}$$

$$v - w - \sum_{j=1}^{i} z_j = 0 \ \textit{in} \ H_0^1\left(\bigcup_{j=i+1}^{m} \Omega_j^c\right), \tag{4.42}$$

$$v - \sum_{j=1}^{i} z_j + \sum_{j=1}^{i} s_j \in K. \tag{4.43}$$

In fact, the above lemma is a consequence of Lemma 4.2 by taking $v_0 = 0$ and $s_0 = 0$. Now, using (4.9), from (4.27) in which the summation index i starts from 1, we obtain the following estimate.

Lemma 4.6. *With $z_i,\ i = 1, 2, \ldots, m$, being defined in* (4.40) *we have*

$$\left\| v - w - \sum_{j=1}^{i} z_j \right\|_0 \le C\Big(|v-w|_1 + \sum_{j=1}^{i} |s_j|_1\Big), \tag{4.44}$$

$$\left| v - w - \sum_{j=0}^{i} z_j \right|_1 \le C(1+\delta^{-1})\Big(|v-w|_1 + \sum_{j=1}^{i} |s_j|_1\Big). \tag{4.45}$$

Consequently, the following result has been proved.

Theorem 4.7. *The estimate* (3.2) *in Assumption* 1 *is valid for the decomposition* (4.6.a) *with*

$$C_1 = C(1+\delta^{-1}). \tag{4.46}$$

An application of Theorem 3.1 indicates that the one-level Schwarz method has the following convergence rate estimate for the obstacle problem:

$$\frac{F(u^{n+1}) - F(u)}{F(u^n) - F(u)} \le 1 - \frac{1}{1 + C(1+\delta^{-2})}$$

or

$$||u^n - u||_1^2 \le \frac{2}{\kappa}\left[1 - \frac{1}{1+C(1+\delta^{-2})}\right]^n [F(u^0) - F(u)].$$

5. Numerical example. To support the convergence theory developed in the previous sections, we present some numerical results here for the obstacle problem approximated by piecewise linear finite elements. To this end, consider the homogeneous problem (4.1) and (4.2) which seeks $u \in H_0^1(\Omega)$ such that

$$\alpha \le u \le \beta : \int_\Omega \nabla u \nabla (v-u) \ge 0 \quad \forall v \in H_0^1(\Omega),\ \alpha \le v \le \beta, \tag{5.1}$$

where $\alpha(x)$ and $\beta(x)$ are two obstacle functions and $\Omega = (0,4) \times (0,3)$.

The two finite element partitions $\mathcal{T}_H$ and $\mathcal{T}_h$ contain right triangles, which are obtained through a uniform refinement of Ω as illustrated in Figure 5.1. In Figure 5.1, the coarse partition $\mathcal{T}_H$ comprises 6×6 rectangles (i.e., 72 triangles) and the fine-level partition $\mathcal{T}_h$ contains 30×30 rectangles (i.e., 1800 triangles). As for the nonoverlapping structure $\{\Omega_i\}_{i=1}^M$ for Ω, we take $M = 9$, and $\{\Omega_i\}_{i=1}^9$ is obtained as a uniform partition of Ω into rectangles. The overlapping decomposition $\{\Omega_i^\delta\}_{i=1}^M$ is constructed by extending each Ω_i with a width of two triangles in $\mathcal{T}_h$. Roughly speaking, the width δ is given by $2h$.

The obstacles $\alpha(x)$ and $\beta(x)$ are shown in Figure 5.2. More precisely, we have

$$\begin{aligned} &\alpha(x,y) = 3 + \sqrt{\left(\tfrac{1}{6}\right)^2 - (x-2)^2 - (y-1.5)^2} \\ &\quad \text{if } (x-2)^2 + (y-1.5)^2 \le \left(\tfrac{1}{6}\right)^2, \text{ or else } \alpha(x,y) = 0; \\ &\beta(x,y) = 1/6 - \sqrt{\left(\tfrac{1}{6}\right)^2 - (x-\tfrac{4}{3})^2 - (y-\tfrac{3}{4})^2} \\ &\quad \text{if } (x-\tfrac{4}{3})^2 + (y-\tfrac{3}{4})^2 \le \left(\tfrac{1}{6}\right)^2, \text{ or else } \beta(x,y) = \tfrac{19}{6}. \end{aligned} \tag{5.2}$$

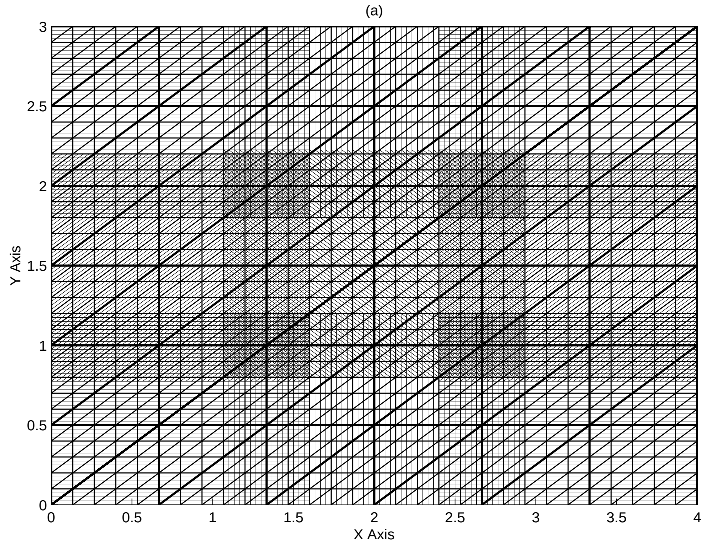


FIG. 5.1. *Meshes $\mathcal{T}_h$ and $\mathcal{T}_H$ and domain decomposition.*

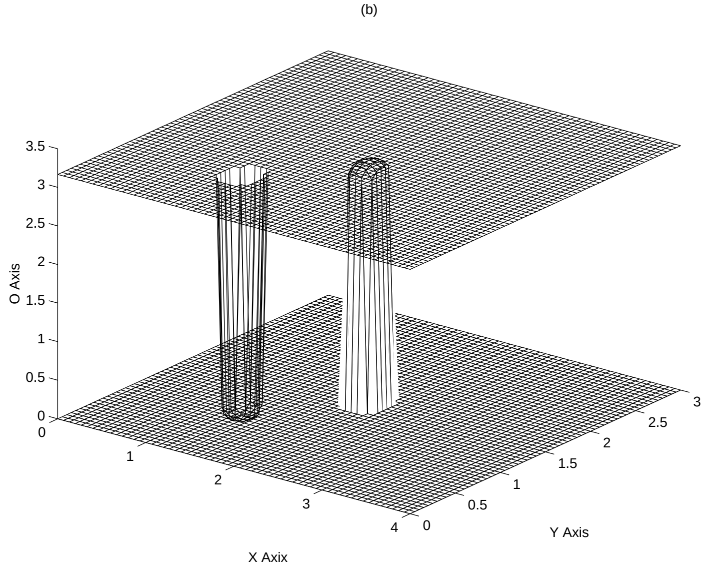


FIG. 5.2. *Obstacle functions $\alpha(x)$ (lower one) and $\beta(x)$ (upper one).*

In the numerical simulations, the obstacles are replaced by their finite element approximations. Corresponding to this obstacle, the finite element solution for (5.1) is as shown in Figure 5.3.

We have seen that the constant C_1 in the convergence estimate of Theorem 4.7 depends on δ^{-1} as given by (4.46) when one-level domain decomposition methods are considered. For two-level domain decomposition methods, the constant C_1 depends on H/h and H/δ as given in (4.39). One of the goals of this section is to numerically verify this dependence by taking various values of H, h, and δ. In all of our numerical tests

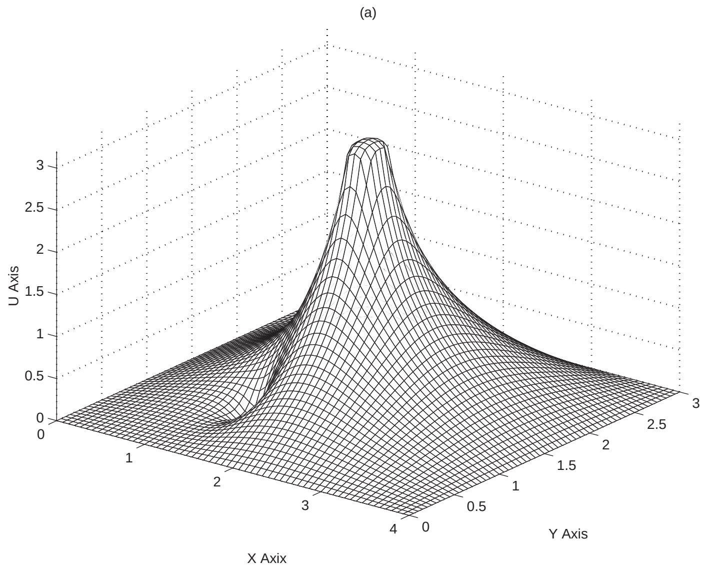


FIG. 5.3. *Solution.*

the iteration is stopped when the maximum error between two consecutive computed solutions is smaller than the tolerance $\epsilon = 0.001$. The solution for each subdomain problem is calculated by using the Gauss–Seidel iteration, which itself is a particular case of the Schwarz domain decomposition method in which each subdomain is merely the support of a nodal basis function of the finite element space. When solving subdomain problems, the calculation is terminated at a relative maximum error of $\epsilon = 10^{-5}$ at the nodes of $\mathcal{T}_h$ between two consecutive computed solutions.

For the results shown in Figure 5.4, the coarse mesh size H varies, while the ratios $H/h = 6$ and $H/\delta = 2$ stay unchanged. The plot shows the total number of iterations in the Schwarz method when the partition $\mathcal{T}_H$ has $20, 18, 16, \ldots, 2$ elements in the x- and y-directions. Starting from six elements, the number of iterations is almost constant for the two-level method, which is in concordance with the convergence theory. It can also be seen that the number of iterations is a decreasing function of H for the one-level method. Since H/δ is constant, it follows that the number of iterations is an increasing function of $1/\delta$, and this is in concordance with the estimate for C_1 in (4.46).

For the results in Figure 5.5, we have taken $H = \frac{5}{12}$, $h = \frac{5}{120}$, and $\delta = h, 2h, \ldots, 10h$. For both one- and two-level methods, the number of iterations is a decreasing function of δ. This observation is in concordance with the estimate on the constant C_1.

For the results shown in Figure 5.6, the values for H, δ are chosen as $H = \frac{5}{6}$ and $\delta = \frac{5}{12}$. The value of h assumes the mesh size of the partition $\mathcal{T}_h$ with $2 \times 6, 4 \times 6, 6 \times 6, \ldots, 20 \times 6$ elements in the x- and y-directions. For the one-level Schwarz method, the number of iterations is constant for $h \leq \frac{5}{12}$, and this confirms the observation that the constant C_1 does not depend on h for the one-level method. For the two-level method, the number of iterations is an decreasing function of h, which is in concordance with the $\log(H/h)$-dependence estimate of C_1 in (4.39).

Finally, we see from the above numerical tests that the number of iterations for the two-level method is significantly less than for the one-level method. We remark

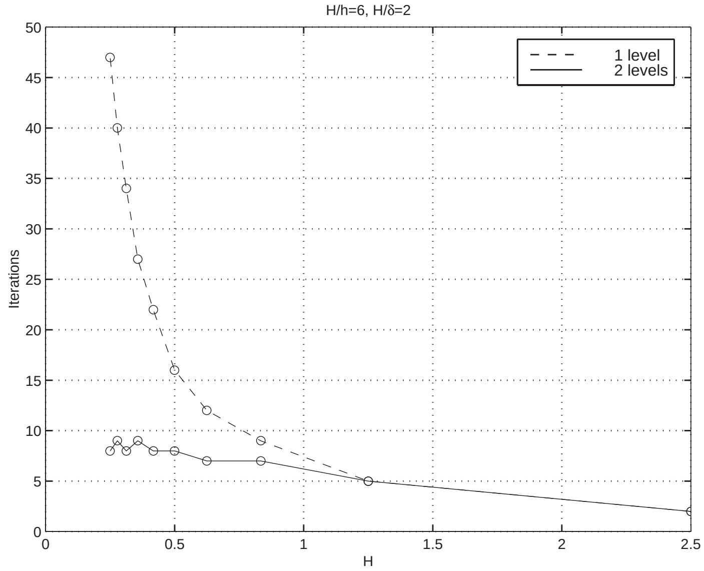


FIG. 5.4. *Number of iterations as a function of H for the Schwarz method when H/h and H/δ are fixed.*

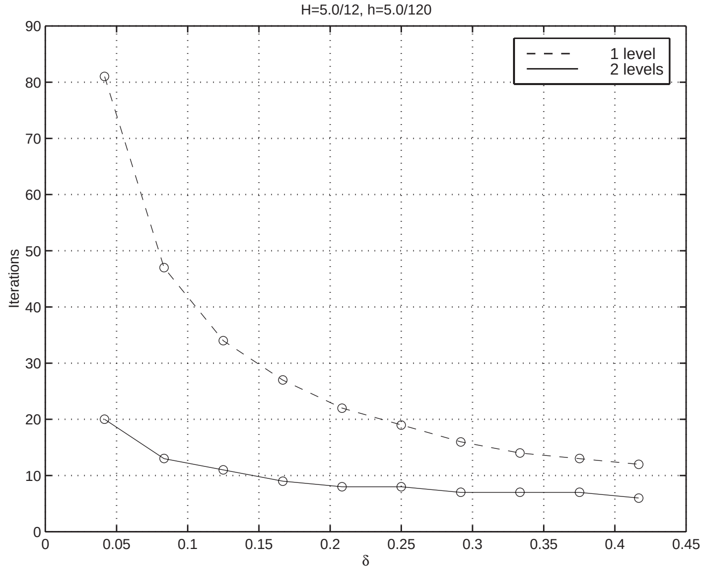


FIG. 5.5. *Number of iterations as a function of δ for the Schwarz method when H and h are fixed.*

that for one-sided obstacle problems, numerical tests using the two-level domain decomposition method have been shown in [28].

In the rest of this section, we make comments on the relaxation method that was used to solve the minimization problem on each subdomain. Notice that in the relaxation method, we have a one-dimensional minimization problem to solve on each support of the nodal basis functions. The solution of these one-dimensional problems

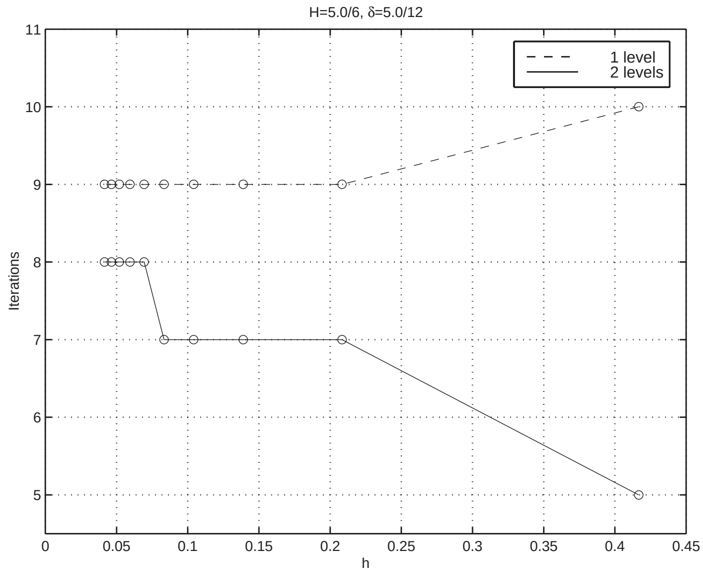


FIG. 5.6. *Number of iterations of the Schwarz method as a function of h when H and δ are fixed.*

was obtained by first solving the one-dimensional problem without constraint and then projecting it to the interval that presents the constraint for this one-dimensional problem. To be more precise, we use two vectors $u(k)$ and $e(k)$, where k runs from 1 to the number of interior nodes in $\mathcal{T}_h$ for the values of $u^{n+\frac{i-1}{m}}$ and e_i^{n+1} obtained from Algorithm 2. Naturally, we have two vectors $\alpha(k)$ and $\beta(k)$ containing the values of the two obstacles at the interior nodal points in $\mathcal{T}_h$. Assume now that we are computing the solution on the subdomain Ω_i and we are seeking the value $e(k)$ of the correction at the node k of $\mathcal{T}_h$. Let $\tilde{e}(k)$ be the value obtained from the one-dimensional problem without constraint. The projection is simply given by

$$e(k) = \min(\beta(k) - u(k),\ \max(\alpha(k) - u(k), \tilde{e}(k))). \tag{5.3}$$

For the problem associated with the coarse mesh, the minimization function (i.e., the correction value e) comes from the coarse mesh finite element space, with constraints imposed on the fine mesh. A relaxation method is employed to solve this problems in which one-dimensional problems associated with interior coarse mesh nodal basis functions $\phi_j^0(x),\ j = 1, 2, \dots, n_0$, are solved. As this is a one-dimensional minimization problem with a constraint, we can first compute the minimizer without constraint and then project this number into the interval which represents the constraint. The computation of the one-dimensional problem without constraint can be done in the same way as for the standard Schwarz method [20, 28]. The computation of the constraint interval can be done similarly as explained in [28, p. 136] for one-sided obstacle problems. To explain the idea more clearly, let us use $u^{old}(x)$ to denote the computed solution, and we need to solve the following problem to get an updated value for a coarse mesh nodal basis function $\phi_j^0(x)$:

$$e(j) = \arg \min_{\{\lambda \in R|\ \alpha(x) \le u^{old} + \lambda\phi_j^0(x) \le \beta(x) \forall x \in supp(\phi_j^0)\}} F(u^{old} + \lambda\phi_j^0), \tag{5.4}$$

where $supp(\phi_j^0)$ is the support set of the function ϕ_j^0. Let $\tilde{e}(j)$ be the minimizer of the one-dimensional unconstrained problem, i.e.,

$$\tilde{e}(j) = \arg\min_{\lambda \in R} F(u^{old} + \lambda\phi_j^0). \tag{5.5}$$

The solution $\tilde{e}(j)$ is found by solving the one-dimensional algebraic equation associated with this minimization problem. Since F is convex, $e(j)$ is the projection of $\tilde{e}(j)$ over the interval

$$[\alpha_j, \beta_j] = \{\lambda \in R \mid \alpha(x) \le u^{old} + \lambda\phi_j^0(x) \le \beta(x)\, \forall x \in supp(\phi_j^0)\},$$

where

$$\alpha_j = \sup_{x \in supp(\phi_j^0)} \frac{\alpha(x) - u^{old}(x)}{\phi_j^0(x)}, \quad \beta_j = \inf_{x \in supp(\phi_j^0)} \frac{\beta(x) - u^{old}(x)}{\phi_j^0(x)}. \tag{5.6}$$

Evidently, we have

$$e(j) = \min(\beta_j, \max(\alpha_j, \tilde{e}(j))). \tag{5.7}$$

We notice that, since $\alpha(x) \le u^{old}(x) \le \beta(x)$, we have $0 \in [\alpha_j, \beta_j]$, and, consequently, this interval is not empty. Naturally, the above $\inf_{x \in supp(\phi_j^0)}$ and $\sup_{x \in supp(\phi_j^0)}$ are calculated only for the mesh nodes of $\mathcal{T}_h$.

Similar relaxation methods have been employed in the domain decomposition for unconstrained minimization problems such as the Dirichlet problem for second order elliptic problems. For the constrained problem, the relaxation method involves an additional step which computes the lower and upper bounds α_j and β_j as given in (5.6) and the projections (5.3) and (5.7).

The projection for the two-level method is more complicated than for the one-level method. However, since the convergence of the two-level method is much faster than the one-level method, the two-level method is more preferable for practical use. For instance, for $H = 5.0/10$, $h = 5.0/60$, and $\delta = 5.0/20$, the number of iterations is 16 for the one-level method, and it is 8 for the two-level method. The computing CPU time on a PC with one processor (Intel Pentium III, 600MHz) is 5.2 minutes for the one-level method, and it is 3.7 minutes for the two-level method. The finite element discretization problem in these numerical tests involves 3481 unknowns.

We shall mention that the subproblems associated with the subdomains and the coarse mesh problem can also be solved by methods other than the relaxation method. In the numerical tests of [26] and [28], the subproblems are solved by the augmented Lagrangian method, which is also rather efficient for handling the constraints both for the subdomain and coarse mesh problems.

6. Proof of Theorem 3.1.

Since e_i^{n+1} minimizes (2.4), it satisfies (see [8])

$$\langle F'(u^{n+\frac{i-1}{m}} + e_i^{n+1}), v_i - e_i^{n+1}\rangle \ge 0 \quad \forall v_i \in V_i \text{ satisfying } v_i + u^{n+\frac{i-1}{m}} \in K. \tag{6.1}$$

Using assumption (3.1), we can prove that (see [23, Lem. 3.2])

$$F(w) - F(v) \ge \langle F'(v), w - v\rangle + \frac{\kappa}{2}\|w - v\|_V^2 \quad \forall v, w \in V. \tag{6.2}$$

Taking $v_i = 0$ in (6.1), we get from the above two inequalities that

$$F(u^{n+\frac{i-1}{m}}) - F(u^{n+\frac{i}{m}}) \ge \frac{\kappa}{2}\|e_i^{n+1}\|_V^2\ . \tag{6.3}$$

It follows from (6.3) that

$$F(u^n) - F(u^{n+1}) = \sum_{i=1}^{m} (F(u^{n+\frac{i-1}{m}}) - F(u^{n+\frac{i}{m}})) \geq \frac{\kappa}{2} \sum_{i=1}^{m} \|e_i^{n+1}\|_V^2. \tag{6.4}$$

Thus, we have

$$F(u^n) \geq F(u^{n+1}).$$

Denote, for a given n,

$$\nu_j^i = \begin{cases} u^n + \sum_{k=1}^{i} e_k^{n+1}, & j \leq i; \\ u^n + \sum_{k=1}^{j} e_k^{n+1}, & j > i. \end{cases}$$

It can be seen that ν_j^i satisfies

$$\begin{aligned} &\nu_j^i - \nu_{j-1}^i = 0, \quad j \leq i; \\ &\nu_j^i - \nu_{j-1}^i = e_j^{n+1}, \quad j > i; \\ &F'(u^{n+1}) - F'(u^{n+\frac{i}{m}}) = \sum_{j=i+1}^{m} \left(F'(\nu_j^i) - F'(\nu_{j-1}^i) \right). \end{aligned} \tag{6.5}$$

As $u,\ u^n,\ u^n + \sum_{j=1}^{i} e_j^{n+1} \in K,\ i = 1, 2, \dots, m$, we get from assumption (3.2) that there exist $z_i^n \in V_i$ such that

$$\begin{cases} \text{(a) } u - u^n = \sum_{i=1}^{m} z_i^n, \qquad \text{(b) } u^n + \sum_{j=1}^{i-1} e_j^{n+1} + z_i^n \in K,\ i = 1, \dots, m, \\ \text{(c) } \left(\sum_{i=1}^{m} \|z_i^n\|_V^2 \right)^{\frac{1}{2}} \leq C_1 \left(\|u^n - u\|_V^2 + \sum_{j=1}^{m} \|e_j^{n+1}\|_V^2 \right)^{\frac{1}{2}}. \end{cases} \tag{6.6}$$

We use (3.3), (6.6), and (6.1) to get

$$\begin{aligned} \langle F'(u^{n+1}), u^{n+1} - u \rangle &= \left\langle F'(u^{n+1}), \sum_{i=1}^{m} e_i^{n+1} + u^n - u \right\rangle \\ &= \sum_{i=1}^{m} \left\langle F'(u^{n+1}), e_i^{n+1} - z_i^n \right\rangle \qquad \text{(using (6.6.a))} \\ &\leq \sum_{i=1}^{m} \left\langle F'(u^{n+1}) - F'(u^{n+i/m}), e_i^{n+1} - z_i^n \right\rangle \\ &\qquad \text{(using (6.6.b) and (6.1))} \\ &= \sum_{i=1}^{m} \sum_{j=i+1}^{m} \left\langle F'(\nu_j^i) - F'(\nu_{j-1}^i), e_i^{n+1} - z_i^n \right\rangle \qquad \text{(using (6.5))} \end{aligned} \tag{6.7}$$

$$\leq C_2\Bigg(\sum_{j=1}^m \|e_j^{n+1}\|_V^2\Bigg)^{\frac{1}{2}}\Bigg(\sum_{i=1}^m \|e_i^{n+1}-z_i^n\|_V^2\Bigg)^{\frac{1}{2}} \qquad \text{(using (3.3))}$$

$$\leq C_2\Bigg(\sum_{j=1}^m \|e_j^{n+1}\|_V^2\Bigg)^{\frac{1}{2}}\Bigg((1+C_1)\Bigg(\sum_{i=1}^m \|e_i^{n+1}\|_V^2\Bigg)^{\frac{1}{2}} + C_1\|u^n-u\|_V\Bigg)$$

(using (6.6.c) and the triangular inequality).

The rest of the proof is the same as in [26, 27]. As u is the unique minimizer for (2.1), we use (6.2) and the optimality condition to obtain

$$F(u^n)-F(u) \geq \langle F'(u), u^n-u\rangle + \frac{\kappa}{2}\|u-u^n\|_V^2 \geq \frac{\kappa}{2}\|u-u^n\|_V^2. \tag{6.8}$$

The following estimate needs to use (6.2), (6.4), (6.7), and (6.8):

$$\begin{aligned}
&F(u^{n+1})-F(u)\\
&\leq \langle F'(u^{n+1}), u^{n+1}-u\rangle \qquad \text{(using (6.2))}\\
&\leq (1+C_1)C_2\frac{2}{\kappa}(F(u^n)-F(u^{n+1})) \qquad \text{(using (6.4) and (6.7))}\\
&\quad + C_1C_2\frac{2}{\kappa}\sqrt{F(u^n)-F(u^{n+1})}\sqrt{F(u^n)-F(u)} \qquad \text{(using (6.8) and (6.7))}.
\end{aligned}$$

Denote $d_n = F(u^n)-F(u)$ for all $n \geq 0$. Let $\mu \in (0,1)$ be a constant to be determined later. Apply the inequality $ab \leq \frac{1}{4\mu}a^2 + \mu b^2$ for all $a, b \in R$ to the last term of the above estimate to get

$$\begin{aligned}
d_{n+1} &\leq (1+C_1)\frac{2C_2}{\kappa}(d_n-d_{n+1}) + C_1C_2\frac{2}{\kappa}\sqrt{d_n-d_{n+1}}\sqrt{d_n}\\
&\leq \bigg((1+C_1)\frac{2C_2}{\kappa} + \frac{[C_1C_2]^2}{\mu\kappa^2}\bigg)(d_n-d_{n+1}) + \mu d_n\\
&\leq C^*\mu^{-1}(d_n-d_{n+1}) + \mu d_n.
\end{aligned}$$

As a consequence, we see that

$$\frac{d_{n+1}}{d_n} \leq \frac{C^*\mu^{-1}+\mu}{1+C^*\mu^{-1}} = 1 - \frac{\mu(1-\mu)}{\mu+C^*}.$$

For a given $C^* > 0$, the function $g(\mu) = \frac{\mu(1-\mu)}{\mu+C^*}$ has a unique maximizer in $[0,1]$, and the maximizer is given by $\mu^* = \sqrt{(C^*)^2+C^*} - C^* \in (0,1)$. Moreover, the maximum value is given by $g(\mu^*) = \frac{1}{(\sqrt{C^*+1}+\sqrt{C^*})^2}$. Consequently, (3.4) holds. The error estimation (3.5) is obtained using (6.8) and (3.4). This completes the proof of the theorem. ∎

REFERENCES

[1] L. Badea, *On the Schwarz alternating method with more than two subdomains for nonlinear monotone problems*, SIAM J. Numer. Anal., 28 (1991), pp. 179–204.

[2] L. Badea, *A generalization of the Schwarz alternating method to an arbitrary number of subdomains*, Numer. Math., 55 (1989), pp. 61–81.

[3] L. Badea and J. Wang, *An additive Schwarz method for variational inequalities*, Math. Comp., 69 (2000), pp. 1341–1354.

[4] J.H. Bramble, J.E. Pasciak, J. Wang, and J. Xu, *Convergence estimates for multigrid algorithms without regularity assumptions*, Math. Comp., 57 (1991), pp. 23–45.

[5] J.H. Bramble and J. Xu, *Some estimate for a weighted L^2 projection*, Math. Comp., 56 (1991), pp. 463–476.

[6] T.F. Chan and T.P. Mathew, *Domain decomposition algorithms*, in Acta Numerica 1994, Cambridge University Press, Cambridge, UK, 1994, pp. 61–143.

[7] P.G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North–Holland, Amsterdam, 1978.

[8] I. Ekeland and R. Temam, *Convex analysis and variational problems*, North–Holland, Amsterdam, 1976.

[9] R.S. Falk, *Error estimate for the approximation of a class of variational inequalities*, Math. Comput., 28 (1974), pp. 963–971.

[10] K. H. Hoffmann and J. Zou, *Parallel algorithms of Schwarz variant for variational inequalities*, Numer. Funct. Anal. Optim., 13 (1992), pp. 449–462.

[11] T. Karkkainen, K. Kunisch, and P. Tarvainen, *Primal-Dual Active Set Methods for Obstacle Problems*, Reports of the Department of Mathematical Information Technology, University of Jyvaskyla, Series B, No. 2, Jyvaskyla, Finland, 2000.

[12] Y. Kuznetsov, P. Neittaanmäki, and P. Tarvainen, *Block relaxation method for algebraic obstacle problems with M-matrices*, East-West J. Numer. Math., 2 (1994), pp. 75–89.

[13] P. Le Tallec, *Domain decomposition methods in computational mechanics*, Comput. Mech. Adv., 1 (1994), pp. 121–220.

[14] P.L. Lions, *On the Schwarz alternating method.* I, in Proceedings of the First International Symposium on Domain Decomposition Methods for Partial Differential Equations, R. Glowinski, G. H. Golub, G. A. Meurant, and J. Periaux, eds., SIAM, Philadelphia, 1988, pp. 2–42.

[15] P.L. Lions, *On the Schwarz alternating method.* II, in Domain Decomposition Methods, T. F. Chan, R. Glowinski, J. Periaux, and O. B. Widlund, eds., SIAM, Philadelphia, 1989, pp. 47–70.

[16] P.L. Lions, *On the Schwarz alternating method.* III, in Proceedings of the Third International Symposium on Domain Decomposition Methods for Partial Differential Equations, T. F. Chan, R. Glowinski, J. Periaux, and O. B. Widlund, eds., SIAM, Philadelphia, 1990, pp. 202–223.

[17] T. Lü, C. Liem, and T. Shih, *Parallel algorithms for variational inequalities based on domain decomposition*, System Sci. Math. Sci., 4 (1991), pp. 341–348.

[18] A. Quarteroni and A. Valli, *Domain Decomposition Methods for Partial Differential Equations*, Oxford University Press, New York, 1999.

[19] I.A. Sharpov, *Multilevel Subspace Correction for Large Scale Optimization Problems*, Technical report CAM-97-31, UCLA, Department of Computational and Applied Mathematics, 1997.

[20] B.F. Smith, P.E. Bjørstad, and W. Gropp, *Domain Decomposition: Parallel Multilevel Methods for Elliptic Differential Equations*, Cambridge University Press, Cambridge, UK, 1996.

[21] X.-C. Tai, *Parallel function decomposition and space decomposition methods with applications to optimisation, splitting and domain decomposition*, preprint 231-1992, Institut für Mathematik, Technische Universität Graz, Austria, 1992. Available online at http://www.mi.uib.no/%7Etai/pub.html.

[22] X.-C. Tai, *Convergence rate analysis of domain decomposition methods for obstacle problems*, East-West J. Numer. Anal., 9 (2001), pp. 233–252.

[23] X.-C. Tai and M. Espedal, *Rate of convergence of some space decomposition methods for linear and nonlinear problems*, SIAM J. Numer. Anal., vol. 35, no. 4 (1998), pp. 1558–1570.

[24] X.-C. Tai, *Parallel function and space decomposition methods. Part* I. *Function decomposition*, Beijing Math., 1 (1991), pp. 104–134.

[25] X.-C. Tai, *Parallel function and space decomposition methods. Part* II. *Space decomposition*, Beijing Math., 1 (1991), pp. 135–152.

[26] X.-C. Tai, *Rate of convergence for some constraint decomposition methods for nonlinear variational inequalities*, Numer. Math., 93 (2003), pp. 755–786.

[27] X.-C. Tai, *Some new domain decomposition and multigrid methods for variational inequalities*, in Proceedings of 14th International Conference on Domain Decomposition Methods, Cocoyoc Morelos, Mexico, 2002, pp. 323–330.

[28] X.-C. Tai, B. ove Heimsund, and J. Xu, *Rate of convergence for parallel subspace correction methods for nonlinear variational inequalities*, in Proceeding of the 13th International

Conference on Domain Decomposition Methods, Lyon, France, 2001, pp. 127–138.

[29] X.-C. Tai and P. Tseng, *Convergence rate analysis of an asynchronous space decomposition method for convex minimization*, Math. Comp., 71 (2001), pp. 1105–1135.

[30] X.-C. Tai and J. Xu, *Global and uniform convergence of subspace correction methods for convex optimization problems*, Math. Comp., 71 (2002), pp. 105–124.

[31] P. Tarvainen, *Two-level Schwarz method for unilateral variational inequalities*, IMA J. Numer. Anal., 14 (1998), pp. 1–18.

[32] J. Xu, *Iterative methods by space decomposition and subspace correction*, SIAM Rev., 34 (1992), pp. 581–613.

[33] J. Zeng and S. Zhou, *On monotone and geometric convergence of Schwarz methods for two-sided obstacle problems*, SIAM J. Numer. Anal., 35 (1998), pp. 600–616.

SIAM J. Numer. Anal.
Vol. 41, No. 3, pp. 1074–1100

ADAPTIVE SOLUTION OF OPERATOR EQUATIONS USING WAVELET FRAMES[*]

ROB STEVENSON[†]

Abstract. In "Adaptive wavelet methods II—Beyond the elliptic case" of Cohen, Dahmen, and DeVore [*Found. Comput. Math.*, 2 (2002), pp. 203–245], an adaptive method has been developed for solving general operator equations. Using a *Riesz basis* of wavelet type for the energy space, the operator equation is transformed into an equivalent matrix-vector system. This system is solved iteratively, where the application of the infinite stiffness matrix is replaced by an adaptive approximation. Assuming that the stiffness matrix is sufficiently compressible, i.e., that it can be sufficiently well approximated by sparse matrices, it was proved that the adaptive method has optimal computational complexity in the sense that it converges with the same rate as the best N-term approximation for the solution, assuming that the latter would be explicitly available. The condition concerning compressibility requires that, dependent on their order, the wavelets have sufficiently many vanishing moments, and that they be sufficiently smooth. However, except on tensor product domains, wavelets that satisfy this smoothness requirement are not easy to construct.

In this paper we write the domain or manifold on which the operator equation is posed as an *overlapping* union of subdomains, each of them being the image under a smooth parametrization of the hypercube. By lifting wavelets on the hypercube to the subdomains, we obtain a *frame* for the energy space. With this frame the operator equation is transformed into a matrix-vector system, after which this system is solved iteratively by an adaptive method similar to the one from the work of Cohen, Dahmen, and DeVore. With this approach, frame elements that have sufficiently many vanishing moments and are sufficiently smooth, something needed for the compressibility, are easily constructed. By handling additional difficulties due to the fact that a frame gives rise to an underdetermined matrix-vector system, we prove that this adaptive method has optimal computational complexity.

Key words. operator equations, adaptive methods, wavelets, frames, optimal computational complexity, best N-term approximation

AMS subject classifications. 41A25, 41A46, 65F10, 65N12, 65N55, 65T60

DOI. 10.1137/S0036142902407988

1. Introduction.

For some boundedly invertible $L : H \to H'$, where H is some Hilbert space with dual H' and some $g \in H'$, we consider the problem of finding $u \in H$ such that

$$Lu = g.$$

As typical examples, we think of linear differential or integral equations in variational form. Although systems of such equations also fit into the framework, for ease of exposition in this introduction let us consider scalar equations, so that H is typically a Sobolev space H^t of some order $t \in \mathbb{R}$.

Assuming that we have a *Riesz basis* Ψ for H^t available, which we formally view as a column vector, by writing $u = \mathbf{u}^T\Psi$, the above problem becomes equivalent to finding $\mathbf{u} \in \ell_2$ satisfying the infinite matrix-vector system

$$\mathbf{M}\mathbf{u} = \mathbf{g},$$

[*] Received by the editors May 21, 2002; accepted for publication (in revised form) November 20, 2002; published electronically June 18, 2003. This work was supported by the Netherlands Organization for Scientific Research.

http://www.siam.org/journals/sinum/41-3/40798.html

[†] Department of Mathematics, Utrecht University, P.O. Box 80.010, NL-3508 TA Utrecht, The Netherlands (stevenson@math.uu.nl).

where $\mathbf{M} := \langle \Psi, L\Psi \rangle : \ell_2 \to \ell_2$ is boundedly invertible and $\mathbf{g} := \langle \Psi, g \rangle \in \ell_2$. Here $\langle \, , \rangle$ denotes the duality product on (H^t, H^{-t}).

Let us denote by $\mathbf{u}_N$ a *best N-term approximation* for $\mathbf{u}$, i.e., a vector with at most N nonzero coefficients that has distance to $\mathbf{u}$ less than or equal to that of any vector with a support of that size. Note that $\|u - \mathbf{u}_N^T \Psi\|_{H^t} \eqsim \|\mathbf{u} - \mathbf{u}_N\|_{\ell_2}$. Considering bases Ψ of sufficiently smooth *wavelet* type, the theory of nonlinear approximation tells us [21, 4] that if both

$$0 < s < \frac{d-t}{n},$$

where d is the *order* of the wavelets and n is the space dimension, and u is in the *Besov space* $B_\tau^{sn+t}(L_\tau)$, with $\tau = (\frac{1}{2} + s)^{-1}$, then

$$\sup_{N \in \mathbb{N}} N^s \|\mathbf{u} - \mathbf{u}_N\|_{\ell_2} < \infty.$$

The condition here involving Besov regularity is much milder that the corresponding condition $u \in H^{sn+t}$ involving Sobolev regularity that would be needed to guarantee the same rate of convergence with linear approximation in the span of N wavelets corresponding to the "coarsest levels." Indeed, assuming a sufficiently smooth right-hand side, for several boundary value problems it has been proved that the solution has a much higher Besov than Sobolev regularity [13, 9]. Note that a rate higher than $\frac{d-t}{n}$ can never be expected with wavelets of order d, except when the solution u happens to be a finite linear combination of wavelets.

So far we have discussed the approximation of $\mathbf{u}$, which, however, is only implicitly given as the solution of $\mathbf{Mu} = \mathbf{g}$. Continuing earlier work in [11], in [5, 6] an iterative adaptive method for solving this system was developed that, given a tolerance $\varepsilon > 0$, yields an approximate solution $\mathbf{u}_\varepsilon$ with $\|\mathbf{u} - \mathbf{u}_\varepsilon\| \leq \varepsilon$, where the number of operations and storage locations it requires is of the same order as the length of the smallest best N-term approximation for $\mathbf{u}$ on distance ε, meaning that the method has *optimal computational complexity.*

When L and thus $\mathbf{M}$ are symmetric and positive definite, the method consists of the application of the simple damped Richardson iteration to the infinite system, where the multiplication of $\mathbf{M}$ with the current finitely supported approximation vector for $\mathbf{u}$ is replaced by an adaptive approximation. In each iteration, each column of $\mathbf{M}$ is replaced by a finitely supported approximation with a tolerance that decreases as a function of the modulus of the corresponding entry in the vector. Note that even for a differential operator the matrix $\mathbf{M}$ is not sparse due to the interaction between wavelets from different levels. A second ingredient of the method is the application after each K steps, with K being some fixed number, of a clean-up or coarsening procedure that removes the smallest entries from the current approximation in order to ensure an optimal work-accuracy balance.

For nonsymmetric or indefinite $\mathbf{M}$, one can simply apply the adaptive method to the normal equations, or alternatively one can apply more advanced iterations which may lead to quantitatively better results [6, 12, 19].

The proof of the optimality of the method requires that $\mathbf{M}$ be sufficiently *compressible*, meaning that, given some tolerance $\delta > 0$, there exists another infinite matrix on distance less than δ, which in each row and column has only a finite and sufficiently small number of nonzero entries. For large classes of differential and integral operators this property can indeed be verified when, dependent on the order d, the wavelets have *sufficiently many vanishing moments* and are *sufficiently smooth*; see [25].

The bottleneck for the application of this adaptive wavelet method is the availability of suitable wavelet bases on general nonrectangular domains or manifolds. An approach to constructing wavelet bases is to write the domain as a *nonoverlapping* union of subdomains, which are the images of the hypercube under smooth parametrizations. Wavelets, or “initial stable completions,” living on the hypercube are lifted to the subdomains. Since in general more than one subdomain is needed, there is the problem of “stitching” functions over the interfaces.

The approach from [17] yields wavelet bases that in principle satisfy all of these requirements. However, since suitable extension operators from one subdomain into neighboring subdomains enter the construction, it seems not easy to implement. The approaches from [16, 3, 8] yield wavelets that over the interfaces between subdomains are only continuous. For example, thinking of a differential equation of order 2 on a two-dimensional domain, with this restricted smoothness only for orders $d \leq 2$ sufficient compressibility of the matrix $\mathbf{M}$ can be shown. Note that it might happen that the solution is smooth everywhere exactly along one of these interfaces on which the wavelets have degenerated properties. However, with these low orders $d \leq 2$ an adaptive method can give at most a small improvement in the order of convergence compared to nonadaptive methods, which in practice might not compensate for the overhead it requires.

Again because of their lack of smoothness beyond continuity, finite element wavelets as constructed in [18, 7, 24] also seem not very well suited for the adaptive method.

The approach followed in this paper is to apply an *overlapping* decomposition of the domain or manifold into subdomains. By lifting wavelets on the hypercube to those subdomains, and by multiplying them by smooth weight functions that vanish at the internal boundaries of these subdomains, a countable set of functions is obtained, which we again denote by Ψ, which is dense in H^t and which for each $u \in H^t$ yields *some* representation $u = \mathbf{u}^T\Psi$ with $\|u\|_{H^t} \eqsim \|\mathbf{u}\|_{\ell_2}$. Such a set Ψ is called a *frame* for H^t. By writing $u = \mathbf{u}^T\Psi$, solving $Lu = g$ is again equivalent to solving $\mathbf{Mu} = \mathbf{g}$, where $\mathbf{M} = \langle \Psi, L\Psi \rangle$ and $\mathbf{g} = \langle \Psi, g \rangle$. However, due to the overlapping decomposition, the representation $u = \mathbf{u}^T\Psi$ will not be unique, and so the system $\mathbf{Mu} = \mathbf{g}$ will have more solutions, which, however, all correspond to the unique solution of $Lu = g$.

When L is symmetric and positive definite, $\mathbf{M}$ is symmetric and *semi*positive definite, and $\mathbf{Mu} = \mathbf{g}$ can be solved by the damped Richardson iteration. In each iteration the norm of the defect is reduced by a constant factor less than 1. For nonsymmetric or indefinite L, the iteration can be applied to the normal equations.

Following [6] for the Riesz basis case, in the practical algorithm the application of $\mathbf{M}$ will be replaced by the adaptive approximation. To be able to prove that the method has optimal computational complexity, it is again needed that $\mathbf{M}$ be sufficiently compressible, i.e., that, dependent on the order d, the wavelets have sufficiently many vanishing moments and be sufficiently smooth. Since its construction does not involve stitching of functions over interfaces, the advantage of this frame approach is that these conditions concerning vanishing moments and smoothness are easily satisfied.

Furthermore, because of the multiplication with the weight functions, boundary conditions at the internal boundaries of the subdomains can be chosen at one’s convenience. In particular, in the case of a closed manifold, this gives the additional advantage that all wavelet bases on the hypercube can be chosen to satisfy *periodic boundary conditions.* Such bases are the most easy to implement, and they have much better quantitative properties than available wavelet bases satisfying other boundary conditions.

A final advantage is that generally an overlapping domain decomposition is much easier to construct with simpler parametrizations than a nonoverlapping one, which might also give a favorable quantitative effect.

The use of a frame instead of a Riesz basis also gives rise to a problem: Since in the adaptive method the matrix-vector product is replaced by an adaptive approximation, each time it is invoked it gives an error that might have a component in the nontrivial kernel of $\mathbf{M}$. Also the clean-up or coarsening step may introduce such components. Just because these components are in the kernel of $\mathbf{M}$, they will not be affected by subsequent Richardson steps, meaning that in the cause of the iteration the component of the current approximation in the kernel of $\mathbf{M}$ may increase. Although this component has no influence on the obtained approximation for the solution of $Lu = g$, that is, after forming the series with the frame elements, it might be responsible for the major part of the computational costs of each iteration. Indeed, recall that in the adaptive approximation of the matrix-vector product the accuracy with which the columns of $\mathbf{M}$ are approximated is determined by the moduli of the corresponding entries in the vector.

Under some technical assumption on the frame, specifically on the projector, called $\mathbf{Q}$, onto the complement of the kernel of $\mathbf{M}$ in ℓ_2, we will prove that the above effect will not occur or only to such an extent that also in the frame case the adaptive method has optimal computational complexity. Unfortunately, although we expect it to hold more generally for our frame construction based on overlapping decompositions, so far we could only give a complete proof that this technical assumption holds in the situation that $t = 0$ and that the wavelet bases on the hypercube are L_2-orthogonal.

The above problem leads us to introduce a modified adaptive algorithm, to which a projection step is added that is applied before each coarsening step. This projector affects only the redundant representation in the overlap regions in a way that the component of the current approximation in the kernel of $\mathbf{M}$ is controlled. The projector itself, called $\mathbf{P}$, is given by an infinite matrix, and in the algorithm, as $\mathbf{M}$, it is only applied approximately using the adaptive matrix-vector product. We show that $\mathbf{P}$ is sufficiently compressible, and prove that this modified algorithm has optimal computational complexity in the general case.

This paper is organized as follows: In section 2, we recall the concept of a frame and show how it can be used to transform an operator equation into an infinite, underdetermined matrix-vector equation. We discuss iterative schemes to solve such equations. Next, we replace those ingredients of such schemes that involve infinite vectors or matrices by practical realizable approximations and show that they, together with a coarsening routine, give rise to a convergent algorithm SOLVE. In addition, we introduce a convergent modified algorithm modSOLVE that contains the inexact application of a projector $\mathbf{P}$ that explicitly controls size of the component of the current approximation in the kernel of $\mathbf{M}$.

In section 3 we study the *rate* of convergence and the computational costs of both algorithms. First we recall some theory dealing with best N-term approximation. We formulate a condition on the compressibility of the stiffness matrix $\mathbf{M}$, and for modSOLVE also of $\mathbf{P}$, that for $\mathbf{M}$ is known to be satisfied for wavelets that, dependent on their order, have sufficiently many vanishing moments and are sufficiently smooth.

Under these conditions, it is proved that both SOLVE and modSOLVE have optimal computational complexity, where for SOLVE we need the aforementioned assumption on the projector $\mathbf{Q}$.

In section 4 we outline the construction of suitable frames using overlapping domain decompositions. Having specified the construction of a frame, we now discuss the condition on $\mathbf{Q}$. Furthermore, we define a suitable $\mathbf{P}$ and show that it is sufficiently compressible.

In order to avoid the repeated use of generic but unspecified constants, in this paper by $C \lesssim D$ we mean that C can be bounded by a multiple of D, independently of parameters on which C and D may depend. Obviously, $C \gtrsim D$ is defined as $D \lesssim C$, and $C \eqsim D$ as $C \lesssim D$ and $C \gtrsim D$.

2. The basic concept.

2.1. Frames.

Let H be a separable real Hilbert space. A countable collection $\Psi \subset H$ is called a *frame* for H when there exist two positive constants A_Ψ, B_Ψ such that

$$A_\Psi \|f\|_{H'}^2 \leq \|f(\Psi)\|^2 \leq B_\Psi \|f\|_{H'}^2 \qquad (f \in H'). \tag{2.1}$$

Here with $f(\Psi)$ we mean the sequence $(f(\psi))_{\psi\in\Psi}$, with $\|f(\Psi)\|$ denoting its ℓ_2-norm. We adapted the definition of a frame given in [20, section 3] by identifying H with its dual H' via the Riesz mapping. As a consequence of (2.1), the frame operators

$$F : H' \to \ell_2 : f \mapsto f(\Psi)$$

and their dual

$$F' : \ell_2 \to H : \mathbf{c} \mapsto \mathbf{c}^T \Psi$$

are bounded with norm less than or equal to $B_\Psi^{\frac{1}{2}}$. Here we have used $\mathbf{c}^T\Psi$ as shorthand notation for $\sum_{\psi\in\Psi} c_\psi \psi$. The composition $F'F : H' \to H$ is boundedly invertible with $\|(F'F)^{-1}\|_{H'\leftarrow H} \leq A_\Psi^{-1}$. The collection $\tilde{\Psi} := (F'F)^{-1}\Psi$ is a frame for H' (the "canonical" dual frame) with frame operators

$$\tilde{F} := F(F'F)^{-1}, \qquad \tilde{F}' = (F'F)^{-1}F'$$

and frame constants B_Ψ^{-1}, A_Ψ^{-1}.

The property of Ψ being a frame for H with constants A_Ψ, B_Ψ can be shown to be equivalent to $\operatorname{clos}\operatorname{span}\Psi = H$ and

$$B_\Psi^{-1}\|u\|_H^2 \leq \inf_{\mathbf{c}\in\ell_2,\, F'\mathbf{c}=u} \|\mathbf{c}\|^2 \leq A_\Psi^{-1}\|u\|_H^2 \qquad (u \in H). \tag{2.2}$$

We have

$$\ell_2 = \operatorname{Ran} F \oplus^\perp \operatorname{Ker} F',$$

and

$$\mathbf{Q} := F(F'F)^{-1}F' : \ell_2 \to \ell_2 \tag{2.3}$$

is the orthogonal projector onto $\operatorname{Ran} F$. The frame Ψ is a Riesz basis for H if and only if $\operatorname{Ker} F' = 0$.

2.2. Transformation of an operator equation to an ℓ_2-problem. Let the linear operator $L : H \to H'$ be boundedly invertible, and let Ψ be a frame for H. Given a $g \in H'$, we consider the problem of finding $u \in H$ such that

$$Lu = g. \tag{2.4}$$

As examples, one may think of L as being a linear differential or integral operator in variational form that defines a homeomorphism between a relevant Sobolev space, or a closed subspace of that, and its dual. A possible construction of a frame will be discussed in section 4.1.

In addition to scalar equations, systems of differential and/or integral equations also fit into this framework. Examples can be found, e.g., in [6, section 3]. In this case, H is a product of relevant Sobolev spaces, and it can be equipped with a frame defined as the product of frames for the coordinate spaces.

Writing $u = F'\mathbf{u}$ for some $\mathbf{u} \in \ell_2$, $\mathbf{u}$ satisfies

$$\mathbf{M}\mathbf{u} = \mathbf{g},$$

where

$$\mathbf{M} := FLF' \quad \text{and} \quad \mathbf{g} := Fg.$$

From

$$\left.\begin{array}{l} \tilde{F}L^{-1}\tilde{F}'FLF' = \tilde{F}F' \\ FLF'\tilde{F}L^{-1}\tilde{F}' = F\tilde{F}' \end{array}\right\} = \mathbf{Q} = \mathbf{id} \quad \text{on } \operatorname{Ran} F,$$

we conclude that $\mathbf{M}|_{\operatorname{Ran} F} : \operatorname{Ran} F \to \operatorname{Ran} F$ is boundedly invertible, in particular with $\|\mathbf{M}\| \le B_\Psi \|L\|_{H' \leftarrow H}$ and $\|\mathbf{M}|_{\operatorname{Ran} F}^{-1}\| \le A_\Psi^{-1} \|L^{-1}\|_{H \leftarrow H'}$, whereas $\operatorname{Ker} \mathbf{M} = \operatorname{Ker} F'$.

2.3. Iterative schemes to solve the infinite-dimensional system $\mathbf{Mu} = \mathbf{g}$. If L is *symmetric* and *positive definite*, i.e., $L' = L$ and $\inf_{0 \ne v \in H} (Lv)(v)/\|v\|^2 > 0$, then $\mathbf{M} = \mathbf{M}^* \ge 0$. With $\lambda_{\max} := \lambda_{\max}(\mathbf{M}) = \|\mathbf{M}\|$ and $\lambda_{\min}^+ := \lambda_{\min}(\mathbf{M}|_{\operatorname{Ran} F}) = \|\mathbf{M}|_{\operatorname{Ran} F}^{-1}\|^{-1}$, for $0 < \alpha < 2/\lambda_{\max}$, we consider the *damped Richardson iteration*

$$\mathbf{u}^{(i+1)} = \mathbf{u}^{(i)} - \alpha(\mathbf{M}\mathbf{u}^{(i)} - \mathbf{g}). \tag{2.5}$$

From $\mathbf{u} - \mathbf{u}^{(i+1)} = (\mathbf{id} - \alpha\mathbf{M})(\mathbf{u} - \mathbf{u}^{(i)})$ and $\mathbf{QM} = \mathbf{MQ}$, we infer that

$$\|\mathbf{Q}(\mathbf{u} - \mathbf{u}^{(i+1)})\| \le \rho \|\mathbf{Q}(\mathbf{u} - \mathbf{u}^{(i)})\|, \tag{2.6}$$

where $\rho := \|(\mathbf{id} - \alpha\mathbf{M})|_{\operatorname{Ran} F}\| = \max\{\alpha\lambda_{\max} - 1, 1 - \alpha\lambda_{\min}^+\} < 1$, with minimum $\frac{\kappa-1}{\kappa+1}$ when $\alpha = 2/(\lambda_{\max} + \lambda_{\min}^+)$, where $\kappa = \lambda_{\max}/\lambda_{\min}^+$. Note that $u - F'\mathbf{u}^{(i)} = F'\mathbf{Q}(\mathbf{u} - \mathbf{u}^{(i)})$.

We will study an inexact version of (2.5) in which the application of the infinite matrix $\mathbf{M}$ is approximated. A difficulty will be that errors made in $\ker F'$ are not reduced in subsequent iterations. Although obviously these errors do not affect $F'\mathbf{u}^{(i)}$, they might hamper a cheap but sufficiently accurate matrix-vector multiplication. Under some condition on $\mathbf{Q}$, i.e., on the frame, we will prove that this is not the case, in the sense that these errors do not pile up too much.

For handling cases in which $\mathbf{Q}$ might not satisfy this condition, we consider a modified algorithm that contains the explicit application of a projector to reduce error components in $\operatorname{Ker} F'$: Let $\mathbf{P} : \ell_2 \to \ell_2$ be some bounded projector with

$$\operatorname{Ker} \mathbf{P} = \operatorname{Ker} F',$$

so that $\ell_2 = \operatorname{Ran}\mathbf{P} \oplus \operatorname{Ker} F'$ is a "stable" splitting. Let $\mathbf{u}^{(i+1)}$ denote the result of applying $\mathbf{P}$ to the result of K damped Richardson iterations starting with $\mathbf{u}^{(i)}$. Using $\mathbf{MPu} = \mathbf{g}$ and $\mathbf{P}(\mathbf{id} - \mathbf{Q}) = 0$, we arrive at

$$\mathbf{Pu} - \mathbf{u}^{(i+1)} = \mathbf{P}(\mathbf{id} - \alpha\mathbf{M})^K(\mathbf{Pu} - \mathbf{u}^{(i)}) = \mathbf{P}(\mathbf{id} - \alpha\mathbf{M})^K\mathbf{Q}(\mathbf{Pu} - \mathbf{u}^{(i)}), \tag{2.7}$$

and so

$$\|\mathbf{Pu} - \mathbf{u}^{(i+1)}\| \leq \|\mathbf{P}\|\rho^K\|\mathbf{Pu} - \mathbf{u}^{(i)}\|,$$

showing convergence when K is chosen such that $\|\mathbf{P}\|\rho^K < 1$. Note that $u - F'\mathbf{u}^{(i)} = F'(\mathbf{Pu} - \mathbf{u}^{(i)})$.

Except when the condition number κ is close to one, the Richardson iteration is known to converge relatively slowly, and quantitatively better results can be expected by more advanced iterations. However, for simplicity we confine the analysis to the easiest algorithm.

The case of L being nonsymmetric or indefinite can be treated by considering the *normal equations*

$$\mathbf{M}^*\mathbf{Mu} = \mathbf{M}^*\mathbf{g}. \tag{2.8}$$

Both $\mathbf{M}|_{\operatorname{Ran} F}$ and $\mathbf{M}^*|_{\operatorname{Ran} F}$ are boundedly invertible on $\operatorname{Ran} F$ and so is $\mathbf{M}^*\mathbf{M}|_{\operatorname{Ran} F}$, whereas $\mathbf{M}^*\mathbf{M}(\operatorname{Ker} F') = 0$. By redefining $\lambda_{\max} = \lambda_{\max}(\mathbf{M}^*\mathbf{M}) = \|\mathbf{M}\|^2$ and $\lambda_{\min}^+ = \lambda_{\min}(\mathbf{M}^*\mathbf{M}|_{\operatorname{Ran} F}) = \|\mathbf{M}|_{\operatorname{Ran} F}^{-1}\|^{-2}$, the damped Richardson iteration, possibly alternated with the projection step, can now be applied to solve (2.8).

In this paper, we confine the analysis to the symmetric positive definite (SPD) case. However, following the lines of [6, section 7], everything that will be said about the SPD case can easily be generalized to the iteration applied to (2.8).

As an alternative for saddle-point problems, in [6, 12, 19] the Uzawa algorithm or a reformulation as a positive definite system is studied, with the aim of obtaining quantitatively better algorithms by avoiding the squaring of the condition number κ. It can be expected that these methods can also be based on frames.

2.4. Approximate iterations. Obviously, since in actual computations we can neither handle the generally infinite vector $\mathbf{g}$ nor apply the infinite matrix $\mathbf{M}$, the damped Richardson iteration, possibly alternated with the projection, is not a practical algorithm. In this section, we study convergence of the iterations in which these ingredients are approximated. Following [6], we assume that we have the following routines at our disposal.

RHS$[\varepsilon, \mathbf{g}] \to \mathbf{g}_\varepsilon$. *This routine determines a finitely supported* $\mathbf{g}_\varepsilon \in \ell_2$ *satisfying*

$$\|\mathbf{g} - \mathbf{g}_\varepsilon\| \leq \varepsilon.$$

APPLY$[\varepsilon, \mathbf{N}, \mathbf{v}] \to \mathbf{w}_\varepsilon$. *This routine determines, for a finitely supported* $\mathbf{v} \in \ell_2$ *and for* $\mathbf{N} = \mathbf{M}$ *(or* $\mathbf{P}$ *or* $\mathbf{M}^*$*), a finitely supported* $\mathbf{w}_\varepsilon$ *satisfying*

$$\|\mathbf{Nv} - \mathbf{w}_\varepsilon\| \leq \varepsilon.$$

COARSE$[\varepsilon, \mathbf{v}] \to \mathbf{v}_\varepsilon$. *This routine creates, for a finitely supported* $\mathbf{v} \in \ell_2$*, a vector* $\mathbf{v}_\varepsilon$ *by replacing all but* N *coefficients of* $\mathbf{v}$ *by zeros such that*

$$\|\mathbf{v} - \mathbf{v}_\varepsilon\| \leq \varepsilon, \tag{2.9}$$

whereas N is at most a constant multiple of the minimal value of N for which (2.9) *is valid.*

In sections 3.1 and 3.2, we will discuss suitable realizations of Coarse and Apply, respectively. The routine Coarse will be necessary for obtaining an optimal work/accuracy balance. The realization of Rhs depends on the problem at hand.

Based on the above routines, we consider the following inexact version of the damped Richardson iteration.

```
    Solve$[\varepsilon, \mathbf{M}, \mathbf{g}] \to \mathbf{u}_\varepsilon$.
Let $\theta < 1/3$ and $K \in \mathbb{N}$ be fixed such that $3\rho^K < \theta$.
$i := 0$, $\mathbf{u}^{(0)} := 0$, $\varepsilon_0 := \|\mathbf{M}|_{\mathrm{Ran}\,F}^{-1}\| \, \|\mathbf{g}\|$.
While $\varepsilon_i > \varepsilon$ do
    $i := i + 1$
    $\varepsilon_i := 3\rho^K \varepsilon_{i-1}/\theta$
    $\mathbf{g}^{(i)} :=$ Rhs$[\frac{\theta\varepsilon_i}{6\alpha K}, \mathbf{g}]$
    $\mathbf{v}^{(i,0)} := \mathbf{u}^{(i-1)}$
    For $j = 1, \dots, K$ do
        $\mathbf{v}^{(i,j)} := \mathbf{v}^{(i,j-1)} - \alpha\left(\text{Apply}[\frac{\theta\varepsilon_i}{6\alpha K}, \mathbf{M}, \mathbf{v}^{(i,j-1)}] - \mathbf{g}^{(i)}\right)$
    od
    $\mathbf{u}^{(i)} :=$ Coarse$[(1-\theta)\varepsilon_i, \mathbf{v}^{(i,K)}]$
od
$\mathbf{u}_\varepsilon := \mathbf{u}^{(i)}$.
```

Proposition 2.1. *Let $\mathbf{u} \in \ell_2$ be some solution of $\mathbf{M}\mathbf{u} = \mathbf{g}$. Then the vectors $\mathbf{u}^{(i)}$, $\mathbf{v}^{(i,K)}$ produced in* Solve$[\varepsilon, \mathbf{M}, \mathbf{g}]$ *satisfy*

$$\|\mathbf{Q}(\mathbf{u} - \mathbf{u}^{(i)})\| \le \varepsilon_i \qquad (i \ge 0), \tag{2.10}$$

and so, in particular, $\|\mathbf{Q}(\mathbf{u} - \mathbf{u}_\varepsilon)\| \le \varepsilon$. Furthermore,

$$\|\mathbf{Q}\mathbf{u} + (\mathbf{id} - \mathbf{Q})\mathbf{u}^{(i-1)} - \mathbf{v}^{(i,K)}\| \le \frac{2}{3}\theta\varepsilon_i \qquad (i \ge 1), \tag{2.11}$$

which will be used in section 3.3.

Proof. For $i = 0$, (2.10) follows from $\mathbf{Q}\mathbf{u} = \mathbf{M}|_{\mathrm{Ran}\,F}^{-1}\mathbf{g}$.

Now for an $i \ge 1$ let $\|\mathbf{Q}(\mathbf{u} - \mathbf{u}^{(i-1)})\| \le \varepsilon_{i-1}$. Since $\mathbf{M}\mathbf{Q}\mathbf{u} = \mathbf{g}$ and $\|\mathbf{id} - \alpha\mathbf{M}\| \le 1$, we have

$$\|\mathbf{Q}\mathbf{u} - \mathbf{v}^{(i,K)} - (\mathbf{id} - \alpha\mathbf{M})^K(\mathbf{Q}\mathbf{u} - \mathbf{u}^{(i-1)})\| \le K\left(\alpha\frac{\theta\varepsilon_i}{6\alpha K} + \alpha\frac{\theta\varepsilon_i}{6\alpha K}\right) = \frac{\theta\varepsilon_i}{3}.$$

From

$$(\mathbf{id} - \alpha\mathbf{M})^K(\mathbf{Q}\mathbf{u} - \mathbf{u}^{(i-1)}) = (\mathbf{id} - \alpha\mathbf{M})^K\mathbf{Q}(\mathbf{u} - \mathbf{u}^{(i-1)}) - (\mathbf{id} - \mathbf{Q})\mathbf{u}^{(i-1)}$$

and $\|(\mathbf{id} - \alpha\mathbf{M})^K\mathbf{Q}(\mathbf{u} - \mathbf{u}^{(i-1)})\| \le \rho^K\|\mathbf{Q}(\mathbf{u} - \mathbf{u}^{(i-1)})\| \le \rho^K\varepsilon_{i-1} = \frac{\theta\varepsilon_i}{3}$, we conclude (2.11). The definition of $\mathbf{u}^{(i)}$ now shows that $\|\mathbf{Q}\mathbf{u} + (\mathbf{id} - \mathbf{Q})\mathbf{u}^{(i-1)} - \mathbf{u}^{(i)}\| \le (\frac{2\theta}{3} + (1 - \theta))\varepsilon_i = (1 - \frac{\theta}{3})\varepsilon_i$, and so $\|\mathbf{Q}(\mathbf{u} - \mathbf{u}^{(i)})\| \le (1 - \frac{\theta}{3})\varepsilon_i \le \varepsilon_i$. □

Remark 2.2. Compared to the Riesz basis setting discussed in [6], we see that in Solve we have to pay for working with a frame. When going from $\mathbf{u}^{(i)}$ to $\mathbf{u}^{(i+1)}$, each of the "evaluation" errors made in the K intermediate steps, and in fact even the sum, should be less than the error that is allowed in $\mathbf{u}^{(i+1)}$. Indeed, since only $\|\mathbf{id} - \alpha\mathbf{M}\| \le 1$, these errors might not be reduced by the iteration.

The inexact version of the damped Richardson iteration, alternated with the inexact application of the projector $\mathbf{P}$, is given by the following.

```
modSOLVE$[\varepsilon, \mathbf{M}, \mathbf{g}] \to \mathbf{u}_\varepsilon$.
Let $\theta < 1/3$ and $K \in \mathbb{N}$ be fixed such that $3\rho^K \|\mathbf{P}\| < \theta$.
$i := 0$, $\mathbf{u}^{(0)} := 0$, $\varepsilon_0 := \|\mathbf{P}\| \, \|\mathbf{M}|_{\operatorname{Ran} F}^{-1}\| \, \|\mathbf{g}\|$.
While $\varepsilon_i > \varepsilon$ do
    $i := i + 1$
    $\varepsilon_i := 3\rho^K \|\mathbf{P}\| \varepsilon_{i-1}/\theta$
    $\mathbf{v}^{(i,0)} := \mathbf{u}^{(i-1)}$
    For $j = 1, \dots, K$ do
        $\mathbf{v}^{(i,j)} := \mathbf{v}^{(i,j-1)} - \alpha(\text{APPLY}[\frac{\rho^j \varepsilon_{i-1}}{2\alpha K}, \mathbf{M}, \mathbf{v}^{(i,j-1)}] - \text{RHS}[\frac{\rho^j \varepsilon_{i-1}}{2\alpha K}, \mathbf{g}])$
    od
    $\mathbf{z}^{(i)} := \text{APPLY}[\frac{\theta \varepsilon_i}{3}, \mathbf{P}, \mathbf{v}^{(i,K)}]$
    $\mathbf{u}^{(i)} := \text{COARSE}[(1-\theta)\varepsilon_i, \mathbf{z}^{(i)}]$
od
$\mathbf{u}_\varepsilon := \mathbf{u}^{(i)}$.
```

PROPOSITION 2.3. *Let $\mathbf{u} \in \ell_2$ be some solution of $\mathbf{Mu} = \mathbf{g}$. Then the vectors $\mathbf{u}^{(i)}$, $\mathbf{z}^{(i)}$ produced in* modSOLVE$[\varepsilon, \mathbf{M}, \mathbf{g}]$ *satisfy*

$$\|\mathbf{Pu} - \mathbf{u}^{(i)}\| \le \varepsilon_i \qquad (i \ge 0), \tag{2.12}$$

and so, in particular, $\|\mathbf{Pu} - \mathbf{u}_\varepsilon\| \le \varepsilon$. Furthermore,

$$\|\mathbf{Pu} - \mathbf{z}^{(i)}\| \le \theta \varepsilon_i \qquad (i \ge 1), \tag{2.13}$$

which will be used in section 3.3.

Proof. For $i = 0$, (2.12) follows from $\mathbf{Pu} = \mathbf{PM}|_{\operatorname{Ran} F}^{-1} \mathbf{g}$.

Now for an $i \ge 1$, let $\|\mathbf{Pu} - \mathbf{u}^{(i-1)}\| \le \varepsilon_{i-1}$. Since $\mathbf{MPu} = \mathbf{g}$, for $1 \le j \le K$, for some δ_j with $\|\delta_j\| \le \frac{\rho^j \varepsilon_{i-1}}{K}$, we have

$$\mathbf{Pu} - \mathbf{v}^{(i,j)} = (\mathbf{id} - \alpha \mathbf{M})(\mathbf{Pu} - \mathbf{v}^{(i,j-1)}) + \delta_j,$$

and so

$$\mathbf{Pu} - \mathbf{v}^{(i,K)} = (\mathbf{id} - \alpha \mathbf{M})^K (\mathbf{Pu} - \mathbf{u}^{(i-1)}) + \sum_{j=1}^{K} (\mathbf{id} - \alpha \mathbf{M})^{K-j} \delta_j.$$

From $(\mathbf{id} - \alpha\mathbf{M})^m = (\mathbf{id} - \alpha\mathbf{M})^m \mathbf{Q} + (\mathbf{id} - \mathbf{Q})$ and $\mathbf{P}(\mathbf{id} - \mathbf{Q}) = 0$, we obtain that $\|\mathbf{P}(\mathbf{id} - \alpha\mathbf{M})^m\| \le \|\mathbf{P}\| \rho^m$. We conclude that

$$\|\mathbf{Pu} - \mathbf{Pv}^{(i,K)}\| \le \|\mathbf{P}\| \left(\rho^K \varepsilon_{i-1} + \sum_{j=1}^{K} \rho^{K-j} \frac{\rho^j \varepsilon_{i-1}}{K} \right) = 2\rho^K \|\mathbf{P}\| \varepsilon_{i-1},$$

$$\|\mathbf{Pu} - \mathbf{z}^{(i)}\| \le 2\rho^K \|\mathbf{P}\| \varepsilon_{i-1} + \frac{\theta \varepsilon_i}{3} = \theta \varepsilon_i,$$

and finally

$$\|\mathbf{Pu} - \mathbf{u}^{(i)}\| \le \theta \varepsilon_i + (1-\theta)\varepsilon_i = \varepsilon_i. \qquad \square$$

3. Convergence rates and computational costs.

3.1. Best *N*-term approximation and COARSE**.** To assess the efficiency of SOLVE or modSOLVE, following [6], we will consider the following benchmark: Suppose that for some solution $\mathbf{u} \in \ell_2$ of $\mathbf{M}\mathbf{u} = \mathbf{g}$ we would have *all* coefficients available. Then the most economical approximation for $\mathbf{u}$ on distance less than ε would be $\mathbf{u}_N$, defined by replacing all but the N largest coefficients in modulus of $\mathbf{u}$ by zeros, with $N = N(\varepsilon, \mathbf{u})$ being the smallest integer such that

$$\|\mathbf{u} - \mathbf{u}_N\| \leq \varepsilon. \tag{3.1}$$

For $N \in \mathbb{N}$, the vector $\mathbf{u}_N$ is called the *best N-term approximation* for $\mathbf{u}$. If for some $s > 0$

$$\|\mathbf{u} - \mathbf{u}_N\| \lesssim N^{-s} \qquad (N \in \mathbb{N}), \tag{3.2}$$

then $N = N(\varepsilon, \mathbf{u})$ from (3.1) would satisfy $N(\varepsilon, \mathbf{u}) \lesssim \varepsilon^{-1/s}$.

Assuming (3.2), in section 3.3 we will prove that for $\mathbf{u}_\varepsilon$ produced by (mod)SOLVE it holds that $\#\operatorname{supp} \mathbf{u}_\varepsilon \lesssim \varepsilon^{-1/s}$, whereas the number of floating point operations to compute it is of the same order. In view of (3.2), we may conclude that (mod)SOLVE is of optimal computational complexity.

The question *whether*, and if so, *for which* s (3.2) is valid, is related to properties of the frame and the (Besov-) regularity of the solution $u \in H$ of the operator equation (2.4). This will be discussed in section 4.2.

Vectors $\mathbf{u} \in \ell_2$ that satisfy (3.2) can be characterized as follows (see [21]): Let $\gamma_n(\mathbf{u})$ denote the nth largest coefficient in modulus of $\mathbf{u}$. For $0 < \tau < 2$, the space ℓ_τ^w is defined by

$$\ell_\tau^w = \left\{ \mathbf{u} \in \ell_2 : |\mathbf{u}|_{\ell_\tau^w} := \sup_n n^{1/\tau} |\gamma_n(\mathbf{u})| < \infty \right\}.$$

It is easily verified that $\ell_\tau \hookrightarrow \ell_\tau^w \hookrightarrow \ell_{\tau+\delta}$ for any $\delta \in (0, 2-\tau]$, justifying why ℓ_τ^w is called *weak* ℓ_τ. The expression $|\mathbf{u}|_{\ell_\tau^w}$ defines only a quasinorm since it does not necessarily satisfy the triangle inequality. Yet, for each $0 < \tau < 2$, there exists a $C_1(\tau) > 0$ with

$$|\mathbf{v} + \mathbf{w}|_{\ell_\tau^w} \leq C_1(\tau) \left(|\mathbf{v}|_{\ell_\tau^w} + |\mathbf{w}|_{\ell_\tau^w} \right) \qquad (\mathbf{v}, \mathbf{w} \in \ell_\tau^w), \tag{3.3}$$

or, equivalently [2, Lemma 3.10.1], for $\mu = \mu(\tau) > 0$ sufficiently small it holds that

$$|\mathbf{v} + \mathbf{w}|_{\ell_\tau^w}^\mu \leq |\mathbf{v}|_{\ell_\tau^w}^\mu + |\mathbf{w}|_{\ell_\tau^w}^\mu \qquad (\mathbf{v}, \mathbf{w} \in \ell_\tau^w).$$

With these ℓ_τ^w-spaces at hand, it can be shown that the property (3.2) is equivalent to $\mathbf{u} \in \ell_\tau^w$, with τ related to s according to $\tau = (\frac{1}{2} + s)^{-1}$. In particular, for each $\tau \in (0, 2)$,

$$\sup_N N^s \|\mathbf{u} - \mathbf{u}_N\| \eqsim |\mathbf{u}|_{\ell_\tau^w}; \tag{3.4}$$

e.g., see [5, Proposition 3.2].

The routine $\mathbf{v}_\varepsilon = \text{COARSE}[\varepsilon, \mathbf{v}]$ might be defined by taking $\mathbf{v}_\varepsilon = \mathbf{v}_N$, with N being the smallest integer such that $\|\mathbf{v} - \mathbf{v}_N\| \leq \varepsilon$. However, since the determination of the best N-term approximation requires sorting all elements of $\mathbf{v}$ by their modulus, this algorithm cannot be implemented in linear time. It requires the order of $(\#\mathrm{supp}\,\mathbf{v}) \times \log(\#\mathrm{supp}\,\mathbf{v})$ operations, with $\#\mathrm{supp}\,\mathbf{v}$ denoting the number of nonzero coefficients of $\mathbf{v}$.

Following ideas from [1, 22], we use a routine COARSE with which this log-*factor* is avoided.

COARSE$[\varepsilon, \mathbf{v}] \to \mathbf{v}_\varepsilon$.

- $q := \lceil \log((\#\mathrm{supp}\,\mathbf{v})^{1/2}\|\mathbf{v}\|/\varepsilon)\rceil$.
- *Divide the elements of* $\mathbf{v}$ *into sets* $V_0, \ldots, V_q$, *where, for* $0 \leq i \leq q-1$, V_i *contains the elements with modulus in* $(2^{-i-1}\|\mathbf{v}\|, 2^{-i}\|\mathbf{v}\|]$, *and possible remaining elements are put into* V_q.
- *Create* $\mathbf{v}_\varepsilon$ *by extracting elements first from* V_0 *and, when it is empty, from* V_1 *and so forth, until* $\|\mathbf{v} - \mathbf{v}_\varepsilon\| \leq \varepsilon$.

The value of q is chosen such that the sum of squares of the elements in V_q is less than or equal to ε^2, meaning that the last element added to $\mathbf{v}_\varepsilon$ (assuming that this vector is nonzero) originates from V_i for some $i < q$. Since then also $\mathbf{v}_N$ with $\|\mathbf{v} - \mathbf{v}_N\| \leq \varepsilon$ must contain elements from this V_i, and since within each V_i the squared values of the elements differ by at most a factor of 4, we obtain the following result.

PROPOSITION 3.1. *For* $\mathbf{v}_\varepsilon$ *yielded by the above routine, it holds that* $\|\mathbf{v} - \mathbf{v}_\varepsilon\| \leq \varepsilon$ *and*

$$\#\mathrm{supp}\,\mathbf{v}_\varepsilon \leq 4\min\{N : \|\mathbf{v} - \mathbf{v}_N\| \leq \varepsilon\}, \tag{3.5}$$

meaning that it defines a valid procedure COARSE. *The number of operations needed for this routine is of the order*

$$\#\mathrm{supp}\,\mathbf{v} + q \lesssim \#\mathrm{supp}\,\mathbf{v} + \log(\varepsilon^{-1}\|\mathbf{v}\|). \tag{3.6}$$

Later, it will appear that the latter log-term is harmless.

Below, in Proposition 3.2, we recall a crucial result proved in [5]. It shows that, for any fixed $\theta < 1/3$, a finitely supported approximation of a target vector in ℓ_τ^w can always be coarsened such that the resulting approximation has an error that is at most $1/\theta$ times the original error, whereas the size of its support is at most some fixed multiple of that of the best N-term approximation with that error. Although this result was proved for best N-term approximations, from (3.5) it is obvious that it is also valid for the current routine COARSE.

PROPOSITION 3.2 (see [5, Corollary 5.2]). *Let* $\theta < 1/3$, $\tau \in (0,2)$, *and* $\tau = (\frac{1}{2} + s)^{-1}$. *Then for any* $\varepsilon > 0$, $\mathbf{v} \in \ell_\tau^w$, *and finitely supported* $\mathbf{w} \in \ell_2$ *with*

$$\|\mathbf{v} - \mathbf{w}\| \leq \theta\varepsilon,$$

for $\overline{\mathbf{w}} = $ COARSE$[(1-\theta)\varepsilon, \mathbf{w}]$ *it holds that*

$$\#\mathrm{supp}\,\overline{\mathbf{w}} \lesssim \varepsilon^{-1/s}|\mathbf{v}|_{\ell_\tau^w}^{1/s},$$

and obviously $\|\mathbf{v} - \overline{\mathbf{w}}\| \leq \varepsilon$.

Remark 3.3. In [5, Corollary 5.2] this result was formulated for $\theta = 1/5$. However, an inspection of the proof, and an easy generalization of [5, (5.4)] concerning thresholding, shows the result for any $\theta < 1/3$. Applying COARSE with a θ larger than 1/5 might give a quantitative improvement of (mod)SOLVE, since then it increases the error with a smaller factor. It is easily seen that in any case Proposition 3.2 can not be valid for $\theta > 1/2$.

Controlling the sizes of the supports of approximations of an ℓ_τ^w-function relative to their errors implies controlling their ℓ_τ^w-(quasi)norms. Indeed, an easy application of the next proposition shows that in the situation of Proposition 3.2, in addition we have that

$$|\overline{\mathbf{w}}|_{\ell_\tau^w} \leq C_2(\tau)|\mathbf{v}|_{\ell_\tau^w} \tag{3.7}$$

for some constant $C_2(\tau)$ independent of ε.

PROPOSITION 3.4 (see [5, Lemma 4.11]). *Let $\tau \in (0,2)$ and $\tau = (\frac{1}{2}+s)^{-1}$. Then for any $\mathbf{v} \in \ell_\tau^w$ and finitely supported $\mathbf{z} \in \ell_2$ we have*

$$|\mathbf{z}|_{\ell_\tau^w} \lesssim |\mathbf{v}|_{\ell_\tau^w} + (\#\operatorname{supp}\mathbf{z})^s \|\mathbf{v}-\mathbf{z}\|.$$

Proof. For convenience we recall the short proof. Let $N = \#\operatorname{supp}\mathbf{z}$; then

$$|\mathbf{z}|_{\ell_\tau^w} \lesssim |\mathbf{z}-\mathbf{v}_N|_{\ell_\tau^w} + |\mathbf{v}_N|_{\ell_\tau^w} \lesssim (2N)^s\|\mathbf{z}-\mathbf{v}_N\| + |\mathbf{v}|_{\ell_\tau^w},$$

where we used $\#\operatorname{supp}(\mathbf{z}-\mathbf{v}_N) \leq 2N$ and (3.4). The proof is completed by

$$\|\mathbf{z}-\mathbf{v}_N\| \leq \|\mathbf{z}-\mathbf{v}\| + \|\mathbf{v}-\mathbf{v}_N\| \leq 2\|\mathbf{z}-\mathbf{v}\|. \qquad \square$$

3.2. Requirements on the infinite-dimensional system. In order to be able to show that (mod)SOLVE has optimal computational complexity, we will have to impose some conditions on the matrix $\mathbf{M}$, and for modSOLVE also on $\mathbf{P}$, as well as on the right-hand side $\mathbf{g}$. Our treatment closely follows [5, 6], except that, following ideas from [1, 22], we avoid some log-factors in the operations count due to sorting.

DEFINITION 3.5. *Let $s^* > 0$. A bounded $\mathbf{N}: \ell_2 \to \ell_2$ is called* s^*-admissible *when for a suitable routine* APPLY, *for each $s \in (0,s^*)$, $\tau = (\frac{1}{2}+s)^{-1}$, for all $\varepsilon > 0$ and finitely supported vectors $\mathbf{v}$, with $\mathbf{w}_\varepsilon = \text{APPLY}[\varepsilon,\mathbf{N},\mathbf{v}]$ the following are valid:*

(I) $\#\operatorname{supp}\mathbf{w}_\varepsilon \lesssim \varepsilon^{-1/s}|\mathbf{v}|_{\ell_\tau^w}^{1/s}$;

(II) *the number of arithmetic operations used to compute $\mathbf{w}_\varepsilon$ is at most a fixed multiple of $\varepsilon^{-1/s}|\mathbf{v}|_{\ell_\tau^w}^{1/s} + \#\operatorname{supp}\mathbf{v}$.*

Remark 3.6. Let $\mathbf{N}$ be s^*-admissible. Then for any $s \in (0,s^*)$, with $\tau = (\frac{1}{2}+s)^{-1}$, $\mathbf{N}: \ell_\tau^w \to \ell_\tau^w$ is bounded. Indeed, let $\mathbf{v} \in \ell_\tau^w$. Part (I) from Definition 3.5 can be written as $\#\operatorname{supp}\mathbf{w}_\varepsilon \leq C\varepsilon^{-1/s}|\mathbf{v}|_{\ell_\tau^w}^{1/s}$ for some constant C. For any $N \in \mathbb{N}$, take $\varepsilon = C^s|\mathbf{v}|_{\ell_\tau^w}N^{-s}$ or, equivalently, $N = C\varepsilon^{-1/s}|\mathbf{v}|_{\ell_\tau^w}^{1/s}$. Let $(\mathbf{N}\mathbf{v})_N$ denote the best N-term approximation for $\mathbf{N}\mathbf{v}$. Then

$$N^s\|\mathbf{N}\mathbf{v}-(\mathbf{N}\mathbf{v})_N\| \leq N^s\|\mathbf{N}\mathbf{v}-\mathbf{w}_\varepsilon\| \leq N^s\varepsilon = C^s|\mathbf{v}|_{\ell_\tau^w},$$

showing $|\mathbf{N}\mathbf{v}|_{\ell_\tau^w} \lesssim |\mathbf{v}|_{\ell_\tau^w}$ by (3.4).

Next, for any $s \in (0,s^*)$ and $\tau = (\frac{1}{2}+s)^{-1}$, the mapping $\mathbf{v} \mapsto \mathbf{w}_\varepsilon := \text{APPLY}[\varepsilon,\mathbf{N},\mathbf{v}]$ is bounded on ℓ_τ^w uniformly in $\varepsilon > 0$. Indeed Proposition 3.4, Definition 3.5(I), and the boundedness of $\mathbf{N}$ demonstrated above show that

$$|\mathbf{w}_\varepsilon|_{\ell_\tau^w} \lesssim |\mathbf{N}\mathbf{v}|_{\ell_\tau^w} + (\#\operatorname{supp}\mathbf{w}_\varepsilon)^s\|\mathbf{N}\mathbf{V}-\mathbf{w}_\varepsilon\| \lesssim |\mathbf{N}\mathbf{v}|_{\ell_\tau^w} + |\mathbf{v}|_{\ell_\tau^w} \lesssim |\mathbf{v}|_{\ell_\tau^w}.$$

It will turn out that a matrix is s^*-admissible when it is s^*-compressible, a property that can be verified for the matrices at hand.

DEFINITION 3.7. *Let $s^* > 0$. A bounded $\mathbf{N}: \ell_2 \to \ell_2$ is called* s^*-compressible, *when for each $j \in \mathbb{N}$ there exist constants α_j and C_j, and an infinite matrix $\mathbf{N}_j$ having at most $\alpha_j 2^j$ nonzero entries in each column, such that*

$$\|\mathbf{N}-\mathbf{N}_j\| \leq C_j, \tag{3.8}$$

$(\alpha_j)_{j\in\mathbb{N}}$ *is summable, and, for any* $s < s^*$, $(C_j 2^{sj})_{j\in\mathbb{N}}$ *is summable.*

For s^*-compressible $\mathbf{N}$, we will make use of the following routine Apply.

Apply$[\varepsilon, \mathbf{N}, \mathbf{v}] \to \mathbf{w}_\varepsilon$.

- $q := \lceil \log((\#\mathrm{supp}\, \mathbf{v})^{1/2} \|\mathbf{v}\| \|\mathbf{N}\| 2/\varepsilon) \rceil$.
- *Divide the elements of* $\mathbf{v}$ *into sets* $V_0, \ldots, V_q$*, where, for* $0 \le i \le q-1$, V_i *contains the elements with modulus in* $(2^{-i-1}\|\mathbf{v}\|, 2^{-i}\|\mathbf{v}\|]$*, and possible remaining elements are put into* V_q.
- *For* $k = 0, 1, \ldots,$ *generate vectors* $\mathbf{v}_{[k]}$ *by subsequently extracting* $2^k - \lfloor 2^{k-1} \rfloor$ *elements from* $\cup_i V_i$*, starting from* V_0 *and when it is empty continuing with* V_1 *and so forth, until for some* $k = \ell$ *either* $\cup_i V_i$ *becomes empty or*

$$\|\mathbf{N}\| \left\| \mathbf{v} - \sum_{k=0}^{\ell} \mathbf{v}_{[k]} \right\| \le \frac{\varepsilon}{2}. \tag{3.9}$$

  In both cases $\mathbf{v}_{[\ell]}$ *may contain less than* $2^\ell - \lfloor 2^{\ell-1} \rfloor$ *elements.*
- *Compute the smallest* $j \ge \ell$ *such that*

$$\sum_{k=0}^{\ell} C_{j-k} \|\mathbf{v}_{[k]}\| \le \frac{\varepsilon}{2}. \tag{3.10}$$

- *For* $0 \le k \le \ell$*, compute the nonzero entries in the matrices* $\mathbf{N}_{j-k}$ *which have a column index in common with one of the entries of* $\mathbf{v}_{[k]}$*, and compute*

$$\mathbf{w}_\varepsilon := \sum_{k=0}^{\ell} \mathbf{N}_{j-k} \mathbf{v}_{[k]}. \tag{3.11}$$

The sizes of the entries of $\mathbf{v}$ determine the accuracy with which the corresponding columns of $\mathbf{N}$ are approximated, which justifies why we speak about an *adaptive* solution method.

Proposition 3.8. *For* $\mathbf{w}_\varepsilon$ *yielded by above routine, we have*

$$\|\mathbf{N}\mathbf{v} - \mathbf{w}_\varepsilon\| \le \varepsilon.$$

Moreover, when $\mathbf{N}$ *is* s^**-compressible, this* Apply *realizes* (I), (II) *of Definition* 3.5*, and so* $\mathbf{N}$ *is* s^**-admissible.*

Proof. From (3.9), (3.8), and (3.10), we have

$$\|\mathbf{N}\mathbf{v} - \mathbf{w}_\varepsilon\| \le \frac{\varepsilon}{2} + \sum_{k=0}^{\ell} C_{j-k} \|\mathbf{v}_{[k]}\| \le \varepsilon.$$

Let $s \in (0, s^*)$ be given and $\tau = (\frac{1}{2} + s)^{-1}$. The number of operations needed for generating the vectors $\mathbf{v}_{[k]}$ is of the order

$$\#\mathrm{supp}\, \mathbf{v} + q \lesssim \#\mathrm{supp}\, \mathbf{v} + \log(\varepsilon^{-1}\|\mathbf{v}\|) \lesssim \#\mathrm{supp}\, \mathbf{v} + \varepsilon^{-1/s} |\mathbf{v}|_{\ell^w_\tau}^{1/s}.$$

By the definition of s^*-compressibility, $\#\mathrm{supp}\, \mathbf{w}_\varepsilon$ and the number of operations needed for the evaluation of (3.11) can be bounded by $\sum_{k=0}^{\ell} \alpha_{j-k} 2^{j-k} 2^k \lesssim 2^j$, meaning that the proof will be completed once we have shown that $2^j \lesssim \varepsilon^{-1/s} |\mathbf{v}|_{\ell^w_\tau}^{1/s}$.

The value of q was chosen such that the sum of squares of elements in V_q is less than or equal to $(\varepsilon/(2\|\mathbf{N}\|))^2$, meaning that for all $k < \ell$, $\mathbf{v}_{[k]}$ contains only elements from V_i for $i < q$. Since within each of these V_i the squared values of the elements differ by at most a factor of 4, for $k \leq \ell$ we obtain that

$$\|\mathbf{v}_{[k]}\| \leq \left\| \mathbf{v} - \sum_{m=0}^{k-1} \mathbf{v}_{[m]} \right\| \leq \|\mathbf{v} - \mathbf{v}_{\lceil 2^{k-1}/4 \rceil}\| \lesssim 2^{-ks} |\mathbf{v}|_{\ell_\tau^w},$$

by (3.4).

Since ℓ is the smallest integer for which (3.9) is valid, assuming it is nonzero, we infer that

$$\frac{\varepsilon}{2} < \|\mathbf{N}\| \left\| \mathbf{v} - \sum_{k=0}^{\ell-1} \mathbf{v}_{[k]} \right\| \lesssim 2^{-\ell s} |\mathbf{v}|_{\ell_\tau^w} \|\mathbf{N}\|$$

or $2^\ell \lesssim \varepsilon^{-1/s} |\mathbf{v}|_{\ell_\tau^w}^{1/s}$.

Now assume that $j > \ell$. Since j is the smallest integer for which (3.10) is valid, we infer that

$$\frac{\varepsilon}{2} < \sum_{k=0}^{\ell} C_{j-1-k} \|\mathbf{v}_{[k]}\| \lesssim \sum_{k=0}^{\ell} C_{j-1-k} 2^{-ks} |\mathbf{v}|_{\ell_\tau^w} \lesssim 2^{-(j-1)s} |\mathbf{v}|_{\ell_\tau^w},$$

by the definition of s^*-compressibility, or $2^j \lesssim \varepsilon^{-1/s} |\mathbf{v}|_{\ell_\tau^w}^{1/s}$. □

We will consider right-hand sides $\mathbf{g}$ that satisfy the following definition.

DEFINITION 3.9. *A vector $\mathbf{g} \in \ell_2$ is called* s^*-optimal *when for a suitable routine* RHS, *for each $s \in (0, s^*)$, $\tau = (\frac{1}{2} + s)^{-1}$, and all $\varepsilon > 0$, with $\mathbf{g}_\varepsilon =$* RHS$[\varepsilon, \mathbf{g}]$ *the following are valid:*

(I) $\#\operatorname{supp} \mathbf{g}_\varepsilon \lesssim \varepsilon^{-1/s} |\mathbf{g}|_{\ell_\tau^w}^{1/s}$,

(II) *the number of arithmetic operations used to compute $\mathbf{g}_\varepsilon$ is at most a multiple of $\varepsilon^{-1/s} |\mathbf{g}|_{\ell_\tau^w}^{1/s}$.*

Remark 3.10. A direct consequence of Proposition 3.4 and Definition 3.9(I) is that

$$|\mathbf{g}_\varepsilon|_{\ell_\tau^w} \lesssim |\mathbf{g}|_{\ell_\tau^w}. \tag{3.12}$$

Implicitly, in the proof of Proposition 3.8 we assumed that each element of $\mathbf{N}_j$ can be computed at unit costs. For a discussion of the circumstances under which this, as well as $\mathbf{g}$ being s^*-optimal, can be expected, we refer to [6, section 6.2].

3.3. The complexity of (mod)SOLVE.

We show that SOLVE and modSOLVE are of optimal computational complexity. We start with modSOLVE, since for this routine the proof follows closely that given in [6] for the Riesz basis case.

THEOREM 3.11. *For some $s^* > 0$, assume that $\mathbf{M}$ and $\mathbf{P}$ are s^*-admissible, $\mathbf{g}$ is s^*-optimal, and that for some $s \in (0, s^*)$, with $\tau = (\frac{1}{2} + s)^{-1}$, $\mathbf{M}\mathbf{u} = \mathbf{g}$ has a solution $\mathbf{u} \in \ell_\tau^w$. Then for all $\varepsilon > 0$, $\mathbf{u}_\varepsilon =$* modSOLVE$[\varepsilon, \mathbf{M}, \mathbf{g}]$ *satisfies the following:*

(I) $\#\operatorname{supp} \mathbf{u}_\varepsilon \lesssim \varepsilon^{-1/s} |\mathbf{u}|_{\ell_\tau^w}^{1/s}$,

(II) *the number of arithmetic operations used to compute $\mathbf{u}_\varepsilon$ is at most a multiple of $\varepsilon^{-1/s} |\mathbf{u}|_{\ell_\tau^w}^{1/s}$.*

Further, as shown in Proposition 2.3, $\|\mathbf{P}\mathbf{u} - \mathbf{u}_\varepsilon\| \leq \varepsilon$, and so $\|u - F'\mathbf{u}_\varepsilon\|_H \leq B_\Psi^{\frac{1}{2}} \varepsilon$.

Proof. It suffices to prove the statements for any $\varepsilon = \varepsilon_i$ with $\varepsilon_i = (3\rho^K \|\mathbf{P}\|/\theta)^i \varepsilon_0$, as in the algorithm modSOLVE.

As noted in Remark 3.6, the fact that $\mathbf{P}$ is s^*-admissible implies that it is bounded on ℓ^w_τ. For any $i \geq 1$, from $\|\mathbf{P}\mathbf{u}-\mathbf{z}^{(i)}\| \leq \theta\varepsilon_i$ proved in Proposition 2.3, Proposition 3.2 and the assumption $\mathbf{u} \in \ell^w_\tau$ show that $\mathbf{u}^{(i)} := \textsc{Coarse}[(1-\theta)\varepsilon_i, \mathbf{z}^{(i)}]$ satisfies

$$\#\mathrm{supp}\,\mathbf{u}^{(i)} \lesssim \varepsilon_i^{-1/s}|\mathbf{P}\mathbf{u}|^{1/s}_{\ell^w_\tau} \lesssim \varepsilon_i^{-1/s}|\mathbf{u}|^{1/s}_{\ell^w_\tau}, \tag{3.13}$$

i.e., (I), and by (3.7), also

$$|\mathbf{u}^{(i)}|_{\ell^w_\tau} \lesssim |\mathbf{P}\mathbf{u}|_{\ell^w_\tau} \lesssim |\mathbf{u}|_{\ell^w_\tau}. \tag{3.14}$$

Here we emphasize that both these results are valid uniformly in i.

To compute $\mathbf{u}^{(i)}$ from $\mathbf{u}^{(i-1)}$, modSolve uses one application of Rhs, K applications of Apply involving $\mathbf{M}$, $2K$ vector updates, one application of Apply involving $\mathbf{P}$, and finally an application of Coarse. From the key estimates (3.13), (3.14), and the fact that K is some *fixed constant*, in the following three paragraphs we show that these computations take not more than a multiple of $\varepsilon_i^{-1/s}|\mathbf{u}|^{1/s}_{\ell^w_\tau}$ operations. Since $(\varepsilon_i)_i$ is a geometrically decreasing sequence, we may therefore conclude (II).

Since $\mathbf{M}$ is s^*-admissible, it is bounded on ℓ^w_τ. As a consequence, $|\mathbf{g}|_{\ell^w_\tau} \lesssim |\mathbf{u}|_{\ell^w_\tau}$, and so the assumption of $\mathbf{g}$ being s^*-optimal gives $\#\mathrm{supp}\,\mathbf{g}^{(i)} \lesssim \varepsilon_i^{-1/s}|\mathbf{u}|^{1/s}_{\ell^w_\tau}$ and $|\mathbf{g}^{(i)}|_{\ell^w_\tau} \lesssim |\mathbf{u}|_{\ell^w_\tau}$ by (3.12), whereas the number of operations used to compute it is at most a multiple of $\varepsilon_i^{-1/s}|\mathbf{u}|^{1/s}_{\ell^w_\tau}$.

Because of $|\mathbf{g}^{(i)}|_{\ell^w_\tau} \lesssim |\mathbf{u}|_{\ell^w_\tau}$ and $|\mathbf{u}^{(i-1)}|_{\ell^w_\tau} \lesssim |\mathbf{u}|_{\ell^w_\tau}$, from the assumption that $\mathbf{M}$ is s^*-admissible we have $|\mathbf{v}^{(i,j)}|_{\ell^w_\tau} \lesssim |\mathbf{u}|_{\ell^w_\tau}$ by Remark 3.6. Again since $\mathbf{M}$ is s^*-admissible, the latter result shows that $\#\mathrm{supp}\,\mathbf{v}^{(i,j)} \lesssim \varepsilon_i^{-1/s}|\mathbf{u}|^{1/s}_{\ell^w_\tau}$ for $1 \leq j \leq K$, whereas by $\#\mathrm{supp}\,\mathbf{v}^{(i,0)} \lesssim \varepsilon_{i-1}^{-1/s}|\mathbf{u}|^{1/s}_{\ell^w_\tau}$ (see (3.13)) and $\#\mathrm{supp}\,\mathbf{g}^{(i)} \lesssim \varepsilon_i^{-1/s}|\mathbf{u}|^{1/s}_{\ell^w_\tau}$, its computation takes not more than a multiple of $\varepsilon_i^{-1/s}|\mathbf{u}|^{1/s}_{\ell^w_\tau}$ operations.

Since $|\mathbf{v}^{(i,K)}|_{\ell^w_\tau} \lesssim |\mathbf{u}|_{\ell^w_\tau}$, $\#\mathrm{supp}\,\mathbf{v}^{(i,K)} \lesssim \varepsilon_i^{-1/s}|\mathbf{u}|^{1/s}_{\ell^w_\tau}$, and $\mathbf{P}$ is s^*-admissible, the computation of $\mathbf{z}^{(i)} := \textsc{Apply}[\theta\varepsilon_i/3, \mathbf{P}, \mathbf{v}^{(i,K)}]$ takes a number of operations that is at most a multiple of $\varepsilon_i^{-1/s}|\mathbf{u}|^{1/s}_{\ell^w_\tau}$, $\#\mathrm{supp}\,\mathbf{z}^{(i)} \lesssim \varepsilon_i^{-1/s}|\mathbf{u}|^{1/s}_{\ell^w_\tau}$, and $|\mathbf{z}^{(i)}|_{\ell^w_\tau} \lesssim |\mathbf{u}|_{\ell^w_\tau}$. Finally, by (3.6), the latter result implies that $\textsc{Coarse}[(1-\theta)\varepsilon_i, \mathbf{z}^{(i)}]$ also needs at most a multiple of $\#\mathrm{supp}\,\mathbf{z}^{(i)} + \log(\varepsilon_i^{-1}\|\mathbf{z}^{(i)}\|) \lesssim \varepsilon_i^{-1/s}|\mathbf{u}|^{1/s}_{\ell^w_\tau}$ operations. □

The key to the proof of Theorem 3.11 is the fact that the iterands produced by modSolve are uniformly bounded in ℓ^w_τ. Unfortunately, generally this will not be the case with Solve. Since Solve does not contain a projection onto a complement space of kerF', it is not capable of reducing errors once made in kerF'. Recall that such errors are not reduced by the Richardson steps since they are in the kernel of $\mathbf{M}$. Although, because of the geometric decrease of the tolerances, these errors are summable in ℓ_2, we cannot show this in ℓ^w_τ, and so boundedness of the iterands in ℓ^w_τ is *not* guaranteed. For example, thinking of Rhs and Apply as being performed exactly, i.e., with zero tolerances, each time $\textsc{Coarse}[(1-\theta)\varepsilon_i, \mathbf{v}^{(i,K)}]$ is invoked it gives an error, which might be completely contained in $\ker F'$, for which we can say not more than that its ℓ^w_τ-norm is less than or equal to $|\mathbf{v}^{(i,K)}|_{\ell^w_\tau}$, i.e., that it is bounded.

For $\mathbf{u}^{(i)}$, ε_i as in Solve, for some $\check{s} \in (s, s^*)$, and with $\check{\tau} = (\frac{1}{2}+\check{s})^{-1}$, in the proof of Theorem 3.12 given below we will show that

$$\varepsilon_i^{(\check{s}/s)-1}|\mathbf{u}^{(i)}|_{\ell^w_{\check{\tau}}} \lesssim |\mathbf{u}|^{\check{s}/s}_{\ell^w_\tau}, \qquad \mathrm{supp}\,\mathbf{u}^{(i)} \lesssim \varepsilon_i^{-1/s}|\mathbf{u}|^{1/s}_{\ell^w_\tau}$$

and, as a consequence of these results, that the method is of optimal complexity. Note that for any $\mathbf{v}$ with finite support,

$$|\mathbf{v}|_{\ell^w_{\check{\tau}}} \le (\#\mathrm{supp}\,\mathbf{v})^{\check{s}-s}|\mathbf{v}|_{\ell^w_\tau}. \tag{3.15}$$

Thus $|\mathbf{u}^{(i)}|_{\ell^w_\tau} \lesssim 1$ and $\#\mathrm{supp}\,\mathbf{u}^{(i)} \lesssim \varepsilon_i^{-1/s}$ would give $\varepsilon_i^{(\check{s}/s)-1}|\mathbf{u}^{(i)}|_{\ell^w_{\check{\tau}}} \lesssim |\mathbf{u}^{(i)}|^{\check{s}/s}_{\ell^w_\tau} \lesssim 1$. However, conversely, under no condition on $\#\mathrm{supp}\,\mathbf{u}^{(i)}$, uniform boundedness of $\varepsilon_i^{(\check{s}/s)-1}|\mathbf{u}^{(i)}|_{\ell^w_{\check{\tau}}}$ implies that of $|\mathbf{u}^{(i)}|_{\ell^w_\tau}$. In other words, it will turn out that uniform boundedness in ℓ^w_τ of the iterands is not a necessary condition for obtaining an optimal complexity result.

THEOREM 3.12. *For some $s^* > 0$, assume that $\mathbf{M}$ is s^*-admissible, $\mathbf{g}$ is s^*-optimal, and that for some $s \in (0, s^*)$, with $\tau = (\frac{1}{2}+s)^{-1}$, $\mathbf{Mu} = \mathbf{g}$ has a solution $\mathbf{u} \in \ell^w_\tau$. In addition, assume that there exists an $\check{s} \in (s, s^*)$ such that, with $\check{\tau} = (\frac{1}{2}+\check{s})^{-1}$,*

$$\mathbf{Q} \text{ is bounded on } \ell^w_{\check{\tau}}. \tag{3.16}$$

Then, if the parameter K in SOLVE *is sufficiently large—sufficient is*

$$3\rho^K < \theta \min\left\{1, \left[C_1(\check{\tau})C_2(\check{\tau})|(\mathbf{id}-\mathbf{Q})|_{\ell^w_{\check{\tau}} \leftarrow \ell^w_{\check{\tau}}}\right]^{s/(\check{s}-s)}\right\}, \tag{3.17}$$

where $C_1(\check{\tau})$, $C_2(\check{\tau})$ are the constants from (3.3), (3.7) *respectively—then for all $\varepsilon > 0$, $\mathbf{u}_\varepsilon = $* SOLVE$[\varepsilon, \mathbf{M}, \mathbf{g}]$ *satisfies the following:*

(I) $\#\mathrm{supp}\,\mathbf{u}_\varepsilon \lesssim \varepsilon^{-1/s}|\mathbf{u}|^{1/s}_{\ell^w_\tau}$,

(II) *the number of arithmetic operations used to compute $\mathbf{u}_\varepsilon$ is at most a multiple of $\varepsilon^{-1/s}|\mathbf{u}|^{1/s}_{\ell^w_\tau}$.*

Further, as shown in Proposition 2.1, *$\|\mathbf{Q}(\mathbf{u}-\mathbf{u}_\varepsilon)\| \le \varepsilon$, and so $\|u - F'\mathbf{u}_\varepsilon\|_H \le B_\Psi^{\frac{1}{2}}\varepsilon$.*

Proof. It suffices to prove the statements for any $\varepsilon = \varepsilon_i$ with $\varepsilon_i = (3\rho^K/\theta)^i\varepsilon_0$, as in the algorithm SOLVE.

Since $\mathbf{Q}$ is bounded on ℓ_2, and by assumption it is bounded on $\ell^w_{\check{\tau}}$, an interpolation argument (cf. [21, (4.24)]) shows that it is bounded on ℓ^w_τ as well. Let N_i be the smallest integer such that $\|\mathbf{Qu} - (\mathbf{Qu})_{N_i}\| \le \theta\varepsilon_i/3$, where $(\mathbf{Qu})_N$ denotes the best N-term approximation for $\mathbf{Qu}$. Then, using the assumption $\mathbf{u} \in \ell^w_\tau$, (3.4) shows that

$$N_i \lesssim \varepsilon_i^{-1/s}|\mathbf{Qu}|^{1/s}_{\ell^w_\tau} \lesssim \varepsilon_i^{-1/s}|\mathbf{u}|^{1/s}_{\ell^w_\tau},$$

and so by (3.15),

$$\varepsilon_i^{(\check{s}/s)-1}|(\mathbf{Qu})_{N_i}|_{\ell^w_{\check{\tau}}} \lesssim |\mathbf{u}|^{(\check{s}/s)-1}_{\ell^w_\tau}|(\mathbf{Qu})_{N_i}|_{\ell^w_{\check{\tau}}} \lesssim |\mathbf{u}|^{(\check{s}/s)-1}_{\ell^w_\tau}|\mathbf{Qu}|_{\ell^w_\tau} \lesssim |\mathbf{u}|^{\check{s}/s}_{\ell^w_\tau}. \tag{3.18}$$

From $\|\mathbf{Qu} + (\mathbf{id}-\mathbf{Q})\mathbf{u}^{(i-1)} - \mathbf{v}^{(i,K)}\| \le 2\theta\varepsilon_i/3$ proved in Proposition 2.1, we get

$$\|(\mathbf{Qu})_{N_i} + (\mathbf{id}-\mathbf{Q})\mathbf{u}^{(i-1)} - \mathbf{v}^{(i,K)}\| \le \theta\varepsilon_i.$$

From (3.7) and then (3.3), it follows that $\mathbf{u}^{(i)} := $ COARSE$[(1-\theta)\varepsilon_i, \mathbf{v}^{(i,K)}]$ satisfies

$$\begin{aligned}|\mathbf{u}^{(i)}|_{\ell^w_{\check{\tau}}} &\le C_2(\check{\tau})|(\mathbf{Qu})_{N_i} + (\mathbf{id}-\mathbf{Q})\mathbf{u}^{(i-1)}|_{\ell^w_{\check{\tau}}} \\ &\le C_1(\check{\tau})C_2(\check{\tau})|(\mathbf{Qu})_{N_i}|_{\ell^w_{\check{\tau}}} + C_1(\check{\tau})C_2(\check{\tau})|(\mathbf{id}-\mathbf{Q})|_{\ell^w_{\check{\tau}} \leftarrow \ell^w_{\check{\tau}}}|\mathbf{u}^{(i-1)}|_{\ell^w_{\check{\tau}}},\end{aligned}$$

and so by (3.18) and $\varepsilon_i = 3\rho^K \varepsilon_{i-1}/\theta$,

$$\begin{aligned}&\big(\varepsilon_i^{(\check{s}/s)-1}|\mathbf{u}^{(i)}|_{\ell^w_{\check{\tau}}}\big)\\ &\qquad\qquad \le C|\mathbf{u}|^{\check{s}/s}_{\ell^w_\tau} + C_1(\check{\tau})C_2(\check{\tau})|(\mathbf{id}-\mathbf{Q})|_{\ell^w_{\check{\tau}}\leftarrow\ell^w_{\check{\tau}}}\left(\frac{3\rho^K}{\theta}\right)^{(\check{s}/s)-1}\big(\varepsilon_{i-1}^{(\check{s}/s)-1}|\mathbf{u}^{(i-1)}|_{\ell^w_{\check{\tau}}}\big)\end{aligned}$$

for some constant $C > 0$. We may conclude that if K satisfies (3.17), then solutions of the homogeneous part of this recursion converge to zero, and so

$$\varepsilon_i^{(\check{s}/s)-1}|\mathbf{u}^{(i)}|_{\ell^w_{\check{\tau}}} \lesssim |\mathbf{u}|^{\check{s}/s}_{\ell^w_\tau}, \tag{3.19}$$

which, as we emphasize here, holds uniformly in i.

Knowing (3.19), Proposition 3.2 and (3.18) show that

$$\begin{aligned}\#\operatorname{supp}\mathbf{u}^{(i)} &\lesssim \varepsilon_i^{-1/\check{s}}|(\mathbf{Qu})_{N_i} + (\mathbf{id}-\mathbf{Q})\mathbf{u}^{(i-1)}|^{1/\check{s}}_{\ell^w_{\check{\tau}}}\\ &\lesssim \varepsilon_i^{-1/s}\big(\varepsilon_i^{(\check{s}/s)-1}\big[|(\mathbf{Qu})_{N_i}|_{\ell^w_{\check{\tau}}} + |\mathbf{id}-\mathbf{Q}|_{\ell^w_{\check{\tau}}\leftarrow\ell^w_{\check{\tau}}}|\mathbf{u}^{(i-1)}|_{\ell^w_{\check{\tau}}}\big]\big)^{1/\check{s}}\\ &\lesssim \varepsilon_i^{-1/s}|\mathbf{u}|^{1/s}_{\ell^w_\tau};\end{aligned}$$

i.e., (I) is valid.

The remainder of the proof resembles that of Theorem 3.11. We have to show that the K intermediate steps that transfer $\mathbf{u}^{(i-1)}$ to $\mathbf{u}^{(i)}$ take a number of operations that is bounded by some multiple of $\varepsilon_i^{-1/s}|\mathbf{u}|^{1/s}_{\ell^w_\tau}$.

As in the proof of Theorem 3.11, since $\mathbf{M}$ is s^*-admissible and $\mathbf{g}$ is s^*-optimal, we have $\#\operatorname{supp}\mathbf{g}^{(i)} \lesssim \varepsilon_i^{-1/s}|\mathbf{u}|^{1/s}_{\ell^w_\tau}$ and $|\mathbf{g}^{(i)}|_{\ell^w_\tau} \lesssim |\mathbf{u}|_{\ell^w_\tau}$, and so, in addition, $\varepsilon_i^{(\check{s}/s)-1}|\mathbf{g}^{(i)}|_{\ell^w_{\check{\tau}}} \lesssim |\mathbf{u}|^{\check{s}/s}_{\ell^w_\tau}$.

Since $\mathbf{M}$ is s^*-admissible, this last result, together with $\varepsilon_{i-1}^{(\check{s}/s)-1}|\mathbf{u}^{(i-1)}|_{\ell^w_{\check{\tau}}} \lesssim |\mathbf{u}|^{\check{s}/s}_{\ell^w_\tau}$, shows that

$$\varepsilon_i^{(\check{s}/s)-1}|\mathbf{v}^{(i,j)}|_{\ell^w_{\check{\tau}}} \lesssim |\mathbf{u}|^{\check{s}/s}_{\ell^w_\tau} \qquad (0 \le j \le K) \tag{3.20}$$

by Remark 3.6 (use $\check{s} < s^*$).

A new element in this proof is the observation that, instead of uniform boundedness in ℓ^w_τ, (3.20) is already sufficient to guarantee that the supports have the appropriate sizes. Indeed, again since $\mathbf{M}$ is s^*-admissible (use $\check{s} < s^*$), it follows that

$$\#\operatorname{supp}\mathbf{v}^{(i,j)} \lesssim \varepsilon_i^{-1/\check{s}}|\mathbf{v}^{(i,j-1)}|^{1/\check{s}}_{\ell^w_{\check{\tau}}} \lesssim \varepsilon_i^{-1/s}|\mathbf{u}|^{1/s}_{\ell^w_\tau} \qquad (1 \le j \le K), \tag{3.21}$$

whereas by $\#\operatorname{supp}\mathbf{v}^{(i,0)} \lesssim \varepsilon_{i-1}^{-1/s}|\mathbf{u}|^{1/s}_{\ell^w_\tau}$ (via (I)), $\#\operatorname{supp}\mathbf{g}^{(i)} \lesssim \varepsilon_i^{-1/s}|\mathbf{u}|^{1/s}_{\ell^w_\tau}$, and the second inequality in (3.21), its computation takes not more than a multiple of $\varepsilon_i^{-1/s}|\mathbf{u}|^{1/s}_{\ell^w_\tau}$ operations.

Finally, by (3.6), the application of COARSE$[(1-\theta)\varepsilon_i, \mathbf{v}^{(i,K)}]$ needs at most a multiple of $\#\operatorname{supp}\mathbf{v}^{(i,K)} + \log(\varepsilon_i^{-1}\|\mathbf{v}^{(i,K)}\|) \lesssim \varepsilon_i^{-1/s}|\mathbf{u}|^{1/s}_{\ell^w_\tau}$ operations. □

Remark 3.13. The conditions imposed in Theorem 3.12 do not exclude the possibility that $\mathbf{u}_\varepsilon \in \operatorname{Ran} F$, and so $\mathbf{Qu}_\varepsilon = \mathbf{u}_\varepsilon$. Then in a manner analogous to Remark 3.6, the estimates $\|\mathbf{Q}(\mathbf{u}-\mathbf{u}_\varepsilon)\| \le \varepsilon$ and $\#\operatorname{supp}\mathbf{u}_\varepsilon \lesssim \varepsilon^{-1/s}|\mathbf{u}|^{1/s}_{\ell^w_\tau}$ *imply* that $\mathbf{Q} : \ell^w_\tau \to \ell^w_\tau$ is bounded. In this sense, the condition imposed in Theorem 3.12 that, for some $\check{\tau} < \tau$, $\mathbf{Q} : \ell^w_{\check{\tau}} \to \ell^w_{\check{\tau}}$ is bounded, is an almost necessary one.

4. Construction of frames.

Recall that $L : H \to H'$ was assumed to be a boundedly invertible operator, where we have in mind a linear differential or integral operator. When L is an operator of order $2t$, typically H is a Sobolev space of order t. We also briefly discussed the case of having systems of such equations, which, however, poses no principal additional difficulties. Thus here we restrict ourselves to scalar equations, where in addition we assume that the equation is imposed on a *domain* $\Omega \subset \mathbb{R}^n$. In particular, in connection with integral equations, it is also relevant to study the case of the equation being formulated on a manifold. However, in that case the construction of a frame may follow the same principles as in the domain case that we outline here.

4.1. Overlapping domain decompositions.

THEOREM 4.1. *For some $\Gamma^D \subset \partial\Omega$, possibly $\Gamma^D = \emptyset$, and $t \in \mathbb{R}$, let*

$$\mathcal{H}^t = \begin{cases} H^t_{0,\Gamma^D}(\Omega) & \text{when } t \geq 0, \\ (H^{-t}_{0,\Gamma^D}(\Omega))' & \text{when } t < 0, \end{cases}$$

where for $t \geq 0$

$$H^t_{0,\Gamma^D}(\Omega) = \operatorname{clos}_{H^t(\Omega)}\{u \in H^t(\Omega) \cap C^\infty(\Omega) : \operatorname{supp} u \cap \Gamma^D = \emptyset\}.$$

Let $\Omega = \cup_{i=1}^m \Omega_i$ be an open covering, by which we mean that the sets Ω_i are open and that there exists a partition of unity $\{\chi_i\}$ relative to $\{\Omega_i\}$; i.e., $\chi_i \in C^\infty(\Omega)$, $0 \leq \chi_i \leq 1$, χ_i vanishes outside Ω_i, and $\sum_i \chi_i = 1$.

With

$$\Gamma_i^D = \begin{cases} \partial\Omega_i \cap (\Omega \cup \Gamma^D) & \text{if } t \geq 0, \\ \partial\Omega_i \cap \Gamma^D & \text{if } t < 0, \end{cases}$$

let

$$\mathcal{H}_i^t = \begin{cases} H^t_{0,\Gamma_i^D}(\Omega_i) & \text{if } t \geq 0, \\ (H^{-t}_{0,\Gamma_i^D}(\Omega_i))' & \text{if } t < 0, \end{cases}$$

and let $\Psi^{(i)}$ be a Riesz basis or, more generally, a frame for $\mathcal{H}_i^t$.

Let $\{\omega_i\}_{1 \leq i \leq m}$ be a collection of nonnegative functions on Ω, with ω_i smooth on Ω_i and zero outside Ω_i, such that there exists an open covering $\Omega = \cup_{i=1}^m \hat{\Omega}_i$ with $\hat{\Omega}_i \subset \Omega_i$ and $\omega_i \eqsim 1$ on $\hat{\Omega}_i$.

Then

$$\cup_i \omega_i \Psi^{(i)} \text{ is a frame for } \mathcal{H}^t.$$

Proof. First we demonstrate that for $u \in \mathcal{H}^t$,

$$\|u\|^2_{\mathcal{H}^t} \eqsim \inf_{\omega_i^{-1} u_i \in \mathcal{H}_i^t, \sum_i u_i = u} \sum_i \|\omega_i^{-1} u_i\|^2_{\mathcal{H}_i^t}. \tag{4.1}$$

Here, by writing $\omega_i^{-1} u_i \in \mathcal{H}_i^t$, in particular we implicitly state that u_i vanishes outside $\operatorname{supp} \omega_i$.

Let $\omega_i^{-1} u_i \in \mathcal{H}_i^t$. Then, since ω_i is smooth on Ω_i, $u_i \in \mathcal{H}_i^t$. Furthermore, the spaces $\mathcal{H}_i^t$ are selected in such a way that the trivial extension with zero of a function on Ω_i extends to a bounded mapping from $\mathcal{H}_i^t \to \mathcal{H}^t$. To see this for $t < 0$, note that the restriction of a function on Ω to Ω_i, which is the adjoint of the zero extension, is a bounded mapping from $H^{-t}_{0,\Gamma_D}(\Omega)$ to $H^{-t}_{0,\partial\Omega_i\cap\Gamma^D}(\Omega_i)$. We conclude that for any $\omega_i^{-1}u_i \in \mathcal{H}_i^t$, the function $u = \sum_i u_i \in \mathcal{H}^t$, with $\|u\|^2_{\mathcal{H}^t} \lesssim \sum_i \|u_i\|^2_{\mathcal{H}_i^t} \lesssim \sum_i \|\omega_i^{-1}u_i\|^2_{\mathcal{H}_i^t}$.

Conversely, let $\{\hat{\chi}_i\}$ be a partition of unity relative to $\{\hat{\Omega}_i\}$. Then any $u \in \mathcal{H}^t$ can be written as $u = \sum_i \hat{\chi}_i u$, where, because of $\omega_i \eqsim 1$ on $\hat{\Omega}_i$, $\omega_i^{-1}\hat{\chi}_i u \in \mathcal{H}_i^t$ and $\|\omega_i^{-1}\hat{\chi}_i u\|_{\mathcal{H}_i^t} \lesssim \|\hat{\chi}_i u\|_{\mathcal{H}_i^t} \lesssim \|u\|_{\mathcal{H}^t}$, completing the proof of (4.1).

Since $\Psi^{(i)}$ is a frame for $\mathcal{H}_i^t$, for $v_i \in \mathcal{H}_i^t$ we have $\|v_i\|^2_{\mathcal{H}_i^t} \eqsim \inf_{\mathbf{c}_i\in\ell_2,\,\mathbf{c}_i^T\Psi^{(i)}=v_i} \|\mathbf{c}_i\|^2_{\ell_2}$. With v_i of the form $\omega_i^{-1}u_i$, $\mathbf{c}_i^T\Psi^{(i)} = v_i$ is equivalent to $\mathbf{c}_i^T\omega_i\Psi^{(i)} = u_i$. Now from (4.1) we conclude that, for $u \in \mathcal{H}^t$,

$$\begin{aligned}\|u\|^2_{\mathcal{H}^t} &\eqsim \inf_{\omega_i^{-1}u_i\in\mathcal{H}_i^t,\,\sum_i u_i=u} \sum_i \inf_{\mathbf{c}_i\in\ell_2,\,\mathbf{c}_i^T\omega_i\Psi^{(i)}=u_i} \|\mathbf{c}_i\|^2 \\ &= \inf_{(\mathbf{c}_1^T,\ldots,\mathbf{c}_m^T)^T\in\ell_2,\,\sum_i \mathbf{c}_i^T\omega_i\Psi^{(i)}=u} \sum_i \|\mathbf{c}_i\|^2,\end{aligned}$$

meaning that $\cup_i\omega_i\Psi^{(i)}$ is a frame for $\mathcal{H}^t$. □

Remark 4.2. If Theorem 4.1 is applied, with ω_i being the characteristic function of $\Omega_i = \hat{\Omega}_i$, then it shows that $\cup_i\Psi^{(i)}$ is a frame for $\mathcal{H}^t$.

If each ω_i is selected such that it *vanishes at* the internal boundary $\partial\Omega_i \cap \Omega$, then the above proof shows that boundary conditions on that part of $\partial\Omega_i$ can actually be chosen at one's convenience; i.e., any $\partial\Omega_i \cap \Gamma^D \subset \Gamma_i^D \subset \partial\Omega_i \cap (\Omega \cup \Gamma^D)$ will do.

To construct collections $\Psi^{(i)}$ that serve as ingredients in Theorem 4.1, we may proceed as follows: Suppose that for each $1 \leq i \leq m$ we have a sufficiently smooth regular parametrization κ_i between $(0,1)^n$, or another reference domain, and Ω_i (see Figure 1). With$\Gamma^D_{i,\square} = \kappa_i^{-1}(\Gamma_i^D)$, let $\Psi^{(i)}_\square$ be a Riesz basis for $H^t_{0,\Gamma^D_{i,\square}}(0,1)^n$ when $t \geq 0$,

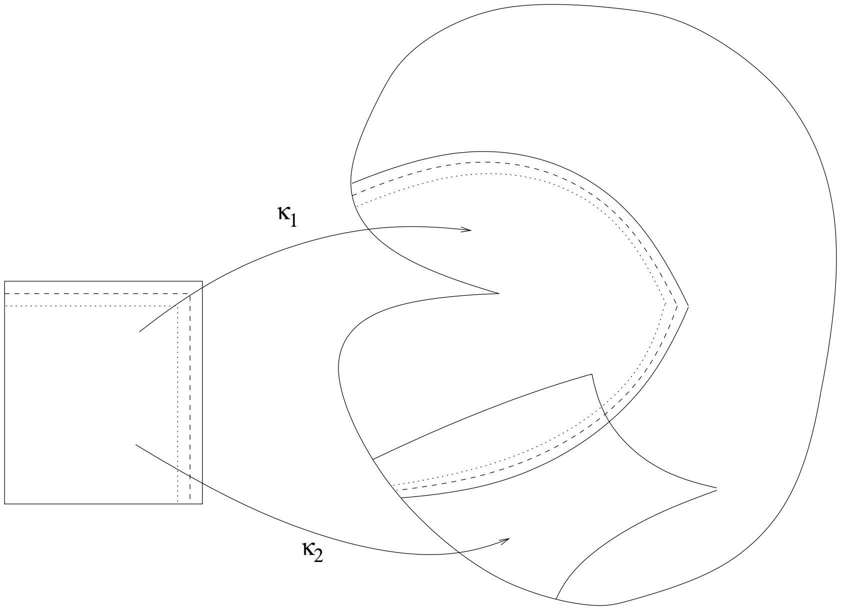


FIG. 1. *Overlapping domain decomposition. (The dashed and dotted lines will be defined in section 4.4.)*

or for $(H^{-t}_{0,\Gamma^D_{i,\square}}(0,1)^n)'$ otherwise. Then we may conclude that $\Psi^{(i)} = \Psi^{(i)}_\square \circ \kappa_i^{-1}$ is a Riesz basis for $\mathcal{H}^t_i$.

At least if the parametrizations are constructed such that the image of a face of $[0,1]^n$ either has empty intersection with Γ_D or is fully contained in Γ_D, then $\Psi^{(i)}_\square$ of wavelet type can easily be constructed by taking tensor products of wavelet bases on the interval with appropriate boundary conditions.

With the construction of spline wavelets on the interval from [15], only wavelets with supports near the endpoints depend on the boundary condition. This means that if the weights ω_i in Theorem 4.1 vanish in a sufficiently large *neighborhood of* the internal boundaries $\partial\Omega_i \cap \Omega$, then boundary conditions at these internal boundaries are irrelevant since they have no influence on the constructed frame.

Compared to the construction of wavelet *bases* for $\mathcal{H}^t$ based on a *nonoverlapping* decomposition of the domain, the frame approach seems to have the following advantages:

- It is easier to construct parametrizations corresponding to an overlapping domain decomposition; only local parametrizations of $\partial\Omega$ are needed. Having less complicated κ_i may also have a favorable quantitative effect on the frame constants A_Ψ and B_Ψ.
- Constructions of wavelet bases based on nonoverlapping domain decompositions all involve a kind of "stitching" of wavelets from different subdomains at the interfaces. The construction from [17] yields wavelets with all desired theoretical properties, but it seems not easy to implement. The constructions proposed in [16, 3, 8] yield near the interfaces wavelets that are only continuous, which restricts the values of s^* for which $\mathbf{M}$ (and $\mathbf{P}$) are s^*-compressible (see [5, Proposition 6.2.2], [25], and section 4.5).
- When a frame construction similar to that in Theorem 4.1 is applied on a *closed manifold* with weights ω_i that vanish at the internal boundaries, then the wavelet bases on $(0,1)^n$ that serve as ingredients may satisfy *periodic boundary conditions*. Not only is the implementation of such bases much easier, but also they are much better conditioned than available wavelet bases that satisfy Dirichlet or Neumann boundary conditions.

4.2. Regularity. As we have seen, under some conditions the routine Solve or modSolve exhibits an error decay of order N^{-s}, with N being the number of operations spent and coefficients stored, *in case* $\mathbf{u} \in \ell^w_\tau$ with $\tau = (\frac{1}{2}+s)^{-1}$. Recall that $\mathbf{u}$ is some solution of $\mathbf{Mu} = \mathbf{g}$; that is, $u = \mathbf{u}^T\Psi$ is the solution of $Lu = g$.

In case Ψ is a Riesz basis for $\mathcal{H}^t$ of biorthogonal wavelet type of *order* d, meaning that $d-1$ is the order of local polynomial reproduction, then it is known that for

$$0 < s < \frac{d-t}{n}$$

it holds that

$$\mathbf{u} \in \ell_\tau \quad \text{if and only if} \quad u \in B^{sn+t}_\tau(L_\tau(\Omega)), \tag{4.2}$$

at least when the wavelets are contained in $B^{sn+t}_\tau(L_\tau(\Omega))$ and $s \leq 1/2$ if $t < -n/2$ (see Figure 2). Recall that $\mathbf{u} \in \ell_\tau$ implies $\mathbf{u} \in \ell^w_\tau$. Here, for $\nu \geq 0$, $B^\nu_p(L_p(\Omega))$ is the usual Besov space, in which possible boundary conditions are incorporated, measuring "ν orders of smoothness in L_p," and for $\nu < 0$, $B^\nu_p(L_p(\Omega)) := (B^{-\nu}_{p'}(L_{p'}(\Omega))'$ with $1/p + 1/p' = 1$, and so necessarily $p \geq 1$. This latter restriction induces the afore-

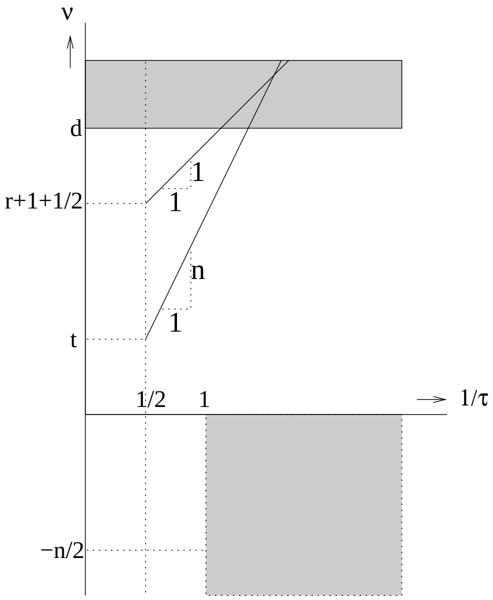


FIG. 2. $B^{\nu}_{\tau}(L_{\tau}(\Omega))$ *with the line* $\nu = sn + t$*, where* $\tau = (\frac{1}{2} + s)^{-1}$*, and the line* $\nu = r + 1 + 1/\tau$*.*

mentioned condition $s \leq 1/2$ if $t < -n/2$. For details about Besov spaces and proofs of (4.2) in various circumstances, we refer to [4].

Remark 4.3. If the wavelets are piecewise smooth globally C^r-functions for some $r \in \mathbb{N} \cup \{-1\}$, with $r = -1$ meaning that they satisfy no global smoothness requirements, then it is known that they are contained in $B^{\nu}_{\tau}(L_{\tau}(\Omega))$ when $\nu < r + 1 + 1/\tau$, whereas they are not contained in this space when $\nu > r+1+1/\tau$. So if for $s = (d-t)/n$ with $\tau = (\frac{1}{2} + s)^{-1}$ it holds that $r + 1 + 1/\tau \geq sn + t$, i.e.,

$$r \geq -\frac{3}{2} + d + \frac{t-d}{n}, \tag{4.3}$$

then smoothness of the wavelets does not limit the range for which (4.2) is valid.

With spline wavelets we have $r = d - 2$, meaning that (4.3) reads as the mild requirement $(d - t)/n \geq 1/2$.

Remark 4.4. Note that to obtain a rate N^{-s} with a linear nonadaptive method it is needed that the solution u be in $H^{sn+t}(\Omega)$, which is a much smaller space than $B^{sn+t}_{\tau}(L_{\tau}(\Omega))$. Recently, a number of regularity proofs have appeared showing that various operators have indeed a much larger regularity in the above Besov scale than in the Sobolev scale (e.g., see [13, 10]). A particular example is the operator corresponding to Poisson's equation on a two-dimensional polygonal domain, which has been shown to have infinity regularity in the Besov scale (see [9]). This means that the Besov regularity of the solution is only limited by the smoothness of the right-hand side, and thus it can be arbitrarily large.

Now let $\Psi = \cup_i \omega_i \Psi^{(i)}$ be a *frame* for $\mathcal{H}^t$, as constructed in Theorem 4.1, where the $\Psi^{(i)}$ are sufficiently smooth biorthogonal wavelet bases of order d for $\mathcal{H}^t_i$. Let $\{\hat{\chi}_i\}$ be a partition of unity relative to $\{\hat{\Omega}_i\}$. Then $u \in B^{sn+t}_{\tau}(L_{\tau}(\Omega))$ implies $\hat{\chi}_i u \in B^{sn+t}_{\tau}(L_{\tau}(\Omega_i))$ and also $\omega_i^{-1}\hat{\chi}_i u \in B^{sn+t}_{\tau}(L_{\tau}(\Omega_i))$. Thus if $0 < s < (d-t)/n$, and $s \leq 1/2$ if $t < -n/2$, then (4.2) demonstrates that each $\omega_i^{-1}\hat{\chi}_i u$ has a unique expansion $\mathbf{u}_i^T\Psi^{(i)}$, and so $\hat{\chi}_i u = \mathbf{u}_i^T\omega_i\Psi^{(i)}$, where $\mathbf{u}_i \in \ell_\tau$. We conclude that u has a representation $\sum_i \mathbf{u}_i^T\omega_i\Psi^{(i)}$ with $(\mathbf{u}_1^T,\dots,\mathbf{u}_m^T)^T \in \ell_\tau \subset \ell_\tau^w$ under the same condition on u as is needed in the Riesz basis case.

4.3. Boundedness of Q, i.e., condition (3.16). Assuming that $\mathbf{Mu} = \mathbf{g}$ has a solution $\mathbf{u} \in \ell_\tau^w$, in Theorem 3.12 the optimal computational complexity of SOLVE is proved when for some $\check{\tau} < \tau$, $\mathbf{Q}$ is bounded on $\ell_{\check{\tau}}^w$. Recall that $\mathbf{Q} = F(F'F)^{-1}F'$, where with the frame construction from Theorem 4.1 and biorthogonal wavelet bases $\Psi^{(i)}$ on the subdomains, $F' : \ell_2 \to \mathcal{H}^t : \mathbf{c} = (\mathbf{c}_1^T,\dots,\mathbf{c}_m^T)^T \mapsto \sum_i \omega_i\mathbf{c}_i^T\Psi^{(i)}$, and so $F : (\mathcal{H}^t)' \to \ell_2 : u \mapsto ((\langle u, \omega_i\Psi^{(i)}\rangle_{L_2(\Omega)})_i)^T = ((\langle \omega_i u, \Psi^{(i)}\rangle_{L_2(\Omega_i)})_i)^T$.

For wavelets that are sufficiently smooth, from (4.2) we know that $\mathbf{c}_i \mapsto \mathbf{c}_i^T\Psi_i$ is bounded from $\ell_{\check{\tau}} \to B_{\check{\tau}}^{\check{s}n+t}(L_{\check{\tau}}(\Omega_i))$ when $0 < \check{s} < (d-t)/n$ and $\check{s} \le 1/2$ if $t < -n/2$.

The mapping $v_i \mapsto \langle v_i, \Psi^{(i)}\rangle_{L_2(\Omega_i)}^T$ is the inverse of $\mathbf{d}_i \mapsto \mathbf{d}_i^T\tilde{\tilde{\Psi}}^{(i)}$, where $\tilde{\Psi}^{(i)}$ is the dual wavelet basis. When the dual wavelets are sufficiently smooth, the latter mapping is boundedly invertible from $\ell_{\check{\tau}}$ to $B_{\check{\tau}}^{\check{s}n-t}(L_{\check{\tau}}(\Omega_i))$ when $0 < \check{s} < (\tilde{d}+t)/n$, where $\tilde{d}$ is the order of the dual multiresolution analysis and $\check{s} \le 1/2$ if $-t < -n/2$.

If we now, in addition, assume that the ω_i vanish at the internal boundaries $\partial\Omega_i \cap \Omega$, and thus that these weights are globally smooth on Ω, then we may conclude that $F' : \ell_{\check{\tau}} \to B_{\check{\mathcal{T}}}^{\check{s}n+t}(L_{\check{\tau}}(\Omega))$ and $F : B_{\check{\tau}}^{\check{s}n-t}(L_{\check{\tau}}(\Omega)) \to \ell_{\check{\tau}}$ are bounded when $0 < \check{s} < \min\{(d-t)/n, (\tilde{d}+t)/n\}$ and $\check{s} \le 1/2$ if $|t| > n/2$.

Unfortunately, so far we can show boundedness of $(F'F)^{-1} : B_{\check{\tau}}^{\check{s}n+t}(L_{\check{\tau}}(\Omega)) \to B_{\check{\tau}}^{\check{s}n-t}(L_{\check{\tau}}(\Omega))$, which in combination with the above boundedness of F' and F would give the desired property of $\mathbf{Q}$, only in the particular situation that $t = 0$ and the $\Psi^{(i)}$ are $L_2(\Omega_i)$-orthonormal bases. In that case, $FF'u = \sum_i \omega_i\langle \omega_i u, \Psi^{(i)}\rangle_{L_2(\Omega_i)}\Psi^{(i)} = (\sum_i \omega_i^2)u$, and so $(FF')^{-1}u = (\sum_i \omega_i^2)^{-1}u$, meaning that, by the global smoothness of the weights, $(FF')^{-1}$ clearly has the above property for any s. Note that, on the other hand, if we do not damp the wavelets near the internal boundaries, i.e., if ω_i is just the characteristic function of Ω_i, then $(F'F)^{-1}$ will be bounded on $B_{\check{\tau}}^{\check{s}n}(L_{\check{\tau}}(\Omega))$ for $\check{s}$ in a limited range only.

Thus at least for $t = 0$ and with sufficiently smooth orthonormal $\Psi^{(i)}$ (which implies $\tilde{d} = d$) and weights ω_i that vanish at internal boundaries $\partial\Omega_i \cap \Omega$, for any τ with $0 < s = 1/\tau - 1/2 < d/n$ (which is the full range for which one may expect that $\mathbf{u} \in \ell_\tau^w$), there exists a $\check{\tau} < \tau$ such that $\mathbf{Q}$ is bounded on $\ell_{\check{\tau}}$. Since $\mathbf{Q}$ is also bounded on ℓ_2, an interpolation argument (cf. [21, (4.24)]) shows that it is bounded on $\ell_{\check{\tau}}^w$.

4.4. Construction of P. The fact that we have no general answer about whether $\mathbf{Q}$ satisfies (3.16) was the motivation to introduce the routine modSOLVE, which contains the inexact application of a suitable projector $\mathbf{P}$.

In the situation of Theorem 4.1, thus with $\{\hat{\chi}_i\}$ a partition of unity relative to $\{\hat{\Omega}_i\}$ and ω_i the weights, and where $\Psi^{(i)}$ are biorthogonal wavelet bases for $\mathcal{H}_i^t$ with duals $\tilde{\Psi}^{(i)}$, let us define $Z : u \mapsto ((\langle \hat{\chi}_i\omega_i^{-1}u, \tilde{\Psi}^{(i)}\rangle_{L_2(\Omega_i)})_i)^T$, which is a bounded mapping from $\mathcal{H}^t \to \ell_2$. It holds that $F'Zu = \sum_i \omega_i\langle \hat{\chi}_i\omega_i^{-1}u, \tilde{\Psi}^{(i)}\rangle_{L_2(\Omega_i)}\Psi^{(i)} = u$. Thus, defining

$$\mathbf{P} = ZF' : (\mathbf{c}_1^T,\dots,\mathbf{c}_m^T)^T \mapsto (((\langle \hat{\chi}_i\omega_i^{-1}{\textstyle\sum_{\breve{i}}}\omega_{\breve{i}}\mathbf{c}_{\breve{i}}^T\Psi^{(\breve{i})}, \tilde{\Psi}^{(i)}\rangle_{L_2(\Omega_i)})_i^T)^T,$$

we infer that $\mathbf{P} : \ell_2 \to \ell_2$ is a bounded projector with $\operatorname{Ker}\mathbf{P} = \operatorname{Ker}F'$, which are the basic requirements on $\mathbf{P}$ imposed in section 2.3 and which guarantee that modSOLVE is convergent, i.e., that Proposition 2.3 is valid. Note that the application of $\mathbf{P}$ may only change coefficients corresponding to wavelets for which the support or the support of the corresponding dual wavelet intersect more than one Ω_i.

To apply the above $\mathbf{P}$, we need a practical construction of the partition of unity $\{\hat{\chi}_i\}$. Apart from this, we discuss here the construction of weights in Theorem 4.1 that vanish at, or even in a neighborhood of, the internal boundaries $\partial\Omega_i \cap \Omega$. As we have seen, the application of weights that vanish at the internal boundaries seems necessary for $\mathbf{Q}$ satisfying (3.16), whereas, even when the application of $\mathbf{P}$ is necessary, such weights may have a favorable quantitative effect. Furthermore, weights that vanish even in a neighborhood of the internal boundaries allow us to ignore boundary conditions at these boundaries with the construction of wavelets on the subdomains.

Let $\Omega = \cup_i \Omega_i$ be an open covering, and $\kappa_i : (0,1)^n \to \Omega_i$ smooth regular parametrizations, where we assume that the image of a face of $[0,1]^n$ under κ_i either has empty intersection with $\partial\Omega$ or is contained in $\partial\Omega$. Then there exist $0 \leq \hat{a}_j^{(i)} \leq \check{a}_j^{(i)} < \check{b}_j^{(i)} \leq \hat{b}_j^{(i)} \leq 1$ such that $\cup_{i=1}^m \kappa_i(\prod_{j=1}^n (\check{a}_j^{(i)}, \check{b}_j^{(i)})) = \Omega$, whereas strict inequalities $0 < \hat{a}_j^{(i)} < \check{a}_j^{(i)}$ or $\check{b}_j^{(i)} < \hat{b}_j^{(i)} < 1$ hold if (and only if) the face corresponds to an internal boundary (cf. dashed and dotted boundaries in Figure 1).

Now let $\eta_i, \phi_i \in C^\infty((0,1)^n)$, with $0 \leq \eta_i, \phi_i \leq 1$, such that $\eta_i \eqsim 1$ on $\prod_j (\hat{a}_j^{(i)}, \hat{b}_j^{(i)})$, whereas it vanishes at, or even in a neighborhood of, faces of $[0,1]^n$ that correspond to internal boundaries, and $\phi_i = 1$ on $\prod_j (\check{a}_j^{(i)}, \check{b}_j^{(i)})$, whereas it vanishes at faces of $\prod_{j=1}^n (\hat{a}_j^{(i)}, \hat{b}_j^{(i)})$ that correspond to internal boundaries.

Defining $\hat{\Omega}_i = \kappa_i(\prod_j (\hat{a}_j^{(i)}, \hat{b}_j^{(i)}))$ and

$$\omega_i = \begin{cases} \eta_i \circ \kappa_i^{-1} & \text{on } \Omega_i, \\ 0 & \text{on } \Omega \backslash \Omega_i \end{cases}$$

as desired, we have that $\omega_i \in C^\infty(\Omega)$, $0 \leq \omega_i \leq 1$, $\omega_i \eqsim 1$ on $\hat{\Omega}_i$, and ω_i vanishes at or even in a neighborhood of $\partial\Omega_i \cap \Omega$.

Defining $\check{\Omega}_i = \kappa_i(\prod_j (\check{a}_j^{(i)}, \check{b}_j^{(i)}))$ and

$$\hat{\chi}_i^{(i)} = \begin{cases} \phi_i \circ \kappa_i^{-1} & \text{on } \Omega_i, \\ 0 & \text{on } \Omega \backslash \Omega_i, \end{cases}$$

we have $0 \leq \hat{\chi}_i^{(i)} \leq 1$, $\hat{\chi}_i^{(i)} = 1$ on $\check{\Omega}_i$, and $\hat{\chi}_i^{(i)}$ vanishes outside $\hat{\Omega}_i$. A partition of unity relative to $\{\hat{\Omega}_i\}$ is now given by $\{\hat{\chi}_i^{(m)} : 1 \leq i \leq m\}$, where for $2 \leq k \leq m$, $\{\hat{\chi}_i^{(k)} : 1 \leq i \leq k-1\}$ is defined by $\hat{\chi}_i^{(k)} := \hat{\chi}_i^{(k-1)}(1 - \hat{\chi}_k^{(k)})$. Indeed, an induction argument shows that $\sum_{i=1}^k \hat{\chi}_i^{(k)} = 1$ on $\cup_{i=1}^k \check{\Omega}_i$, and so in particular $\sum_{i=1}^m \hat{\chi}_i^{(m)} = 1$ on $\cup_{i=1}^m \check{\Omega}_i = \Omega$. Furthermore, $\hat{\chi}_i^{(m)} \in C^\infty(\Omega)$, $0 \leq \hat{\chi}_i^{(m)} \leq 1$, and, since $\operatorname{supp} \hat{\chi}_i^{(m)} \subset \operatorname{supp} \hat{\chi}_i^{(i)}$, $\hat{\chi}_i^{(m)}$ vanishes outside $\hat{\Omega}_i$.

4.5. Compressibility, i.e., the value of s^*. Let $\Psi = \cup_{i=1}^m \omega_i \Psi^{(i)}$ be a frame for $\mathcal{H}^t$ as constructed in Theorem 4.1, where the $\Psi^{(i)}$ are biorthogonal wavelet bases for $\mathcal{H}_i^t$ of order d, with dual bases $\tilde{\Psi}^{(i)}$ of order $\tilde{d}$.

We write $\Psi^{(i)} = \{\psi_\lambda^{(i)} : \lambda \in J^{(i)}\}$ and $\tilde{\Psi}^{(i)} = \{\tilde{\psi}_\lambda^{(i)} : \lambda \in J^{(i)}\}$, where we think of λ as consisting of two coordinates referring to scale and location, respectively. Denoting the scale associated with λ by $|\lambda| \in \mathbb{N}$, we assume that the primal wavelets are *local* in the sense that

$$\operatorname{diam}(\operatorname{supp} \psi_\lambda^{(i)}) \lesssim 2^{-|\lambda|} \quad \text{and} \quad \sup_{x \in \Omega_i, \ell \in \mathbb{N}} \#\{|\lambda| = \ell : x \in \operatorname{supp} \psi_\lambda^{(i)}\} < \infty.$$

We set

$$\gamma = \sup\{s \in \mathbb{R} : \|\psi_\lambda^{(i)}\|_{H^s(\Omega_i)} \lesssim 2^{|\lambda| s} \|\psi_\lambda^{(i)}\|_{L_2(\Omega_i)}, \lambda \in J^{(i)}, 1 \leq i \leq m\},$$

with an analogous definition of $\tilde{\gamma}$ involving dual wavelets. Necessarily, it holds that $t \in (-\tilde{\gamma}, \gamma)$. It is known that if the primal wavelets are piecewise smooth globally C^r-functions for some $r \in \mathbb{N} \cup \{-1\}$, then $\gamma = r + \frac{3}{2}$. It holds that $r \leq d - 2$, with the equality sign for spline wavelets.

Now let $L : \mathcal{H}^t \to (\mathcal{H}^t)'$ be boundedly invertible. Then $\mathbf{M} = FLF'$ is represented by an $m \times m$ blockmatrix with its $(i, \breve{\imath})$th block equal to the infinite matrix $\langle \omega_i \Psi^{(i)}, L\omega_{\breve{\imath}} \Psi^{(\breve{\imath})} \rangle_{L_2(\Omega)}$. Assuming that the weights ω_i vanish at the internal boundaries so that they are globally smooth, the analysis of the compressibility of each of these blocks can follow exactly the same lines as that of the compressibility of L with respect to a biorthogonal wavelet *basis* characterized by the same tuple $(d, \gamma, \tilde{d}, \tilde{\gamma})$.

If for some $\sigma > 0$, $L, L' : \mathcal{H}^{t+\sigma} \to \mathcal{H}^{-t+\sigma}$ are bounded, then, by substituting the estimates [14, (9.4.5), (9.4.8)] into [5, Proposition 6.6.2], we infer that $\mathbf{M}$ is s^*-compressible with

$$s^* = \frac{\min\{\sigma, \gamma - t, t + \tilde{d}\}}{n} - \frac{1}{2}, \tag{4.4}$$

at least when this value is positive. (We have used the fact that the condition $\sigma < t + \tilde{\gamma}$ imposed for [14, (9.4.8)] can actually be relaxed to $\sigma \leq t + \tilde{d}$.) The above result holds true for local operators L, i.e., $\langle v, Lw \rangle_{L_2(\Omega)} = 0$ when $\operatorname{supp} v \cap \operatorname{supp} w = \emptyset$, as well as for nonlocal L of the form

$$(Lv)(x) = \int_\Omega K(x, y) v(y) dy,$$

with a Schwarz kernel that has the Calderon–Zygmund property

$$|\partial_x^\alpha \partial_y^\beta K(x, y)| \lesssim |x - y|^{-(n+2t+|\alpha|+|\beta|)}, \qquad n + 2t + |\alpha| + |\beta| > 0.$$

The spaces $\mathcal{H}^r$ used to formulate the above continuity assumptions on L and L' are defined for $r \geq 0$ by

$$\mathcal{H}^r = \begin{cases} [L_2(\Omega), H^{|t|}_{0,\Gamma^D}(\Omega)]_{r/t}, & r \leq |t|, \\ H^{|t|}_{0,\Gamma^D}(\Omega) \cap H^r(\Omega), & r \geq |t|, \end{cases}$$

and $\mathcal{H}^{-r} = (\mathcal{H}^r)'$.

The result given in (4.4) is not completely satisfactory. Indeed, in any case when the primal wavelets are sufficiently smooth, we learned in section 4.2 that if the solution u is in $B^{sn+t}_\tau(L_\tau(\Omega))$ for some $s \in (0, \frac{d-t}{n})$, then it has a representation $u = \mathbf{u}^T \Psi$ such that the best N-term approximation for $\mathbf{u}$ converges with a rate N^{-s}. On the other hand, the convergence rate of the solutions yielded by (mod)SOLVE is bounded not only by the above value of s but also by s^*. Since s^* given in (4.4) is less than or equal to $\frac{\gamma - t}{n} - \frac{1}{2}$, and since, moreover, $\gamma < d$, on the basis of this result we may conclude only that (mod)SOLVE has optimal computational complexity for solutions *with limited regularity.*

However, in a forthcoming paper [25] it will be shown that (4.4) is actually too pessimistic and that, when $\sigma > d - t$, with suitable wavelets for local as well as for nonlocal operators, s^*-compressibility with $s^* > \frac{d-t}{n}$ can be shown.

Since the use of the projector $\mathbf{P}$ applied in modSOLVE seems restricted to the frame construction from this paper, we discuss its compressibility here. Recall that $\mathbf{P}$ is given by an $m \times m$ block matrix with its $(i, \breve{\imath})$th block being equal to the infinite matrix $\mathbf{P}^{(i,\breve{\imath})} = \langle \hat{\chi}_i \omega_i \tilde{\Psi}^{(i)}, \omega_{\breve{\imath}} \Psi^{(\breve{\imath})} \rangle_{L_2(\Omega)}$. Thus it is sufficient to investigate the compressibility of any of these blocks.

For biorthogonal wavelet bases it can be shown that for $r \in [-\tilde{d}, \gamma)$, $s < \gamma$,

$$\| \cdot \|_{\mathcal{H}^r_{\breve{i}}} \lesssim 2^{\breve{\ell}(r-s)} \| \cdot \|_{\mathcal{H}^s_{\breve{i}}} \quad \text{on} \quad W^{(\breve{i})}_{\breve{\ell}} := \operatorname{span}\{\psi^{(\breve{i})}_{\breve{\lambda}} : |\breve{\lambda}| = \breve{\ell}\}, \tag{4.5}$$

and analogously that for $r \in [-d, \tilde{\gamma})$, $s < \tilde{\gamma}$,

$$\| \cdot \|_{\mathcal{H}^r_{i}} \lesssim 2^{\ell(r-s)} \| \cdot \|_{\mathcal{H}^s_{i}} \quad \text{on} \quad \tilde{W}^{(i)}_{\ell} = \operatorname{span}\{\tilde{\psi}^{(i)}_{\lambda} : |\lambda| = \ell\}. \tag{4.6}$$

Here $\mathcal{H}^r_{\breve{i}}$ is defined as $\mathcal{H}^r$ with (Ω, Γ^D) replaced by $(\Omega_{\breve{i}}, \Gamma^D_{\breve{i}})$.

Thus, assuming that the weights vanish at the internal boundaries, for $\tilde{w}^{(i)}_{\ell} \in \tilde{W}^{(i)}_{\ell}$, $w^{(\breve{i})}_{\breve{\ell}} \in W^{(\breve{i})}_{\breve{\ell}}$, and $-d \leq s - t < \tilde{\gamma}$, $-\tilde{d} \leq t - s < \gamma$, i.e., $s \in (t - \gamma, t + \tilde{\gamma})$, we have

$$\begin{aligned} |\langle \hat{\chi}_i \omega_i^{-1} \tilde{w}^{(i)}_{\ell}, \omega_{\breve{i}} w^{(\breve{i})}_{\breve{\ell}} \rangle_{L_2(\Omega)}| &\lesssim \|\hat{\chi}_i \omega_i^{-1} \tilde{w}^{(i)}_{\ell}\|_{\mathcal{H}^{s-t}} \|\omega_{\breve{i}} w^{(\breve{i})}_{\breve{\ell}}\|_{\mathcal{H}^{t-s}} \\ &\lesssim \|\tilde{w}^{(i)}_{\ell}\|_{\mathcal{H}^{s-t}_i} \|w^{(\breve{i})}_{\breve{\ell}}\|_{\mathcal{H}^{t-s}_{\breve{i}}} \lesssim 2^{s(\ell - \breve{\ell})} \|\tilde{w}^{(i)}_{\ell}\|_{\mathcal{H}^{-t}_i} \|w^{(\breve{i})}_{\breve{\ell}}\|_{\mathcal{H}^t_{\breve{i}}}. \end{aligned}$$

Let us now define $\hat{\mathbf{P}}^{(i,\breve{i})}_j$ by removing from $\mathbf{P}^{(i,\breve{i})}$ all blocks $[\langle \hat{\chi}_i \omega_i \tilde{\psi}^{(i)}_{\lambda}, \omega_{\breve{i}} \psi^{(\breve{i})}_{\breve{\lambda}} \rangle_{L_2(\Omega)}]_{|\lambda|=\ell, |\breve{\lambda}|=\breve{\ell}}$ for which $\ell > \breve{\ell} + k_1(j,n)$ or $\breve{\ell} > \ell + \frac{\gamma - t}{\tilde{\gamma} + t} k_1(j,n)$, where $k_1(j,n)$ is an integer that will be determined below. Then, using the fact that the $\tilde{\Psi}^{(i)}$ or $\Psi^{(\breve{i})}$ are Riesz bases for $\mathcal{H}^{-t}_i$ or $\mathcal{H}^t_{\breve{i}}$, respectively, we infer that for any $0 < s < \gamma - t$

$$\|\mathbf{P}^{(i,\breve{i})} - \hat{\mathbf{P}}^{(i,\breve{i})}_j\| \lesssim 2^{-s k_1(j,n)}. \tag{4.7}$$

For the next step, we will assume also that *the dual wavelets are local.* Furthermore, we assume that the primal wavelets are *piecewise smooth.* By this we mean that that $\operatorname{supp} \psi^{(\breve{i})}_{\breve{\lambda}} \backslash \operatorname{sing\,supp} \psi^{(\breve{i})}_{\breve{\lambda}}$ is the disjoint union of m open "uniformly Lipschitz" domains $\Xi^{(\breve{i},1)}_{\breve{\lambda}}, \ldots, \Xi^{(\breve{i},k)}_{\breve{\lambda}}$, with $\cup_{q=1}^{k} \overline{\Xi^{(\breve{i},q)}_{\breve{\lambda}}} = \operatorname{supp} \psi^{(\breve{i})}_{\breve{\lambda}}$, and that $\psi^{(\breve{i})}_{\breve{\lambda}}|_{\Xi^{(\breve{i},q)}_{\breve{\lambda}}}$ is smooth with

$$\sup_{x \in \Xi^{(\breve{i},q)}_{\breve{\lambda}}} |\partial^{\beta} \psi^{(\breve{i})}_{\breve{\lambda}}(x)| \lesssim 2^{(|\beta| + \frac{n}{2} - t)|\lambda|}, \quad \beta \in \mathbb{N}^n.$$

From [23] we learn that $\psi^{(\breve{i})}_{\breve{\lambda}}|_{\Xi^{(\breve{i},q)}_{\breve{\lambda}}}$ has an extension to a smooth function $\xi^{(\breve{i},q)}_{\breve{\lambda}}$ with

$$\|\xi^{(\breve{i},q)}_{\breve{\lambda}}\|_{H^d(\mathbb{R}^n)} \lesssim 2^{|\breve{\lambda}|(d-t)} \tag{4.8}$$

(cf. also [25, Remark 4.5]).

Given $\breve{\lambda}$, $1 \leq q \leq k$, and $\ell > |\breve{\lambda}|$, let $A^{(i,\breve{i})}_{\ell,\breve{\lambda},q} = \{|\lambda| = \ell : \operatorname{supp} \tilde{\psi}^{(i)}_{\lambda} \subset \overline{\Xi^{(\breve{i},q)}_{\breve{\lambda}}}\}$. For some $k_2(j,n) \leq k_1(j,n)$ that will be determined below, we define $\mathbf{P}^{(i,\breve{i})}_j$ by removing all entries $\langle \hat{\chi}_i \omega_i \tilde{\psi}^{(i)}_{\lambda}, \omega_{\breve{i}} \psi^{\breve{i}}_{\breve{\lambda}} \rangle_{L_2(\Omega)}$ from $\hat{\mathbf{P}}^{(i,\breve{i})}_j$ when $|\lambda| - |\breve{\lambda}| > k_2(j,n)$ and $\lambda \in A^{(i,\breve{i})}_{|\lambda|,\breve{\lambda},q}$ for some $1 \leq q \leq k$. Then, by using (4.8) and (4.6) with $(r,s) = (-d,-t)$, for any

$\mathbf{c}, \mathbf{d} \in \ell_2$ we have

$$\begin{aligned}
\langle \mathbf{c}, (\hat{\mathbf{P}}_j^{(i,\breve{\imath})} - \mathbf{P}_j^{(i,\breve{\imath})})\mathbf{d}\rangle_{\ell_2} &= \left| \sum_{\ell - \breve{\ell} > k_2(j,n)} \sum_{|\breve{\lambda}| = \breve{\ell}} \mathbf{d}_{\breve{\lambda}} \left\langle \sum_{q=1}^{k} \sum_{\lambda \in A^{(i,\breve{\imath})}_{\ell,\breve{\lambda},q}} \mathbf{c}_\lambda \hat{\chi}_i \omega_i^{-1} \tilde{\psi}_\lambda^{(i)}, \omega_{\breve{\imath}} \xi_{\breve{\lambda}}^{(\breve{\imath},q)} \right\rangle \right| \\
&\lesssim \sum_{\ell - \breve{\ell} > k_2(j,n)} 2^{-(d-t)(\ell - \breve{\ell})} \sum_{|\breve{\lambda}| = \breve{\ell}} |\mathbf{d}_{\breve{\lambda}}| \sum_{q=1}^{k} \left\| \sum_{\lambda \in A^{(i,\breve{\imath})}_{\ell,\breve{\lambda},q}} \mathbf{c}_\lambda \tilde{\psi}_\lambda^{(i)} \right\|_{\mathcal{H}_i^{-t}} \\
&\lesssim \sum_{\ell - \breve{\ell} > k_2(j,n)} 2^{-(d-t)(\ell - \breve{\ell})} \sqrt{\sum_{|\breve{\lambda}| = \breve{\ell}} |\mathbf{d}_\lambda|^2} \sqrt{\sum_{|\breve{\lambda}| = \breve{\ell}} \left(\sum_{q=1}^{k} \sqrt{\sum_{\lambda \in A^{(i,\breve{\imath})}_{\ell,\breve{\lambda},q}} |\mathbf{c}_\lambda|^2} \right)^2} \\
&\lesssim \sum_{\ell - \breve{\ell} > k_2(j,n)} 2^{-(d-t)(\ell - \breve{\ell})} \sqrt{\sum_{|\breve{\lambda}| = \breve{\ell}} |\mathbf{d}_\lambda|^2} \sqrt{\sum_{|\lambda| = \ell} |\mathbf{c}_\lambda|^2} \lesssim 2^{-(d-t)k_2(j,n)} \|\mathbf{d}\| \|\mathbf{c}\|,
\end{aligned}$$

where for the last line we have used that, by the locality of the primal wavelets, each λ is contained in at most a uniformly bounded number of sets $A^{(i,\breve{\imath})}_{|\lambda|,\breve{\lambda},q}$. We conclude that

$$\|\hat{\mathbf{P}}_j^{(i,\breve{\imath})} - \mathbf{P}_j^{(i,\breve{\imath})}\| \lesssim 2^{(d-t)k_2(j,n)}. \tag{4.9}$$

By the locality of both primal and dual wavelets and the piecewise smoothness of the primal wavelets, the number of nonzeros in each *column* of $\mathbf{P}_j^{(i,\breve{\imath})}$ is of the order $2^{nk_2(j,n)} + 2^{(n-1)k_1(j,n)}$. By substituting $k_2(j,n) = \frac{j + \log(\alpha_j)}{n}$ and $k_1(j,n) = \frac{j + \log(\alpha_j)}{\max\{n-1,1\}}$ with, for example, $\alpha_j = j^{-(1+\varepsilon)}$ for some $\varepsilon > 0$, from (4.7) and (4.9) we infer that $\mathbf{P}$ is s^*-compressible with

$$s^* = \begin{cases} \min\{\frac{\gamma - t}{n-1}, \frac{d-t}{n}\} & \text{when } n > 1, \\ d - t & \text{when } n = 1. \end{cases}$$

Thus, if, when $n > 1$, $r \geq -\frac{3}{2} + d + (t-d)/n$, which was also assumed in (4.3), then $s^* = \frac{d-t}{n}$.

Acknowledgments. The author would like to thank Prof. Karsten Urban (Universität Ulm) for many inspiring discussions and for his valuable comments on an earlier version of this manuscript. Thanks also to Prof. Hans Duistermaat (Universiteit Utrecht) for suggestions that lead the author to consider frames in the first place.

REFERENCES

[1] A. Barinka, *Fast Evaluation Tools for Adaptive Wavelet Schemes*, Ph.D. thesis, Rheinisch-Westfälische Technische Hochschule, Aachen, Germany, 2003.

[2] J. Bergh and J. Löfström, *Interpolation Spaces*, Grundlehren Math. Wiss. 223, Springer-Verlag, Berlin, 1976.

[3] C. Canuto, A. Tabacco, and K. Urban, *The wavelet element method part* I: *Construction and analysis*, Appl. Comput. Harmon. Anal., 6 (1999), pp. 1–52.

[4] A. Cohen, *Wavelet methods in numerical analysis*, in Handbook of Numerical Analysis, Vol. 7, P. Ciarlet and J. L. Lions, eds., North–Holland, Amsterdam, 2000, pp. 417–711.

[5] A. Cohen, W. Dahmen, and R. DeVore, *Adaptive wavelet methods for elliptic operator equations—Convergence rates*, Math. Comp., 70 (2001), pp. 27–75.

[6] A. Cohen, W. Dahmen, and R. DeVore, *Adaptive wavelet methods* II*—Beyond the elliptic case*, Found. Comput. Math., 2 (2002), pp. 203–245.

[7] A. Cohen, L. Echeverry, and Q. Sun, *Finite Element Wavelets*, Technical report, Laboratoire d'Analyse Numérique, Université Pierre et Marie Curie, Paris, 2000.

[8] A. Cohen and R. Masson, *Wavelet adaptive method for second order elliptic problems: Boundary conditions and domain decomposition*, Numer. Math., 86 (2000), pp. 193–238.

[9] S. Dahlke, *Besov regularity for elliptic boundary value problems on polygonal domains*, Appl. Math. Lett., 12 (1999), pp. 31–36.

[10] S. Dahlke, *Besov regularity for the Stokes problem*, in Advances in Multivariate Approximation, W. Haussmann, K. Jetter, and M. Reimer, eds., Math. Res. 107, Wiley-VCH, Berlin, 1999, pp. 129–138.

[11] S. Dahlke, W. Dahmen, R. Hochmuth, and R. Schneider, *Stable multiscale bases and local error estimation for elliptic problems*, Appl. Numer. Math., 23 (1997), pp. 21–48.

[12] S. Dahlke, W. Dahmen, and K. Urban, *Adaptive wavelet methods for saddle point problems—Optimal convergence rates*, SIAM J. Numer. Anal., 40 (2002), pp. 1230–1262.

[13] S. Dahlke and R. DeVore, *Besov regularity for elliptic boundary value problems*, Comm. Partial Differential Equations, 22 (1997), pp. 1–16.

[14] W. Dahmen, *Wavelet and multiscale methods for operator equations*, Acta Numer., 6 (1997), pp. 55–228.

[15] W. Dahmen, A. Kunoth, and K. Urban, *Biorthogonal spline-wavelets on the interval—Stability and moment conditions*, Appl. Comput. Harmon. Anal., 6 (1999), pp. 132–196.

[16] W. Dahmen and R. Schneider, *Composite wavelet bases for operator equations*, Math. Comp., 68 (1999), pp. 1533–1567.

[17] W. Dahmen and R. Schneider, *Wavelets on manifolds* I: *Construction and domain decomposition*, SIAM J. Math. Anal., 31 (1999), pp. 184–230.

[18] W. Dahmen and R. Stevenson, *Element-by-element construction of wavelets satisfying stability and moment conditions*, SIAM J. Numer. Anal., 37 (1999), pp. 319–352.

[19] W. Dahmen, K. Urban, and J. Vorloeper, *Adaptive wavelet methods—Basic concepts and applications to the Stokes problem*, in Wavelet Analysis, D.-X. Zhou, ed., World Scientific, River Edge, NJ, 2002.

[20] I. Daubechies, *Ten Lectures on Wavelets*, CBMS-NSF Reg. Conf. Ser. in Appl. Math. 6, SIAM, Philadelphia, 1992.

[21] R. DeVore, *Nonlinear approximation*, Acta Numer., 7 (1998), pp. 51–150.

[22] A. Metselaar, *Handling Wavelet Expansions in Numerical Methods*, Ph.D. thesis, University of Twente, Enschede, The Netherlands, 2002.

[23] E. Stein, *Singular Integrals and Differentiability Properties of Functions*, Princeton University Press, Princeton, NJ, 1970.

[24] R. Stevenson, *Locally supported, piecewise polynomial biorthogonal wavelets on non-uniform meshes*, Constr. Approx., to appear.

[25] R. Stevenson, *On the Compressibility of Operators in Wavelet Coordinates*, Technical report 1249, University of Utrecht, Utrecht, The Netherlands, 2002.

SIAM J. Numer. Anal.
Vol. 41, No. 3, pp. 1101–1130

A SMALL EDDY CORRECTION METHOD FOR NONLINEAR DISSIPATIVE EVOLUTIONARY EQUATIONS[*]

YANREN HOU[†] AND KAITAI LI[†]

Abstract. Considering the interaction between the large and small eddy components of solutions and using the idea of the Newton iteration, a small eddy correction method is proposed for approximating and numerically solving nonlinear dissipative PDEs of parabolic type, in particular the Navier–Stokes equations (NSE). We assume that the large eddy approximation to the solution is known. Formally applying the Newton iterative procedure to the small eddy equation, we then generate approximate systems. It is shown that the first two steps in fact lead to the standard Galerkin method (SGM) and the so-called optimum nonlinear Galerkin method (ONG), and therefore the small eddy correction method is actually a certain generalization of SGM and ONG. The boundedness and convergence analysis are presented in the framework of the two-dimensional NSE. The results show that the small eddy correction method can greatly improve the accuracy of SGM approximate solutions.

Key words. dissipative equations, spectral methods, Newton iteration, multilevel method

AMS subject classifications. 65N30, 76D05

DOI. 10.1137/S0036142901396375

1. Introduction.

Considering an evolutionary dissipative nonlinear PDE system of parabolic type, for example, the Navier–Stokes equation (NSE), despite the considerable increase in the available computing power during the past few years, numerically solving this kind of system, especially the integration of evolutionary NSE on large time intervals under physically realistic situations, still remains a difficult problem whose solution is not close at hand. We thereby intend to solve dissipative evolution PDEs in dynamically nontrivial situations, i.e., when the long-term behavior is not merely convergent to a steady state. In this case, the solution to be simulated remains time dependent, and as time goes to infinity, it converges to a set, an attractor, which can be a complicated set (a fractal). Studying the complicated structure of this set, which to some extent is reflected by the long-term behavior of the solution, is of great importance to understanding the nature of turbulent phenomena. That is one of the main reasons why people are so interested in the long-term behavior of the solution and the construction of more accurate and effective numerical schemes.

We present this work in the context of the functional form of the two-dimensional NSE in divergence-free Hilbert space H defined on an open bounded domain $\Omega \subset \mathcal{R}^2$ with smooth boundary $\partial\Omega$,

$$\begin{cases} \dfrac{du}{dt} + \nu A u + B(u,u) = f, \\ u(0) = a. \end{cases} \tag{1.1}$$

[*] Received by the editors October 11, 2001; accepted for publication (in revised form) February 18, 2003; published electronically July 30, 2003. This work was subsidized by the Special Funds for Major State Basic Research Projects G1999032801-01-5, NSFC (10101020, 10001028), and Tian Yuan Mathematical Funds (TY10126004).

http://www.siam.org/journals/sinum/41-3/39637.html

[†] College of Science, Xi'an Jiaotong University, Xi'an 710049, China (yrhou@mail.xjtu.edu.cn, ktli@mail.xjtu.edu.cn).

Here u stands for the velocity field, f is the external force which drives the flow, $\nu > 0$ is the viscosity, and a is the initial velocity field. A and $B(\cdot,\cdot)$ are the Stokes operator and bilinear operator whose detailed definitions will be given in section 2.

For a given positive integer $m \in \mathcal{N}$, let us denote by H_m and P_m a finite-dimensional subspace of H and an associated orthogonal projection from H onto H_m, that is, $H_m = P_m H$. Then the SGM approximate system of (1.1) reads

$$\left\{\begin{array}{l} \dfrac{du_m}{dt} + \nu A u_m + P_m B(u_m, u_m) = P_m f, \\ u_m(0) = P_m a. \end{array}\right. \tag{1.2}$$

It follows from [6] that, for sufficiently large $t_0 > 0$,

$$|u(t) - u_m(t)|_{\mathcal{L}^2(\Omega)} \le c(t) L_m \lambda_{m+1}^{-1} \quad \forall t \ge t_0, \tag{1.3}$$

where $\lambda_{m+1} > 0$ is the $(m+1)$th eigenvalue of the Stokes operator A, which tends to $+\infty$ when m tends to infinity and

$$L_m \sim \left(1 + \log \frac{\lambda_m}{\lambda_1}\right)^{\frac{1}{2}}. \tag{1.4}$$

On the other hand, if we denote $Q_m = I - P_m$ and

$$u = p + q \quad \text{with} \quad p = P_m u,\ q = Q_m u,$$

we can rewrite (1.1) in the following coupled system:

$$\frac{dp}{dt} + \nu A p + P_m B(p+q, p+q) = P_m f, \tag{1.5}$$

$$\frac{dq}{dt} + \nu A q + Q_m B(p+q, p+q) = Q_m f. \tag{1.6}$$

Set

$$e = u - u_m, \quad e_p = P_m e = p - u_m, \quad e_q = Q_m e = q.$$

Simple calculation shows that

$$|e_p(t)|_{\mathcal{L}^2(\Omega)} \le c e^{ct} \sup_{0 \le s \le t} |e_q(s)|_{\mathcal{L}^2(\Omega)}. \tag{1.7}$$

That is to say that the large eddy error and thus the total error of the standard Galerkin method (SGM) can be controlled by the small eddy error. In other words, to improve the accuracy of the SGM approximation, we have only to improve the approximation of the small eddy components, which is approximated by 0 in the case of SGM. This is the basic idea of this work.

Indeed, many authors have already applied this idea to developing new methods and techniques. One of these new methods arises in connection with the concept of the approximate inertial manifold (AIM) (see [7]) and is called the nonlinear Galerkin method (NGM). For example, Marion and Temam proposed the first and the most frequently discussed NGM in [15]:

$$\left\{\begin{array}{l} \dfrac{dp_m}{dt} + \nu A p_m + P_m B(p_m + q_m, p_m + q_m) = P_m f, \\ q_m = \Phi(p_m), \end{array}\right. \tag{1.8}$$

where $\Phi : H_m \to Q_m H$ provides an approach to approximating the small eddy components, which are determined by solving the following steady Stokes problem in $Q_m H$:

$$\nu A q_m + Q_m B(p_m, p_m) = Q_m f.$$

It is shown that

$$|u(t) - (p_m(t) + \Phi(p_m(t)))|_{\mathcal{L}^2(\Omega)} \leq c(t) L_m \lambda_{m+1}^{-\frac{3}{2}} \quad \forall t \geq t_0. \tag{1.9}$$

That is, $p_m + \Phi(p_m)$ is a much better approximation to u than the SGM approximation u_m. From that point on, many authors investigated this version of the NGM (see [1], [5], [6], [14], [16], [17], [18], etc.) and its applications (see [3], [19], etc.). To improve the effectiveness of numerical schemes, Garcia-Archilla, Novo, and Titi proposed a kind of postprocessing procedure to the Galerkin method (PPGM) in [8] based on the same mapping Φ in (1.8). For any prescribed time T, they use $\Phi(u_m(T))$ to approximate $q(T)$ and show that

$$|u(T) - (u_m(T) + \Phi(u_m(T)))|_{\mathcal{L}^2(\Omega)} \leq c(T) L_m^4 \lambda_{m+1}^{-\frac{3}{2}} \quad \forall T \geq t_0. \tag{1.10}$$

Equations (1.9) and (1.10) indicate that both the NGM (1.8) and PPGM can greatly improve the accuracy of the SGM approximation and be more effective than the SGM. But there are some problems which have already been discussed in [12]. For example, they used a steady Stokes problem to approximate the small eddy equation (1.6) and overlooked the self evolution of the small eddy components. This is only valid when the small eddy part of the solution varies very slowly.

To overcome this defect of NGM (1.8), some people have already given certain modifications (see [9], [10], [11], [13], [20]). One of these modifications is the so-called optimal nonlinear Galerkin method (ONG) (see [10], [11]):

$$\frac{d\bar{p}_m}{dt} + \nu A \bar{p}_m + P_m B(\bar{p}_m + \bar{q}_m, \bar{p}_m + \bar{q}_m) = P_m f, \tag{1.11}$$

$$\frac{d\bar{q}_m}{dt} + \nu A \bar{q}_m + Q_m B(\bar{p}_m, \bar{q}_m) + Q_m B(\bar{q}_m, \bar{p}_m) = Q_m[f - B(\bar{p}_m, \bar{p}_m)], \tag{1.12}$$

where the small eddy equation in (1.8) is replaced by a generalized unsteady Stokes problem (1.12), and therefore the self evolution of the small eddy components is involved. Furthermore, it is shown in this paper that

$$|u(t) - (\bar{p}_m + \bar{q}_m)|_{\mathcal{L}^2(\Omega)} \leq c(t) L_m^4 \lambda_{m+1}^{-2} \quad \forall t \geq t_0. \tag{1.13}$$

To get a reliable long-term simulation of the NSE, we still have to consider the problem of error accumulation. Unfortunately, almost all numerical methods for the NSE in general (including the SGM, NGM, PPGM, and ONG) will lead to an exponentially increasing error, which makes the long-term simulation almost meaningless. A possible remedy is to improve the accuracy of the approximate solution without too much computational cost. Meanwhile, we have to mention that the rapid increase of available computing power in the past few years as well as the rapid development of computer networks and parallel computing techniques, which can integrate many CPUs in a system, make the large scale computation possible. If computing facilities are not a problem, we want to know whether it is possible to get a reliable simulation in a large time interval.

From the point of view of our basic idea mentioned above (the small eddy correction), the NGM, PPGM, and ONG schemes all do a small eddy correction once, and their analysis shows that they can improve the convergence rate of the SGM without too much computational cost. It seems that one cannot improve the convergence rate any more if the correction is performed only once. Then it is very natural for us to think about certain iteration methods for improving the convergence rate further. And it is also very natural for us to think about the Newton iteration because of its fast convergence speed. Based upon these considerations, we propose a small eddy correction method in this paper (see the later definition (3.1)–(3.3)). Moreover, the analysis shows that

$$|u(t) - u^l(t)|_{\mathcal{L}^2(\Omega)} \leq c(t) L_m^2 (L_m \lambda_{m+1}^{-1})^{2^l}, \tag{1.14}$$

where u^l is the small eddy correction approximation and $l \in \mathcal{N}$ is the number of iteration steps used.

This paper is arranged as follows. In section 2, the detail of a two-dimensional NSE is given. In section 3, we introduce a small eddy correction method by formally applying the Newton iteration to the small eddy equation (1.6). Then we prove that this procedure is a bounded procedure and can generate a bounded approximate solution of (1.1). In section 4, we present some error estimates of the proposed small eddy correction method and show that it can lead to a very accurate approximation of the solution of the NSE. Finally in section 5, for a one-dimensional Burger's equation and a two-dimensional NSE, we give a full discrete form of the small eddy correction scheme and then present some numerical results to check the high accuracy and high effectiveness of this method.

2. Navier–Stokes equations.

We consider the NSE on a smooth bounded domain $\Omega \in \mathcal{R}^2$:

$$\begin{cases} \dfrac{\partial u}{\partial t} - \nu \Delta u + (u \cdot \nabla) u + \nabla p = F, & (x,t) \in \Omega \times (0, +\infty), \\ \nabla \cdot u = 0, & (x,t) \in \Omega \times [0, +\infty), \\ u|_{t=0} = a(x), & x \in \Omega, \\ \text{Dirichlet or periodic boundary conditions.} \end{cases} \tag{2.1}$$

Here $u(x,t)$ and $p(x,t)$ stand for the velocity field and pressure, respectively; $F(x,t)$ the external force, which drives the flow; $a(x)$ the initial velocity, which satisfies $\nabla \cdot a = 0$; and $\nu > 0$ the kinetic viscosity.

Now, we define the Hilbert space H:

$$H = \{v \in (L^2(\Omega))^2 : \nabla \cdot v = 0,\ v \cdot n|_{\partial\Omega} = 0\}$$

in the case of the homogeneous Dirichlet boundary condition, where n denotes the unit outward normal vector to $\partial\Omega$, or

$$H = \left\{ v \in (L^2_{per}(\Omega))^2 : \nabla \cdot v = 0,\ \int_\Omega v dx = 0 \right\}$$

in the case of a periodic boundary condition. The space H is equipped with the usual L^2-inner product $(\cdot,\cdot)$ and norm $|\cdot| = (\cdot,\cdot)^{\frac{1}{2}}$ and is a closed subspace of $(L^2(\Omega))^2$.

We also define a Hilbert space

$$V = \{v \in (H_0^1(\Omega))^2 : \ \nabla \cdot v = 0\}$$

or

$$V = \left\{ v \in (H_{per}^1(\Omega))^2 : \ \nabla \cdot v = 0, \int_\Omega v dx = 0 \right\}$$

depending on the use of boundary conditions. Equipped with the usual H^1-inner product and norm, it is a Hilbert space. Let us denote by P the orthogonal projection from $(L^2(\Omega))^2$ onto H and project (2.1) onto H. Then we can get the functional form (1.1), where $A = -P\Delta$ is the Stokes operator which will be $-\Delta$ in the case of a periodical boundary condition, $B(u,u) = P[(u \cdot \nabla)u]$ and $f = PF$, which is assumed to be time independent or in $L^\infty(\mathcal{R}^+, H)$.

For the Stokes operator A, it is well known that it is a symmetric, positive definite, self-adjoint, and unbounded operator in H with compact inverse. Therefore, its eigenvalues and the associated eigenfunctions admit

$$A\phi_i = \lambda_i \phi_i, \quad 0 < \lambda_1 \le \lambda_2 \le \cdots \to +\infty,$$

and the set of eigenfunctions $\{\phi_1, \phi_2, \ldots\}$ forms a unit orthonormal basis of H. For any $\alpha \in \mathcal{R}$, the space

$$D(A^\alpha) = \left\{ v = \sum_{i=1}^\infty v_i \phi_i : \sum_{i=1}^\infty v_i^2 \lambda_i^{2\alpha} < +\infty \right\}$$

is a Hilbert space if it is equipped with the natural inner product and norm

$$(\cdot,\cdot)_{D(A^\alpha)} = (A^\alpha \cdot, A^\alpha \cdot), \qquad |\cdot|_{D(A^\alpha)} = |A^\alpha \cdot|.$$

It is known that $D(A^{\frac{1}{2}}) = V$, and $\{\phi_1, \phi_2, \ldots\}$ is also an orthonormal basis of $D(A^\alpha)$. Here and later on, we denote by $|\cdot|_s$ the $(H^s(\Omega))^2$ norm, and by $|\cdot|_\infty$ the $(L^\infty(\Omega))^2$ norm. Especially, we use $|\cdot|$ to denote $|\cdot|_0$. For the spatial periodic case, we know that $|A^\alpha \cdot|$ and $|\cdot|_{2\alpha}$ are equivalent norms for any $\alpha \in \mathcal{R}$, and this equivalence property holds for the nonslip case at least for $\alpha \le 1$.

To investigate the interaction between the large and small eddies, we define the finite-dimensional subspace H_m as in the Introduction for given $m \in \mathcal{N}$ as

$$H_m = \text{span}\{\phi_1, \phi_2, \ldots, \phi_m\},$$

and define the orthogonal projection P_m from H onto H_m as

$$u = \sum_{i=1}^\infty u_i \phi_i \in H, \quad P_m u = \sum_{i=1}^m u_i \phi_i, \quad u_i \in \mathcal{R}.$$

Also, we define $Q_m = I - P_m$. Projecting (1.1) by P_m and Q_m, we can get the coupled system (1.5)–(1.6).

For the later analysis, we recall the Agmon inequality, the Brézis–Gallouet inequality [2], and the Sobolev interpolation inequality in two dimensions: there exists a constant $c_0 > 0$ such that

$$(2.2)\quad \begin{cases} |u|_\infty \le c_0 |u|^{\frac{1}{2}}\, |Au|^{\frac{1}{2}} \ \text{ or } \ c_0 |A^{\frac{1}{4}} u|^{\frac{2}{3}} |Au|^{\frac{1}{3}} & \forall u \in D(A), \\ |u|_\infty \le c_0 |A^{\frac{1}{2}} u| \left(1 + \log \frac{|Au|}{\lambda_1^{\frac{1}{2}} |A^{\frac{1}{2}} u|}\right)^{\frac{1}{2}} & \forall u \in D(A), \\ |A^s u| \le c_0 |u|^{(m-s)/m} |A^m u|^{s/m} & \forall u \in D(A^m),\ 0 < s < m \le 1. \end{cases}$$

As a result of the Brézis–Gallouet inequality, we have

$$(2.3)\qquad |u|_\infty \le c_0 L_m |A^{\frac{1}{2}} u| \quad \forall u \in H_m,$$

where L_m is defined in (1.4). To avoid having too many symbols, we always regard c_0 as 1 in the rest of this paper, and this will not cause any significant difference.

For the projections P_m and Q_m, the following properties hold (see [4]):

$$(2.4)\quad |A^\beta P_m u| \le \lambda_m^{\beta-\mu} |A^\mu u|, \quad |A^\mu Q_m u| \le \lambda_{m+1}^{\mu-\beta} |A^\beta u| \quad \forall \mu < \beta,\ u \in D(A^\beta).$$

In what follows, we often use the following orthogonal property:

$$(2.5)\qquad |A^s(P_m u + Q_m u)|^2 = |A^s P_m u|^2 + |A^s Q_m u|^2 \quad \forall u \in D(A^s), s \in \mathcal{R}.$$

3. Small eddy correction method and its boundedness. As shown in (1.7), the large eddy error or the total error of the SGM approximation can be controlled by the small eddy error, the truncation error of u_m. This suggests that we get a more accurate approximation by paying more attention to correcting the small eddy approximation successively. In this section, we will construct a small eddy correction method to NSE (1.1) by formally applying the Newton iteration to its small eddy equation (1.6).

First of all, we define

$$F(p,q) = \frac{dq}{dt} + \nu A q + Q_m B(q,q) + Q_m B(q,p) + Q_m B(p,q) + Q_m B(p,p) - Q_m f.$$

Then (1.6) is equivalent to

$$F(p,q) = 0.$$

Suppose that the large eddy component p in the above abstract equation is known. Formally applying the Newton iteration to it, we can get the following iterative procedure: supposing the initial guess of the small eddy component $q^0 = 0$ and the kth approximation $q^k(t) \in Q_m H$ is known for some $k \in \mathcal{N}$, find the $(k+1)$th approximation $q^{k+1}(t) \in Q_m H$, which should be a more accurate approximation of $q(t)$, such that

$$D_q F(p, q^k)(q^{k+1} - q^k) = -F(p, q^k).$$

Simple calculation shows that it is the following small eddy correction procedure:

$$\begin{aligned} \frac{dq^{k+1}}{dt} &+ \nu A q^{k+1} + Q_m B(p,p) + Q_m B(p, q^{k+1}) + Q_m B(q^{k+1}, p) \\ &+ Q_m B(q^k, q^{k+1}) + Q_m B(q^{k+1}, q^k) = Q_m[f + B(q^k, q^k)]. \end{aligned}$$

Combining the above small eddy correction with the large eddy equation (1.5), we can obtain the desired small eddy correction method: with $w^0 = 0$,

$$\frac{dv}{dt} + \nu Av + P_m B(v + w^l, v + w^l) = P_m f \quad \text{for any fixed integer } l \geq 0, \tag{3.1}$$

$$\begin{aligned}\frac{dw^k}{dt} &+ \nu Aw^k + Q_m B(v, v) + Q_m B(v, w^k) + Q_m B(w^k, v) + Q_m B(w^{k-1}, w^k) \\ &+ Q_m B(w^k, w^{k-1}) = Q_m[f + B(w^{k-1}, w^{k-1})] \quad \forall 1 \leq k \leq l,\end{aligned} \tag{3.2}$$

$$v(0) = P_m a, \quad w^i(0) = Q_m a, \quad i = 1, 2, \ldots, l, \quad l \geq 1. \tag{3.3}$$

Remark 1. The small eddy correction scheme (3.1)–(3.3) can be regarded as a certain generalization of the SGM and ONG. In fact, if we take $l = 0$ in the above approximate procedure, (3.1) admits the SGM approximate equation (1.2). And if we take $l = 1$, (3.1)–(3.3) are the ONG approximate equations (1.11)–(1.12). For the sake of simplicity of expression, we give a symbol $u^l = v + w^l$ and call it the lth approximate solution to the NSE (1.1). Thus the SGM approximate solution is the 0th approximate solution, and the ONG approximate solution is the 1st approximate solution. Observing the above small eddy correction method, it is analogous to the classical Newton iteration for elliptic problems. We expect that it can have the second order convergence rate just as the usual Newton method does. Fortunately, our later analysis gives us a positive answer. It is worth mentioning that directly applying the Newton method to NSE (1.1) cannot reach such a second order convergence rate. The reason is that it does not take the advantage of the fast decay property of small eddies. For example, we refer readers to our estimates (4.11), (4.15), etc. Here (4.11) and (4.15) show that the accuracy of the small eddy approximation is always a half order higher ($\lambda_{m+1}^{-\frac{1}{2}}$) than that of the large eddy approximation, and the small eddy approximation converges very quickly, which is very important for us to derive the high order convergence rate, while the direct application of Newton iteration to NSE does not distinguish the small and large eddy components and cannot take this advantage.

In the rest of this section, we will show that the small eddy correction method (3.1)–(3.3) is a bounded procedure and can generate a bounded approximate solution. Before that, we define a trilinear form

$$b(u, v, w) = ((u \cdot \nabla)v, w) \quad \forall u, v, w \in (H^1(\Omega))^2$$

and state some properties (e.g., see [21]):

$$b(u, v, w) = -b(u, w, v) \quad \forall v, w \in (H^1(\Omega))^2,\ u \in V, \tag{3.4}$$

$$2|b(u, v, w)| \leq c_1 |u|_{s_1} |A^{\frac{1}{2}} v|_{s_2} |w|_{s_3}, \tag{3.5}$$

where $u \in (H^{s_1}(\Omega))^2$, $v \in (H^{s_2+1}(\Omega))^2$, $w \in (H^{s_3}(\Omega))^2$, and $s_1, s_2, s_3 \geq 0$, $s_1 + s_2 + s_3 \geq 1$, $(s_1, s_2, s_3) \neq (1,0,0)$,$(0,1,0)$,$(0,0,1)$, and $c_1 > 0$ is a constant independent of u, v, w. Especially, (3.5) is valid if we substitute the L^2-norm for any two of the three norms on the right-hand side and replace the rest of the norm by L^∞-norms.

Hereafter, we use $|f|$ to denote $|f|_{\mathcal{L}(\mathcal{R}^+, H)}$ and set some dimensionless constants throughout this section:

Reynold's number $\quad Re = \nu^{-1}|a| \quad$ if $\ |a| \neq 0$,
Grashof's number $\quad Gr = \lambda_1 \nu^{-2} |f|$.

Theorem 3.1. *For any given nonnegative integer l, there exist constants M_0, M_1 independent of m, l, t and a constant $M_2(T)$ (for any given $T > 0$) independent of m and l such that, if m is so large that*

$$\lambda_{m+1} \geq 128 c_1^2 \lambda_1^{-1} \nu^{-2} L_m^2 M_1^2, \tag{3.6}$$

then (3.1)–(3.3) *has a bounded solution on* $[0, +\infty)$

$$U_m(t) = (v(t), w^1(t), \ldots, w^l(t))$$

and, for any $0 \leq k \leq l$ *and* $T > 0$,

$$|v(t) + w^k(t)| \leq M_0, \quad |A^{\frac{1}{2}}(v(t) + w^k(t))| \leq M_1 \quad \forall t \geq 0,$$

$$\int_0^T |A(v(s) + w^k(s))|^2 ds \leq M_2(T),$$

where

$$M_0 = \sqrt{2\nu^2 Re^2 + 4\nu^2 Gr^2}, \qquad M_1 = 4\lambda_1^{\frac{1}{2}} \nu (Re + 2Gr) \exp(c_1^4 (2Re^2 + 5Gr^2)^2),$$

and

$$M_2(T) = 2\nu^{-1} M_1^2 (6 + c_1^4 \nu^{-3} M_0^2 M_1^2 T) + 32 \lambda_1^2 \nu^2 T Gr^2.$$

The complex form of the constant M_1 comes from the fact that we do not distinguish the types of boundary conditions. Actually, for the spatial periodic case, the corresponding M_1 will have a much simpler appearance (without the exponential factor).

Compared with NGM, PPGM, and ONG, in which the boundedness of the approximate solution is quite easy to obtain and the existence of the approximate solution is a direct result, the boundedness and the existence of the approximate solution is no longer obvious in our case, especially for $l \geq 2$.

To prove this theorem, we first consider the following Galerkin approximation of (3.1)–(3.3): for any given positive integer $M > m$, constant $T > 0$, and $k = 1, 2, \ldots, l$ find

$$v_m = \sum_{i=1}^{m} g_{im}(t) \phi_i \in H_m, \qquad w_M^k = \sum_{i=m+1}^{M} g_{im}^k(t) \phi_i \in (P_M - P_m)H \overset{\triangle}{=} P_{mM} H$$

on time interval $[0, T]$ such that for any fixed integer $l \geq 0$ and $w_M^0 = 0$

$$\frac{dv_m}{dt} + \nu A v_m + P_m B(v_m + w_M^l, v_m + w_M^l) = P_m f, \tag{3.7}$$

$$\begin{aligned} \frac{dw_M^k}{dt} &+ \nu A w_M^k + P_{mM} B(v_m, v_m) + P_{mM} B(v_m, w_M^k) \\ &+ P_{mM} B(w_M^k, v_m) + P_{mM} B(w_M^{k-1}, w_M^k) + P_{mM} B(w_M^k, w_M^{k-1}) \\ &= P_{mM}[f + B(w_M^{k-1}, w_M^{k-1})] \quad \forall 1 \leq k \leq l, \end{aligned} \tag{3.8}$$

$$v_m(0) = P_m a, \quad w_M^i(0) = P_{mM} a, \quad i = 1, 2, \ldots, l, \quad l \geq 1. \tag{3.9}$$

For simplicity of expression, we introduce the following notation for the rest of this section:

$$u_{mM}^k = v_m + w_M^k, \quad k = 0, \dots, l.$$

Noticing the orthogonal properties of $\{\phi_1, \dots, \phi_m, \dots\}$ in H and V, that is,

$$(\phi_i, \phi_j) = \delta_{ij}, \quad (A^{\frac{1}{2}}\phi_i, A^{\frac{1}{2}}\phi_j) = (A\phi_i, \phi_j) = \lambda_i \delta_{ij} \quad \forall i, j = 1, \dots, m, \dots,$$

we find that the system (3.7)–(3.9) is actually a first order ODE of $\{g_{im}(t)\}_{i=1}^m$ and $\{g_{im}^k(t)\}_{i=m+1}^M$, $k = 1, \dots, l$:

$$\begin{aligned}
&\dot{g}_{im} + \nu\lambda_i g_{im} + \sum_{j,n=1}^m \alpha_{nji} g_{jm} g_{nm} + \sum_{j=1}^m \sum_{n=m+1}^M \beta_{nji} g_{jm} g_{nm}^l \\
&\qquad + \sum_{j,n=m+1}^M \alpha_{nji} g_{jm}^l g_{nm}^l = f_i, \quad i = 1, \dots, m, \\
&\dot{g}_{im}^k + \nu\lambda_i g_{im}^k + \sum_{j,n=1}^m \alpha_{nji} g_{jm} g_{nm} + \sum_{j=1}^m \sum_{n=m+1}^M \beta_{nji} g_{jm} g_{nm}^k + \sum_{j,n=m+1}^M \beta_{nji} g_{jm}^k g_{nm}^{k-1} \\
&\qquad - \sum_{j,n=m+1}^M \alpha_{nji} g_{jm}^{k-1} g_{nm}^{k-1} = f_i \quad \forall i = m+1, \dots, M, \\
&g_{im}(0) = (a, \phi_i), \quad g_{jm}^k(0) = (a, \phi_j), \quad i = 1, \dots, m, \ j = m+1, \dots, M.
\end{aligned}$$

Here

$$\begin{cases} \alpha_{nji} = b(\phi_n, \phi_j, \phi_i), \ \beta_{nji} = b(\phi_n, \phi_j, \phi_i) + b(\phi_j, \phi_n, \phi_i), \\ f_i = (f, \phi_i), \ g_{km}^0 = 0 \ \text{ for } k = m+1, \dots, M. \end{cases}$$

Thanks to the theory of ODEs, the above nonlinear ODE has a maximal continuous solution defined on some interval $[0, T_M)$. If $T_M < T$, then $|A^{\frac{1}{2}} u_{mM}^k(t)|$, for some $0 \le k \le l$, must tend to $+\infty$ as $t \to T_M$. But we will show that this does not happen. Therefore $T_M \ge T$. Furthermore, both $|u_{mM}^k(t)|$ and $|A^{\frac{1}{2}} u_{mM}^k|$ are bounded by some constants independent of m, M, T, and l. To prove this, we need the following result.

Lemma 3.1. *For any given $t_M > 0$, suppose*

$$|A^{\frac{1}{2}} u_{mM}^k(t)| \le M_1 \quad \forall 1 \le k \le l, \ t \in [0, t_M].$$

If the condition (3.6) *is valid, we have, for $k = 1, \dots, l$,*

$$|u_{mM}^k(t)| \le M_0 \quad \forall t \in [0, t_M],$$

$$\int_t^{t+r} |A^{\frac{1}{2}} u_{mM}^k(s)|^2 ds \le \frac{4M_0^2}{\nu} + \frac{2r|f|^2}{\lambda_1 \nu^2} \qquad \textit{for any } t \ge 0, r > 0 \textit{ with } t + r \le t_M,$$

where $M_1, M_0 > 0$ are defined in Theorem 3.1.

Proof. The summation of (3.7) and (3.8) implies

$$\begin{aligned}
&\frac{du_{mM}^k}{dt} + \nu A u_{mM}^k + P_M B(v_m, v_m) + P_m[B(v_m, w_M^l) + B(w_M^l, v_m) + B(w_M^l, w_M^l)] \\
&\qquad + P_{mM}[B(v_m, w_M^k) + B(w_M^k, v_m) + B(w_M^{k-1}, w_M^k) \\
&\qquad\qquad + B(w_M^k, w_M^{k-1}) - B(w_M^{k-1}, w_M^{k-1})] = P_M f. \qquad (3.10)
\end{aligned}$$

Multiplying (3.10) by $2u_{mM}^k$, integrating it on Ω, and using the property (3.4) of the trilinear form, we get

$$\begin{aligned}\frac{d|u_{mM}^k|^2}{dt} &+ 2\nu|A^{\frac{1}{2}}u_{mM}^k|^2 \le 2|b(v_m, v_m, w_M^k)| + 2|b(v_m, w_M^l, v_m)| + 2|b(w_M^l, w_M^l, v_m)| \\ &+ 2|b(w_M^k, v_m, w_M^k)| + 2|b(w_M^k, w_M^{k-1}, w_M^k)| + 2|b(w_M^{k-1}, w_M^{k-1}, w_M^k)| + 2|(f, u_{mM}^k)|.\end{aligned}$$

Due to (2.5) and (3.6), we know $|A^{\frac{1}{2}}v_m|, |A^{\frac{1}{2}}w_M^k| \le M_1$, and $c_1M_1L_m\lambda_{m+1}^{-\frac{1}{2}} \le \frac{\nu}{10}$. By using (2.2)–(2.4) and (3.4)–(3.5), we summarize the estimates of the seven terms on the right-hand side of the above inequality as follows:

$$\begin{aligned}2|b(v_m, v_m, w_M^k)| &\le c_1|v_m|_\infty|A^{\frac{1}{2}}v_m|\,|w_M^k| \le \frac{c_1M_1L_m}{\lambda_{m+1}^{\frac{1}{2}}}|A^{\frac{1}{2}}v_m|^2 \le \frac{\nu}{10}|A^{\frac{1}{2}}v_m|^2, \\ 2|b(v_m, w_M^l, v_m)| &= 2|b(v_m, v_m, w_M^l)| \le c_1|v_m|_\infty|A^{\frac{1}{2}}v_m|\,|w_M^l| \\ &\le \frac{c_1M_1L_m}{\lambda_{m+1}^{\frac{1}{2}}}|A^{\frac{1}{2}}v_m|^2 \le \frac{\nu}{10}|A^{\frac{1}{2}}v_m|^2, \\ 2|b(w_M^l, w_M^l, v_m)| &= 2|b(w_M^l, v_m, w_M^l)| \le c_1|A^{\frac{1}{4}}w_M^l|^2|A^{\frac{1}{2}}v_m| \\ &\le \frac{c_1}{\lambda_{m+1}^{\frac{1}{2}}}|A^{\frac{1}{2}}v_m|\,|A^{\frac{1}{2}}w_M^l|^2, \\ 2|b(w_M^k, v_m, w_M^k)| &\le c_1|A^{\frac{1}{4}}w_M^k|^2|A^{\frac{1}{2}}v_m| \le \frac{c_1M_1}{\lambda_{m+1}^{\frac{1}{2}}}|A^{\frac{1}{2}}w_M^k|^2 \le \frac{\nu}{10}|A^{\frac{1}{2}}w_M^k|^2, \\ 2|b(w_M^k, w_M^{k-1}, w_M^k)| &\le c_1|A^{\frac{1}{4}}w_M^k|^2|A^{\frac{1}{2}}w_M^{k-1}| \le \frac{c_1M_1}{\lambda_{m+1}^{\frac{1}{2}}}|A^{\frac{1}{2}}w_M^k|^2 \le \frac{\nu}{10}|A^{\frac{1}{2}}w_M^k|^2, \\ 2|b(w_M^{k-1}, w_M^{k-1}, w_M^k)| &= 2|b(w_M^{k-1}, w_M^k, w_M^{k-1})| \le \frac{c_1}{\lambda_{m+1}^{\frac{1}{2}}}|A^{\frac{1}{2}}w_M^k|\,|A^{\frac{1}{2}}w_M^{k-1}|^2, \\ 2|(f, u_{mM}^k)| &\le 2|A^{-\frac{1}{2}}P_Mf|\,|A^{\frac{1}{2}}u_{mM}^k| \le \nu|A^{\frac{1}{2}}u_{mM}^k|^2 + \frac{|f|^2}{\lambda_1\nu}.\end{aligned}$$

Finally, we see that

$$\begin{aligned}\frac{d|u_{mM}^k|^2}{dt} &+ \frac{4\nu}{5}|A^{\frac{1}{2}}u_{mM}^k|^2 \\ &\le \frac{|f|^2}{\lambda_1\nu} + c_1\lambda_{m+1}^{-\frac{1}{2}}(|A^{\frac{1}{2}}w_m^k|\,|A^{\frac{1}{2}}w_M^{k-1}|^2 + |A^{\frac{1}{2}}v_m|\,|A^{\frac{1}{2}}w_M^l|^2).\end{aligned} \tag{3.11}$$

Because of $|A^{\frac{1}{2}}u_{mM}^k|^2 \ge \lambda_1|u_{mM}^k|^2$, we have

$$\begin{aligned}\frac{d|u_{mM}^k|^2}{dt} &+ \frac{11\lambda_1\nu}{20}|u_{mM}^k|^2 + \frac{\nu}{4}|A^{\frac{1}{2}}w_M^k|^2 \\ &\le \frac{|f|^2}{\lambda_1\nu} + c_1\lambda_{m+1}^{-\frac{1}{2}}(|A^{\frac{1}{2}}w_m^k|\,|A^{\frac{1}{2}}w_M^{k-1}|^2 + |A^{\frac{1}{2}}v_m|\,|A^{\frac{1}{2}}w_M^l|^2).\end{aligned}$$

Thanks to (2.5), $\max_{0\le t\le t_M}|A^{\frac{1}{2}}u_{mM}^k(t)| \le M_1$ implies

$$\max_{0\le t\le t_M}\left(\max_{1\le k\le l}|A^{\frac{1}{2}}w_M^k(t)|^2 + |A^{\frac{1}{2}}v_m(t)|^2\right) \le M_1^2.$$

Therefore

$$\max_{0\le t\le t_M}\left(\max_{1\le k\le l}|A^{\frac{1}{2}}w_M^k(t)|+|A^{\frac{1}{2}}v_m(t)|\right)\le\sqrt{2}M_1. \tag{3.12}$$

Integrating the above inequality on time interval $[0,t]$ and noticing (3.12), we obtain

$$\begin{aligned}&|u_{mM}^k(t)|^2+\frac{\nu}{4}\int_0^t e^{-\frac{11\lambda_1\nu}{20}(t-s)}|A^{\frac{1}{2}}w_M^k(s)|^2ds\le|a|^2+\frac{20}{11\lambda_1^2\nu^2}|f|^2\\&\quad+\frac{\sqrt{2}c_1M_1}{\lambda_{m+1}^{\frac{1}{2}}}\max\left\{\int_0^t e^{-\frac{11\lambda_1\nu}{20}(t-s)}|A^{\frac{1}{2}}w_M^{k-1}(s)|^2ds,\int_0^t e^{-\frac{11\lambda_1\nu}{20}(t-s)}|A^{\frac{1}{2}}w_M^l(s)|^2ds\right\}.\end{aligned}$$

Because both terms on the left-hand side of the above inequality are positive, we have

$$\begin{aligned}&|u_{mM}^k(t)|^2\le|a|^2+\frac{20}{11\lambda_1^2\nu^2}|f|^2\\&\quad+\frac{\sqrt{2}c_1M_1}{\lambda_{m+1}^{\frac{1}{2}}}\max\left\{\int_0^t e^{-\frac{11\lambda_1\nu}{20}(t-s)}|A^{\frac{1}{2}}w_M^{k-1}(s)|^2ds,\int_0^t e^{-\frac{11\lambda_1\nu}{20}(t-s)}|A^{\frac{1}{2}}w_M^l(s)|^2ds\right\},\end{aligned}$$

$$\begin{aligned}&\frac{\nu}{4}\int_0^t e^{-\frac{11\lambda_1\nu}{20}(t-s)}|A^{\frac{1}{2}}w_M^k(s)|^2ds\le|a|^2+\frac{20}{11\lambda_1^2\nu^2}|f|^2\\&\quad+\frac{\sqrt{2}c_1M_1}{\lambda_{m+1}^{\frac{1}{2}}}\max\left\{\int_0^t e^{-\frac{11\lambda_1\nu}{20}(t-s)}|A^{\frac{1}{2}}w_M^{k-1}(s)|^2ds,\int_0^t e^{-\frac{11\lambda_1\nu}{20}(t-s)}|A^{\frac{1}{2}}w_M^l(s)|^2ds\right\}.\end{aligned}$$

Taking the maximum value with respect to $1\le k\le l$ on both sides of the above two inequalities and noticing $w_M^0=0$, the summation of the results admits

$$\begin{aligned}&\max_{1\le k\le l}|u_{mM}^k(t)|^2+\frac{\nu}{4}\max_{1\le k\le l}\int_0^t e^{-\frac{11\lambda_1\nu}{20}(t-s)}|A^{\frac{1}{2}}w_M^k(s)|^2ds\\&\qquad\le 2|a|^2+\frac{4}{\lambda_1^2\nu^2}|f|^2+2\sqrt{2}c_1M_1\lambda_{m+1}^{-\frac{1}{2}}\max_{1\le k\le l}\int_0^t e^{-\frac{11\lambda_1\nu}{20}(t-s)}|A^{\frac{1}{2}}w_M^k(s)|^2ds.\end{aligned} \tag{3.13}$$

Since we know $2\sqrt{2}c_1M_1\lambda_{m+1}^{-\frac{1}{2}}\le\frac{\nu}{4}$ from the condition (3.6), then (3.13) implies

$$|u_{mM}^k(t)|\le\sqrt{2|a|^2+\frac{4|f|^2}{\lambda_1^2\nu^2}}=\sqrt{2\nu^2Re^2+4\nu^2Gr^2}=M_0\quad\forall t\in[0,t_M],\ 1\le k\le l.$$

From (3.11) and the result we just obtained, we have for any $t\ge0, r>0$, and $t+r\le t_M$

$$\frac{4\nu}{5}\int_t^{t+r}|A^{\frac{1}{2}}u_{mM}^k|^2ds\le 2M_0^2+\frac{|f|^2r}{\lambda_1\nu}+2\sqrt{2}c_1M_1\lambda_{m+1}^{-\frac{1}{2}}\max_{1\le i\le l}\int_t^{t+r}|A^{\frac{1}{2}}w_M^i|^2ds.$$

Noticing that $|A^{\frac{1}{2}}w_M^k|^2\le|A^{\frac{1}{2}}u_{mM}^k|^2$ and that (3.6) guarantees $2\sqrt{2}c_1M_1\lambda_{m+1}^{-\frac{1}{2}}\le\frac{\nu}{4}$, we can easily get from the above inequality that

$$\max_{1\le k\le l}\int_t^{t+r}|A^{\frac{1}{2}}u_{mM}^k(s)|^2ds\le\frac{4M_0^2}{\nu}+\frac{2r|f|^2}{\lambda_1\nu^2}\quad\forall t\ge0,\ r>0,\ t+r\le t_M.\qquad\square$$

Now we are ready to prove that for the ODE system (3.7)–(3.9) there exists a global uniformly bounded solution.

Lemma 3.2. *For any given* $0 < T < \infty$, *if* m *is so large that the condition* (3.6) *holds, we have*

$$|A^{\frac{1}{2}}u^k_{mM}(t)| \leq M_1 \quad \forall t \in [0,T],\ 0 \leq k \leq l,$$

and

$$\int_0^T |Au^k_{mM}(s)|^2 ds \leq M_2(T), \quad k = 0, 1, \ldots, l,$$

where M_1 *and* $M_2(T)$ *are given in Theorem* 3.1.

Proof. Noticing what was mentioned before, that the ODE system (3.7)–(3.9) has a maximal continuous solution defined on some interval $[0, T_M)$, we will prove this lemma by contradiction.

Assuming that

$$T_M < T, \tag{3.14}$$

we then assert that $|A^{\frac{1}{2}}u^k_{mM}(t)|$ goes to ∞ as t tends to T_M for certain $0 \leq k \leq l$. Since $|A^{\frac{1}{2}}u^k_{mM}(t)|$ is a continuous function of t and $M_1 > |A^{\frac{1}{2}}a|$, there exists a constant $t_M < T_M$ such that

$$\max_{0\leq k\leq l} |A^{\frac{1}{2}}u^k_{mM}(t)| < M_1 \text{ on } [0, t_M) \quad \text{and} \quad \max_{0\leq k\leq l} |A^{\frac{1}{2}}u^k_{mM}(t_M)| = M_1. \tag{3.15}$$

Multiplying (3.10) by $2Au^k_{mM}$ and integrating it on Ω, we have

$$\begin{aligned}\frac{d|A^{\frac{1}{2}}u^k_{mM}|^2}{dt} + 2\nu|Au^k_{mM}|^2 &\leq 2|b(v_m, v_m, Au^k_{mM})| + 2|b(v_m, w^l_M, Av_m)| \\ &+ 2|b(w^l_M, v_m, Av_m)| + 2|b(w^l_M, w^l_M, Av_m)| + 2|b(v_m, w^k_M, Aw^k_M)| \\ &+ 2|b(w^k_M, v_m, Aw^k_M)| + 2|b(w^{k-1}_M, w^k_M, Aw^k_M)| + 2|b(w^k_M, w^{k-1}_M, Aw^k_M)| \\ &+ 2|b(w^{k-1}_M, w^{k-1}_M, Aw^k_M)| + 2|(f, Au^k_{mM})|.\end{aligned}$$

Under the conditions of (3.6) and (3.15), Lemma 3.1 asserts that $|u^k_{mM}| \leq M_0$ on $[0, t_M]$ for any $0 \leq k \leq l$. Of course, we know $|v_m|, |w^k_{mM}| \leq M_0$ from (2.5). By using (2.2)–(2.4), (3.4)–(3.5), and the result of Lemma 3.1, we have the following estimates of the ten right-hand-side terms in the above inequality on $[0, t_M]$:

$$\begin{aligned}2|b(v_m, v_m, Au^k_{mM})| &\leq c_1|v_m|_\infty|A^{\frac{1}{2}}v_m|\,|Au^k_{mM}| \leq c_1|v_m|^{\frac{1}{2}}|A^{\frac{1}{2}}v_m|\,|Au^k_{mM}|^{\frac{3}{2}} \\ &\leq \frac{3\nu}{4}|Au^k_{mM}|^2 + \frac{c_1^4M_0^2}{4\nu^3}|A^{\frac{1}{2}}v_m|^4, \\ 2|b(v_m, w^l_M, Av_m)| &\leq c_1|v_m|_\infty|A^{\frac{1}{2}}w^l_M|\,|Av_m| \leq \frac{c_1L_m}{\lambda^{\frac{1}{2}}_{m+1}}|A^{\frac{1}{2}}v_m|\,|Aw^l_M|\,|Av_m| \\ &\leq \frac{c_1L_m}{4\lambda^{\frac{1}{2}}_{m+1}}|A^{\frac{1}{2}}v_m|\,|Aw^l_M|^2 + \frac{c_1L_m}{\lambda^{\frac{1}{2}}_{m+1}}|A^{\frac{1}{2}}v_m|\,|Av_m|^2, \\ 2|b(w^l_M, v_m, Av_m)| &\leq c_1|w^l_M|\,|A^{\frac{1}{2}}v_m|_\infty|Av_m| \leq \frac{c_1L_m}{\lambda^{\frac{1}{2}}_{m+1}}|A^{\frac{1}{2}}w^l_M|\,|Av_m|^2,\end{aligned}$$

$$\begin{aligned}
2|b(w_M^l, w_M^l, Av_m)| &\le c_1|w_M^l|_\infty|A^{\frac{1}{2}}w_M^l|\,|Av_m| \\
&\le c_1|w_M^l|^{\frac{1}{2}}|Aw_M^l|^{\frac{1}{2}}|w_M^l|^{\frac{1}{2}}|Aw_M^l|^{\frac{1}{2}}|Av_m| \\
&= c_1|w_M^l|\,|Aw_M^l|\,|Av_m| \le \frac{c_1}{\lambda_{m+1}^{\frac{1}{2}}}|A^{\frac{1}{2}}w_M^l|\,|Aw_M^l|\,|Av_m| \\
&\le \frac{c_1L_m}{4\lambda_{m+1}^{\frac{1}{2}}}|A^{\frac{1}{2}}w_M^l|\,|Aw_M^l|^2 + \frac{c_1L_m}{\lambda_{m+1}^{\frac{1}{2}}}|A^{\frac{1}{2}}w_M^l|\,|Av_m|^2, \\
2|b(v_m, w_M^k, Aw_M^k)| &\le c_1|v_m|_\infty|A^{\frac{1}{2}}w_M^k|\,|Aw_M^k| \le \frac{c_1L_m}{\lambda_{m+1}^{\frac{1}{2}}}|A^{\frac{1}{2}}v_m|\,|Aw_M^k|^2, \\
2|b(w_M^k, v_m, Aw_M^k)| &\le c_1|w_M^k|_\infty|A^{\frac{1}{2}}v_m|\,|Aw_M^k| \le c_1|w_M^k|^{\frac{1}{2}}|Aw_M^k|^{\frac{1}{2}}|A^{\frac{1}{2}}v_m|\,|Aw_M^k| \\
&\le \frac{c_1}{\lambda_{m+1}^{\frac{1}{2}}}|A^{\frac{1}{2}}v_m|\,|Aw_M^k|^2, \\
2|b(w_M^{k-1}, w_M^k, Aw_M^k)| &\le c_1|A^{\frac{1}{4}}w_M^{k-1}|\,|A^{\frac{3}{4}}w_M^k|\,|Aw_M^k| \le \frac{c_1}{\lambda_{m+1}^{\frac{1}{2}}}|A^{\frac{1}{2}}w_M^{k-1}|\,|Aw_M^k|^2, \\
2|b(w_M^k, w_M^{k-1}, Aw_M^k)| &\le c_1|w_M^k|_\infty|A^{\frac{1}{2}}w_M^{k-1}|\,|Aw_M^k| \le c_1|w_M^k|^{\frac{1}{2}}|A^{\frac{1}{2}}w_M^{k-1}|\,|Aw_M^k|^{\frac{3}{2}} \\
&\le \frac{c_1}{\lambda_{m+1}^{\frac{1}{2}}}|A^{\frac{1}{2}}w_M^{k-1}|\,|Aw_M^k|^2, \\
2|b(w_M^{k-1}, w_M^{k-1}, Aw_M^k)| &\le c_1|w_M^{k-1}|_\infty|A^{\frac{1}{2}}w_M^{k-1}|\,|Aw_M^k| \\
&\le c_1|w_M^{k-1}|^{\frac{1}{2}}|Aw_M^{k-1}|^{\frac{1}{2}}|w_M^{k-1}|^{\frac{1}{2}}|Aw_M^{k-1}|^{\frac{1}{2}}|Aw_M^k| \\
&= c_1|w_M^{k-1}|\,|Aw_M^{k-1}|\,|Aw_M^k| \le \frac{c_1}{\lambda_{m+1}^{\frac{1}{2}}}|A^{\frac{1}{2}}w_M^{k-1}|\,|Aw_M^{k-1}|\,|Aw_M^k| \\
&\le \frac{c_1}{2\lambda_{m+1}^{\frac{1}{2}}}|A^{\frac{1}{2}}w_M^{k-1}|\,|Aw_M^k|^2 + \frac{c_1}{2\lambda_{m+1}^{\frac{1}{2}}}|A^{\frac{1}{2}}w_M^{k-1}|\,|Aw_M^{k-1}|^2, \\
2|(f, Au_{mM}^k)| &\le \frac{4}{\nu}|f|^2 + \frac{\nu}{4}|Au_{mM}^k|^2.
\end{aligned}$$

Combining the above estimates yields

$$\begin{aligned}
\frac{d|A^{\frac{1}{2}}u_{mM}^k|^2}{dt} + \nu|Au_{mM}^k|^2 \le{}& \frac{4}{\nu}|f|^2 + \frac{c_1^4M_0^2}{4\nu^3}|A^{\frac{1}{2}}v_m|^4 \\
&+ \frac{1}{4}c_1L_m\lambda_{m+1}^{-\frac{1}{2}}(|A^{\frac{1}{2}}v_m| + |A^{\frac{1}{2}}w_M^l|)|Aw_M^l|^2 \\
&+ c_1L_m\lambda_{m+1}^{-\frac{1}{2}}(|A^{\frac{1}{2}}v_m|\,|Av_m|^2 + 2|A^{\frac{1}{2}}w_M^l|\,|Av_m|^2 \\
&\qquad + 2|A^{\frac{1}{2}}v_m|\,|Aw_M^k|^2 + 2|A^{\frac{1}{2}}w_M^{k-1}|\,|Aw_M^k|^2) \\
&+ \frac{c_1}{2}\lambda_{m+1}^{-\frac{1}{2}}|A^{\frac{1}{2}}w_M^{k-1}|\,|Aw_M^k|^2 + \frac{c_1}{2}\lambda_{m+1}^{-\frac{1}{2}}|A^{\frac{1}{2}}w_M^{k-1}|\,|Aw_M^{k-1}|^2.
\end{aligned}$$

Noticing (3.15) and (3.12), we have for $t \in [0, t_M]$

$$\begin{aligned}
&|A^{\frac{1}{2}}v_m|\,|Av_m|^2 + |A^{\frac{1}{2}}w_M^l|\,|Av_m|^2 + |A^{\frac{1}{2}}v_m|\,|Aw_M^k|^2 + |A^{\frac{1}{2}}w_M^{k-1}|\,|Aw_M^k|^2 \\
&\qquad \le \sqrt{2}M_1|Au_{mM}^k|^2.
\end{aligned}$$

The combination of the above estimations gives

$$\frac{d|A^{\frac{1}{2}}u^k_{mM}|^2}{dt}+\nu|Au^k_{mM}|^2\le\frac{4}{\nu}|f|^2+\frac{c_1^4M_0^2}{4\nu^3}|A^{\frac{1}{2}}v_m|^4+\frac{\sqrt{2}}{4}c_1M_1L_m\lambda_{m+1}^{-\frac{1}{2}}|Aw^l_M|^2$$
$$+\frac{2\sqrt{2}c_1M_1L_m}{\lambda_{m+1}^{\frac{1}{2}}}|Au^k_{mM}|^2+\frac{c_1M_1}{2\lambda_{m+1}^{\frac{1}{2}}}|Aw^k_M|^2+\frac{c_1M_1}{2\lambda_{m+1}^{\frac{1}{2}}}|Aw^{k-1}_M|^2.$$

Thanks to (3.6), we have $2\sqrt{2}c_1M_1L_m\lambda_{m+1}^{-\frac{1}{2}}\le\frac{\nu}{4}$. Then

$$\frac{d|A^{\frac{1}{2}}u^k_{mM}|^2}{dt}+\frac{3\nu}{4}|Au^k_{mM}|^2\le\frac{4}{\nu}|f|^2+\frac{c_1^4M_0^2}{4\nu^3}|A^{\frac{1}{2}}v_m|^4$$
$$+\frac{c_1}{2}M_1L_m\lambda_{m+1}^{-\frac{1}{2}}(|Aw^l_M|^2+|Aw^k_M|^2+|Aw^{k-1}_M|^2).$$

For any $t\ge 0$, $r>0$ satisfying $t+r\le t_M$ and $\tau\in[t,t+r]$, integrating the above inequality on $[\tau,t+r]$ shows that for all $1\le k\le l$

$$|A^{\frac{1}{2}}u^k_{mM}(t+r)|^2+\frac{3\nu}{4}\int_\tau^{t+r}|Au^k_{mM}(s)|^2ds\le\frac{4r}{\nu}|f|^2+\int_\tau^{t+r}\frac{c_1^4M_0^2}{4\nu^3}|A^{\frac{1}{2}}v_m(s)|^4ds$$
$$+\frac{c_1}{2}M_1L_m\lambda_{m+1}^{-\frac{1}{2}}\int_\tau^{t+r}[|Aw^l_M(s)|^2+|Aw^k_M(s)|^2+|Aw^{k-1}_M(s)|^2]ds+|A^{\frac{1}{2}}u^k_{mM}(\tau)|^2.$$

Now, we define

$$y(s)=\max_{1\le i\le l}|A^{\frac{1}{2}}u^i_{mM}(s)|^2.$$

By noticing that (3.6) implies $3c_1M_1L_m\lambda_{m+1}^{-\frac{1}{2}}\le\frac{\nu}{2}$ and using a similar method for deriving (3.13),

$$y(t+r)+\frac{\nu}{4}\max_{1\le i\le l}\int_\tau^{t+r}|Au^i_{mM}(s)|^2ds \tag{3.16}$$
$$\le 2y(\tau)+\frac{8r}{\nu}|f|^2+\int_\tau^{t+r}\frac{c_1^4M_0^2}{2\nu^3}y^2(s)ds.$$

Thanks to Lemma 3.1, we know

$$\int_t^{t+r}y(s)ds\le\frac{4M_0^2}{\nu}+\frac{2r|f|^2}{\lambda_1\nu^2}\triangleq a_3(r)\quad\forall t\ge 0,\ r>0,\ t+r\le t_M.$$

Also, we define

$$a_2(r)=\frac{8r}{\nu}|f|^2,\qquad a_1(r)=\frac{c_1^4M_0^2}{2\nu^3}a_3(r).$$

It is obvious that $a_1(r)$, $a_2(r)$, and $a_3(r)$ are functions of r and independent of l and M_1. Omitting the second term on the left-hand side of (3.16) and combining it with the above three inequalities, we can use the idea of the proof of the uniform Gronwall inequality (see [22]) to get

$$y(t)\le\left(\frac{2a_3(r)}{r}+a_2(r)\right)\exp(a_1(r))\triangleq B_{11}(r)\quad\forall r\le t\le t_M. \tag{3.17}$$

For $t \in [0, r]$, by using the ordinary Gronwall inequality, (3.16) implies

$$y(t) \le (2|A^{\frac{1}{2}}a|^2 + a_2(r)) \exp(a_1(r)) \overset{\triangle}{=} B_{12}(r). \tag{3.18}$$

Now let us complete the proof in two cases.

Case 1. If $t_M > \frac{1}{\lambda_1 \nu}$, we take $r = \frac{1}{\lambda_1 \nu}$ in (3.17)–(3.18). This admits

$$\max_{1 \le i \le l} |A^{\frac{1}{2}} u^i_{mM}(t)| \le \sqrt{\max\left\{ B_{11}\left(\frac{1}{\lambda_1 \nu}\right), B_{12}\left(\frac{1}{\lambda_1 \nu}\right) \right\}} \quad \forall t \in [0, t_M].$$

It is easy to verify that

$$\sqrt{\max\left\{ B_{11}\left(\frac{1}{\lambda_1 \nu}\right), B_{12}\left(\frac{1}{\lambda_1 \nu}\right) \right\}} < 4\lambda_1^{\frac{1}{2}} \nu (Re + 2Gr) \exp(c_1^4 (2Re^2 + 5Gr^2)^2) = M_1,$$

and this is a contradiction of (3.15).

Case 2. If $t_M \le \frac{1}{\lambda_1 \nu}$, it is quite easy to verify that (3.16) is still valid with $t = \tau = 0$ and any $r \le t_M$. Then we can use the ordinary Gronwall inequality on (3.16) to get

$$\max_{1 \le i \le l} |A^{\frac{1}{2}} u^i_{mM}(r)| \le \sqrt{B_{12}\left(\frac{1}{\lambda_1 \nu}\right)} < M_1 \quad \forall r \in [0, t_M].$$

In particular, $\max_{1 \le i \le l} |A^{\frac{1}{2}} u^i_{mM}(t_M)| < M_1$. This also leads to a contradiction with (3.15).

Therefore the assumption (3.14) is invalid. We can thus get $T_M \ge T$; that is, the solution of (3.7)–(3.9) will not blow up in any bounded time interval. And the proof shows that its solution is uniformly (with respect to m, M, l, and T) bounded with bound M_1.

Furthermore, from (3.16) we can get, for $k = 1, \ldots, l$,

$$\begin{aligned} \int_0^T |Au^k_{mM}(s)|^2 ds &\le 12\nu^{-1} M_1^2 + 32T\nu^{-2}|f|^2 + \frac{2c_1^4 M_0^2 M_1^4 T}{\nu^4} \\ &\le 2\nu^{-1} M_1^2 (6 + c_1^4 \nu^{-3} M_0^2 M_1^2 T) + 32\lambda_1^2 \nu^2 T Gr^2 = M_2^2(T). \qquad \square \end{aligned}$$

Proof of Theorem 3.1. By the results of Lemma 3.2, we claim that for any fixed m the sequence

$$U_{mM} \overset{\triangle}{=} \{v_m, w^1_M, \ldots, w^l_M\}, \quad M > m, \tag{3.19}$$

remains in a bounded set of $\mathcal{L}^2(0, T; \mathbf{D(A)}) \cap \mathcal{L}^\infty(0, T; \mathbf{V})$,

where $\mathbf{D(A)} = P_m D(A) \times (Q_m D(A))^l$, $\mathbf{V} = P_m V \times (Q_m V)^l$. And it is easy for us to verify that

$$U'_{mM} = \frac{dU_{mM}}{dt} \quad \text{remains in a bounded set of } \mathcal{L}^2(0, T; \mathbf{H}). \tag{3.20}$$

Here $\mathbf{H} = H_m \times (Q_m H)^l$.

Now we define $X_0 = \mathbf{D(A)}$, $X = \mathbf{V}$, $X_1 = \mathbf{H}$, and

$$\mathcal{Y} = \{U \in \mathcal{L}^2(0, T; X_0),\ U' \in \mathcal{L}^2(0, T; X_1)\}.$$

From (3.20) and the second result of Lemma 3.2, we have that the sequence $\{U_{mM}\}_{M>m}$ remains in a bounded set of $\mathcal{Y}$. Thanks to (3.19) and the compactness theorem (see Theorem 2.1 in Chapter III of [21]), we can assert the existence of an element $U_m = \{\tilde{v}_m, \tilde{w}^1, \dots, \tilde{w}^l\} \in \mathcal{L}^\infty(0,T;\mathbf{V})$ and a subsequence $\{U_{mM'}\}_{M'>m}$ such that

$$\left\{\begin{array}{l} U_{mM'} \to U_m \text{ in } \mathcal{L}^2(0,T;\mathbf{D(A)}) \text{ weakly, in} \\ \mathcal{L}^\infty(0,T;\mathbf{V}) \text{ weak star, and in} \\ \mathcal{L}^2(0,T;\mathbf{V}) \text{ strongly, as } M' \to \infty. \end{array}\right.$$

We can certainly get $U_{mM'} \to U_m$ weakly in $\mathcal{L}^2(0,T;\mathbf{V})$, weak star in $\mathcal{L}^\infty(0,T;\mathbf{H})$, and strongly in $\mathcal{L}^2(0,T;\mathbf{H})$. Noticing the continuity property of the trilinear form $b(\cdot,\cdot,\cdot)$ (see Lemma 3.2 in Chapter III of [21]), the passage to the limit shows that U_m is the solution of system (3.1)–(3.3), which shares the same bound of U_{mM}. □

In the following theorem, we will show that the small eddy components, namely $|w_M^k|$ and $|A^{\frac{1}{2}} w_M^k|$, are bounded by some small quantities. This result is very important for the error estimates in the next section.

THEOREM 3.2. *If $|A^{\frac{1}{2}} u_{mM}^k(t)| \le M_1$ for any $t \ge 0$ and $1 \le k \le l$ and m is large enough such that* (3.6) *holds, then there exists a positive constant $T_0 = T_0(a,f,\nu)$ such that*

$$|w_M^k(t)| \le \frac{K_0 L_m}{\lambda_{m+1}}, \qquad |A^{\frac{1}{2}} w_M^k(t)| \le \frac{K_1 L_m}{\lambda_{m+1}^{\frac{1}{2}}} \quad \forall t \ge T_0, 1 \le k \le l,$$

where

$$K_0 = \frac{2(c_1 M_1^2 + 2|P_{mM} f| L_m^{-1})}{\nu}, \qquad K_1 = \frac{5(c_1 M_1^2 + |P_{mM} f| L_m^{-1})}{\nu}.$$

Proof. Multiplying (3.8) with $2w_M^k$, integrating it on Ω, and using (3.4), we get

$$\begin{aligned} &\frac{d|w_M^k|^2}{dt} + 2\nu|A^{\frac{1}{2}} w_M^k|^2 + 2b(v_m, v_m, w_M^k) + 2b(w_M^k, v_m + w_M^{k-1}, w_M^k) \\ &\quad = 2(f, w_M^k) + 2b(w_M^{k-1}, w_M^{k-1}, w_M^k). \end{aligned}$$

For the trilinear forms and the force term above, using (2.2)–(2.5), (3.4)–(3.5), and the assumption $|A^{\frac{1}{2}}(v_m + w_M^k)| \le M_1$, we have

$$\begin{aligned} 2b(v_m, v_m, w_M^k) &\le c_1 |v_m|_\infty |A^{\frac{1}{2}} v_m|\,|w_M^k| \le c_1 L_m \lambda_{m+1}^{-\frac{1}{2}} |A^{\frac{1}{2}} v_m|^2 |A^{\frac{1}{2}} w_M^k| \\ &\le \frac{\nu}{4}|A^{\frac{1}{2}} w_M^k|^2 + \frac{c_1^2 L_m^2 |A^{\frac{1}{2}} v_m|^4}{\nu\lambda_{m+1}}, \\ 2b(w_M^k, v_m + w_M^{k-1}, w_M^k) &\le c_1 |A^{\frac{1}{4}} w_M^k|\,|A^{\frac{1}{2}}(v_m + w_M^{k-1})|\,|A^{\frac{1}{4}} w_M^k| \le \frac{c_1 M_1}{\lambda_{m+1}^{\frac{1}{2}}} |A^{\frac{1}{2}} w_M^k|^2, \\ 2b(w_M^{k-1}, w_M^{k-1}, w_M^k) &= -2b(w_M^{k-1}, w_M^k, w_M^{k-1}) \le c_1 |A^{\frac{1}{4}} w_M^{k-1}|\,|A^{\frac{1}{2}} w_M^k|\,|A^{\frac{1}{4}} w_M^{k-1}| \\ &\le c_1 \lambda_{m+1}^{-\frac{1}{2}} |A^{\frac{1}{2}} w_M^{k-1}|^2 |A^{\frac{1}{2}} w_M^k| \le \frac{\nu}{4}|A^{\frac{1}{2}} w_M^k|^2 + \frac{c_1^2 |A^{\frac{1}{2}} w_M^{k-1}|^4}{\nu\lambda_{m+1}}, \\ 2(f, w_M^k) &\le 2\lambda_{m+1}^{-\frac{1}{2}} |P_{mM} f|\,|A^{\frac{1}{2}} w_M^k| \le \frac{\nu}{4}|A^{\frac{1}{2}} w_M^k|^2 + \frac{4}{\nu\lambda_{m+1}}|P_{mM} f|^2. \end{aligned}$$

Thanks to (2.5), $|A^{\frac{1}{2}} v_m|^4 + |A^{\frac{1}{2}} w_M^{k-1}|^4 \le |A^{\frac{1}{2}}(v_m + w_M^{k-1})|^4 \le M_1^4$. By (3.6), we see that $c_1 M_1 \lambda_{m+1}^{-\frac{1}{2}} \le \frac{\nu}{4}$. Then

$$\frac{d|w_M^k|^2}{dt} + \nu\lambda_{m+1}|w_M^k|^2 \le \frac{c_1^2 M_1^4 L_m^2 + 4|P_{mM} f|^2}{\nu\lambda_{m+1}}.$$

Integrating the above inequality on $[0,t]$ yields

$$|w_M^k(t)| \le \frac{L_m}{\lambda_{m+1}} \cdot \frac{c_1 M_1^2 + 2|P_{mM} f| L_m^{-1}}{\nu} + e^{-\nu\lambda_{m+1}t/2}|P_{mM}a| \quad \forall t \ge 0. \tag{3.21}$$

Similarly, by using equation (3.8) again, we have

$$\begin{aligned}\frac{d|A^{\frac{1}{2}}w_M^k|^2}{dt} &+ 2\nu|Aw_M^k|^2 + 2b(v_m, v_m, Aw_M^k) + 2b(v_m + w_M^{k-1}, w_M^k, Aw_M^k)\\ &+ 2b(w_M^k, v_m + w_M^{k-1}, Aw_M^k) - 2b(w_M^{k-1}, w_M^{k-1}, Aw_M^k) = 2(f, Aw_M^k).\end{aligned}$$

Applying (2.2)–(2.5) and (3.5), the following estimates hold:

$$\begin{aligned}
2b(v_m, v_m, Aw_M^k) &\le c_1|v_m|_\infty |A^{\frac{1}{2}}v_m|\,|Aw_M^k|\\
&\le c_1 L_m |A^{\frac{1}{2}}v_m|^2 |Aw_M^k| \le \frac{\nu}{8}|Aw_M^k|^2 + \frac{4c_1^2 L_m^2 M_1^4}{\nu},\\
2b(v_m + w_M^{k-1}, w_M^k, Aw_M^k) &\le c_1|v_m|_\infty |A^{\frac{1}{2}}w_M^k|\,|Aw_M^k| + c_1|A^{\frac{1}{4}}w_M^{k-1}|\,|A^{\frac{3}{4}}w_M^k|\,|Aw_M^k|\\
&\le \frac{c_1}{\lambda_{m+1}^{\frac{1}{2}}}(L_m|A^{\frac{1}{2}}v_m| + |A^{\frac{1}{2}}w_M^{k-1}|)|Aw_M^k|^2\\
&\le \frac{\sqrt{2}c_1 M_1 L_m}{\lambda_{m+1}^{\frac{1}{2}}}|Aw_M^k|^2,\\
2b(w_M^k, v_m + w_M^{k-1}, Aw_M^k) &\le c_1|w_M^k|_\infty |A^{\frac{1}{2}}(v_m + w_M^{k-1})|\,|Aw_M^k|\\
&\le c_1 M_1 |w_M^k|^{\frac{1}{2}}|Aw_M^k|^{\frac{3}{2}} \le \frac{c_1 M_1}{\lambda_{m+1}^{\frac{1}{2}}}|Aw_M^k|^2,\\
2b(w_M^{k-1}, w_M^{k-1}, Aw_M^k) &\le c_1|w_M^{k-1}|_\infty |A^{\frac{1}{2}}w_M^{k-1}|\,|Aw_M^k|\\
&\le c_1|w_M^{k-1}|^{\frac{1}{2}}|Aw_M^{k-1}|^{\frac{1}{2}}|A^{\frac{1}{2}}w_M^{k-1}|\,|Aw_M^k|\\
&\le \frac{c_1}{\lambda_{m+1}^{\frac{1}{2}}}|A^{\frac{1}{2}}w_M^{k-1}|\,|Aw_M^{k-1}|\,|Aw_M^k|\\
&\le \frac{\nu}{8}|Aw_M^k|^2 + \frac{4c_1^2 M_1^2}{\nu\lambda_{m+1}}|Aw_M^{k-1}|^2,\\
2(f, Aw_M^k) &\le 2|P_{mM}f|\,|Aw_M^k| \le \frac{\nu}{4}|Aw_M^k|^2 + \frac{4}{\nu}|P_{mM}f|^2.
\end{aligned}$$

Thanks again to (3.6), $\sqrt{2}c_1 M_1 L_m \lambda_{m+1}^{-\frac{1}{2}} \le \frac{\nu}{4}$. Then we get

$$\frac{d|A^{\frac{1}{2}}w_M^k|^2}{dt} + \nu|Aw_M^k|^2 \le \frac{4}{\nu}|P_{mM}f|^2 + \frac{4c_1^2 M_1^2}{\nu\lambda_{m+1}}|Aw_M^{k-1}|^2 + \frac{4c_1^2 M_1^4 L_m^2}{\nu}. \tag{3.22}$$

By using (2.4), we have $|Aw_M^k|^2 \ge \frac{1}{2}|Aw_M^k|^2 + \frac{\lambda_{m+1}}{2}|A^{\frac{1}{2}}w_M^k|^2$. Then it follows that

$$\begin{aligned}\frac{d|A^{\frac{1}{2}}w_M^k|^2}{dt} &+ \frac{\nu\lambda_{m+1}}{2}|A^{\frac{1}{2}}w_M^k|^2 + \frac{\nu}{2}|Aw_M^k|^2\\ &\le \frac{4}{\nu}|P_{mM}f|^2 + \frac{4c_1^2 M_1^2}{\nu\lambda_{m+1}}|Aw_M^{k-1}|^2 + \frac{4c_1^2 M_1^4 L_m^2}{\nu}.\end{aligned}$$

Integrating this inequality on $[0,t]$ admits for $1 \le k \le l$

$$|A^{\frac{1}{2}}w_M^k(t)|^2 + \frac{\nu}{2}\int_0^t e^{-\nu\lambda_{m+1}(t-s)/2}|Aw_M^k(s)|^2 ds \le \frac{8(|P_{mM}f|^2 + c_1^2M_1^4L_m^2)}{\nu^2\lambda_{m+1}}$$
$$+ e^{-\nu\lambda_{m+1}t/2}|A^{\frac{1}{2}}P_{mM}a|^2 + \frac{4c_1^2M_1^2}{\nu\lambda_{m+1}}\int_0^t e^{-\nu\lambda_{m+1}(t-s)/2}|Aw_M^{k-1}(s)|^2 ds.$$

By using the same method for deriving (3.15) again, we have

$$|A^{\frac{1}{2}}w_M^k(t)|^2 + \frac{\nu}{2}\max_{1\le i\le l}\int_0^t e^{-\nu\lambda_{m+1}(t-s)/2}|Aw_M^i(s)|^2 ds \le \frac{16(|P_{mM}f|^2 + c_1^2M_1^4L_m^2)}{\nu^2\lambda_{m+1}}$$
$$+ 2e^{-\nu\lambda_{m+1}t/2}|A^{\frac{1}{2}}P_{mM}a|^2 + \frac{8c_1^2M_1^2}{\nu\lambda_{m+1}}\max_{1\le i\le l}\int_0^t e^{-\nu\lambda_{m+1}(t-s)/2}|Aw_M^i(s)|^2 ds.$$

If m is so large that (3.6) is satisfied, we have $\frac{8c_1^2M_1^2}{\nu\lambda_{m+1}} \le \frac{\nu}{2}$. Therefore, we can obtain for $t \ge 0$

$$(3.23) \quad |A^{\frac{1}{2}}w_M^k(t)| \le \frac{L_m}{\lambda_{m+1}^{\frac{1}{2}}}\cdot\frac{4(|P_{mM}f|L_m^{-1} + c_1M_1^2)}{\nu} + \sqrt{2}e^{-\nu\lambda_{m+1}t/4}|A^{\frac{1}{2}}P_{mM}a|.$$

Taking into account (3.21) and (3.23), there must exist a constant $T_0(a,f,\nu) > 0$ such that the results of the theorem are valid. □

Comparing this result with the small eddy estimates of the NSE given in [8], we find that the small eddy components obtained in the small eddy correction method share the same properties as those of the NSE, which will be listed in the next section.

4. Convergence analysis. First of all, we recall the following property of the NSE (see [8]). There exist constants, which will be also denoted by M_0, M_1, K_0, K_1, and $T_0 = T_0(a,f,\nu)$ appearing in Theorem 3.1 and 3.2, such that for any solution $u = p + q$ of (1.1) or (1.5)–(1.6),

$$(4.1) \qquad |u(t)| \le M_0, \quad |A^{\frac{1}{2}}u(t)| \le M_1 \quad \forall t \ge 0,$$

and

$$(4.2) \qquad |q(t)| \le K_0L_m\lambda_{m+1}^{-1}, \quad |A^{\frac{1}{2}}q| \le K_1L_m\lambda_{m+1}^{-\frac{1}{2}} \quad \forall t \ge T_0.$$

LEMMA 4.1. *Under the conditions of Theorem* 3.1, *there exists a constant* $T_0' = T_0'(a,f,\nu) > 0$ *such that for any* $t \ge T_0'$

$$\nu\int_0^t e^{-\nu\lambda_{m+1}(t-s)/2}|Aw^1(s)|^2 ds \le \frac{K_1^2L_m^2}{\lambda_{m+1}},$$
$$\nu\int_0^t e^{-\nu(t-s)/2}|Aw^1(s)|^2 ds \le K_1^2L_m^2.$$

Proof. Consider $k = 1$ in (3.2). From (3.10), we can get

$$\frac{d|A^{\frac{1}{2}}w^1|^2}{dt} + \nu|Aw^1|^2 \le \frac{3}{\nu}|Q_mf|^2 + \frac{3c_1^2L_m^2M_1^4}{\nu}.$$

Since $\nu|Aw^1|^2 \ge \frac{\nu\lambda_{m+1}}{2}|A^{\frac{1}{2}}w^1|^2 + \frac{\nu}{2}|Aw^1|^2$, we have

$$(4.3) \qquad \frac{d|A^{\frac{1}{2}}w^1|^2}{dt} + \frac{\nu}{2}|Aw^1|^2 + \frac{\nu\lambda_{m+1}}{2}|A^{\frac{1}{2}}w^1|^2 \le \frac{3}{\nu}|Q_mf|^2 + \frac{3c_1^2L_m^2M_1^4}{\nu}.$$

Integrating the above inequality leads to

$$\begin{aligned}
&\nu\int_0^t e^{-\nu\lambda_{m+1}(t-s)/2}|Aw^1(s)|^2ds\\
&\leq 2e^{-\nu\lambda_{m+1}t/2}|A^{\frac{1}{2}}Q_ma|^2+\frac{6|Q_mf|^2}{\nu^2\lambda_{m+1}}+\frac{6c_1^2L_m^2M_1^4}{\nu^2\lambda_{m+1}}\\
&\leq 2e^{-\nu\lambda_{m+1}t/2}|A^{\frac{1}{2}}Q_ma|^2+\frac{L_m^2}{6\lambda_{m+1}}\times\frac{36(L_m^{-2}|Q_mf|^2+c_1^2M_1^4)}{\nu^2}\\
&\leq 2e^{-\nu\lambda_{m+1}t/2}|A^{\frac{1}{2}}Q_ma|^2+\frac{L_m^2}{6\lambda_{m+1}}\left(\frac{6(L_m^{-1}|Q_mf|+c_1M_1^2)}{\nu}\right)^2\\
&= 2e^{-\nu\lambda_{m+1}t/2}|A^{\frac{1}{2}}Q_ma|^2+\frac{L_m^2}{6\lambda_{m+1}}K_1^2.
\end{aligned}$$

Not enlarging the third term on the left-hand side of (4.3) to $\frac{\nu\lambda_{m+1}}{2}|A^{\frac{1}{2}}w^1|^2$, we can use the same procedure to get

$$\nu\int_0^t e^{-\nu(t-s)/2}|Aw^1(s)|^2ds\leq 2e^{-\nu t/2}|A^{\frac{1}{2}}Q_ma|^2+\frac{L_m^2}{6}K_1^2.$$

Certainly, there is a $T_0'(a,f,\nu)>0$ such that the results are valid. □

Now, let us introduce some notation:

$$e(t)=u(t)-u^l(t),\quad \varepsilon^k(t)=w^k(t)-w^{k-1}(t),\quad |||\cdot(t)|||=\sup_{0\leq s\leq t}|\cdot(s)|.$$

In particular, $\varepsilon^1(t)=w^1(t)$.

THEOREM 4.1. *Under the condition of Theorem* 3.1*, that is, with m large enough such that* (3.6) *is valid, we assume that the results of Theorem* 3.2*,* (4.1)–(4.2)*, and Lemma* 4.1 *hold for $T_0=0$ and $T_0'=0$. Then we have for $l\geq 1$*

$$|e(t)|\leq\frac{\nu K_0L_m^2}{2^{\frac{l-4}{2}}c_1K_1}\left(\frac{2^{\frac{1}{4}}c_1K_1L_m}{\nu\lambda_{m+1}}\right)^{2^l}\exp(c_1^2M_1^2\nu^{-1}t/4)\quad\forall t\geq 0,$$

where the constants M_1, K_0, and K_1 are defined in Theorems 3.1 *and* 3.2.

Proof. Combine (3.1) and (3.2) to see that

$$\frac{du^l}{dt}+\nu Au^l+B(u^l,u^l)-Q_mB(\varepsilon^l,\varepsilon^l)=f. \tag{4.4}$$

Subtracting (4.4) from (1.1) yields

$$\frac{de}{dt}+\nu Ae+B(e,u)+B(u^l,e)+Q_mB(\varepsilon^l,\varepsilon^l)=0. \tag{4.5}$$

Next, we give a rough estimate of $|e|^2$ in terms of $|\varepsilon^l|^2$ and $|A^{\frac{1}{2}}\varepsilon^l|^2$. Multiplying (4.5) by $2e$ and integrating it on Ω, we have

$$\frac{d|e|^2}{dt}+2\nu|A^{\frac{1}{2}}e|^2\leq 2|b(e,u,e)|+2|b(\varepsilon^l,\varepsilon^l,Q_me)|.$$

For the two terms on the right-hand side of the above inequality, we have

$$\begin{aligned}2|b(e,u,e)| &\le c_1|A^{\frac14}e|^2|A^{\frac12}u| \le c_1|e|\,|A^{\frac12}e|\,|A^{\frac12}u|\\ &\le c_1M_1|e|\,|A^{\frac12}e| \le \nu|A^{\frac12}e|^2+\frac{c_1^2M_1^2}{4\nu}|e|^2,\end{aligned}$$

$$2|b(\varepsilon^l,\varepsilon^l,Q_me)| = 2|b(\varepsilon^l,Q_me,\varepsilon^l)| \le c_1|\varepsilon^l|\,|A^{\frac12}\varepsilon^l|\,|A^{\frac12}e| \le \nu|A^{\frac12}e|^2+\frac{c_1^2}{4\nu}|\varepsilon^l|^2|A^{\frac12}\varepsilon^l|^2.$$

Here we applied the Sobolev interpolation inequality (2.2) (the third equation) to the second inequality of both of the above two expressions with $s=\frac14$, $m=\frac12$. Therefore, we get

$$\frac{d|e|^2}{dt} \le \frac{c_1^2M_1^2}{4\nu}|e|^2+\frac{c_1^2}{4\nu}|\varepsilon^l|^2|A^{\frac12}\varepsilon^l|^2.$$

Integrate the above inequality on $[0,t]$ to obtain

$$|e(t)|^2 \le M_1^{-2}e^{(c_1^2M_1^2\nu^{-1}t/4)}|||\varepsilon^l(t)|||^2|||A^{\frac12}\varepsilon^l(t)|||^2. \tag{4.6}$$

To take advantage of the fast decay property of the small eddy components, we estimate the large and small eddy errors $|e_p|^2$ and $|e_q|^2$ in terms of $|\varepsilon^l|^2$ and $|A^{\frac12}\varepsilon^l|^2$ by using (4.6).

Projecting (4.5) onto H_m and Q_mH, respectively, we have

$$\frac{de_p}{dt}+\nu Ae_p+P_mB(e_p+e_q,u)+P_mB(u^l,e_p+e_q)=0, \tag{4.7}$$

$$\frac{de_q}{dt}+\nu Ae_q+Q_mB(e_p+e_q,u)+Q_mB(u^l,e_p+e_q)+Q_mB(\varepsilon^l,\varepsilon^l)=0. \tag{4.8}$$

Multiplying (4.7) by $2e_p$, integrating it on Ω, and noticing (3.4) lead to

$$\frac{d|e_p|^2}{dt}+2\nu|A^{\frac12}e_p|^2 \le 2|b(e_p,u,e_p)|+2|b(e_q,u,e_p)|+2|b(u^l,e_q,e_p)|.$$

Thanks to (2.3)–(2.4), (3.5), and Lemma 1 in [8], we know that

$$2b(u^l,e_p,e_q)=2b(P_mu^l,e_p,e_q)+2b(Q_mu^l,e_p,e_q) \le c_1M_1L_m|A^{\frac12}e_p|\,|e_q|. \tag{4.9}$$

By using (2.2)–(2.4), (3.4)–(3.5), and (4.9), we have

$$\begin{aligned}2b(e_p,u,e_p) &\le c_1M_1|e_p|\,|A^{\frac12}e_p| \le \nu|A^{\frac12}e_p|^2+\frac{c_1^2M_1^2}{4\nu}|e_p|^2,\\ 2b(e_q,u,e_p) &\le c_1M_1|e_q|\,|e_p|_\infty \le c_1M_1L_m|e_q|\,|A^{\frac12}e_p| \le \frac{\nu}{2}|A^{\frac12}e_p|^2+\frac{c_1^2M_1^2L_m^2}{2\nu}|e_q|^2,\\ 2b(u^l,e_q,e_p) &\le c_1M_1L_m|A^{\frac12}e_p|\,|e_q| \le \frac{\nu}{2}|A^{\frac12}e_p|^2+\frac{c_1^2M_1^2L_m^2}{2\nu}|e_q|^2.\end{aligned}$$

Finally, we obtain

$$\frac{d|e_p|^2}{dt} \le \frac{c_1^2M_1^2}{4\nu}|e_p|^2+\frac{c_1^2M_1^2L_m^2}{\nu}|e_q|^2.$$

Integrating the above inequality on $[0,t]$ admits

$$|e_p(t)|^2 \le 4L_m^2 \exp(c_1^2 M_1^2 \nu^{-1} t/4) |||e_q(t)|||^2. \tag{4.10}$$

Multiplying (4.8) with $2e_q$ and integrating it on Ω, we get

$$\frac{d|e_q|^2}{dt} + 2\nu |A^{\frac{1}{2}} e_q|^2 \le 2|b(e_p, u, e_q)| + 2|b(e_q, u, e_q)| + 2|b(u^l, e_p, e_q)| + 2|b(\varepsilon^l, \varepsilon^l, e_q)|.$$

For the right-hand side terms, we have

$$\begin{aligned}
2b(e_p, u, e_q) &\le c_1 M_1 |A^{\frac{1}{4}} e_p|\,|A^{\frac{1}{4}} e_q| \le c_1 M_1 |A^{\frac{1}{2}} e_q|\,|e_p| \le \frac{\nu}{4} |A^{\frac{1}{2}} e_q|^2 + \frac{c_1^2 M_1^2}{\nu} |e_p|^2, \\
2b(e_q, u, e_q) &\le c_1 M_1 |A^{\frac{1}{4}} e_q|^2 \le c_1 M_1 |A^{\frac{1}{2}} e_q|\,|e_q| \le \frac{\nu}{4} |A^{\frac{1}{2}} e_q|^2 + \frac{c_1^2 M_1^2}{\nu} |e_q|^2, \\
2b(u^l, e_p, e_q) &\le c_1 M_1 L_m |A^{\frac{1}{2}} e_p|\,|e_q| \le c_1 M_1 L_m |e_p|\,|A^{\frac{1}{2}} e_q| \\
&\le \frac{\nu}{4} |A^{\frac{1}{2}} e_q|^2 + \frac{c_1^2 M_1^2 L_m^2}{\nu} |e_p|^2, \\
2b(\varepsilon^l, \varepsilon^l, e_q) &\le c_1 |A^{\frac{1}{4}} \varepsilon^l|^2 |A^{\frac{1}{2}} e_q| \le c_1 |\varepsilon^l|\,|A^{\frac{1}{2}} \varepsilon^l|\,|A^{\frac{1}{2}} e_q| \le \frac{\nu}{4} |A^{\frac{1}{2}} e_q|^2 + \frac{c_1^2}{\nu} |\varepsilon^l|^2 |A^{\frac{1}{2}} \varepsilon^l|^2.
\end{aligned}$$

Then we derive

$$\frac{d|e_q|^2}{dt} + \nu \lambda_{m+1} |e_q|^2 \le \frac{3c_1^2 M_1^2 L_m^2}{\nu} |e_p|^2 + \frac{c_1^2}{\nu} |\varepsilon^l|^2 |A^{\frac{1}{2}} \varepsilon^l|^2.$$

Integrating this inequality on $[0,t]$ yields

$$|e_q(t)|^2 \le \frac{3c_1^2 M_1^2 L_m^2}{\nu^2 \lambda_{m+1}} |||e_p(t)|||^2 + \frac{c_1^2}{\nu^2 \lambda_{m+1}} |||\varepsilon^l(t)|||^2 |||A^{\frac{1}{2}} \varepsilon^l(t)|||^2. \tag{4.11}$$

Thanks to (4.6), we obtain

$$|e_q(t)|^2 \le \frac{4c_1^2 L_m^2 e^{(c_1^2 M_1^2 \nu^{-1} t)/4}}{\nu^2 \lambda_{m+1}} |||\varepsilon^l(t)|||^2 |||A^{\frac{1}{2}} \varepsilon^l(t)|||^2. \tag{4.12}$$

Then the combination of (4.12) and (4.10) admits

$$|e(t)|^2 \le \frac{20c_1^2 L_m^4 e^{(c_1^2 M_1^2 \nu^{-1} t)/2}}{\nu^2 \lambda_{m+1}} |||\varepsilon^l(t)|||^2 |||A^{\frac{1}{2}} \varepsilon^l(t)|||^2. \tag{4.13}$$

To complete the proof, we have to estimate the two factors on the right-hand side of (4.13) for $2 \le k \le l$. First of all, from (3.2) we find that $\varepsilon^k(t)$ satisfies

$$\begin{aligned}
&\frac{d\varepsilon^k}{dt} + \nu A \varepsilon^k + Q_m B(v, \varepsilon^k) + Q_m B(\varepsilon^k, v) + Q_m B(w^{k-1}, \varepsilon^k) \\
&\qquad + Q_m B(\varepsilon^k, w^{k-1}) + Q_m B(\varepsilon^{k-1}, \varepsilon^{k-1}) = 0 \quad \forall 2 \le k \le l.
\end{aligned} \tag{4.14}$$

Multiplying (4.14) by $2\varepsilon^k$ and integrating it on Ω gives

$$\frac{d|\varepsilon^k|^2}{dt} + 2\nu |A^{\frac{1}{2}} \varepsilon^k|^2 \le 2|b(\varepsilon^k, v, \varepsilon^k)| + 2|b(\varepsilon^k, w^{k-1}, \varepsilon^k)| + 2|b(\varepsilon^{k-1}, \varepsilon^{k-1}, \varepsilon^k)|.$$

We can summarize the estimates of the right-hand-side terms of this inequality as

$$\begin{aligned}
2|b(\varepsilon^k, v, \varepsilon^k)| &\le c_1|A^{\frac14}\varepsilon^k|^2|A^{\frac12}v| \le c_1M_1\lambda_{m+1}^{-\frac12}|A^{\frac12}\varepsilon^k|^2,\\
2|b(\varepsilon^k, w^{k-1}, \varepsilon^k)| &\le c_1|A^{\frac14}\varepsilon^k|^2|A^{\frac12}w^{k-1}| \le c_1M_1\lambda_{m+1}^{-\frac12}|A^{\frac12}\varepsilon^k|^2,\\
2|b(\varepsilon^{k-1}, \varepsilon^{k-1}, \varepsilon^k)| &\le c_1|\varepsilon^{k-1}|\,|A^{\frac12}\varepsilon^{k-1}|\,|A^{\frac12}\varepsilon^k| \le \frac{\nu}{3}|A^{\frac12}\varepsilon^k|^2 + \frac{3c_1^2}{4\nu}|\varepsilon^{k-1}|^2|A^{\frac12}\varepsilon^{k-1}|^2,
\end{aligned}$$

where, in the first inequality of the last expression above, we used the interpolation inequality in (2.2) with $s = \frac14$ and $m = \frac12$.

By choosing m large enough such that (3.6) is valid, we have

$$\frac{d|\varepsilon^k|^2}{dt} + \nu|A^{\frac12}\varepsilon^k|^2 \le \frac{3c_1^2}{4\nu}|\varepsilon^{k-1}|^2|A^{\frac12}\varepsilon^{k-1}|^2.$$

Finally, we get

$$(4.15)\qquad |\varepsilon^k(t)|^2 \le \frac{3c_1^2}{4\nu^2\lambda_{m+1}}|||\varepsilon^{k-1}(t)|||^2|||A^{\frac12}\varepsilon^{k-1}(t)|||^2 \quad \forall t \ge 0,\ 2 \le k \le l.$$

Now, let us estimate $|A^{\frac12}\varepsilon^k(t)|^2$. Multiplying (4.14) by $2A\varepsilon^k$ and integrating it on Ω, we obtain

$$\begin{aligned}
&\frac{d|A^{\frac12}\varepsilon^k|^2}{dt} + 2\nu|A\varepsilon^k|^2\\
&\qquad \le 2|b(v + w^{k-1}, \varepsilon^k, A\varepsilon^k)| + 2|b(\varepsilon^k, v + w^{k-1}, A\varepsilon^k)| + 2|b(\varepsilon^{k-1}, \varepsilon^{k-1}, A\varepsilon^k)|.
\end{aligned}$$

For the three terms on the right-hand side, we have

$$\begin{aligned}
2|b(v + w^{k-1}, \varepsilon^k, A\varepsilon^k)| &\le c_1(|v|_\infty|A^{\frac12}\varepsilon^k| + |A^{\frac14}w^{k-1}|\,|A^{\frac34}\varepsilon^k|)|A\varepsilon^k| \le \frac{c_1M_1L_m}{\lambda_{m+1}^{\frac12}}|A\varepsilon^k|^2,\\
2|b(\varepsilon^k, v + w^{k-1}, A\varepsilon^k)| &\le c_1|\varepsilon^k|_\infty|A^{\frac12}(v + w^{k-1})|\,|A\varepsilon^k| \le c_1M_1\lambda_{m+1}^{-\frac12}|A\varepsilon^k|^2,\\
2|b(\varepsilon^{k-1}, \varepsilon^{k-1}, A\varepsilon^k)| &\le c_1|\varepsilon^{k-1}|_\infty|A^{\frac12}\varepsilon^{k-1}|\,|A\varepsilon^k| \le c_1\lambda_{m+1}^{-\frac12}|A^{\frac12}\varepsilon^{k-1}|\,|A\varepsilon^{k-1}|\,|A\varepsilon^k|\\
&\le \frac{\nu}{3}|A\varepsilon^k|^2 + \frac{3c_1^2}{4\nu\lambda_{m+1}}|A^{\frac12}\varepsilon^{k-1}|^2|A\varepsilon^{k-1}|^2.
\end{aligned}$$

Thanks to (3.6) again, we derive

$$(4.16)\qquad \frac{d|A^{\frac12}\varepsilon^k|^2}{dt} + \frac{\nu\lambda_{m+1}}{2}|A^{\frac12}\varepsilon^k|^2 + \frac{\nu}{2}|A\varepsilon^k|^2 \le \frac{3c_1^2}{4\nu\lambda_{m+1}}|A^{\frac12}\varepsilon^{k-1}|^2|A\varepsilon^{k-1}|^2.$$

Integrating the above inequality on $[0, t]$ yields

$$\begin{aligned}
&|A^{\frac12}\varepsilon^k(t)|^2 + \frac{\nu}{2}\int_0^t e^{-\nu\lambda_{m+1}(t-s)/2}|A\varepsilon^k(s)|^2ds\\
&\quad \le e^{-\nu\lambda_{m+1}t/2}|A^{\frac12}\varepsilon^k(0)|^2 + \frac{3c_1^2|||A^{\frac12}\varepsilon^{k-1}(t)|||^2}{4\nu\lambda_{m+1}}\int_0^t e^{-\nu\lambda_{m+1}(t-s)/2}|A\varepsilon^{k-1}(s)|^2ds.
\end{aligned}$$

Thanks to (3.3), we have for $2 \le k \le l$

$$\left\{\begin{aligned} &|A^{\frac{1}{2}}\varepsilon^k(t)|^2 \le \frac{c_1^2 |||A^{\frac{1}{2}}\varepsilon^{k-1}(t)|||^2}{\nu\lambda_{m+1}} \int_0^t e^{-\nu\lambda_{m+1}(t-s)/2} |A\varepsilon^{k-1}(s)|^2 ds,\\ &\frac{\nu}{2}\int_0^t e^{-\nu\lambda_{m+1}(t-s)/2}|A\varepsilon^k(s)|^2 ds\\ &\qquad \le \frac{c_1^2 |||A^{\frac{1}{2}}\varepsilon^{k-1}(t)|||^2}{\nu\lambda_{m+1}} \int_0^t e^{-\nu\lambda_{m+1}(t-s)/2}|A\varepsilon^{k-1}(s)|^2 ds. \end{aligned}\right. \tag{4.17}$$

By (4.15) and (4.17), if we define

$$\begin{aligned} &a_k = |||\varepsilon^k|||^2, \quad b_k = |||A^{\frac{1}{2}}\varepsilon^k|||^2,\\ &c_k = \nu\int_0^t e^{-\nu\lambda_{m+1}(t-s)/2}|A\varepsilon^k(s)|^2 ds, \quad \alpha = \frac{c_1^2}{\nu^2\lambda_{m+1}} < 1, \end{aligned}$$

we have for $2 \le k \le l$

$$a_k \le \alpha a_{k-1}b_{k-1}, \quad b_k \le \alpha b_{k-1}c_{k-1}, \quad c_k \le 2\alpha b_{k-1}c_{k-1}. \tag{4.18}$$

From the last two inequalities in (4.18), we see that

$$b_k c_k \le 2\alpha^2 (b_{k-1}c_{k-1})^2.$$

So we get for $2 \le k \le l$

$$b_k \le \left\{\begin{array}{ll} \alpha b_1 c_1 & (k=2)\\ \alpha\left[\prod_{i=0}^{k-3}(2\alpha^2)^{2^i}\right](b_1c_1)^{2^{k-2}} & (k \ge 3) \end{array}\right\} \le 2^{-\frac{1}{2}}(\sqrt{2}\alpha)^{2^{k-1}-1}(b_1c_1)^{2^{k-2}}. \tag{4.19}$$

The inequality for b_2 is obvious, and the one for b_k is obtained as follows: for $3 \le k \le l$

$$\begin{aligned} b_k &\le \alpha b_{k-1}c_{k-1} \le \alpha\cdot(2\alpha^2)^{2^0}(b_{k-2}c_{k-2})^{2^1} \le \alpha\cdot(2\alpha^2)^{2^0}\cdot(2\alpha^2)^{2^1}(b_{k-3}c_{k-3})^{2^2} \le \cdots\\ &\le \alpha\cdot(2\alpha^2)^{2^0+2^1+\cdots+2^{k-3}}(b_1c_1)^{2^{k-2}} = \alpha\left[\prod_{i=0}^{k-3}(2\alpha^2)^{2^i}\right](b_1c_1)^{2^{k-2}}. \end{aligned}$$

Thanks to inequality (4.18), we have for $2 \le k \le l$

$$a_k b_k \le \alpha a_{k-1}b_{k-1}b_k \le \alpha^2 a_{k-2}b_{k-2}b_{k-1}b_k \le \cdots \le \alpha^{k-1}\left[\prod_{i=2}^{k} b_i\right](a_1b_1).$$

Then by using (4.19), we can finally get

$$\begin{aligned} a_k b_k &\le \alpha^{k-1}\prod_{i=2}^{k}\left[2^{-\frac{1}{2}}(\sqrt{2}\alpha)^{2^{i-1}-1}(b_1c_1)^{2^{i-2}}\right](a_1b_1)\\ &= \alpha^{k-1}\cdot 2^{-\frac{k-1}{2}}\cdot(\sqrt{2}\alpha)^{\sum_{i=2}^{k}(2^{i-1}-1)}\\ &\quad\times(b_1c_1)^{\sum_{i=2}^{k}2^{i-2}}\cdot(a_1b_1) \le 2^{1-k}(\sqrt{2}\alpha)^{2^k-2}a_1b_1(b_1c_1)^{2^{k-1}-1}. \end{aligned} \tag{4.20}$$

From the results of Theorem 3.2 and Lemma 4.1, we can finally get the result of the theorem from (4.13) and (4.18)–(4.20) by simple calculations. ∎

For H^1-error estimates, we have the following result.

THEOREM 4.2. *Under the same conditions of Theorem* 4.1*, we have for* $l \geq 1$

$$|A^{\frac{1}{2}}e(t)| \leq 4M_1(c_1K_1L_m^5\nu^{-1})^{\frac{1}{2}}\left(\frac{2^{\frac{1}{4}}c_1K_1L_m}{\nu\lambda_{m+1}}\right)^{2^l-\frac{1}{2}}\exp{(c_3t)} \quad \forall t \geq 0,$$

where $c_3 = 2c_1^3\nu^{-2}M_1^3 + 2^{-1}\nu$ *and* M_1, K_1 *are constants defined in Theorems* 3.1 *and* 3.2.

Proof. Multiplying (4.5) by $2Ae$ and integrating it on Ω yields

$$\frac{d|A^{\frac{1}{2}}e|^2}{dt} + 2\nu|Ae|^2 \leq 2|b(e,u,Ae)| + 2|b(u^l,e,Ae)| + 2|b(\varepsilon^l,\varepsilon^l,Ae)|.$$

Thanks to (2.2), (2.4), and (3.5), we have

$$\begin{aligned}2|b(e,u,Ae)| + 2|b(u^l,e,Ae)| &\leq c_1M_1|e|_\infty|Ae| + c_1M_1|A^{\frac{1}{2}+\frac{1}{3}}e|\,|Ae|\\ &\leq 2c_1M_1|A^{\frac{1}{2}}e|^{\frac{2}{3}}|Ae|^{\frac{4}{3}} \leq \nu|Ae|^2 + \frac{2c_1^3M_1^3}{\nu^2}|A^{\frac{1}{2}}e|^2,\\ 2|b(\varepsilon^l,\varepsilon^l,Ae)| &\leq c_1|\varepsilon^l|_\infty|A^{\frac{1}{2}}\varepsilon^l|\,|Ae| \leq c_1|\varepsilon^l|^{\frac{1}{2}}|A\varepsilon^l|^{\frac{1}{2}}|A^{\frac{1}{2}}\varepsilon^l|\,|Ae|\\ &\leq c_1\lambda_{m+1}^{-\frac{1}{2}}|A^{\frac{1}{2}}\varepsilon^l|\,|A\varepsilon^l|\,|Ae| \leq \nu|Ae|^2 + \frac{c_1^2}{4\nu\lambda_{m+1}}|A^{\frac{1}{2}}\varepsilon^l|^2|A\varepsilon^l|^2.\end{aligned}$$

Therefore

$$\frac{d|A^{\frac{1}{2}}e|^2}{dt} - \frac{2c_1^3M_1^3}{\nu^2}|A^{\frac{1}{2}}e|^2 \leq \frac{c_1^2}{4\nu\lambda_{m+1}}|||A^{\frac{1}{2}}\varepsilon^l(t)|||^2|A\varepsilon^l(t)|^2.$$

Integrating the above inequality admits

$$|A^{\frac{1}{2}}e(t)|^2 \leq \frac{c_1^2|||A^{\frac{1}{2}}\varepsilon^l(t)|||^2}{4\nu\lambda_{m+1}}\int_0^t e^{\frac{2c_1^3M_1^3(t-s)}{\nu^2}}|A\varepsilon^l(s)|^2ds.$$

Defining $c_3 = 2c_1^3\nu^{-2}M_1^3 + 2^{-1}\nu$, we can deduce from the above inequality that

$$|A^{\frac{1}{2}}e(t)|^2 \leq \frac{c_1^2e^{c_3t}|||A^{\frac{1}{2}}\varepsilon^l(t)|||^2}{4\nu\lambda_{m+1}}\int_0^t e^{-\frac{\nu(t-s)}{2}}|A\varepsilon^l(s)|^2ds. \tag{4.21}$$

Similarly to the proof of Theorem 4.1, we estimate $|A^{\frac{1}{2}}e_p|^2$ and $|A^{\frac{1}{2}}e_q|^2$, respectively. Noticing (2.4) and (4.10), we have

$$\begin{aligned}|A^{\frac{1}{2}}e_p(t)|^2 \leq \lambda_{m+1}|e_p|^2 &\leq 4L_m^2\lambda_{m+1}e^{\frac{c_1^2M_1^2\nu^{-1}t}{4}}|||e_q(t)|||^2\\ &\leq 4L_m^2e^{\frac{c_1^2M_1^2\nu^{-1}t}{4}}|||A^{\frac{1}{2}}e_q(t)|||^2.\end{aligned} \tag{4.22}$$

Multiplying (4.8) with $2Ae_q$, integrating it on Ω, doing some estimates of the corresponding trilinear terms, and then integrating the final differential inequality of $|A^{\frac{1}{2}}e_q|^2$ on $[0,t]$, we can finally get

$$\begin{aligned}|A^{\frac{1}{2}}e_q|^2 \leq{}& \frac{5c_1^2L_m^2M_1^2}{\nu^2\lambda_{m+1}}|||A^{\frac{1}{2}}e_p(t)|||^2\\ &+ \frac{5c_1^2|||A^{\frac{1}{2}}\varepsilon(t)^l|||^2}{\nu\lambda_{m+1}}\int_0^t e^{-\nu\lambda_{m+1}(t-s)/2}|A\varepsilon^l(s)|^2ds.\end{aligned} \tag{4.23}$$

Combination of (4.21)–(4.23) admits

$$(4.24)\quad |A^{\frac{1}{2}}e(t)|^2 \le \frac{10c_1^4 M_1^2 L_m^4 e^{2c_3 t} |||A^{\frac{1}{2}}\varepsilon^l(t)|||^2}{\nu^3\lambda_{m+1}^2} \times \left(\int_0^t e^{-\frac{\nu(t-s)}{2}} |A\varepsilon^l(s)|^2 ds + \lambda_{m+1}\int_0^t e^{-\frac{\nu\lambda_{m+1}(t-s)}{2}} |A\varepsilon^l(s)|^2 ds \right).$$

Let us define

$$b_k = |||A^{\frac{1}{2}}\varepsilon^l(t)|||^2, \quad c_k = \nu\int_0^t e^{-\frac{\nu\lambda_{m+1}(t-s)}{2}} |A\varepsilon^k(s)|^2 ds, \quad \alpha = \frac{c_1^2}{\nu^2\lambda_{m+1}},$$

and

$$d_k = \nu\int_0^t e^{-\nu(t-s)/2} |A\varepsilon^k(s)|^2 ds.$$

We have from (4.17)–(4.18) and (4.24) that

$$|A^{\frac{1}{2}}e(t)|^2 \le 10M_1^2 L_m^4 e^{2c_3 t}\alpha^2 b_l(d_l + \lambda_{m+1}c_l),$$

$$b_k \le \alpha b_{k-1}c_{k-1}, \quad c_k \le 2\alpha b_{k-1}c_{k-1}, \quad d_k \le 2\alpha b_{k-1}d_{k-1}.$$

The last inequality is obvious if we substitute $\frac{\nu\lambda_{m+1}}{2}|A^{\frac{1}{2}}\varepsilon^k|^2$ with $\frac{\nu}{2}|A^{\frac{1}{2}}\varepsilon^k|^2$ in (4.16). From the last two inequalities we have for $l \ge 1$

$$\begin{aligned} d_l + \lambda_{m+1}c_l &\le 2\alpha b_{l-1}d_{l-1} + 2\alpha\lambda_{m+1}b_{l-1}c_{l-1} = 2\alpha b_{l-1}(d_{l-1} + \lambda_{m+1}c_{l-1}) \\ &\le (2\alpha)^2 b_{l-1}b_{l-2}(d_{l-2} + \lambda_{m+1}c_{l-2}) \le \cdots \le (2\alpha)^{l-1}\left[\prod_{i=1}^{l-1} b_i\right](d_1 + \lambda_{m+1}c_1). \end{aligned}$$

Now it is easy to get for $l \ge 1$

$$(4.25)\qquad b_l(d_l + \lambda_{m+1}c_l) \le (2\alpha)^{l-1}\left(\prod_{i=1}^{l} b_i\right)(d_1 + \lambda_{m+1}c_1).$$

Thanks to (4.19) and (4.25) we have for $l \ge 1$

$$\begin{aligned} b_l(d_l + \lambda_{m+1}c_l) &\le (2\alpha)^{l-1} b_1 \left[\prod_{i=2}^{l} 2^{-\frac{1}{2}}(\sqrt{2}\alpha)^{2^{i-1}-1}(b_1c_1)^{2^{i-2}}\right](d_1 + \lambda_{m+1}c_1) \\ &\le \sqrt{2}(2\alpha)^{l-1} b_1 (b_1c_1)^{-2^{-1}} \left[\prod_{i=1}^{l} 2^{-\frac{1}{2}}(\sqrt{2}\alpha)^{2^{i-1}-1}(b_1c_1)^{2^{i-2}}\right](d_1 + \lambda_{m+1}c_1) \\ &\le 2^{-\frac{1}{2}}(\sqrt{2}\alpha)^{2^l-2}(b_1c_1)^{2^{l-1}-\frac{1}{2}}(d_1 + \lambda_{m+1}c_1). \end{aligned}$$

Combining the above inequality with the results of Theorem 3.2, Lemma 4.1, and (4.24) leads to the result of the theorem. □

Remark 2. Compared with NGM (1.8) and PPGM (1.10), the small eddy correction method (3.1)–(3.3) involves the self evolution of the small eddy components as well as the interaction between the large and small eddies. Therefore it is suitable for approximating the NSE whenever the small eddy components change slowly or rapidly (in this case, approximating the small eddy equation with the steady Stokes equation in both NGM (1.8) and PPGM (1.10) is not suitable) and can be expected to yield a more accurate approximation. Actually, the result of Theorem 4.1 shows that u^1 can improve the L^2-accuracy of both NGM and PPGM approximations for almost half order ($\lambda_{m+1}^{-\frac{1}{2}}$).

As is said in Remark 1, u^1 is the ONG-approximate solution given in [11], in which the authors imposed some rigorous conditions on the data of the NSE, for example, $a \in D(A)$ and $f \in L^\infty(\mathcal{R}^+, V)$, and for the periodic case proved for the semidiscrete formula (3.7)–(3.9) (see (101) and (102) in [11]) that

$$|u - u_{mM}^1| \leq c(t)(\lambda_{m+1}^{-2} + \lambda_{M+1}^{-1}).$$

From the result of Theorem 4.1, it is obvious that our conditions on the data are much weaker. Indeed we demand only that $a \in V$ and $f \in L^\infty(\mathcal{R}^+, H)$. And we get almost the same estimate for both periodic and nonslip boundary conditions except for the logarithm term L_m^4, which increases very slowly as $m \to \infty$ and can be regarded as a constant in general circumstances. A more important thing is that we provide a successive correction procedure which can double the convergence rate of the previous approximate solution just as the Newton method does for elliptic problems. Of course, the larger l is, the more accurate the approximate solution is. On the other hand, Theorem 4.1 tells us that for any fixed $T > 0$ and $t \in [0, T]$

$$|u(t) - u^l(t)| \to 0 \qquad \text{as } l \to \infty.$$

That is, to ensure that $u^l(t)$ approximates $u(t)$ uniformly (with respect to certain prescribed error bound) in certain fixed time interval $[0, T]$, we have two choices: enlarge m or choose a large l.

Remark 3. In this paper, we consider only the continuous small eddy correction method. But some intermediate steps in the proofs of Theorems 4.1 and 4.2, i.e., (4.10)–(4.11) and (4.22)–(4.23), give us some suggestions for constructing its full discrete form. Actually, the four inequalities listed above show that the total error $|e|$ (or $|A^{\frac{1}{2}}e|$) of the scheme can be controlled by its small eddy error $|e_q|$ (or $|A^{\frac{1}{2}}e_q|$), and the accuracy of $|e_q|$ (or $|A^{\frac{1}{2}}e_q|$) is always a half-order higher than the large eddy error $|e_p|$ (or $|A^{\frac{1}{2}}e_p|$). To balance this kind of difference and make the full discrete algorithm more effective, one possible choice is to use different time step lengths for large eddy and small eddy equations.

5. Numerical examples. In this section, we will present some numerical examples of the small eddy correction method for dissipative evolutionary PDEs.

The small eddy correction method (3.1)–(3.3), proposed in section 3, is a time continuous scheme and is defined in an infinite-dimensional Hilbert space. In practice, we have to construct its full discrete formulations, that is, to restrict it in a finite-dimensional subspace (for example, see (3.7)–(3.9)) and do time discretization by a finite difference scheme. Of course, we have to investigate its corresponding numerical stability, error analysis, and possible multilevel scheme in time discretization, as was pointed out in Remark 3. We will address these questions elsewhere.

As the first numerical example, we integrate the following one-dimensional Burger equation with the homogeneous Dirichlet boundary condition on $[0, \pi]$. Using notation similar to that of the NSE, we have

$$\frac{du}{dt} + \nu A u + B(u, u) = f, \qquad u(0) = u_0, \tag{5.1}$$

where, in this case, $A = -\frac{\partial^2}{\partial x^2}$ with domain $D(A) = H^2(0,\pi) \cap H_0^1(0,\pi)$ and $B(u,v) = \frac{2}{3}uv_x + \frac{1}{3}u_x v$ for $u, v \in H_0^1(0,\pi)$. The eigenfunctions of A are $\phi_i = \sqrt{2/\pi}\sin(ix)$ with corresponding eigenvalues $\lambda_i = i^2$, $i = 1, 2, \dots$.

We choose an exact solution $u_e(x,t)$ and then compute the time dependent forcing term $f(x,t)$. This makes it possible to compare the numerical solutions with the exact solution without computing a large Galerkin approximation as an "exact" solution. We choose

$$u_e(x,t) = \sum_{j=1}^{\infty} \frac{a_j(t)}{j^3}\sin(jx), \quad a_j(t) = 1 + \gamma\sin(j^2 t).$$

To give a numerical implementation of the small eddy correction for (5.1), we use a spectral method for the space discretization, and the Euler backward scheme for the time discretization, with time step length $\tau > 0$. For any two positive integers $M \gg m$, we have the following two finite-dimensional subspaces:

$$\begin{aligned} H_m &= P_m H = \{\phi_1, \phi_2, \dots, \phi_m\} \quad \text{and} \\ \hat{H}_{mM} &= P_{mM} H = \{\phi_{m+1}, \phi_{m+2}, \dots, \phi_M\}. \end{aligned}$$

Then the corresponding numerical scheme for (5.1) reads: find $v_{n+1} \in H_m$ and $w_{n+1}^k \in \hat{H}_{mM}$, for $k = 1, 2, \dots, l$, such that

$$v_{n+1} - v_n + \nu\tau A v_{n+1} + \tau P_m B(v_{n+1} + w_n^l, v_{n+1} + w_n^l) = \tau P_m f((n+1)\tau), \tag{5.2}$$

$$\begin{aligned} w_{n+1}^k - w_n^k &+ \nu\tau A w_{n+1}^k + \tau P_{mM}[B(v_{n+1}, v_{n+1}) + B(v_{n+1}, w_{n+1}^k) \\ &+ B(w_{n+1}^k, v_{n+1}) + B(w_{n+1}^{k-1}, w_{n+1}^k) + B(w_{n+1}^k, w_{n+1}^{k-1})] \\ &= \tau P_{mM}[f((n+1)\tau) + B(w_{n+1}^{k-1}, w_{n+1}^{k-1})] \qquad \forall 1 \le k \le l, \end{aligned} \tag{5.3}$$

$$v_0 = P_m u_0,\ w_0^i = P_{mM} u_0,\ i = 1, 2, \dots, l,\ l \ge 1, \ \text{ and } \ w_n^0 = 0 \ \text{ for } \ n \ge 0. \tag{5.4}$$

Owing to our limited computing resources, we compute only the u^0 on H_m, u^1 and u^2 on $[H_m, H_M]$. For fixed $m = 2$, we choose a suitable $M > m$ according to the principle that the error of u^1 on $[H_m, H_M]$ will decrease no further when M increases. Here we choose $M = 254$. The time step length is $\tau = 0.001$, and it is determined in a similar manner. That is, for fixed m and M, decreasing k will not improve the accuracy of u^1 any further. Then we can regard the error as mainly determined by the space discretization. We computed the u^0, u^1, and u^2 in time interval $[0, 2]$ for $\nu = 1.0$. Furthermore, as a comparison, we also computed another u^0 on H_{m+M} in this interval. Following are the L^2-error comparison graph (Figure 5.1) and the CPU time table (Table 5.1), which indicates the CPU time used for deriving u^0, u^1, u^2 on H_m, $[H_m, H_M]$ and a large scale u^0 on H_{m+M}.

Note that the curves which represent the error of u^2 on $[H_m, H_M]$ and u^0 on H_{m+M} coincide.

As the second numerical example, we consider the following two-dimensional NSEs in the bounded domain $\Omega = [0,1] \times [0,1]$:

$$\begin{cases} \dfrac{\partial u}{\partial t} - \nu\Delta u + (u \cdot \nabla)u + \nabla p = f, \\ \text{div } u = 0, \quad u|_{t=0} = 0, \\ \text{periodic boundary condition,} \end{cases} \tag{5.5}$$

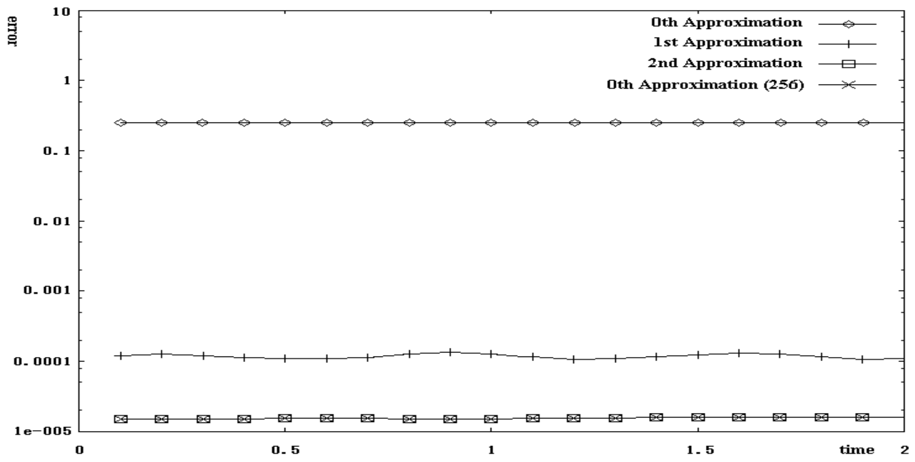


FIG. 5.1. *L^2-error comparison.*

TABLE 5.1
CPU time comparison.

u^0 on H_m	u^1 on $[H_m, H_M]$	u^2 on $[H_m, H_M]$	u^0 on H_{m+M}
0.44 sec	1328.83 sec	7469.48 sec	10054.39 sec

where $\nu = 0.01$ is the kinetic viscosity and $f(t, x, y) = f_1(x, y)(2 + \cos(t))/3$ is the density of the external forces, where $f_1(x, y) = f_1(r, \phi) = (0, f_\phi)^T$,

$$f_\phi(r,\phi) = \begin{cases} \dfrac{1}{8r_+}\displaystyle\int_0^{r_+} \rho(1+\cos(4\rho))^2 d\rho & \text{if } r_+ < \frac{1}{8}, \\ -\dfrac{1}{8r_-}\displaystyle\int_0^{r_-} \rho(1+\cos(4\rho))^2 d\rho & \text{if } r_- < \frac{1}{8}, \\ 0 & \text{otherwise,} \end{cases} \tag{5.6}$$

$r_\pm = |x + iy - (\frac{1}{2}(1+i) \pm \frac{1}{4}e^{i\theta})|$, and $\theta = 0.7$. This external force f represents stirring the fluid in opposite directions at the locations $\frac{1}{2}(1+i) \pm \frac{1}{4}e^{i\theta}$.

In this particular case, if we denote $k = (k_1, k_2)^T \in Z^2$,

$$L^2(\Omega)^2 = \left\{ \phi = \sum_{k\in Z^2, k\neq 0} c_k e^{2\pi i k\cdot x},\ c_k = \overline{c_{-k}},\ \sum_{k\in Z^2, k\neq 0} |c_k|^2 < \infty \right\},$$

$H = PL^2(\Omega)^2$, and $H_m = P_m H$, where, for any $\phi = \sum_{k\in Z^2, k\neq 0} c_k e^{2\pi i k\cdot x} \in L^2(\Omega)^2$,

$$P\phi = \sum_{k\in Z^2, k\neq 0} \left(I - \frac{k\cdot k^T}{|k|^2} \right) c_k e^{2\pi i k\cdot x} \quad \text{and} \quad P_m P\phi = \sum_{0<|k|^2\leq m} \left(I - \frac{k\cdot k^T}{|k|^2} \right) c_k e^{2\pi i k\cdot x}.$$

Projecting the above equations onto H, we can get its functional form (1.1), and its fully discrete form is completely the same as (5.2)–(5.4). Considering the computing scale, we take

$$m = 9^2, \quad M = 19^2, \quad \text{and} \quad \tau = 0.005.$$

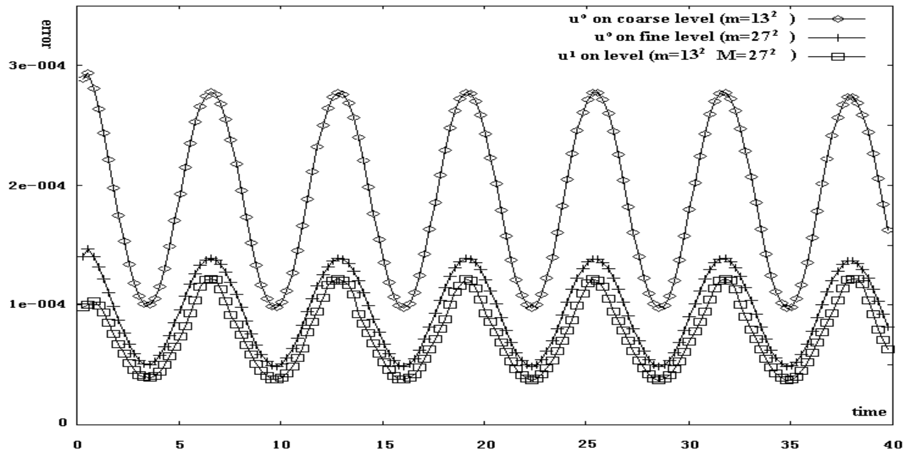
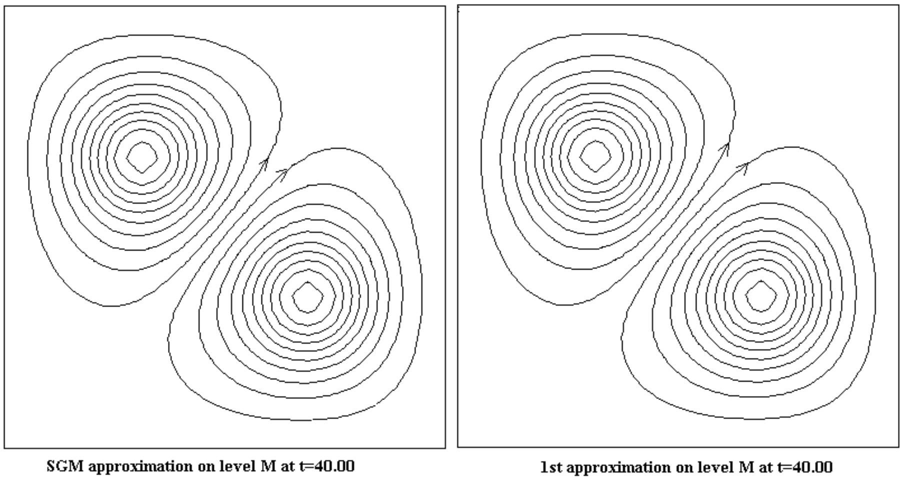


FIG. 5.2. *H^1-error comparison.*


SGM approximation on level M at t=40.00

1st approximation on level M at t=40.00

FIG. 5.3. *H^1-streamline graphs.*

We compute only u^0 (with $|k_1|, |k_2| \leq \sqrt{m}$) and u^1 (with $|k_1|, |k_2| \leq \sqrt{M}$). To get the error of these numerical results, we compute another u^0 (with $|k_1|, |k_2| \leq \sqrt{\tilde{M}}$) on a larger finite-dimensional subspace $H_{\tilde{M}}$ with $\tilde{M} = 39^2$ and regard it as the "exact" solution.

In Figures 5.2 and 5.3, we give the H^1-error curves of u^0 on H_m and H_M and u^1 on $[H_m, H_M]$, and the streamline graphs of u^0 on H_M and u^1 on $[H_m, H_M]$ at $t = 40$. Here the CPU time used by u^1 on $[H_m, H_M]$ is less than one-half of the CPU time used by u^0 on H_M.

From the error comparison of Figures 5.1 and 5.2, we can easily find that both u^1 and u^2 can greatly improve the accuracy of u^0, the SGM approximation, with less CPU time than the large scale SGM approximation. If the CPU time is what we care about, we would prefer the 1st approximation u^1. However, if we care more about the accuracy of the approximate solution, we prefer a higher order approximation like u^2.

Acknowledgment. We are grateful to the two anonymous referees for their many helpful corrections, suggestions, and comments.

REFERENCES

[1] A. Ait Ou Ammi and M. Marion, *Nonlinear Galerkin methods and mixed finite elements: Two-grid algorithms for the Navier-Stokes equations*, Numer. Math., 68 (1994), pp. 189–213.

[2] H. Brézis and T. Gallouet, *Nonlinear Schrödinger evolution equations*, Nonlinear Anal., 4 (1980), pp. 677–681.

[3] J. R. Cannon, R. E. Ewing, Y. N. He, and Y. P. Lin, *A modified nonlinear Galerkin method for the viscoelastic fluid motion equations*, Internat. J. Engrg. Sci., 37 (1999), pp. 1643–1662.

[4] C. Canuto, M. Y. Hussaini, A. Quarteroni, and T. A. Zang, *Spectral Methods in Fluid Dynamics*, Springer-Verlag, New York, Heidelberg, Berlin, 1987.

[5] C. Devulder and M. Marion, *A class of numerical algorithms for large time integration: The nonlinear Galerkin methods*, SIAM J. Numer. Anal., 29 (1992), pp. 462–483.

[6] C. Devulder, M. Marion, and E. S. Titi, *On the rate of convergence of nonlinear Galerkin methods*, Math. Comp., 60 (1993), pp. 495–514.

[7] C. Foias, O. Manley, and R. Temam, *Modeling of the interaction of small and large eddies in two-dimensional turbulent flow*, Math. Model. Numer. Anal., 22 (1988), pp. 93–114.

[8] B. Garcia-Archilla, J. Novo, and E. S. Titi, *An approximate inertial manifolds approach to postprocessing the Galerkin method for the Navier-Stokes equations*, Math. Comp., 68 (1999), pp. 893–911.

[9] Y. He and K. Li, *Convergence and stability of finite element nonlinear Galerkin method for the Navier-Stokes equations*, Numer. Math., 79 (1998), pp. 77–106.

[10] Y. He and K. Li, *Taylor expansion algorithm for the nonlinear operator equations*, Acta Math. Sinica, 41 (1998), pp. 317–326.

[11] Y. He, Y. Hou, and K. Li, *Stability and convergence of optimum spectral non-linear Galerkin methods*, Math. Methods Appl. Sci., 24 (2001), pp. 298–317.

[12] J. G. Heywood and R. Rannacher, *On the question of turbulence modeling by approximate inertial manifolds and the nonlinear Galerkin method*, SIAM J. Numer. Anal., 30 (1993), pp. 1603–1621.

[13] K. Li and Y. Hou, *An AIM and one step Newton method for the Navier-Stokes equations*, Comput. Methods Appl. Mech. Engrg., 190 (2001), pp. 6141–6155.

[14] K. Li and Y. Hou, *Fourier nonlinear Galerkin method for N-S equations*, Discrete Contin. Dynam. Systems, 2 (1996), pp. 497–524.

[15] M. Marion and R. Temam, *Nonlinear Galerkin methods*, SIAM J. Numer. Anal., 26 (1989), pp. 1139–1157.

[16] M. Marion and R. Temam, *Nonlinear Galerkin methods: The finite element case*, Numer. Math., 57 (1990), pp. 205–226.

[17] M. Marion and J. Xu, *Error estimates on a new nonlinear Galerkin method based on two-grid finite elements*, SIAM J. Numer. Anal., 32 (1995), pp. 1170–1184.

[18] J. Shen, *Long time stability and convergence for fully discrete nonlinear Galerkin methods*, Appl. Anal., 38 (1990), pp. 201–229.

[19] J. Shen and R. Temam, *Nonlinear Galerkin method using Chebyshev and Legendre polynomials* I. *The one-dimensional case*, SIAM J. Numer. Anal., 32 (1995), pp. 215–234.

[20] R. Temam, *Stability analysis of the nonlinear Galerkin method*, Math. Comp., 57 (1991), pp. 477–505.

[21] R. Temam, *Navier-Stokes Equations, Theory and Numerical Analysis*, 3rd ed., North–Holland, Amsterdam, 1984.

[22] R. Temam, *Infinite-Dimensional Dynamical Systems in Mechanics and Physics*, Springer-Verlag, New York, 1988.

SIAM J. NUMER. ANAL.
Vol. 41, No. 3, pp. 1131–1162

STABILIZED FINITE ELEMENTS ON ANISOTROPIC MESHES: A PRIORI ERROR ESTIMATES FOR THE ADVECTION-DIFFUSION AND THE STOKES PROBLEMS*

STEFANO MICHELETTI†, SIMONA PEROTTO†, AND MARCO PICASSO‡

Abstract. Stabilized finite elements on strongly anisotropic meshes are considered. The design of the stability coefficients is addressed for both the advection-diffusion and the Stokes problems when using continuous piecewise linear finite elements on triangles. Using the polar decomposition of the Jacobian of the affine mapping from the reference triangle to the current one, K, and from a priori error estimates, a new definition of the stability coefficients is proposed. Our analysis shows that these coefficients do not depend on the element diameter h_K but on a characteristic length associated with K via the polar decomposition. A numerical assessment of the theoretical analysis is carried out.

Key words. anisotropic error estimates, advection-diffusion problems, Stokes problem, stabilized finite elements

AMS subject classifications. 65N12, 65N15, 65N30

DOI. 10.1137/S0036142902403759

1. Introduction.

Stabilized finite elements like the Galerkin least squares (GLS) method are currently widely used in the finite element community. The goal is to employ simple finite element approximations (for instance, continuous piecewise linear) while ensuring stability of the method through extra (consistent) terms in the weak formulation.

Stabilized finite elements have been used, for instance, in [21] for solving the Stokes problem and in [16, 22, 29] for the approximation of the scalar advection-diffusion problem. Extensions to the Navier–Stokes equations have been proposed in [11, 23, 24]. Finally, stabilized finite elements have been also successfully applied to other complex problems such as viscoelastic flows [3, 6, 7], shells [12], magnetohydrodynamics [26], and semiconductors [33].

The critical issue in stabilized finite elements is the design of the so-called stability coefficients weighting the extra terms added to the weak formulation. These coefficients typically depend on some dimensionless numbers usually tuned on benchmark problems and on the local mesh size h_K, K being a mesh element. A theoretical estimation of these quantities is proposed in [25, 28] for isotropic meshes. An alternative approach consists of stabilizing via the residual-free bubble theory [9, 10, 39]. This method has the advantage of providing a self-consistent expression for the stability coefficients with no tuning parameters.

As far as we know, however, few papers have dealt with the design of the stability coefficients for strongly anisotropic meshes. In all cases, these quantities are related to some minimum size associated with each element and whose definition varying from work to work seems to be relevant for the anisotropic analysis. In [2] an anisotropic

*Received by the editors March 11, 2002; accepted for publication (in revised form) January 2, 2003; published electronically July 30, 2003.
http://www.siam.org/journals/sinum/41-3/40375.html

†MOX–Modeling and Scientific Computing, Dipartimento di Matematica "F. Brioschi," Politecnico di Milano, Via Bonardi 9, 20133 Milano, Italy (stefano.micheletti@mate.polimi.it, simona.perotto@mate.polimi.it).

‡Institut de Mathématiques, École Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland (marco.picasso@epfl.ch).

a priori error analysis is provided for the advection–diffusion-reaction problem. It is shown that the height, say $\widetilde{h}_K$, with respect to the diameter of each element K, should be used for the design of the stability coefficient in the case of external boundary layers. The analysis is carried out for finite elements of arbitrary order but is restricted by a maximal angle condition and a coordinate system condition which may be a limitation in the framework of adaptivity. In [31] an alternative approach is proposed showing that $\widetilde{h}_K$ is again the correct choice. In this analysis, however, the interpolation constants depend on the alignment between the stretching direction of the mesh and the solution. In [40] an a posteriori error estimator for anisotropically refined grids as well as interpolation estimates are introduced, in which the minimum size of the elements parallel to the coordinate system plays a major role. In [4, 5] the authors consider the stabilization of the Stokes problem in the case of the Q_1/Q_1 pair of finite elements on anisotropic quadrilateral meshes aligned with the Cartesian coordinate axes. The stabilizing term is a variant to the one considered in the present work and takes into account both mesh spacings along the axes of the coordinate system; see section 4. Furthermore, in [36] numerical experiments show that good results can be obtained when using the minimum edge length instead of h_K.

In this paper an alternative technique is introduced based on the anisotropic interpolation error estimates derived in [18]. The maximal angle and the coordinate system conditions of [2] are avoided. Moreover, the interpolation constants depend only on the reference element (or alternatively on the reference patch), and the information about the alignment between the stretching direction of the mesh and the solution appears explicitly in the right-hand side of the estimates through some anisotropic weightings of the first or second order derivatives of the solution.

Namely, the scalar advection-diffusion and the Stokes problems are addressed with approximations based on continuous piecewise linear finite elements. Following the a priori error analysis of [20, 21, 22] and using the anisotropic interpolation estimates of [18, 19], new definitions of the stability coefficients are proposed. Numerical results confirm the theoretical predictions. Notice that the recipe derived in the present paper has already been employed in anisotropic a posteriori error analyses for both the advection-diffusion and the Stokes problem [37, 17].

The outline of the paper is as follows. In section 2 we introduce the anisotropic framework of [18]. Next, after recalling some of the results derived in [18], we prove some further anisotropic estimates upon which we develop the a priori analysis in sections 3 and 4 for the advection-diffusion and the Stokes problems, respectively. Finally, numerical results are presented in section 5.

2. Anisotropic and functional setting.

Let Ω be a polygonal domain of $\mathbb{R}^2$. For any $0 < h \leq 1$, let $\{\mathcal{T}_h\}_h$ be a family of conforming triangulations of $\overline{\Omega}$ into triangles K of diameter $h_K \leq h$. Since we are working with strongly anisotropic meshes, however, the standard regularity assumption in [13] does not hold in our analysis. Let $T_K : \widehat{K} \to K$ be the affine transformation which maps the reference triangle $\widehat{K}$ into K, where $\widehat{K}$ can be, e.g., the right triangle $(0,0), (1,0), (0,1)$ or the equilateral one $(-1/2, 0), (1/2, 0), (0, \sqrt{3}/2)$. In either case let $M_K \in \mathbb{R}^{2\times 2}$ be the Jacobian of T_K, that is,

$$\mathbf{x} = T_K(\widehat{\mathbf{x}}) = M_K \widehat{\mathbf{x}} + \mathbf{t}_K,$$

with $\mathbf{t}_K \in \mathbb{R}^2$, $\mathbf{x} = (x_1, x_2)^T \in K$, and $\widehat{\mathbf{x}} = (\widehat{x}_1, \widehat{x}_2)^T \in \widehat{K}$. Since M_K is invertible, it admits a unique polar decomposition $M_K = B_K Z_K$, where B_K and Z_K are symmetric

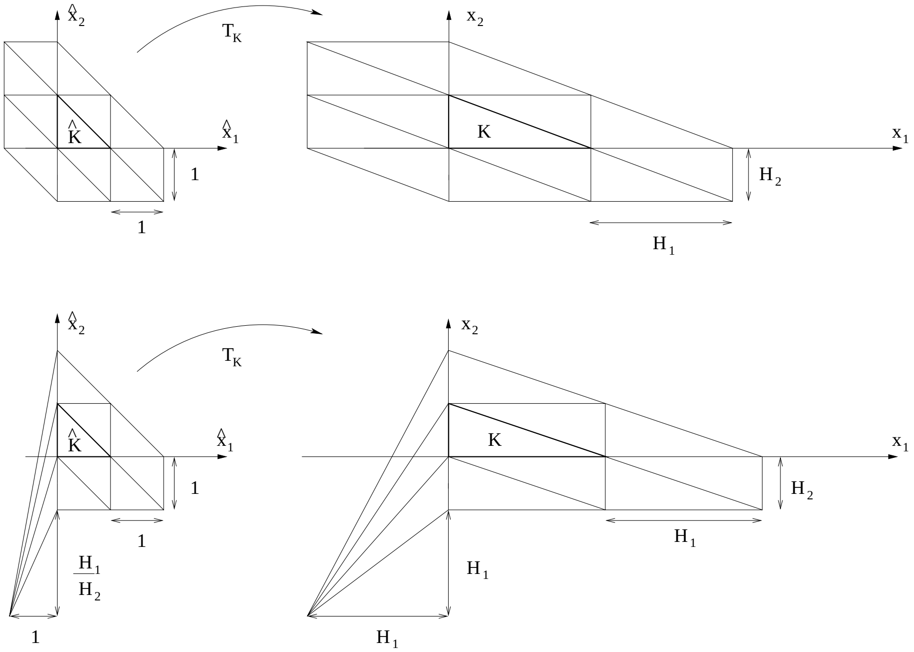


FIG. 2.1. *Example of an acceptable patch (top): the size of the reference patch $\Delta_{\widehat{K}}$ does not depend on the aspect ratio H_1/H_2. Example of a nonacceptable patch (bottom): the size of the reference patch $\Delta_{\widehat{K}}$ now depends on the aspect ratio H_1/H_2. Notice that in both cases $\mathbf{r}_{1,K} = (1,0)^T$, $\mathbf{r}_{2,K} = (0,1)^T$, $\lambda_{1,K} = H_1$, and $\lambda_{2,K} = H_2$.*

positive definite and orthogonal matrices, respectively (see, e.g., [27]). Moreover, B_K can be factorized as $B_K = R_K^T \Lambda_K R_K$, where Λ_K is diagonal with positive entries and R_K is orthogonal. Thus let

$$\Lambda_K = \begin{bmatrix} \lambda_{1,K} & 0 \\ 0 & \lambda_{2,K} \end{bmatrix} \quad \text{and} \quad R_K = \begin{bmatrix} \mathbf{r}_{1,K}^T \\ \mathbf{r}_{2,K}^T \end{bmatrix},$$

where $\lambda_{1,K}$, $\lambda_{2,K}$ (with $\lambda_{1,K} \geq \lambda_{2,K}$) and $\mathbf{r}_{1,K}$, $\mathbf{r}_{2,K}$ are the eigenvalues and the eigenvectors of B_K, respectively.

For any $K \in \mathcal{T}_h$, let us define the stretching factor $s_K = \lambda_{1,K}/\lambda_{2,K} (\geq 1)$, measuring the deformation of K with respect to $\widehat{K}$ for which $s_{\widehat{K}} = 1$, and let Δ_K be the union of all the elements sharing a vertex with K. In view of the use of Clément-type interpolation operators, we assume throughout that the cardinality of any patch Δ_K as well as the diameter of the reference patch $\Delta_{\widehat{K}} = T_K^{-1}(\Delta_K)$ are uniformly bounded independently of the geometry of the mesh; i.e., for any $K \in \mathcal{T}_h$,

$$\text{card}(\Delta_K) \leq \Gamma \quad \text{and} \quad \text{diam}(\Delta_{\widehat{K}}) \leq \widehat{C} \simeq O(1), \tag{2.1}$$

where $\widehat{C} \geq h_{\widehat{K}}$. In particular, the latter hypothesis rules out some too distorted reference patches (see Figure 2.1, where examples of acceptable and nonacceptable patches are shown).

Remark 1. Throughout we express the dependence of any constant through an explicit list. For example, $C = C(\widehat{K})$ is a constant depending only on the geometry of the reference triangle $\widehat{K}$, and $C = C(\Gamma, \widehat{C})$ is a constant taking into account the assumptions (2.1), while C is a number not depending on any geometrical quantity whatsoever. Moreover, notice that in what follows any constant C can take on different values at different occurrences.

Finally, for any function v defined on K, we let $\widehat{v}$ be the corresponding function defined on $\widehat{K}$ via the map T_K, i.e., $\widehat{v}(\widehat{\mathbf{x}}) = v(T_K(\widehat{\mathbf{x}}))$.

Let us introduce now the functional spaces used in what follows. First, let $C^k(\overline{\Omega})$, where the integer $k \geq 0$, be the space of functions with continuous derivatives in $\overline{\Omega}$ up to the kth order. We denote $L^2(\Omega)$ the space of the Lebesgue square-integrable functions with norm $\|\cdot\|_{L^2(\Omega)}$ and scalar product $(\cdot,\cdot)$. Moreover, the space $L^2_0(\Omega)$ denotes the subspace of $L^2(\Omega)$ of functions with zero average over Ω.

Let $W^{k,p}(\Omega)$ be the classical Sobolev spaces of Lebesgue-measurable functions, with k a nonnegative integer and $1 \leq p \leq \infty$ [32].

In the case of scalar valued functions, we denote $H^k(\Omega)$ the space $W^{k,2}(\Omega)$ with norm and seminorm $\|\cdot\|_{H^k(\Omega)}$ and $|\cdot|_{H^k(\Omega)}$, respectively. Then the norm of the space $W^{0,\infty}(\Omega)$, i.e., $L^\infty(\Omega)$, is denoted $\|\cdot\|_{L^\infty(\Omega)}$. When the norms or seminorms refer to some subspace S of Ω, they are written as $\|\cdot\|_{H^k(S)}$, $|\cdot|_{H^k(S)}$, $\|\cdot\|_{L^\infty(S)}$, while the scalar products are denoted $(\cdot,\cdot)_S$.

In the case of functions with values in $\mathbb{R}^2$, we use the same notation for the norms and scalar products as those for the scalar case by replacing the corresponding Sobolev spaces with $(H^k(\Omega))^2$, $(L^2(\Omega))^2$, etc.

Finally, in the case $k=1$ we let $H^1_0(\Omega)$ and $(H^1_0(\Omega))^2$ be the subspaces of $H^1(\Omega)$ and $(H^1(\Omega))^2$, respectively, satisfying homogeneous Dirichlet boundary conditions on the boundary $\partial\Omega$ of Ω.

2.1. Anisotropic interpolation estimates. This subsection provides a continuation of the analysis developed in [18], where anisotropic interpolation error estimates are derived starting from the decompositions described above. In more detail, after recalling some of the results in [18] (see Lemmas 2.1–2.4), we prove some further anisotropic inequalities in view of the convergence analysis of sections 3 and 4 (see Propositions 2.5 and 2.6 and Corollary 2.7).

The results below can be found in the proofs of Lemmas 2.2 and 2.1 in [18].

LEMMA 2.1. *For any function $v \in H^1(\Omega)$ and for any $K \in \mathcal{T}_h$, the relations*

$$|\widehat{v}|_{H^1(\widehat{K})} = \left[s_K \,\|\nabla v \cdot \mathbf{r}_{1,K}\|^2_{L^2(K)} + \frac{1}{s_K}\,\|\nabla v \cdot \mathbf{r}_{2,K}\|^2_{L^2(K)} \right]^{1/2}, \tag{2.2}$$

$$|v|_{H^1(K)} \leq s_K^{1/2} |\widehat{v}|_{H^1(\widehat{K})} \tag{2.3}$$

can be proved.

LEMMA 2.2. *For any function $v \in H^2(\Omega)$ and for any $K \in \mathcal{T}_h$, the following identity holds:*

$$\begin{aligned}|\widehat{v}|^2_{H^2(\widehat{K})} = {} & \frac{\lambda^3_{1,K}}{\lambda_{2,K}}\, L_K(\mathbf{r}_{1,K},\, \mathbf{r}_{1,K};\, v) + \frac{\lambda^3_{2,K}}{\lambda_{1,K}}\, L_K(\mathbf{r}_{2,K},\, \mathbf{r}_{2,K};\, v) \\ & + 2\,\lambda_{1,K}\,\lambda_{2,K} L_K(\mathbf{r}_{1,K},\, \mathbf{r}_{2,K};\, v),\end{aligned}$$

where

$$L_K(\mathbf{r}_{i,K},\, \mathbf{r}_{j,K};\, v) = \int_K \left(\mathbf{r}^T_{i,K}\, H_K(v)\, \mathbf{r}_{j,K}\right)^2 d\mathbf{x}, \quad \textit{with } i,\, j = 1,\, 2, \tag{2.4}$$

and $H_K(v)$ is the Hessian matrix associated with the function v (restricted to K)

$$H_K(v) = \begin{bmatrix} \dfrac{\partial^2 v}{\partial x_1^2} & \dfrac{\partial^2 v}{\partial x_1 \partial x_2} \\ \dfrac{\partial^2 v}{\partial x_1 \partial x_2} & \dfrac{\partial^2 v}{\partial x_2^2} \end{bmatrix}.$$

After introducing the finite element space W_h comprising continuous affine elements, let $r_h : C^0(\overline{\Omega}) \to W_h$ and $R_h : L^2(\Omega) \to W_h$ be the standard Lagrange and Clément linear interpolants, respectively, and let their restrictions to each element $K \in \mathcal{T}_h$ be denoted r_K and R_K. Then the results below can be proved (see Propositions 3.2, 3.1, and 2.1 in [18]).

LEMMA 2.3. *Let $v \in H^1(\Omega)$. Then there exists a constant $C = C(\Gamma, \widehat{C})$ such that, for any $K \in \mathcal{T}_h$,*

$$\|v - R_K(v)\|_{L^2(K)} \leq C\, [\lambda_{1,K}^2 (\mathbf{r}_{1,K}^T\, G_K(v)\, \mathbf{r}_{1,K}) + \lambda_{2,K}^2 (\mathbf{r}_{2,K}^T\, G_K(v)\, \mathbf{r}_{2,K})]^{1/2},$$

$G_K(v)$ being the symmetric positive semidefinite matrix in $\mathbb{R}^{2\times 2}$ given by

$$G_K(v) = \sum_{T\in\Delta_K} \begin{bmatrix} \displaystyle\int_T \left(\frac{\partial v}{\partial x_1}\right)^2 d\mathbf{x} & \displaystyle\int_T \frac{\partial v}{\partial x_1}\frac{\partial v}{\partial x_2}\, d\mathbf{x} \\ \displaystyle\int_T \frac{\partial v}{\partial x_1}\frac{\partial v}{\partial x_2}\, d\mathbf{x} & \displaystyle\int_T \left(\frac{\partial v}{\partial x_2}\right)^2 d\mathbf{x} \end{bmatrix}. \tag{2.5}$$

LEMMA 2.4. *Let $v \in H^2(\Omega)$. Then there exist two constants $C_1 = C_1(\widehat{K})$ and $C_2 = C_2(\widehat{K})$ such that, for any $K \in \mathcal{T}_h$,*

$$\begin{aligned} \|v - r_K(v)\|_{L^2(K)} \leq C_1\, [&\lambda_{1,K}^4\, L_K(\mathbf{r}_{1,K}, \mathbf{r}_{1,K}; v) + \lambda_{2,K}^4\, L_K(\mathbf{r}_{2,K}, \mathbf{r}_{2,K}; v) \\ &+ 2\, \lambda_{1,K}^2 \lambda_{2,K}^2\, L_K(\mathbf{r}_{1,K}, \mathbf{r}_{2,K}; v)]^{1/2}, \end{aligned} \tag{2.6}$$

$$\begin{aligned} |v - r_K(v)|_{H^1(K)} \leq C_2 \Bigg[&\frac{\lambda_{1,K}^4}{\lambda_{2,K}^2}\, L_K(\mathbf{r}_{1,K}, \mathbf{r}_{1,K}; v) + \lambda_{2,K}^2\, L_K(\mathbf{r}_{2,K}, \mathbf{r}_{2,K}; v) \\ &+ 2\, \lambda_{1,K}^2\, L_K(\mathbf{r}_{1,K}, \mathbf{r}_{2,K}; v)\Bigg]^{1/2}, \end{aligned} \tag{2.7}$$

the quantities $L_K(\mathbf{r}_{i,K}, \mathbf{r}_{j,K}; v)$ being defined in (2.4).

We are now ready to prove the new anisotropic results used in the convergence analysis of the advection-diffusion and the Stokes problems.

PROPOSITION 2.5. *For any $v \in H^1(\Omega)$, there exists a constant $C = C(\Gamma, \widehat{C})$ such that, for any $K \in \mathcal{T}_h$,*

$$|v - R_K(v)|_{H^1(K)} \leq C \left[\frac{\lambda_{1,K}^2}{\lambda_{2,K}^2} (\mathbf{r}_{1,K}^T\, G_K(v)\, \mathbf{r}_{1,K}) + (\mathbf{r}_{2,K}^T\, G_K(v)\, \mathbf{r}_{2,K})\right]^{1/2}, \tag{2.8}$$

$G_K(v)$ being the matrix defined in (2.5).

Proof. Applying relation (2.3) to the interpolation error $(v - R_K(v))$, we get

$$|v - R_K(v)|_{H^1(K)} \leq s_K^{1/2}\, |\widehat{v} - R_{\widehat{K}}(\widehat{v})|_{H^1(\widehat{K})}, \tag{2.9}$$

where $[R_K(v)]\hat{} = R_{\widehat{K}}(\widehat{v})$ (see [18] for a proof of this last equality). Exploiting the result

$$|v - R_T(v)|_{H^1(T)} \le C\, |v|_{H^1(\Delta_T)}, \quad \text{for any } v \in H^1(\Omega) \text{ and for any } T \in \mathcal{T}_h,$$

of the theory of Clément [14] in the right-hand side of (2.9) (identifying T with $\widehat{K}$) yields

$$|v - R_K(v)|^2_{H^1(K)} \le C\, s_K\, |\widehat{v}|^2_{H^1(\Delta_{\widehat{K}})} = C\, s_K \sum_{\widehat{T}\in\Delta_{\widehat{K}}} |\widehat{v}|^2_{H^1(\widehat{T})}. \tag{2.10}$$

By applying (2.2) to each seminorm $|\widehat{v}|_{H^1(\widehat{T})}$ in (2.10), we get

$$|v - R_K(v)|^2_{H^1(K)} \le C\, s_K \sum_{T\in\Delta_K} \left(s_K\, \|\nabla v\cdot \mathbf{r}_{1,K}\|^2_{L^2(T)} + \frac{1}{s_K}\, \|\nabla v\cdot \mathbf{r}_{2,K}\|^2_{L^2(T)} \right),$$

that is, observing that

$$\sum_{T\in\Delta_K} \|\nabla v\cdot \mathbf{r}_{i,\,K}\|^2_{L^2(T)} = \mathbf{r}^T_{i,\,K}\, G_K(v)\, \mathbf{r}_{i,\,K} \quad \text{for } i = 1,\, 2, \tag{2.11}$$

the desired inequality (2.8). □

Proposition 2.6 establishes an anisotropic relation between the H^2-seminorms of v and $\widehat{v}$ defined on K and $\widehat{K}$, respectively.

PROPOSITION 2.6. *For any function $v \in H^2(K)$ and for any $K \in \mathcal{T}_h$, the following inequalities hold:*

$$\frac{\lambda^{1/2}_{1,K}\,\lambda^{1/2}_{2,K}}{(\lambda^2_{1,K} + \lambda^2_{2,K})}\, |\widehat{v}|_{H^2(\widehat{K})} \le |v|_{H^2(K)} \le \frac{(\lambda^2_{1,K} + \lambda^2_{2,K})}{\lambda^{3/2}_{1,K}\,\lambda^{3/2}_{2,K}}\, |\widehat{v}|_{H^2(\widehat{K})}. \tag{2.12}$$

Proof. Let us begin by proving the upper bound of (2.12). Let $H_{\widehat{K}}(\widehat{v})$ and $H_K(v)$ be the Hessian matrices associated with $\widehat{v}$ and v and referred to as the elements $\widehat{K}$ and K, respectively. Then the following relation holds:

$$H_K(v) = (M_K^{-1})^T\, H_{\widehat{K}}(\widehat{v})\, M_K^{-1}.$$

Let us consider the Frobenius norm $\|\cdot\|_F$ of $H_K(v)$ while exploiting the decompositions introduced above for the matrices M_K and B_K. We get

$$\|H_K(v)\|_F = \|B_K^{-1}\, Z_K\, H_{\widehat{K}}(\widehat{v})\, Z_K^T\, B_K^{-1}\|_F = \|\Lambda_K^{-1}\, R_K\, Z_K\, H_{\widehat{K}}(\widehat{v})\, Z_K^T\, R_K^T\, \Lambda_K^{-1}\|_F,$$

as the Frobenius norm is invariant with respect to orthogonal matrices. Using the relation between the Frobenius norm of the Hessian matrix $H_K(v)$ and the H^2-seminorm of v on K yields

$$\begin{aligned} |v|^2_{H^2(K)} &= \int_K \|H_K(v)\|^2_F\, d\mathbf{x} = \int_K \|\Lambda_K^{-1}\, R_K\, Z_K\, H_{\widehat{K}}(\widehat{v})\, Z_K^T\, R_K^T\, \Lambda_K^{-1}\|_F\, d\mathbf{x} \\ &= \int_K \|\Lambda_K^{-1}\, P_K\, H_{\widehat{K}}(\widehat{v})\, P_K^T\, \Lambda_K^{-1}\|_F\, d\mathbf{x}, \end{aligned} \tag{2.13}$$

where the orthogonal matrix $P_K = R_K\, Z_K$ has been introduced in order to simplify the notation. Let us also denote Q_K the matrix

$$Q_K = \Lambda_K^{-1}\, P_K = \begin{bmatrix} \mathbf{q}_{1,K}^T \\ \mathbf{q}_{2,K}^T \end{bmatrix}, \quad \text{with } \mathbf{q}_{i,K} = \frac{1}{\lambda_{i,K}}\, \mathbf{p}_{i,K} \tag{2.14}$$

and where $\mathbf{p}_{i,K}$ are the columns of the matrix P_K, with $i = 1, 2$. Notice that, from a geometric viewpoint, matrix Q_K simply identifies a rotation followed by a rescaling of the coordinate axes. Introducing (2.14) into (2.13), the chain of equalities below may be inferred:

(2.15)

$$\begin{aligned}
|v|^2_{H^2(K)} &= \int_K \|Q_K\, H_{\widehat{K}}(\widehat{v})\, Q_K^T\|_F^2\, d\mathbf{x} \\
&= \int_K \left\{(\mathbf{q}_{1,K}^T\, H_{\widehat{K}}(\widehat{v})\, \mathbf{q}_{1,K})^2 + (\mathbf{q}_{2,K}^T\, H_{\widehat{K}}(\widehat{v})\, \mathbf{q}_{2,K})^2 + 2(\mathbf{q}_{1,K}^T\, H_{\widehat{K}}(\widehat{v})\mathbf{q}_{2,K})^2\right\} d\mathbf{x} \\
&= \frac{1}{\lambda_{1,K}^4} \int_K (\mathbf{p}_{1,K}^T\, H_{\widehat{K}}(\widehat{v})\, \mathbf{p}_{1,K})^2\, d\mathbf{x} + \frac{1}{\lambda_{2,K}^4} \int_K (\mathbf{p}_{2,K}^T\, H_{\widehat{K}}(\widehat{v})\, \mathbf{p}_{2,K})^2\, d\mathbf{x} \\
&+ \frac{2}{\lambda_{1,K}^2\, \lambda_{2,K}^2} \int_K (\mathbf{p}_{1,K}^T\, H_{\widehat{K}}(\widehat{v})\, \mathbf{p}_{2,K})^2\, d\mathbf{x} \\
&= \frac{\lambda_{2,K}}{\lambda_{1,K}^3} \int_{\widehat{K}} (\mathbf{p}_{1,K}^T\, H_{\widehat{K}}(\widehat{v})\, \mathbf{p}_{1,K})^2\, d\hat{\mathbf{x}} + \frac{\lambda_{1,K}}{\lambda_{2,K}^3} \int_{\widehat{K}} (\mathbf{p}_{2,K}^T\, H_{\widehat{K}}(\widehat{v})\, \mathbf{p}_{2,K})^2\, d\hat{\mathbf{x}} \\
&+ \frac{2}{\lambda_{1,K}\, \lambda_{2,K}} \int_{\widehat{K}} (\mathbf{p}_{1,K}^T\, H_{\widehat{K}}(\widehat{v})\, \mathbf{p}_{2,K})^2\, d\hat{\mathbf{x}},
\end{aligned}$$

the last sum having been obtained by expressing the integrals over the generic element K in terms of the integrals over the reference triangle $\widehat{K}$.

Let us bound the integrals of the last sum in (2.15), suitably rewriting the terms $(\mathbf{p}_{i,K}^T H_{\widehat{K}}(\widehat{v})\, \mathbf{p}_{j,K})$ for i, $j = 1, 2$. These terms are scalar products between the vectors $\mathbf{p}_{i,K}$ and $H_{\widehat{K}}(\widehat{v})\, \mathbf{p}_{j,K}$; then we have

$$\begin{aligned}
\mathbf{p}_{i,K}^T H_{\widehat{K}}(\widehat{v})\, \mathbf{p}_{j,K} &\le \|\mathbf{p}_{i,K}\|\, \|H_{\widehat{K}}(\widehat{v})\, \mathbf{p}_{j,K}\| \\
&\le \|\mathbf{p}_{i,K}\|\, \|H_{\widehat{K}}(\widehat{v})\|_F\, \|\mathbf{p}_{j,K}\| = \|H_{\widehat{K}}(\widehat{v})\|_F,
\end{aligned} \tag{2.16}$$

where $\|\cdot\|$ is the Euclidean norm and the relations $\|\mathbf{p}_{i,K}\| = 1$ for $i = 1, 2$, together with the compatibility of the Frobenius norm with the Euclidean one, have been used. Thus, by exploiting the relation between the Frobenius norm of $H_{\widehat{K}}(\widehat{v})$ and the H^2-seminorm $|\widehat{v}|_{H^2(\widehat{K})}$ as in (2.13), we obtain

$$\int_{\widehat{K}} (\mathbf{p}_{i,K}^T H_{\widehat{K}}(\widehat{v})\, \mathbf{p}_{j,K})^2\, d\hat{\mathbf{x}} \le \int_{\widehat{K}} \|H_{\widehat{K}}(\widehat{v})\|_F^2\, d\hat{\mathbf{x}} = |\widehat{v}|^2_{H^2(\widehat{K})} \tag{2.17}$$

for i, $j = 1$, 2. Substituting this result into (2.15), we deduce that

$$|v|^2_{H^2(K)} \le \left[\frac{\lambda_{2,K}}{\lambda_{1,K}^3} + \frac{\lambda_{1,K}}{\lambda_{2,K}^3} + \frac{2}{\lambda_{1,K}\, \lambda_{2,K}}\right] |\widehat{v}|^2_{H^2(\widehat{K})} = \frac{(\lambda_{1,K}^2 + \lambda_{2,K}^2)^2}{\lambda_{1,K}^3\, \lambda_{2,K}^3}\, |\widehat{v}|^2_{H^2(\widehat{K})},$$

that is, the upper bound in (2.12).

Let us verify now the lower bound of (2.12) moving from Lemma 2.2. As the calculations in (2.16) and (2.17) can be repeated on the terms $L_K(\mathbf{r}_{i,K}, \mathbf{r}_{j,K}; v)$ (with $i, j = 1, 2$) simply by identifying $\mathbf{r}_{i,K}$ with $\mathbf{p}_{i,K}$, $H_K(v)$ with $H_{\widehat{K}}(\widehat{v})$, and K with $\widehat{K}$, we obtain

$$|\widehat{v}|^2_{H^2(\widehat{K})} \le \left[\frac{\lambda^3_{1,K}}{\lambda_{2,K}} + \frac{\lambda^3_{2,K}}{\lambda_{1,K}} + 2\lambda_{1,K}\,\lambda_{2,K}\right] |v|^2_{H^2(K)} \le \frac{(\lambda^2_{1,K} + \lambda^2_{2,K})^2}{\lambda_{1,K}\,\lambda_{2,K}}\,|v|^2_{H^2(K)},$$

which immediately provides the inequality in the left-hand side of (2.12). □

We are now in a position to bound the H^2-seminorm of a generic function $v \in H^2(K)$ in terms of the anisotropic quantities $L_K(\mathbf{r}_{i,K}, \mathbf{r}_{j,K}; v)$ defined in (2.4).

COROLLARY 2.7. *For any function $v \in H^2(K)$ and for any $K \in \mathcal{T}_h$, we have*

$$(2.18)\qquad \begin{aligned} |v|_{H^2(K)} \le (\lambda^2_{1,K} + \lambda^2_{2,K}) \Bigg[&\frac{1}{\lambda^4_{2,K}} L_K(\mathbf{r}_{1,K}, \mathbf{r}_{1,K}; v) \\ &+ \frac{1}{\lambda^4_{1,K}} L_K(\mathbf{r}_{2,K}, \mathbf{r}_{2,K}; v) + \frac{2}{\lambda^2_{1,K}\,\lambda^2_{2,K}} L_K(\mathbf{r}_{1,K}, \mathbf{r}_{2,K}; v) \Bigg]^{1/2}, \end{aligned}$$

where the quantities $L_K(\mathbf{r}_{i,K}, \mathbf{r}_{j,K}; v)$ are defined in (2.4).

Proof. Let us start from the upper bound in Proposition 2.6. By adding the anisotropic information provided by Lemma 2.2, we get

$$\begin{aligned} |v|^2_{H^2(K)} \le\ & \frac{(\lambda^2_{1,K} + \lambda^2_{2,K})^2}{\lambda^4_{2,K}} L_K(\mathbf{r}_{1,K}, \mathbf{r}_{1,K}; v) + \frac{(\lambda^2_{1,K} + \lambda^2_{2,K})^2}{\lambda^4_{1,K}} L_K(\mathbf{r}_{2,K}, \mathbf{r}_{2,K}; v) \\ &+ 2\,\frac{(\lambda^2_{1,K} + \lambda^2_{2,K})^2}{\lambda^2_{1,K}\,\lambda^2_{2,K}} L_K(\mathbf{r}_{1,K}, \mathbf{r}_{2,K}; v), \end{aligned}$$

which is exactly result (2.18) after simple algebraic manipulations. □

Remark 2 (anisotropic interpolation estimates for vector valued functions). All of the interpolation estimates obtained above can be easily generalized to the case when $\mathbf{v} : \Omega \to \mathbb{R}^2$. In this case, the above results still hold formally, provided that the terms $L_K(\mathbf{r}_{i,K}, \mathbf{r}_{j,K}; v)$ and $G_K(v)$ defined in (2.4) and (2.5), respectively, are replaced by the new ones, $L_K(\mathbf{r}_{i,K}, \mathbf{r}_{j,K}; \mathbf{v})$ and $G_K(\mathbf{v})$, given by

$$(2.19)\qquad L_K(\mathbf{r}_{i,K}, \mathbf{r}_{j,K}; \mathbf{v}) = \sum_{l=1,2} \int_K \left(\mathbf{r}^T_{i,K}\, H_K(v_l)\, \mathbf{r}_{j,K}\right)^2 d\mathbf{x}, \quad \text{with } i, j = 1, 2,$$

$$\begin{aligned} G_K(\mathbf{v}) &= \sum_{l=1,2} \sum_{T \in \Delta_K} \begin{bmatrix} \displaystyle\int_T \left(\frac{\partial v_l}{\partial x_1}\right)^2 d\mathbf{x} & \displaystyle\int_T \frac{\partial v_l}{\partial x_1}\frac{\partial v_l}{\partial x_2}\, d\mathbf{x} \\ \displaystyle\int_T \frac{\partial v_l}{\partial x_1}\frac{\partial v_l}{\partial x_2}\, d\mathbf{x} & \displaystyle\int_T \left(\frac{\partial v_l}{\partial x_2}\right)^2 d\mathbf{x} \end{bmatrix} \\ &= \sum_{T \in \Delta_K} \begin{bmatrix} \displaystyle\int_T \frac{\partial \mathbf{v}}{\partial x_1} \cdot \frac{\partial \mathbf{v}}{\partial x_1}\, d\mathbf{x} & \displaystyle\int_T \frac{\partial \mathbf{v}}{\partial x_1} \cdot \frac{\partial \mathbf{v}}{\partial x_2}\, d\mathbf{x} \\ \displaystyle\int_T \frac{\partial \mathbf{v}}{\partial x_1} \cdot \frac{\partial \mathbf{v}}{\partial x_2}\, d\mathbf{x} & \displaystyle\int_T \frac{\partial \mathbf{v}}{\partial x_2} \cdot \frac{\partial \mathbf{v}}{\partial x_2}\, d\mathbf{x} \end{bmatrix}, \end{aligned}$$

where v_l, for $l = 1, 2$, are the Cartesian components of $\mathbf{v}$.

For alternative anisotropic interpolation error estimates see [1, 4, 30, 40].

3. The advection-diffusion problem.

In this section we focus our attention on the scalar advective-diffusive model. Starting from the formulation of stabilized methods presented in [22], we readdress the question of a careful design for the stability coefficients, crucial for the good performance of these methods, in the framework of anisotropic meshes.

In particular, we extend the results obtained in [22], where an expression for the stability coefficients in the whole range varying from advective to diffusive dominated flows is introduced, to the case of (possibly) highly stretched elements. In [2] an anisotropic a priori error analysis is provided for the advection–diffusion-reaction problem in the case of finite elements of arbitrary order, but the results obtained are limited by a maximal angle condition and a coordinate system condition. We consider only the case of affine finite elements so that the stabilized methods, such as GLS, SUPG [11, 29], and the method proposed in [15], do coincide with each other.

We study the convergence of the stabilized method in a mesh dependent norm, taking into account also the stability coefficients. The optimal value for these is obtained by error analysis considerations by requiring that the convergence rate, in both the advective and the diffusive dominated regimes, is of maximal order.

Theorem 3.1 is the main result of this section.

3.1. Problem statement and finite element discretization.

Let us consider the standard advection-diffusion problem for the scalar field $u = u(\mathbf{x})$

$$\begin{cases} -\mu\,\Delta u + \mathbf{a}\cdot\nabla u = f & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega, \end{cases} \tag{3.1}$$

where $\mu = \text{const} > 0$ is the diffusivity, $\mathbf{a} = \mathbf{a}(\mathbf{x}) \in (C^1(\overline{\Omega}))^2$ is the given flow velocity with $\nabla\cdot\mathbf{a} = 0$ in Ω, and $f = f(\mathbf{x}) \in L^2(\Omega)$ is the source term.

The variational formulation of problem (3.1) reads as follows: find $u \in H_0^1(\Omega)$ which satisfies

$$B(u, v) = F(v) \quad \text{for any } v \in H_0^1(\Omega), \tag{3.2}$$

where $B(\cdot,\cdot)$ and $F(\cdot)$ define the bilinear and linear forms

$$B(u, v) = (\mu\,\nabla u,\, \nabla v) + (\mathbf{a}\cdot\nabla u,\, v) \quad \text{and} \quad F(v) = (f,\, v), \tag{3.3}$$

respectively, for any u and $v \in H_0^1(\Omega)$.

As we are interested in advection dominated problems, we discretize problem (3.2) by a stabilized finite element approach (GLS) [29]. The discrete problem thus reads as follows: find $u_h \in W_{h,0}$ such that

$$B_h(u_h,\, v_h) = F_h(v_h) \quad \text{for any } v_h \in W_{h,0}, \tag{3.4}$$

with

$$B_h(u_h,\, v_h) = B(u_h,\, v_h) + \sum_{K\in\mathcal{T}_h} (-\mu\,\Delta u_h + \mathbf{a}\cdot\nabla u_h,\, \tau_K(-\mu\,\Delta v_h + \mathbf{a}\cdot\nabla v_h))_K, \tag{3.5}$$

$$F_h(v_h) = F(v_h) + \sum_{K\in\mathcal{T}_h} (f,\, \tau_K(-\mu\,\Delta v_h + \mathbf{a}\cdot\nabla v_h))_K, \tag{3.6}$$

where we let $W_{h,0} = W_h \cap H_0^1(\Omega)$. In particular, as we are dealing with continuous affine finite elements, the terms $\Delta u_h\big|_K$ and $\Delta v_h\big|_K$ in (3.5) and (3.6) are identically equal to zero. Finally, concerning the stability coefficients τ_K, we define it as

$$\tau_K = \frac{\delta_K}{2} \, \frac{\xi(\mathrm{Pe}_K)}{\|\mathbf{a}\|_{L^\infty(K)}}, \tag{3.7}$$

where δ_K is a characteristic dimension of element K and the function $\xi(\cdot)$ is defined as

$$\xi(\mathrm{Pe}_K) = \begin{cases} \mathrm{Pe}_K & \text{if} \quad \mathrm{Pe}_K < 1, \\ 1 & \text{if} \quad \mathrm{Pe}_K \geq 1. \end{cases} \tag{3.8}$$

This choice corresponds to considering a locally advection dominated flow when the element Péclet number Pe_K, given by

$$\mathrm{Pe}_K = \delta_K \, \frac{\|\mathbf{a}\|_{L^\infty(K)}}{6\,\mu}, \tag{3.9}$$

is greater than or equal to one. Notice that recipe (3.7) generalizes the corresponding one in [22], where $\delta_K = h_K$. Nevertheless, in the case of anisotropic meshes, this choice turns out not to be the optimal one (see also, e.g., [2]). We provide in what follows an alternative determination of δ_K based on the error analysis.

3.2. Error analysis. To begin with, let us recall that the stabilized scheme (3.4) is a *consistent* method in the sense that if additional regularity is demanded for the solution u of the variational problem (3.2), that is, $u \in H^2(\Omega) \cap H^1_0(\Omega)$, then the following relation holds:

$$B_h(u, \, v_h) = F_h(v_h) \quad \text{for any } v_h \in W_{h,0}. \tag{3.10}$$

As a consequence, the well-known Galerkin orthogonality property

$$B_h(u - u_h, \, v_h) = 0 \quad \text{for any } v_h \in W_{h,0} \tag{3.11}$$

follows. In the convergence analysis provided below we endow the space $H^1_0(\Omega)$ with the discrete norm $\|\cdot\|_h$ defined, for any $w \in H^1_0(\Omega)$, by

$$\|w\|_h^2 = \mu \, \|\nabla w\|^2_{L^2(\Omega)} + \sum_{K \in \mathcal{T}_h} \|\tau_K^{1/2} \, \mathbf{a} \cdot \nabla w\|^2_{L^2(K)}. \tag{3.12}$$

We anticipate the main result of this section.

THEOREM 3.1 (convergence in norm $\|\cdot\|_h$). *Let $u \in H^2(\Omega) \cap H^1_0(\Omega)$ be the solution to (3.2), and let $u_h \in W_{h,0}$ be the solution to (3.4). Then the anisotropic definitions of the stability coefficients and of the local Péclet number are*

$$\tau_K = \frac{\lambda_{2,K}}{2} \, \frac{\xi(\mathrm{Pe}_K)}{\|\mathbf{a}\|_{L^\infty(K)}}, \tag{3.13}$$

$$\mathrm{Pe}_K = \lambda_{2,K} \, \frac{\|\mathbf{a}\|_{L^\infty(K)}}{6\,\mu}, \tag{3.14}$$

respectively, where $\xi(\cdot)$ is the same as in (3.8). Moreover, under this choice there exists a constant $C = C(\widehat{K})$ such that it holds that

$$\begin{aligned} \|u - u_h\|_h^2 \leq C \sum_{K \in \mathcal{T}_h} \Big\{ & \lambda_{2,K}^2 \Big(\lambda_{2,K} \|\mathbf{a}\|_{L^\infty(K)} \mathcal{H}(\mathrm{Pe}_K - 1) + \mu \mathcal{H}(1 - \mathrm{Pe}_K) \Big) \\ & \big[s_K^4 L_K(\mathbf{r}_{1,K}, \, \mathbf{r}_{1,K}; \, u) + L_K(\mathbf{r}_{2,K}, \, \mathbf{r}_{2,K}; \, u) + 2 s_K^2 L_K(\mathbf{r}_{1,K}, \, \mathbf{r}_{2,K}; \, u) \big] \Big\}, \end{aligned} \tag{3.15}$$

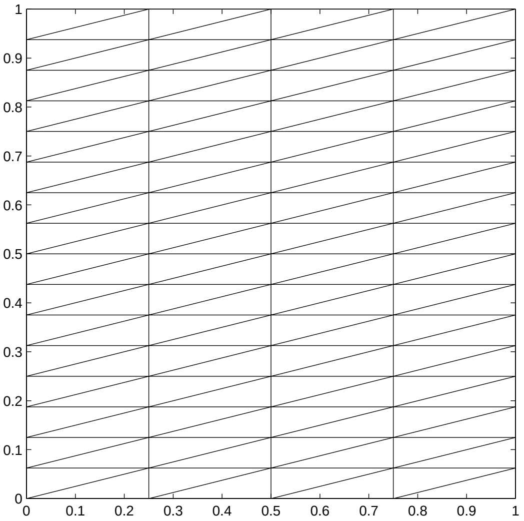


FIG. 3.1. *An example of mesh.*

where the quantities $L_K(\mathbf{r}_{i,K}, \mathbf{r}_{j,K}; u)$ *are defined by* (2.4) *and where* $\mathcal{H}(\cdot)$ *is the Heaviside function given by*

$$\mathcal{H}(s) = \begin{cases} 0 & \textit{if } s < 0, \\ 1 & \textit{if } s > 0. \end{cases} \tag{3.16}$$

Before proving Theorem 3.1, some remarks are in order. In the isotropic case when $\lambda_{1,K} \simeq \lambda_{2,K}$, (3.15) recovers Theorem 3.1 in [22] in the case of affine elements. It is interesting to observe that the interpolation error estimates in section 2.1 and in turn the convergence result (3.15) contain quantities related to the stretching factors s_K of the elements, but, as pointed out also in Remark 5.1 in [4], if the anisotropic refinement is along the correct direction, i.e., where high derivatives of the solution occur, then these terms are of smaller size than the other terms. Actually, estimate (3.15) guarantees convergence when s_K is bounded, giving a convergence rate of the order of $\lambda_{2,K}^{3/2}$ and $\lambda_{2,K}$ in the discrete norm $\|\cdot\|_h$ for the cases of $\mathrm{Pe}_K \geq 1$ and $\mathrm{Pe}_K < 1$, respectively, but convergence can be achieved even if s_K is unbounded, as we discuss in the following two examples.

Let us introduce a parameter $t < 1$, and let $\lambda_{1,K} = t$ and $\lambda_{2,K} = t^j$ (with $j > 1$) so that $s_K > 1$. In order to study convergence, we let $t \to 0$ (i.e., $s_K \to \infty$). This implies that the dominant term in the second row of (3.15) is the one of the order of $s_K^4 = t^{4(1-j)}$. Thus, the right-hand side in (3.15) behaves like $\lambda_{2,K}^3 s_K^4 = t^{4-j}$ or $\lambda_{2,K}^2 s_K^4 = t^{4-2j}$ according to whether the problem is convection dominated or not, respectively, and convergence can still be expected, provided that $1 < j < 4$ in the first case or $1 < j < 2$ in the second one. This corresponds to a "directional limit," where $\lambda_{1,K}$ and $\lambda_{2,K}$ tend to zero in a constrained manner.

Example 1. Let us provide an instance of this situation. Consider a mesh such as the one in Figure 3.1, referred to as $\Omega = (0,1)^2$, and suppose that the problem is advection dominated, i.e., $\mathrm{Pe}_K \geq 1$. Let N_x and N_y be the number of subdivisions in the x- and y-direction, respectively. Using for $\widehat{K}$ the right triangle $(0,0), (1,0), (0,1)$, it can be checked that $\lambda_{1,K} = 1/N_x$ and $\lambda_{2,K} = 1/N_y$. After introducing the parameter $t < 1$ such that $\lambda_{1,K} = t$ and $\lambda_{2,K} = t^j$, we have that $N_x = 1/t$ and $N_y = N_x^j$ for a fixed $j > 1$. Notice that the parameter t should be chosen such that $1/t$ is an integer, e.g., $t = 1/n$ with $n \in \mathbb{N}^+$.

Finally, let us introduce a family of triangulations depending on the parameter t by recalling from the above discussion that $1 < j < 4$ as the problem is convection dominated. The sequence of numerical solutions computed on this family of triangulations, as t decreases, converges at a rate $t^{2-j/2} = \lambda_{1,K}^{2-j/2}$ in the discrete norm $\|\cdot\|_h$ while s_K diverges like t^{1-j} for any $K \in \mathcal{T}_h$.

Example 2. Another case where convergence occurs while s_K may be unbounded is when $L_K(\mathbf{r}_{1,K}, \mathbf{r}_{1,K};\, u) = L_K(\mathbf{r}_{1,K}, \mathbf{r}_{2,K};\, u) = 0$ while $L_K(\mathbf{r}_{2,K}, \mathbf{r}_{2,K};\, u)$ is bounded in (3.15). Starting from the same type of meshes as in Figure 3.1 and using for $\widehat{K}$ the same right triangle, it can be checked that $\mathbf{r}_{1,K} = (1,0)^T$ and $\mathbf{r}_{2,K} = (0,1)^T$, provided that the number of subdivisions $N_x < N_y$. In this case the term $L_K(\mathbf{r}_{1,K},\, \mathbf{r}_{1,K};\, u) = \|\partial^2 u/\partial x_1^2\|^2_{L^2(K)} = 0$ and $L_K(\mathbf{r}_{1,K},\, \mathbf{r}_{2,K};\, u) = \|\partial^2 u/\partial x_1 \partial x_2\|^2_{L^2(K)} = 0$. This implies that whenever $u = u(x_2)$ with $\|\partial^2 u/\partial x_2^2\|_{L^2(\Omega)}$ bounded, convergence is guaranteed, provided that the mesh size tends to zero in the x_2-direction while no constraint is required for the mesh size in the x_1-direction (see section 5).

To summarize, in a loose sense, if the mesh is aligned with the solution, then convergence occurs independently of the stretching factor. Otherwise, convergence may not occur. Moving from the residual-free bubble theory [9, 10, 39], we propose an alternative recipe taking into account the orientation of the convective field with respect to the mesh in [34].

In order to prove Theorem 3.1 we analyze in turn the stability and the continuity of the bilinear form $B_h(\cdot, \cdot)$ (see also [2, 22]). Let us begin with the stability result.

LEMMA 3.2 (stability in norm $\|\cdot\|_h$). *For any $v_h \in W_{h,0}$,*

$$B_h(v_h,\, v_h) = \|v_h\|_h^2. \tag{3.17}$$

Thus (3.4) *has a unique solution.*

Proof. Set $u_h = v_h$ in (3.5). □

The next result can be obtained from the anisotropic interpolation error estimates provided in section 2.1.

LEMMA 3.3. *Let us assume that the solution u to* (3.2) *satisfies $u \in H^2(\Omega) \cap H_0^1(\Omega)$. Then for any $K \in \mathcal{T}_h$*

(i) *if* $\mathrm{Pe}_K \geq 1$, *then*

(3.18)
$$\|\tau_K^{-1/2}\,(u - r_K(u))\|^2_{L^2(K)} + \mu\,\|\nabla(u - r_K(u))\|^2_{L^2(K)} + \|\tau_K^{1/2}\,\mathbf{a}\cdot\nabla(u - r_K(u))\|^2_{L^2(K)}$$
$$+ \|\tau_K^{1/2}\,\mu\,\Delta(u - r_K(u))\|^2_{L^2(K)} \leq C\left[\frac{1}{\delta_K} + \frac{\delta_K}{\lambda_{2,K}^2} + \delta_K^3\,\frac{(\lambda_{1,K}^2 + \lambda_{2,K}^2)^2}{\lambda_{1,K}^4\,\lambda_{2,K}^4}\right]\|\mathbf{a}\|_{L^\infty(K)}$$
$$[\lambda_{1,K}^4\, L_K(\mathbf{r}_{1,K},\, \mathbf{r}_{1,K};\, u) + \lambda_{2,K}^4\, L_K(\mathbf{r}_{2,K},\, \mathbf{r}_{2,K};\, u) + \, 2\lambda_{1,K}^2\,\lambda_{2,K}^2\, L_K(\mathbf{r}_{1,K},\, \mathbf{r}_{2,K};\, u)];$$

(ii) *if* $\mathrm{Pe}_K < 1$*, then*

(3.19)
$$\|\tau_K^{-1/2}\,(u - r_K(u))\|^2_{L^2(K)} + \mu\,\|\nabla(u - r_K(u))\|^2_{L^2(K)} + \|\tau_K^{1/2}\,\mathbf{a}\cdot\nabla(u - r_K(u))\|^2_{L^2(K)}$$
$$+ \|\tau_K^{1/2}\,\mu\,\Delta(u - r_K(u))\|^2_{L^2(K)} \le C\left[\frac{1}{\delta_K^2} + \frac{1}{\lambda_{2,K}^2} + \delta_K^2\,\frac{(\lambda_{1,K}^2 + \lambda_{2,K}^2)^2}{\lambda_{1,K}^4\,\lambda_{2,K}^4}\right]\mu$$
$$[\lambda_{1,K}^4\,L_K(\mathbf{r}_{1,K},\,\mathbf{r}_{1,K};\,u) + \lambda_{2,K}^4\,L_K(\mathbf{r}_{2,K},\,\mathbf{r}_{2,K};\,u) + 2\lambda_{1,K}^2\,\lambda_{2,K}^2\,L_K(\mathbf{r}_{1,K},\,\mathbf{r}_{2,K};\,u)],$$

where the quantities $L_K(\mathbf{r}_{i,K},\,\mathbf{r}_{j,K};\,u)$ *are defined by* (2.4) *and* $C = C(\widehat{K})$*.*

Proof. Let us start by analyzing separately the four terms in the left-hand side of both (3.18) and (3.19), independently of the particular values of Pe_K. Concerning the first one, the inequality (2.6) in Lemma 2.4, together with the definition (3.7) of the stability coefficients τ_K, yields

$$\|\tau_K^{-1/2}\,(u - r_K(u))\|^2_{L^2(K)} \le C\,\frac{\|\mathbf{a}\|_{L^\infty(K)}}{\delta_K\,\xi(\mathrm{Pe}_K)}[\lambda_{1,K}^4\,L_K(\mathbf{r}_{1,K},\,\mathbf{r}_{1,K};\,u)$$
$$+ \lambda_{2,K}^4 L_K(\mathbf{r}_{2,K},\,\mathbf{r}_{2,K};\,u) + 2\lambda_{1,K}^2\,\lambda_{2,K}^2\,L_K(\mathbf{r}_{1,K},\,\mathbf{r}_{2,K};\,u)]. \tag{3.20}$$

The anisotropic estimate (2.7) in Lemma 2.4 derived for the H^1-seminorm of the interpolation error allows us to bound the second quantity. Indeed, we have

$$\begin{aligned}\mu\,\|\nabla(u - r_K(u))\|^2_{L^2(K)} &\le C\,\mu\,\left[\frac{\lambda_{1,K}^4}{\lambda_{2,K}^2}\,L_K(\mathbf{r}_{1,K},\,\mathbf{r}_{1,K};\,u)\right.\\ &\qquad\left. + \lambda_{2,K}^2 L_K(\mathbf{r}_{2,K},\,\mathbf{r}_{2,K};\,u) + 2\lambda_{1,K}^2\,L_K(\mathbf{r}_{1,K},\,\mathbf{r}_{2,K};\,u)\right]\\ &= C\,\frac{\delta_K\,\|\mathbf{a}\|_{L^\infty(K)}}{\lambda_{2,K}^2\,\mathrm{Pe}_K}[\lambda_{1,K}^4\,L_K(\mathbf{r}_{1,K},\,\mathbf{r}_{1,K};\,u)\\ &\qquad + \lambda_{2,K}^4 L_K(\mathbf{r}_{2,K},\,\mathbf{r}_{2,K};\,u) + 2\lambda_{1,K}^2\,\lambda_{2,K}^2\,L_K(\mathbf{r}_{1,K},\,\mathbf{r}_{2,K};\,u)],\end{aligned} \tag{3.21}$$

where, thanks to (3.9), the diffusivity μ has been expressed, within a constant, in terms of the Péclet number associated with element K. Let us consider now the third term. We have

(3.22)
$$\|\tau_K^{1/2}\,\mathbf{a}\cdot\nabla(u - r_K(u))\|^2_{L^2(K)} = \frac{\delta_K\,\xi(\mathrm{Pe}_K)}{2\,\|\mathbf{a}\|_{L^\infty(K)}}\int_K (\mathbf{a}\cdot\nabla(u - r_K(u)))^2\,d\mathbf{x}$$
$$\le C\,\delta_K\,\|\mathbf{a}\|_{L^\infty(K)}\xi(\mathrm{Pe}_K)\,\|\nabla(u - r_K(u))\|^2_{L^2(K)} \le C\,\frac{\delta_K\,\|\mathbf{a}\|_{L^\infty(K)}\,\xi(\mathrm{Pe}_K)}{\lambda_{2,K}^2}$$
$$[\lambda_{1,K}^4\,L_K(\mathbf{r}_{1,K},\,\mathbf{r}_{1,K};\,u) + \lambda_{2,K}^4 L_K(\mathbf{r}_{2,K},\,\mathbf{r}_{2,K};\,u) + 2\lambda_{1,K}^2\,\lambda_{2,K}^2\,L_K(\mathbf{r}_{1,K},\,\mathbf{r}_{2,K};\,u)].$$

In the chain of inequalities above, the definition of the stability coefficients τ_K and the anisotropic estimate (2.7) in Lemma 2.4 have been used. Finally, let us bound suitably the last term in the left-hand side of (3.18) and (3.19), again using the expressions of τ_K and of the diffusivity μ in terms of the Péclet number Pe_K together with Corollary

2.7. Thus, it can be deduced that

(3.23)
$$\|\tau_K^{1/2}\,\mu\,\Delta(u-r_K(u))\|^2_{L^2(K)} \le \tau_K\,\mu^2\,|u|^2_{H^2(K)} \le \tau_K\,\mu^2(\lambda_{1,K}^2+\lambda_{2,K}^2)^2$$
$$\left[\frac{1}{\lambda_{2,K}^4}\,L_K(\mathbf{r}_{1,K},\,\mathbf{r}_{1,K};\,u)+\frac{1}{\lambda_{1,K}^4}L_K(\mathbf{r}_{2,K},\,\mathbf{r}_{2,K};\,u)+\,\frac{2}{\lambda_{1,K}^2\,\lambda_{2,K}^2}\,L_K(\mathbf{r}_{1,K},\,\mathbf{r}_{2,K};\,u)\right]$$
$$\le C\,\frac{\delta_K^3\,\|\mathbf{a}\|_{L^\infty(K)}\xi(\mathrm{Pe}_K)}{\mathrm{Pe}_K^2}\,\frac{(\lambda_{1,K}^2+\lambda_{2,K}^2)^2}{\lambda_{1,K}^4\,\lambda_{2,K}^4}$$
$$[\lambda_{1,K}^4\,L_K(\mathbf{r}_{1,K},\,\mathbf{r}_{1,K};\,u)+\lambda_{2,K}^4L_K(\mathbf{r}_{2,K},\,\mathbf{r}_{2,K};\,u)+\,2\lambda_{1,K}^2\,\lambda_{2,K}^2\,L_K(\mathbf{r}_{1,K},\,\mathbf{r}_{2,K};\,u)].$$

Let us deal now with the case $\mathrm{Pe}_K \ge 1$, and consider each term (3.20)–(3.23) in turn. In the first of these terms we use definition (3.8); in the second one we use $1/\mathrm{Pe}_K \le 1$; in the third we employ definition (3.8); and in the fourth one we use again (3.8) and $1/\mathrm{Pe}_K^2 \le 1$. This allows us to further bound the left-hand side of (3.18) as

$$\|\tau_K^{-1/2}(u-r_K(u))\|^2_{L^2(K)}+\mu\,\|\nabla(u-r_K(u))\|^2_{L^2(K)}+\|\tau_K^{1/2}\,\mathbf{a}\cdot\nabla(u-r_K(u))\|^2_{L^2(K)}$$
$$+\,\|\tau_K^{1/2}\,\mu\,\Delta(u-r_K(u))\|^2_{L^2(K)} \le C\left[\frac{1}{\delta_K}+\frac{\delta_K}{\lambda_{2,K}^2}+\delta_K^3\,\frac{(\lambda_{1,K}^2+\lambda_{2,K}^2)^2}{\lambda_{1,K}^4\,\lambda_{2,K}^4}\right]\|\mathbf{a}\|_{L^\infty(K)}$$
$$[\lambda_{1,K}^4\,L_K(\mathbf{r}_{1,K},\,\mathbf{r}_{1,K};\,u)+\lambda_{2,K}^4\,L_K(\mathbf{r}_{2,K},\,\mathbf{r}_{2,K};\,u)+\,2\lambda_{1,K}^2\,\lambda_{2,K}^2\,L_K(\mathbf{r}_{1,K},\,\mathbf{r}_{2,K};\,u)].$$

Let us now consider the case $\mathrm{Pe}_K < 1$, where we follow a different path with respect to the previous case. Let us analyze the four terms (3.20)–(3.23) above in turn. We first use the definition (3.8) and then (3.9) in (3.20); on the term (3.21) we employ (3.9); in the third one we use (3.8) and then express, within a constant, the term $\|\mathbf{a}\|_{L^\infty(K)}$ as $\mu\mathrm{Pe}_K/\delta_K$ from (3.9); and finally we use $\mathrm{Pe}_K^2 < 1$. In the fourth term we employ definition (3.8) and then compute, within a constant, $\mathrm{Pe}_K/\|\mathbf{a}\|_{L^\infty(K)}$ as δ_K/μ. We can then bound the left-hand side in (3.19) as

$$\|\tau_K^{-1/2}(u-r_K(u))\|^2_{L^2(K)}+\mu\,\|\nabla(u-r_K(u))\|^2_{L^2(K)}+\|\tau_K^{1/2}\,\mathbf{a}\cdot\nabla(u-r_K(u))\|^2_{L^2(K)}$$
$$+\,\|\tau_K^{1/2}\,\mu\,\Delta(u-r_K(u))\|^2_{L^2(K)} \le C\left[\frac{1}{\delta_K^2}+\frac{1}{\lambda_{2,K}^2}+\delta_K^2\frac{(\lambda_{1,K}^2+\lambda_{2,K}^2)^2}{\lambda_{1,K}^4\lambda_{2,K}^4}\right]\mu$$
$$[\lambda_{1,K}^4\,L_K(\mathbf{r}_{1,K},\,\mathbf{r}_{1,K};\,u)+\lambda_{2,K}^4\,L_K(\mathbf{r}_{2,K},\,\mathbf{r}_{2,K};\,u)+\,2\lambda_{1,K}^2\,\lambda_{2,K}^2\,L_K(\mathbf{r}_{1,K},\,\mathbf{r}_{2,K};\,u)],$$

and this concludes the proof. □

Let us now prove the continuity of the bilinear form $B_h(\cdot,\,\cdot)$.

LEMMA 3.4 (continuity of $B_h(\cdot,\,\cdot)$). *For any $u \in H^2(\Omega)\cap H_0^1(\Omega)$ and for any $v_h \in W_{h,0}$, there exists a constant C such that*

$$|B_h(u,v_h)| \le C\Bigg[\mu\|\nabla u\|^2_{L^2(\Omega)}+\sum_{K\in\mathcal{T}_h}\Big(\|\tau_K^{-1/2}u\|^2_{L^2(K)}$$
$$+\,\|\tau_K^{1/2}\mathbf{a}\cdot\nabla u\|^2_{L^2(K)}+\|\tau_K^{1/2}\mu\Delta u\|^2_{L^2(K)}\Big)\Bigg]^{1/2}\|v_h\|_h.$$

Proof. From the definition (3.5) of the bilinear form $B_h(\cdot,\,\cdot)$ and since $\Delta v_h\big|_K = 0$ for any $K\in\mathcal{T}_h$ and for any $v_h \in W_{h,0}$, we have

$$B_h(u,v_h) = (\mu\nabla u,\nabla v_h)+(\mathbf{a}\cdot\nabla u,v_h)+\sum_{K\in\mathcal{T}_h}(-\mu\Delta u+\mathbf{a}\cdot\nabla u,\tau_K(\mathbf{a}\cdot\nabla v_h))_K.$$

Then integrating by parts the second term and using the Cauchy–Schwarz inequality, we get

$$|B_h(u,v_h)| \leq \mu \|\nabla u\|_{L^2(\Omega)} \|\nabla v_h\|_{L^2(\Omega)} + \sum_{K\in\mathcal{T}_h} \|\tau_K^{1/2}\mathbf{a}\cdot\nabla v_h\|_{L^2(K)} \|\tau_K^{-1/2} u\|_{L^2(K)}$$
$$+ \sum_{K\in\mathcal{T}_h} \|\tau_K^{1/2}\mathbf{a}\cdot\nabla v_h\|_{L^2(K)} \Big(\|\tau_K^{1/2}\mathbf{a}\cdot\nabla u\|_{L^2(K)} + \|\tau_K^{1/2}\mu\Delta u\|_{L^2(K)} \Big).$$

Finally, thanks to the discrete Cauchy–Schwarz inequality and some simple calculations, we obtain

$$|B_h(u,v_h)| \leq C \Bigg(\mu\|\nabla u\|^2_{L^2(\Omega)} + \sum_{K\in\mathcal{T}_h} \|\tau_K^{-1/2} u\|^2_{L^2(K)}$$
$$+ \sum_{K\in\mathcal{T}_h} \Big(\|\tau_K^{1/2}\mathbf{a}\cdot\nabla u\|^2_{L^2(K)} + \|\tau_K^{1/2}\mu\Delta u\|^2_{L^2(K)} \Big) \Bigg)^{1/2}$$
$$\Bigg(\mu\|\nabla v_h\|^2_{L^2(\Omega)} + \sum_{K\in\mathcal{T}_h} \|\tau_K^{1/2}\mathbf{a}\cdot\nabla v_h\|^2_{L^2(K)} \Bigg)^{1/2},$$

that is, the desired result. □

We are now in a position to state the following anisotropic *a priori error estimate* with respect to the norm $\|\cdot\|_h$ defined in (3.12).

PROPOSITION 3.5 (a priori error estimate in norm $\|\cdot\|_h$). *Let* $u \in H^2(\Omega)\cap H^1_0(\Omega)$ *be the solution to* (3.2), *and let* $u_h \in W_{h,0}$ *be the solution to* (3.4). *Then there exists a constant* $C = C(\widehat{K})$ *such that the a priori estimate*

(3.24)
$$\|u-u_h\|_h^2 \leq C \sum_{K\in\mathcal{T}_h} \Bigg\{ \Bigg(\Bigg[\frac{1}{\delta_K} + \frac{\delta_K}{\lambda^2_{2,K}} + \delta_K^3 \frac{(\lambda^2_{1,K}+\lambda^2_{2,K})^2}{\lambda^4_{1,K}\,\lambda^4_{2,K}} \Bigg] \|\mathbf{a}\|_{L^\infty(K)}\, \mathcal{H}(\mathrm{Pe}_K - 1)$$
$$+ \Bigg[\frac{1}{\delta_K^2} + \frac{1}{\lambda^2_{2,K}} + \delta_K^2 \frac{(\lambda^2_{1,K}+\lambda^2_{2,K})^2}{\lambda^4_{1,K}\lambda^4_{2,K}} \Bigg] \mu\mathcal{H}(1-\mathrm{Pe}_K) \Bigg)$$
$$\big[\lambda^4_{1,K} L_K(\mathbf{r}_{1,K},\, \mathbf{r}_{1,K};\, u) + \lambda^4_{2,K}\, L_K(\mathbf{r}_{2,K},\, \mathbf{r}_{2,K};\, u) + 2\lambda^2_{1,K}\,\lambda^2_{2,K}\, L_K(\mathbf{r}_{1,K},\, \mathbf{r}_{2,K};\, u) \big] \Bigg\}$$

holds true, with $L_K(\mathbf{r}_{i,K},\, \mathbf{r}_{j,K};\, u)$ *defined in* (2.4) *and where* $\mathcal{H}(\cdot)$ *is the Heaviside function defined in* (3.16).

Proof. The stability result (3.17), combined with the Galerkin orthogonality property (3.11), yields the relations

$$\|u_h - r_h(u)\|_h^2 = B_h(u_h - r_h(u),\, u_h - r_h(u)) = B_h(u_h - u + u - r_h(u),\, u_h - r_h(u))$$
$$= B_h(u - r_h(u),\, u_h - r_h(u)) \leq |B_h(u - r_h(u),\, u_h - r_h(u))|.$$

From Lemma 3.4 we immediately get

(3.25)
$$\|u_h - r_h(u)\|_h \leq C \Bigg(\sum_{K\in\mathcal{T}_h} \Big[\|\tau_K^{-1/2}(u - r_K(u))\|^2_{L^2(K)} + \mu\|\nabla(u - r_K(u))\|^2_{L^2(K)}$$
$$+ \|\tau_K^{1/2}\mathbf{a}\cdot\nabla(u - r_K(u))\|^2_{L^2(K)} + \|\tau_K^{1/2}\mu\Delta(u - r_K(u))\|^2_{L^2(K)} \Big] \Bigg)^{1/2}.$$

Using the triangle and Young inequalities, we have

$$\|u-u_h\|_h^2 \le 2(\|u-r_h(u)\|_h^2 + \|u_h - r_h(u)\|_h^2).$$

As, trivially,

$$\begin{aligned}\|u-r_h(u)\|_h^2 \le \sum_{K\in\mathcal{T}_h} \Big[&\|\tau_K^{-1/2}\,(u-r_K(u))\|^2_{L^2(K)} + \mu\,\|\nabla(u-r_K(u))\|^2_{L^2(K)} \\ &+ \|\tau_K^{1/2}\,\mathbf{a}\cdot\nabla(u-r_K(u))\|^2_{L^2(K)} + \|\tau_K^{1/2}\,\mu\,\Delta(u-r_K(u))\|^2_{L^2(K)}\Big],\end{aligned}$$

and from (3.25), it follows that

$$\begin{aligned}\|u-u_h\|_h^2 \le C \sum_{K\in\mathcal{T}_h} \Big[&\|\tau_K^{-1/2}\,(u-r_K(u))\|^2_{L^2(K)} + \mu\,\|\nabla(u-r_K(u))\|^2_{L^2(K)} \\ &+ \|\tau_K^{1/2}\,\mathbf{a}\cdot\nabla(u-r_K(u))\|^2_{L^2(K)} + \|\tau_K^{1/2}\,\mu\,\Delta(u-r_K(u))\|^2_{L^2(K)}\Big].\end{aligned} \tag{3.26}$$

Result (3.24) eventually follows by applying Lemma 3.3 to the right-hand side of (3.26). □

Proof of Theorem 3.1. Let us consider the error estimate obtained in Proposition 3.5, and, after using the definition of stretching factor, let us rewrite it as follows:

$$\begin{aligned}\|u-u_h\|_h^2 \le C \sum_{K\in\mathcal{T}_h} \Bigg\{ \Bigg(&\underbrace{\left[\frac{\lambda_{2,K}^4}{\delta_K} + \delta_K \lambda_{2,K}^2 + \delta_K^3\,\frac{(\lambda_{1,K}^2+\lambda_{2,K}^2)^2}{\lambda_{1,K}^4}\right]}_{\text{(I)}} \|\mathbf{a}\|_{L^\infty(K)}\,\mathcal{H}(\mathrm{Pe}_K - 1) \\ &+ \underbrace{\left[\frac{\lambda_{2,K}^4}{\delta_K^2} + \lambda_{2,K}^2 + \delta_K^2 \frac{(\lambda_{1,K}^2+\lambda_{2,K}^2)^2}{\lambda_{1,K}^4}\right]}_{\text{(II)}} \mu\mathcal{H}(1-\mathrm{Pe}_K)\Bigg) \\ &\underbrace{\left[s_K^4\, L_K(\mathbf{r}_{1,K},\,\mathbf{r}_{1,K};\,u) + L_K(\mathbf{r}_{2,K},\,\mathbf{r}_{2,K};\,u) + 2s_K^2\, L_K(\mathbf{r}_{1,K},\,\mathbf{r}_{2,K};\,u)\right]}_{\text{(III)}}\Bigg\},\end{aligned} \tag{3.27}$$

where the term (III) is now equivalent to the H^2-norm of u on K, on recalling the definition (2.4) and that s_K is a dimensionless quantity. Moreover, $1 < \frac{(\lambda_{1,K}^2+\lambda_{2,K}^2)^2}{\lambda_{1,K}^4} \le 4$ so that it does not play any role in the convergence analysis. Let us first deal with the isotropic case, where $\lambda_{1,K} \simeq \lambda_{2,K} \simeq h_K$. By looking at term (I) of (3.27), it turns out that the maximal order of convergence is obtained when all these terms are of the same order. This occurs when $\delta_K \simeq h_K$; i.e., we recover the recipe in [22]. In the anisotropic case, letting $\delta_K \simeq \lambda_{1,K}^m \lambda_{2,K}^n$ for some m, $n \in \mathbb{Q}$, we find these values by requiring that all the three terms in (I) be of the same order with respect to both $\lambda_{1,K}$ and $\lambda_{2,K}$. By doing so, it turns out that $m = 0$ and $n = 1$, which yields $\delta_K \simeq \lambda_{2,K}$. It also turns out that, under this choice, (I) behaves like $\lambda_{2,K}^3$.

By a similar line of reasoning, it can be checked that the same value for δ_K is obtained also for term (II). However, this behaves like $\lambda_{2,K}^2$.

Having computed the value of δ_K, relations (3.13)–(3.14) follow immediately on recalling (3.7) and (3.9). □

Notice that in the above proof the parameter δ_K is determined up to a constant. The definitions (3.13)–(3.14) are consistent with a choice of this constant equal to 1.

4. The Stokes problem.

In this section we extend the result obtained in section 3 to the case of the Stokes problem.

In the very same spirit as in the advection-diffusion case, starting from the GLS formulation presented in [20, 21], we readdress the question of a careful design for the stability coefficients in the anisotropic framework. With regard to this, in [4, 5] the authors consider the stabilization of the Stokes problem in the case of the Q_1/Q_1 pair of finite elements on anisotropic quadrilateral meshes aligned with the Cartesian coordinate axes.

In this section, we provide a possible generalization of the convergence results obtained in Theorem 3.1 in [20], in the case of continuous piecewise linear finite elements for both the velocity and the pressure, to the situation of a general anisotropic mesh. We study the convergence of the stabilized method in a mesh dependent norm, taking into account also the stability coefficients τ_K. The optimal value for this is obtained by error analysis considerations and turns out to depend on $\lambda_{2,K}$, as should be expected after the analysis leading to Theorem 3.1.

The main contribution of this section is Theorem 4.1.

4.1. Problem statement and finite element discretization.

Given the viscosity $\mu = \text{const} > 0$ and the source term $\mathbf{f} = \mathbf{f}(\mathbf{x}) \in (L^2(\Omega))^2$, we seek $\mathbf{u} = \mathbf{u}(\mathbf{x})$ and $p = p(\mathbf{x})$ such that

$$\begin{cases} -\mu\,\Delta\mathbf{u} + \nabla p = \mathbf{f} & \text{in } \Omega, \\ \nabla\cdot\mathbf{u} = 0 & \text{in } \Omega, \\ \mathbf{u} = \mathbf{0} & \text{on } \partial\Omega. \end{cases} \tag{4.1}$$

The variational formulation for the above problem consists of finding $(\mathbf{u}, p) \in V \times Q$ such that

$$B(\mathbf{u}, p; \mathbf{v}, q) = F(\mathbf{v}, q) \quad \text{for any } (\mathbf{v}, q) \in V \times Q. \tag{4.2}$$

Here $V = (H_0^1(\Omega))^2$, $Q = L_0^2(\Omega)$, while $B(\cdot\,;\cdot)$ and $F(\cdot)$ now are the symmetric bilinear and linear forms

$$B(\mathbf{u}, p; \mathbf{v}, q) = \mu(\nabla\mathbf{u}, \nabla\mathbf{v}) - (p, \nabla\cdot\mathbf{v}) - (q, \nabla\cdot\mathbf{u}) \quad \text{and} \quad F(\mathbf{v}, q) = (\mathbf{f}, \mathbf{v}),$$

respectively, for any $(\mathbf{u}, p)$, $(\mathbf{v}, q) \in V \times Q$. Let us notice that we are using the same notation to address the bilinear and linear forms $B(\cdot\,;\cdot)$ and $F(\cdot)$ for both the advection-diffusion and the Stokes problems.

As is done in the advection-diffusion case, we discretize problem (4.2) by using the GLS method with affine finite elements [20, 21]. The discrete problem is as follows: find $(\mathbf{u}_h, p_h) \in V_h \times Q_h$ such that

$$B_h(\mathbf{u}_h, p_h; \mathbf{v}_h, q_h) = F_h(\mathbf{v}_h, q_h) \quad \text{for any } (\mathbf{v}_h, q_h) \in V_h \times Q_h, \tag{4.3}$$

where $V_h \times Q_h \subset V \times Q$ is the approximation space for velocity and pressure comprising continuous affine functions over $\mathcal{T}_h$, namely $V_h = (W_{h,0})^2$ and $Q_h = W_h \cap L_0^2(\Omega)$. Here the symmetric bilinear form $B_h(\cdot\,;\cdot)$ and the linear form $F_h(\cdot)$ are defined by

$$B_h(\mathbf{u}_h, p_h; \mathbf{v}_h, q_h) = B(\mathbf{u}_h, p_h; \mathbf{v}_h, q_h) - \sum_{K\in\mathcal{T}_h} (-\mu\,\Delta\mathbf{u}_h + \nabla p_h, \tau_K(-\mu\,\Delta\mathbf{v}_h + \nabla q_h))_K,$$

(4.4)

$$F_h(\mathbf{v}_h, q_h) = F(\mathbf{v}_h, q_h) - \sum_{K\in\mathcal{T}_h} (\mathbf{f}, \tau_K(-\mu\,\Delta\mathbf{v}_h + \nabla q_h))_K,$$

with τ_K stability coefficients to be suitably chosen. The terms $\Delta \mathbf{u}_h\big|_K$ and $\Delta \mathbf{v}_h\big|_K$ in (4.4) are identically equal to zero due to the choice made for the finite element space V_h.

We point out that the technique suggested in [4, 5] is a variant with respect to the GLS method where the stabilizing term, proportional to $(\nabla p_h, \nabla q_h)_K$, is replaced by $(\nabla p_h, S\,\nabla q_h)_K$, with $S = \mathrm{diag}(h_{x_1}^2, h_{x_2}^2)$, h_{x_1}, h_{x_2} being the mesh spacings in the coordinate directions. This term is shown to produce more satisfactory numerical results in the case of anisotropic meshes among several choices of the stabilization term.

4.2. Error analysis. As in [20, 21], the GLS scheme (4.3) is said to be *consistent* in the following sense. If the solution $(\mathbf{u}, p) \in V \times Q$ of the weak formulation (4.2) is regular enough, that is, if $(\mathbf{u}, p) \in (V \cap (H^2(\Omega))^2) \times (Q \cap H^1(\Omega))$, then $(\mathbf{u}, p)$ satisfies

$$B_h(\mathbf{u}, p; \mathbf{v}_h, q_h) \;=\; F_h(\mathbf{v}_h, q_h) \quad \text{for any } (\mathbf{v}_h, q_h) \in V_h \times Q_h.$$

Consequently, if $(\mathbf{u}_h, p_h) \in V_h \times Q_h$ is the solution to (4.3), we obtain the Galerkin orthogonality property

$$B_h(\mathbf{u} - \mathbf{u}_h, p - p_h; \mathbf{v}_h, q_h) \;=\; 0 \quad \text{for any } (\mathbf{v}_h, q_h) \in V_h \times Q_h. \tag{4.5}$$

Likewise, as was done in section 3.2, we introduce the discrete norm $\|\cdot\|_h$ defined for any $(\mathbf{v}, q) \in V \times (Q \cap H^1(\Omega))$ by

$$\|(\mathbf{v}, q)\|_h^2 = \mu \|\nabla \mathbf{v}\|^2_{L^2(\Omega)} + \sum_{K \in \mathcal{T}_h} \|\tau_K^{1/2} \nabla q\|^2_{L^2(K)}. \tag{4.6}$$

We now state the main result of this section, which is an anisotropic counterpart of Theorem 3.1 in [20] restricted to the case of (continuous) affine elements for both the velocity and the pressure. Moreover, we provide estimates in a different norm, namely the discrete norm $\|\cdot\|_h$ in (4.6), while in [20] the errors $\|\mathbf{u} - \mathbf{u}_h\|_{(H^1(\Omega))^2}$, $\|\mathbf{u} - \mathbf{u}_h\|_{(L^2(\Omega))^2}$, and $\|p - p_h\|_{L^2(\Omega)}$ are considered.

THEOREM 4.1 (convergence in norm $\|\cdot\|_h$). *Let $(\mathbf{u}, p) \in (V \cap (H^2(\Omega))^2) \times (Q \cap H^1(\Omega))$ be the solution to* (4.2)*, and let $(\mathbf{u}_h, p_h) \in V_h \times Q_h$ be the solution to* (4.3)*. Then the anisotropic definition of the stability coefficients is*

$$\tau_K = \alpha \frac{\lambda_{2,K}^2}{\mu}, \tag{4.7}$$

where $\alpha \simeq O(1)$ is any positive constant. Moreover, under this choice there exists a constant $C = C(\Gamma, \widehat{C}, \widehat{K})$ such that it holds that

$$\begin{aligned}
&\|(\mathbf{u} - \mathbf{u}_h, p - p_h)\|_h^2 \le \\
&C \sum_{K \in \mathcal{T}_h} \Bigg\{ \lambda_{2,K}^2 \Bigg(\mu \Big[s_K^4\, L_K(\mathbf{r}_{1,K}, \,\mathbf{r}_{1,K}; \,\mathbf{u}) + L_K(\mathbf{r}_{2,K}, \mathbf{r}_{2,K}; \,\mathbf{u}) + 2 s_K^2 L_K(\mathbf{r}_{1,K}, \,\mathbf{r}_{2,K}; \,\mathbf{u}) \Big] \\
&\qquad + \frac{1}{\mu} \Big[s_K^2 (\mathbf{r}_{1,K}^T G_K(p)\, \mathbf{r}_{1,K}) + (\mathbf{r}_{2,K}^T G_K(p)\, \mathbf{r}_{2,K}) \Big] \Bigg) \Bigg\}.
\end{aligned} \tag{4.8}$$

Remark 3. Notice that in the case when $\lambda_{1,K} \simeq \lambda_{2,K}$, i.e., in the isotropic case, (4.8) recovers Theorem 3.1 in [20] in the case of affine elements. As far as the term depending on the velocity in (4.8) is concerned, we note that it has the same structure as the corresponding one in Theorem 3.1. Thus the same considerations discussed in the case of the advection-diffusion problem carry over to the case of the Stokes problem, provided that the contribution of the terms depending on the pressure in (4.8) are negligible. Under this assumption we may have convergence even if s_K is unbounded. Likewise, in the case when the terms depending on $\mathbf{u}$ are negligible, convergence may be achieved independently of the stretching factor, provided that $\mathbf{r}_{1,K}^T G_K(p)\,\mathbf{r}_{1,K} \simeq 0$; i.e., the gradients of the pressure in the direction $\mathbf{r}_{1,K}$ are small. Thus, as in the case of the advection-diffusion problem, roughly speaking, if the mesh is aligned with the solution, then convergence occurs independently of the stretching factor. Otherwise, convergence may not occur.

Theorem 4.1 is proved by analyzing both the stability and the continuity of the bilinear form $B_h(\cdot;\cdot)$. First let us provide the stability result.

Lemma 4.2 (stability in norm $\|\cdot\|_h$). *For any* $(\mathbf{v}_h, q_h) \in V_h \times Q_h$, *we have*

$$B_h(\mathbf{v}_h, -q_h; \mathbf{v}_h, q_h) = \|(\mathbf{v}_h, q_h)\|_h^2. \tag{4.9}$$

Therefore (4.3) *has a unique solution.*

Proof. Set $\mathbf{u}_h = \mathbf{v}_h$ and $p_h = -q_h$ in $(4.4)_1$. □

With the next lemma we analyze the continuity of the bilinear form $B_h(\cdot\,;\cdot)$.

Lemma 4.3 (continuity of $B_h(\cdot\,;\cdot)$). *For any* $(\mathbf{u}, p) \in (V \cap (H^2(\Omega))^2) \times (Q \cap H^1(\Omega))$ *and for any* $(\mathbf{v}_h, q_h) \in V_h \times Q_h$, *there exists a constant C such that*

$$|B_h(\mathbf{u}, p; \mathbf{v}_h, q_h)| \le C\Bigg[\mu\|\nabla\mathbf{u}\|_{L^2(\Omega)}^2 + \frac{1}{\mu}\|p\|_{L^2(\Omega)}^2 + \sum_{K\in\mathcal{T}_h}\Big(\|\tau_K^{-1/2}\mathbf{u}\|_{L^2(K)}^2 + \|\tau_K^{1/2}\mu\Delta\mathbf{u}\|_{L^2(K)}^2 + \|\tau_K^{1/2}\nabla p\|_{L^2(K)}^2\Big)\Bigg]^{1/2} \|(\mathbf{v}_h, q_h)\|_h.$$

Proof. From the definition of $B_h(\cdot\,;\cdot)$ we have

$$\begin{aligned} B_h(\mathbf{u}, p; \mathbf{v}_h, q_h) &= \mu(\nabla\mathbf{u}, \nabla\mathbf{v}_h) - (p, \nabla\cdot\mathbf{v}_h) - (q_h, \nabla\cdot\mathbf{u}) \\ &\quad - \sum_{K\in\mathcal{T}_h}(-\mu\,\Delta\mathbf{u} + \nabla p, \tau_K(-\mu\,\Delta\mathbf{v}_h + \nabla q_h))_K. \end{aligned}$$

Integrating by parts the third term in the right-hand side of the above equality and since $\Delta\mathbf{v}_h\big|_K = 0$ on each triangle $K \in \mathcal{T}_h$, we have

$$\begin{aligned} B_h(\mathbf{u}, p; \mathbf{v}_h, q_h) &= \mu(\nabla\mathbf{u}, \nabla\mathbf{v}_h) - (p, \nabla\cdot\mathbf{v}_h) + \sum_{K\in\mathcal{T}_h}(\mathbf{u}, \nabla q_h)_K \\ &\quad - \sum_{K\in\mathcal{T}_h}(-\mu\,\Delta\mathbf{u} + \nabla p, \tau_K\nabla q_h)_K. \end{aligned}$$

Using the Cauchy–Schwarz inequality and the straightforward relation $\|\nabla\cdot\mathbf{v}_h\|_{L^2(\Omega)} \le \sqrt{2}\|\nabla\mathbf{v}_h\|_{L^2(\Omega)}$, we obtain

$$\begin{aligned} &|B_h(\mathbf{u}, p; \mathbf{v}_h, q_h)| \le \mu\|\nabla\mathbf{u}\|_{L^2(\Omega)}\|\nabla\mathbf{v}_h\|_{L^2(\Omega)} + \sqrt{2}\,\|p\|_{L^2(\Omega)}\|\nabla\mathbf{v}_h\|_{L^2(\Omega)} \\ &+\sum_{K\in\mathcal{T}_h}\|\mathbf{u}\|_{L^2(K)}\|\nabla q_h\|_{L^2(K)} +\sum_{K\in\mathcal{T}_h}\|\tau_K^{1/2}\nabla q_h\|_{L^2(K)}\big(\|\tau_K^{1/2}\mu\Delta\mathbf{u}\|_{L^2(K)}+\|\tau_K^{1/2}\nabla p\|_{L^2(K)}\big). \end{aligned}$$

Using the discrete Cauchy–Schwarz inequality, we thus have

$$|B_h(\mathbf{u},p;\mathbf{v}_h,q_h)| \le C\biggl(\mu\|\nabla\mathbf{u}\|^2_{L^2(\Omega)} + \frac{1}{\mu}\|p\|^2_{L^2(\Omega)} + \sum_{K\in\mathcal{T}_h}\Bigl(\|\tau_K^{-1/2}\mathbf{u}\|^2_{L^2(K)} + \|\tau_K^{1/2}\mu\Delta\mathbf{u}\|^2_{L^2(K)} + \|\tau_K^{1/2}\nabla p\|^2_{L^2(K)}\Bigr)\biggr)^{1/2} \biggl(\mu\|\nabla\mathbf{v}_h\|^2_{L^2(\Omega)} + \sum_{K\in\mathcal{T}_h}\|\tau_K^{1/2}\nabla q_h\|^2_{L^2(K)}\biggr)^{1/2},$$

which yields the result. □

The proofs of Lemmas 4.2 and 4.3 allow us to derive an anisotropic *a priori error estimate* for the discretization error of both the velocity and the pressure with respect to the discrete norm $\|\cdot\|_h$ defined in (4.6).

PROPOSITION 4.4 (a priori error estimate in norm $\|\cdot\|_h$). *Let* $(\mathbf{u},p)\in(V\cap(H^2(\Omega))^2)\times(Q\cap H^1(\Omega))$ *be the solution to* (4.2), *and let* $(\mathbf{u}_h,p_h)\in V_h\times Q_h$ *be the solution to* (4.3). *Then there exists a constant* $C=C(\Gamma,\widehat{C},\widehat{K})$ *such that*

$$\|(\mathbf{u}-\mathbf{u}_h,p-p_h)\|_h^2 \le C\Biggl\{\sum_{K\in\mathcal{T}_h}\biggl[\frac{1}{\tau_K}+\frac{\mu}{\lambda_{2,K}^2}+\frac{(\lambda_{1,K}^2+\lambda_{2,K}^2)^2}{\lambda_{1,K}^4\,\lambda_{2,K}^4}\tau_K\mu^2\biggr] \Bigl[\lambda_{1,K}^4\,L_K(\mathbf{r}_{1,K},\,\mathbf{r}_{1,K};\,\mathbf{u})+\lambda_{2,K}^4\,L_K(\mathbf{r}_{2,K},\,\mathbf{r}_{2,K};\,\mathbf{u})+2\lambda_{1,K}^2\,\lambda_{2,K}^2\,L_K(\mathbf{r}_{1,K},\,\mathbf{r}_{2,K};\,\mathbf{u})\Bigr] + \sum_{K\in\mathcal{T}_h}\biggl[\frac{1}{\mu}+\frac{\tau_K}{\lambda_{2,K}^2}\biggr]\Bigl[\lambda_{1,K}^2(\mathbf{r}_{1,K}^T G_K(p)\,\mathbf{r}_{1,K})+\,\lambda_{2,K}^2(\mathbf{r}_{2,K}^T G_K(p)\,\mathbf{r}_{2,K})\Bigr]\Biggr\}, \tag{4.10}$$

where the quantities $L_K(\mathbf{r}_{i,K},\,\mathbf{r}_{j,K};\,\mathbf{u})$ *and the matrix* G_K *are defined in* (2.19) *and* (2.5)*, respectively.*

Proof. We point out that the hypothesis $(\mathbf{u},p)\in(V\cap(H^2(\Omega))^2)\times(Q\cap H^1(\Omega))$, which amounts to requiring the elliptic regularity, holds, e.g., when Ω is convex. We have

$$\|(\mathbf{u}-\mathbf{u}_h,p-p_h)\|_h \le \|(\mathbf{u}-r_h(\mathbf{u}),p-R_h(p))\|_h + \|(r_h(\mathbf{u})-\mathbf{u}_h,R_h(p)-p_h)\|_h.$$

Thanks to Lemma 4.2 and to the Galerkin orthogonality property (4.5), we have

$$\begin{aligned}\|(r_h(\mathbf{u})-\mathbf{u}_h,R_h(p)-p_h)\|_h^2 &= B_h(r_h(\mathbf{u})-\mathbf{u}_h,R_h(p)-p_h;r_h(\mathbf{u})-\mathbf{u}_h,p_h-R_h(p))\\ &= B_h(r_h(\mathbf{u})-\mathbf{u},R_h(p)-p;r_h(\mathbf{u})-\mathbf{u}_h,p_h-R_h(p)).\end{aligned}$$

With Lemma 4.3 and since $\Delta(r_h(\mathbf{u}))\big|_K=0$ on each triangle $K\in\mathcal{T}_h$, we have

$$\|(r_h(\mathbf{u})-\mathbf{u}_h,R_h(p)-p_h)\|_h \le C\biggl(\mu\|\nabla(\mathbf{u}-r_h(\mathbf{u}))\|^2_{L^2(\Omega)}+\frac{1}{\mu}\,\|p-R_h(p)\|^2_{L^2(\Omega)} + \sum_{K\in\mathcal{T}_h}\Bigl(\|\tau_K^{-1/2}(\mathbf{u}-r_K(\mathbf{u}))\|^2_{L^2(K)}+\|\tau_K^{1/2}\mu\Delta\mathbf{u}\|^2_{L^2(K)}+\|\tau_K^{1/2}\nabla(p-R_K(p))\|^2_{L^2(K)}\Bigr)\biggr)^{1/2}.$$

By the additivity of the norms, it then suffices to use the interpolation results of Lemmas 2.3 and Proposition 2.5, together with the vectorial extensions of Lemma 2.4 and Corollary 2.7 (see Remark 2), to conclude. □

Proof of Theorem 4.1. We proceed by mimicking what was done in the proof of Theorem 3.1. Thus, using in (4.10) the definition of stretching factor in the terms depending on $L_K(\mathbf{r}_{i,K},\, \mathbf{r}_{j,K};\, \mathbf{u})$ and on $\mathbf{r}_{i,K}^T G_K(p)\, \mathbf{r}_{i,K}$, we obtain

$$(4.11)\quad \|(\mathbf{u}-\mathbf{u}_h, p-p_h)\|_h^2 \le C \sum_{K\in\mathcal{T}_h} \Bigg\{ \underbrace{\left(\frac{\lambda_{2,K}^4}{\tau_K} + \mu\lambda_{2,K}^2 + \frac{(\lambda_{1,K}^2+\lambda_{2,K}^2)^2}{\lambda_{1,K}^4}\mu^2\tau_K \right)}_{\text{(I)}}$$

$$\Big[s_K^4\, L_K(\mathbf{r}_{1,K},\, \mathbf{r}_{1,K};\, \mathbf{u}) + L_K(\mathbf{r}_{2,K},\, \mathbf{r}_{2,K};\, \mathbf{u}) + 2s_K^2 L_K(\mathbf{r}_{1,K},\, \mathbf{r}_{2,K};\, \mathbf{u}) \Big]$$

$$+ \underbrace{\left(\frac{\lambda_{2,K}^2}{\mu} + \tau_K \right)}_{\text{(II)}} \Big[s_K^2 (\mathbf{r}_{1,K}^T G_K(p)\, \mathbf{r}_{1,K}) + (\mathbf{r}_{2,K}^T G_K(p)\, \mathbf{r}_{2,K}) \Big] \Bigg\}.$$

Now, by inspecting terms (I) and (II) above, it turns out that the optimal value for τ_K is of the order of $\lambda_{2,K}^2/\mu$ in both cases; i.e., $\tau_K = \alpha\lambda_{2,K}^2/\mu$ for some $\alpha > 0$. Moreover, under this choice terms (I) and (II) behave like $\mu\lambda_{2,K}^2$ and $\lambda_{2,K}^2/\mu$, respectively. This concludes the proof. □

Remark 4. Under the same choice of the stability coefficients τ_K as obtained in the previous analysis, i.e., $\tau_K = \alpha\,\lambda_{2,K}^2/\mu$, we can further prove a convergence result with respect to the new norm $\|\cdot\|_{V\times Q}$ defined, for any $(\mathbf{v}, q) \in V \times Q$, by

$$(4.12)\qquad \|(\mathbf{v}, q)\|_{V\times Q}^2 = \mu\|\nabla\mathbf{v}\|_{L^2(\Omega)}^2 + \frac{1}{\mu}\, \frac{1}{\left(1 + \max\limits_{K\in\mathcal{T}_h} s_K^2\right)} \|q\|_{L^2(\Omega)}^2,$$

provided that the stability result in Lemma 4.2 is replaced by the new one stated in Lemma A.2 in the appendix. We remark that we have not been able to prove an optimal estimate yet with respect to the norm in (4.12), since our result still depends on the stretching factors s_K while the numerical results do not. This is the reason why we confine the proof of the new stability result to the appendix.

5. Numerical results.

In this section we first show numerically that on strongly anisotropic meshes better results are derived by using the definitions (3.13) and (4.7) of the stability coefficients τ_K when compared with the ones in [20, 21, 22]. Then we check that the convergence rate proved in Theorems 3.1 and 4.1 is confirmed by the numerical results. We also provide a numerical assessment of an adaptive anisotropic a posteriori error procedure based on standard gradient recovery techniques [37]. Two test cases for the advection-diffusion problem with parabolic internal and boundary layers, as well as with outflow boundary layers, are carried out using (3.13).

5.1. The advection-diffusion problem.

Let us consider the standard one-dimensional boundary layer problem

$$\begin{cases} -\mu\,\dfrac{d^2 v}{dx_1^2} + \dfrac{dv}{dx_1} = 1 & \text{in } \Omega \equiv (0,\,1), \\ v = 0 & \text{on } \partial\Omega, \end{cases}$$

with solution

$$v(x_1) = x_1 - \frac{1-\exp\left(\dfrac{x_1}{\mu}\right)}{1-\exp\left(\dfrac{1}{\mu}\right)}.$$

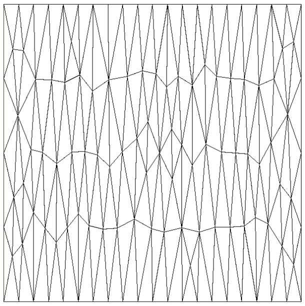


FIG. 5.1. *The advection-diffusion problem. A* 20×4 *anisotropic mesh of the domain* Ω. *All the meshes have been obtained using the* BL2D *library* [8].

To extend this test case in the two-dimensional framework, it suffices to set $\Omega = (0,1)^2$, $f = 1$, $\mathbf{a} = (1,0)^T$ in (3.1), while imposing homogeneous Dirichlet and homogeneous Neumann boundary conditions on the vertical and horizontal sides of Ω, respectively. The solution of (3.1) is thus given by $u(x_1, x_2) = v(x_1)$.

Validation of the stability coefficients. We consider the stabilized formulation (3.4), τ_K being defined, for any $K \in \mathcal{T}_h$, by

$$\tau_K = \frac{\lambda_{2,K}}{2} \frac{\xi(\mathrm{Pe}_K)}{\|\mathbf{a}\|_{L^\infty(K)}} \quad \text{with} \quad \mathrm{Pe}_K = \lambda_{2,K} \frac{\|\mathbf{a}\|_{L^\infty(K)}}{6\,\mu} \tag{5.1}$$

and

$$\tau_K = \frac{|K|^{1/2}}{2} \frac{\xi(\mathrm{Pe}_K)}{\|\mathbf{a}\|_{L^\infty(K)}} \quad \text{with} \quad \mathrm{Pe}_K = |K|^{1/2} \frac{\|\mathbf{a}\|_{L^\infty(K)}}{6\,\mu}, \tag{5.2}$$

respectively. Formula (5.2) corresponds to the stabilized method in [22] where the triangle diameter h_K is replaced by $|K|^{1/2}$, this replacement aiming at reducing the numerical dissipation. On the other hand, (5.1) represents our anisotropic design. Moreover, in all the numerical results, the reference element $\widehat{K}$ has been chosen as the right triangle $(0,0), (1,0), (0,1)$, and the values of $\lambda_{2,K}$ have been computed using a singular value decomposition routine.

In order to show that (5.1) yields better results than (5.2) in the presence of strongly anisotropic meshes, let us consider a 1000×4 anisotropic mesh of Ω of the type shown in Figure 5.1. In Figure 5.2 we have reported u_h along the bottom side of Ω for several values of the diffusivity μ. The choice (5.1) clearly turns out to be better than (5.2), with numerical diffusion being lower inside the boundary layer.

Convergence rate. In order to check the convergence rate of the GLS method with (5.1), we set $\mu = 10^{-2}$ in (3.1), select the anisotropic mesh in Figure 5.1, and refine the mesh size in the horizontal direction only. From Table 5.1 we observe that the order of convergence of the discretization error is equal to one with respect to the discrete norm $\|\cdot\|_h$ defined in (3.12). This agrees with the theoretical predictions (see Theorem 3.1). Moreover, the exact and the computed solutions along the bottom side of Ω are shown in Figure 5.3 in the presence of three different meshes.

A posteriori validation. Let us assess the behavior of recipe (5.1) on test cases characterized by two-dimensional anisotropic features (see also [34]). With this aim, we have carried out an adaptive iterative procedure based on the a posteriori analysis of [37]. With reference to (3.1), the data for the first test case (T_1) are $\mu = 10^{-4}$, $\mathbf{a} = (2,1)^T$, $f = 0$, $\Omega = (0,1)^2$, completed with Dirichlet boundary conditions, i.e.,

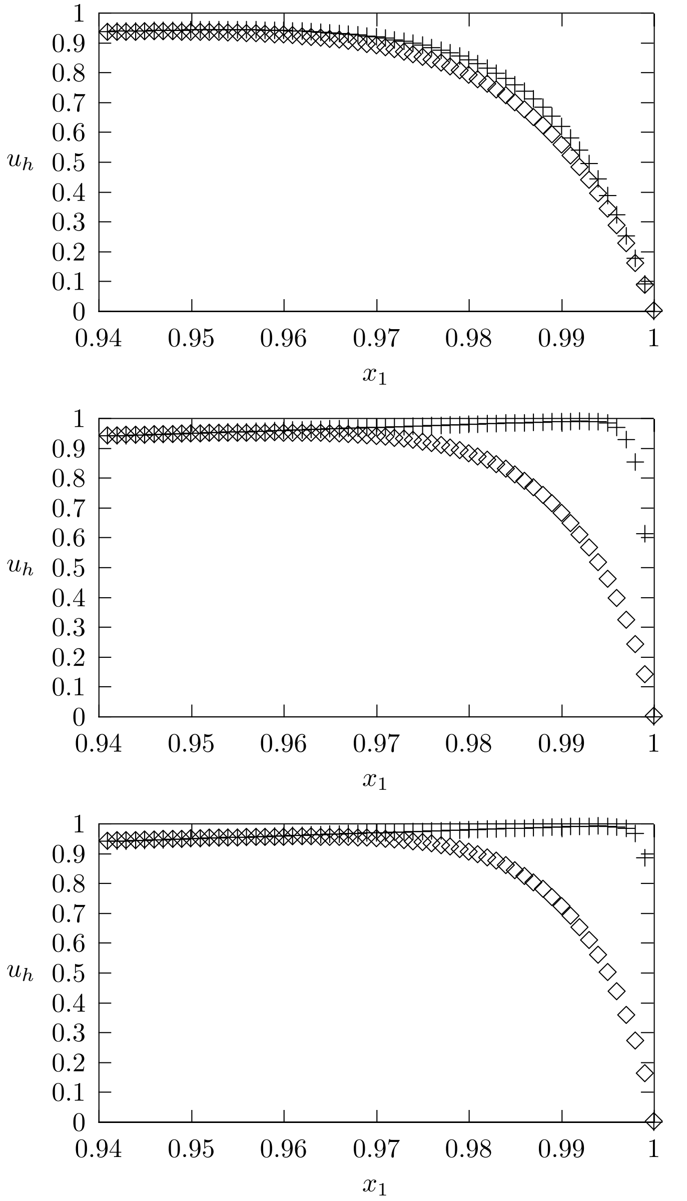


FIG. 5.2. *The advection-diffusion problem. Plots of* u_h *along the bottom side of* Ω *when using a* 1000×4 *anisotropic mesh. Only the part of the plot corresponding to* $0.94 \leq x_1 \leq 1$ *is displayed. Top:* $\mu = 10^{-2}$*; middle:* $\mu = 10^{-3}$*; bottom:* $\mu = 10^{-4}$*. Crosses: GLS plus* (5.1)*; diamonds: GLS plus* (5.2)*.*

$u = 1$ on the left and top sides and $u = 0$ on the remaining ones. For the second test case (T_2) we set $\mu = 10^{-4}$, $\mathbf{a} = (-1, 0)^T$, $f = 1$, $\Omega = (0,1)^2$, in addition to homogeneous Dirichlet boundary conditions on $\partial\Omega$. Notice that we expect the solutions of test cases (T_1) and (T_2) to show an internal layer and a boundary layer, and parabolic boundary layers, respectively.

TABLE 5.1
The advection-diffusion problem. Errors with respect to the discrete norm (3.12) *in the presence of different anisotropic meshes and with* $\mu = 10^{-2}$ *in* (3.1).

Mesh	$\|u - u_h\|_h$
20×4	0.95
40×4	0.38
80×4	0.17
160×4	0.091
320×4	0.045
640×4	0.022

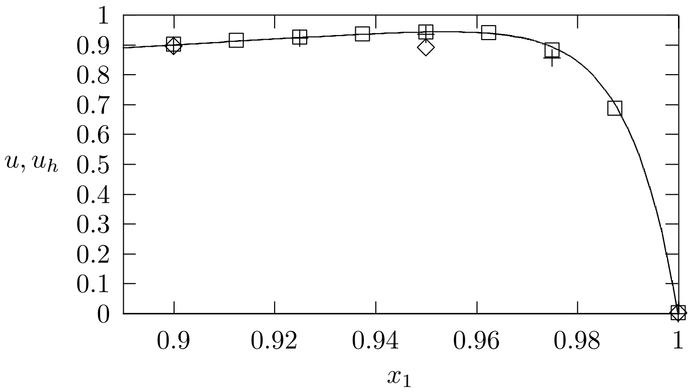


FIG. 5.3. *The advection-diffusion problem. Continuous line: profile of* u *along the bottom side of* Ω *with* $\mu = 10^{-2}$ *in* (3.1). *Solution obtained by GLS plus* (5.1) *on a:* 20×4 *mesh (diamonds);* 40×4 *mesh (crosses);* 80×4 *mesh (squares).*

Figure 5.4 shows the isolines between 0.1 and 0.9 (on the left) along with the adapted meshes (on the right) for both test cases. We remark that all the internal and boundary layers are very well captured by the adapted meshes consisting of less than 300 nodes. Moreover, it can be checked that the thickness of the internal and parabolic boundary layers is $O(\sqrt{\mu}) \simeq 10^{-2}$, while the one of the outflow boundary layers is $O(\mu) \simeq 10^{-4}$. Consequently, the aspect ratio of the two adapted meshes reaches large values of the order of 10^5.

5.2. Stokes problem. First we consider the classical Poiseuille test case in order to prove the superiority of the GLS method plus (4.7) compared with the one in [21]. Then we analyze a second test case to verify the convergence rate predicted by Theorem 4.1.

Validation of the stability coefficients. Due to the symmetry of the solution, let us choose $\Omega = (0, 0.15) \times (0, 0.03)$ (corresponding to half the physical domain) so that the Poiseuille exact solution is given by

$$\mathbf{u}(x_1, x_2) = \begin{bmatrix} (0.03 - x_2)(0.03 + x_2) \\ 0 \end{bmatrix} \quad \text{and} \quad p(x_1, x_2) = -0.02x_1 + 0.003.$$

Let us choose in (4.1) $\mathbf{f} = \mathbf{0}$ and $\mu = 10^{-2}$, while imposing Dirichlet and homogeneous Neumann boundary conditions for the velocity $\mathbf{u}$ along the left-top and bottom-right sides of Ω, respectively. Then we consider the stabilized formulation (4.3) with the

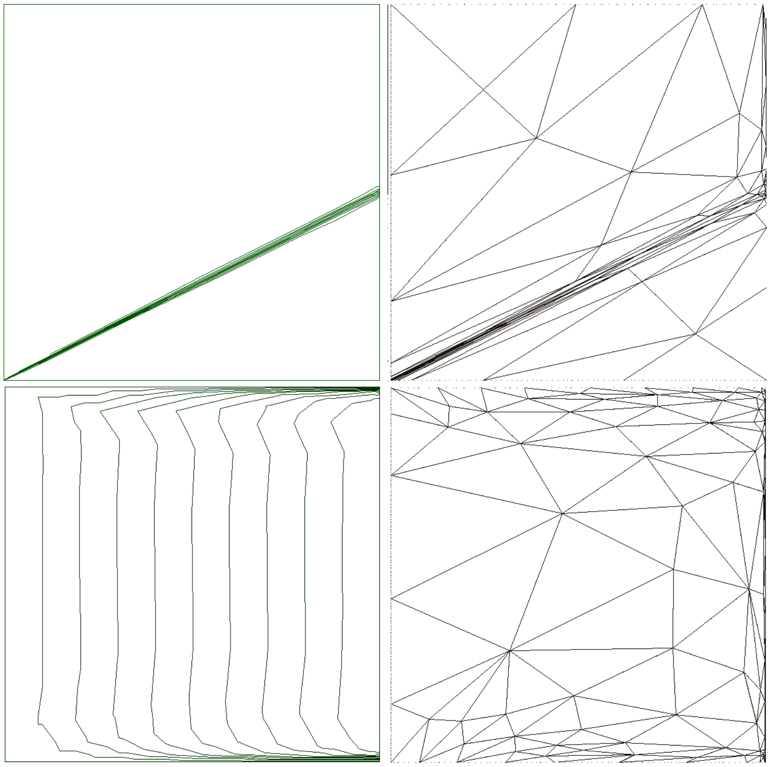


FIG. 5.4. *Isolines between* 0.1 *and* 0.9 *(left column) and adapted mesh (right column) for the test case* (T_1) *(top row) and for* (T_2) *(bottom row).*

coefficients τ_K defined, for any $K \in \mathcal{T}_h$, by

$$\tau_K = \alpha \, \frac{\lambda_{2,K}^2}{\mu} \tag{5.3}$$

and

$$\tau_K = \tilde{\alpha} \, \frac{|K|}{\mu}. \tag{5.4}$$

While (5.3) represents our design, formula (5.4) corresponds to the stabilized method of [21] with the triangle diameter h_K replaced by $|K|^{1/2}$ as was done in section 5.1. The two dimensionless coefficients α and $\tilde{\alpha}$ are tuned on the mesh of Figure 5.5 so that both computations give a reasonably precise solution.

In Figure 5.6 (top) the exact pressure profile along the bottom side of Ω is shown together with the stabilized approximate solutions with $\alpha = 0.1$ and $\tilde{\alpha} = 0.01$ in (5.3) and (5.4), respectively.

Then the same calculations are performed on the computational domain Ω stretched in the horizontal direction by a factor 10 and 100, that is, on $\Omega = (0, 1.5) \times (0, 0.03)$ and $\Omega = (0, 15) \times (0, 0.03)$. These new choices for the domain Ω allow us to preserve the total number of the elements and of the nodes of the mesh while increasing the maximal aspect ratio of the triangles (close to 1000 for the last mesh). The corresponding results are summarized in Figure 5.6 (middle and bottom). We

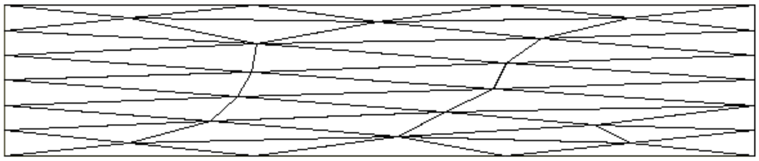


FIG. 5.5. *Stokes problem: Poiseuille test case. A* 3×6 *mesh of the domain* Ω.

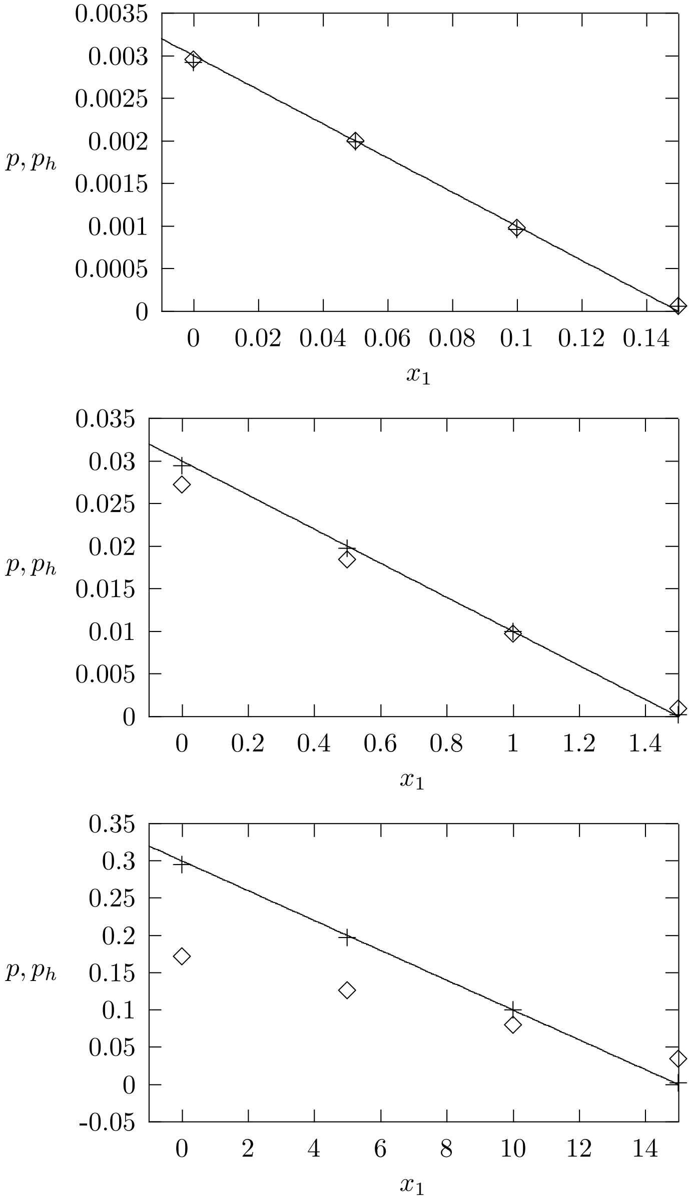


FIG. 5.6. *Stokes problem: Poiseuille test case. Continuous line: exact pressure profile along the bottom side of* Ω*; diamonds: solution obtained by GLS plus* (5.4) *with* $\tilde{\alpha} = 0.01$*; crosses: solution obtained by GLS plus* (5.3) *with* $\alpha = 0.1$. *Top:* $\Omega = (0, 0.15) \times (0, 0.03)$*; middle:* $\Omega = (0, 1.5) \times (0, 0.03)$*; bottom:* $\Omega = (0, 15) \times (0, 0.03)$.

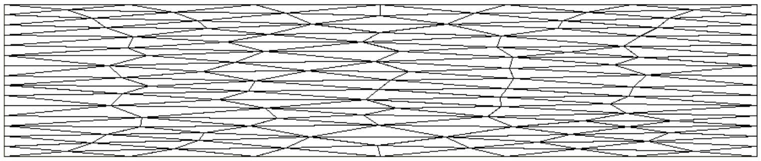


FIG. 5.7. *Stokes problem: second test case. A* 6×15 *mesh of the domain* Ω.

TABLE 5.2
Stokes problem: second test case. Errors with respect to the discrete norm (4.6) *in the presence of different anisotropic meshes and for the choice* $\mu = 10^{-2}$ *in* (4.1).

Mesh	$\|(\mathbf{u}-\mathbf{u}_h, p-p_h)\|_h$	$\|p-p_h\|_{L^2(\Omega)}$
6×30	0.83	0.9
6×60	0.31	0.37
6×120	0.15	0.18

observe that in all three cases the GLS method plus (5.3) yields better results than the GLS method plus (5.4).

Convergence rate. We consider the domain $\Omega = (0, 0.15) \times (0, 0.03)$ and choose in (4.1) $\mu = 10^{-2}$ and $\mathbf{f}$ so that the solution is given by

$$\mathbf{u}(x_1, x_2) = \left[\begin{array}{c} u_1(x_2) \\ 0 \end{array} \right] \quad \text{and} \quad p(x_1, x_2) = 0,$$

with

$$u_1(x_2) = \begin{cases} 1 & \text{if } \ 0 < x_2 \le 0.01, \\ \exp\left(\dfrac{(x_2 - 0.01)^2}{(x_2-0.01)^2 - 0.0001} \right) & \text{if } \ 0.01 < x_2 < 0.02, \\ 0 & \text{if } \ \ 0.02 \le x_2 < 0.03. \end{cases}$$

Thus the velocity is horizontal, smooth, zero, or one valued except in a region of width 0.01. We consider the same boundary conditions as in the previous example. Starting from the grid in Figure 5.7, we refine such a mesh in the vertical direction only. Then we approximate the solution of the Stokes problem by choosing as stability coefficients those defined in (5.3) with $\alpha = 0.1$. From Table 5.2 we observe that the error in the discrete norm $\| \cdot \|_h$ defined in (4.6) is of order one in agreement with the theoretical predictions (see Theorem 4.1). Moreover, it seems that the pressure error in the L^2-norm is of order one too. Finally, the exact and approximate velocity profiles computed along the vertical right side of the domain Ω are shown in Figure 5.8 in the presence of two different meshes.

We point out that the numerical results seem to show that the theory developed for homogeneous Dirichlet boundary conditions in (3.1) and (4.1) covers also the case of more general boundary conditions.

6. Conclusions. In this paper a theoretically sound design of the stability coefficients is proposed for the scalar advection-diffusion and the Stokes problems solved on strongly anisotropic meshes. Only continuous piecewise linear stabilized finite elements are considered.

Anisotropic adaptive finite elements have been developed with the present design of the stability coefficients (see [17, 37] for details). The numerical method seems to be robust since boundary layers of order 10^{-4} can be obtained with no more than

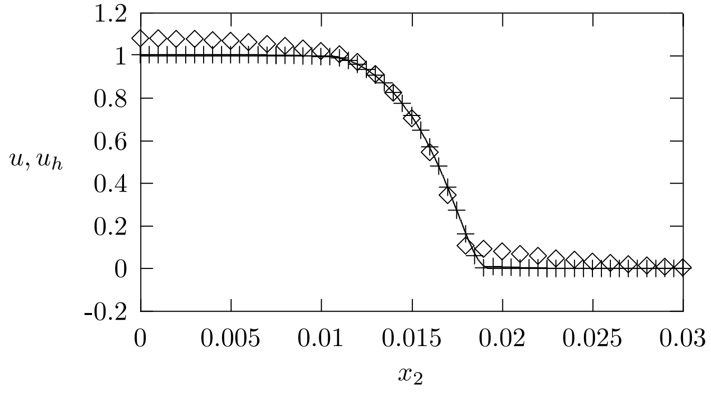


FIG. 5.8. *Stokes problem: second test case. Continuous line: profile of* $\mathbf{u}$ *along the vertical right side of* Ω *for the choice* $\mu = 10^{-2}$ *in* (4.1). *Solution obtained by GLS plus* (5.3) *on a:* 6×30 *mesh (diamonds);* 6×60 *mesh (crosses).*

300 vertices, the maximum aspect ratio being greater than 10^5. Even if the values obtained for both the aspect ratio and the number of vertices are no surprise in the literature (see, e.g., [35, 38]), we point out that our results have been obtained by a completely automatic adaptive procedure (see Figure 5.4).

Appendix. Let us assume that in the definition of the stability coefficients (4.7) $\alpha = 1$. First let us recall the standard inf-sup condition.

LEMMA A.1 (standard inf-sup condition). *There exists a constant* $\beta > 0$ *(depending only on the domain* Ω*) such that, for any* $p \in Q$*, we have*

$$\sup_{\substack{\mathbf{v}\in V\\ \mathbf{v}\neq\mathbf{0}}} \frac{-(p, \nabla\cdot\mathbf{v})}{\|\nabla\mathbf{v}\|_{L^2(\Omega)}} \geq \beta\|p\|_{L^2(\Omega)}.$$

Now we prove the stability of the bilinear form $B_h(\cdot,\cdot)$ associated with the Stokes problem with respect to the norm $\|\cdot\|_{V\times Q}$ defined in (4.12).

LEMMA A.2 (stability in norm $\|\cdot\|_{V\times Q}$). *There exists a constant* $C = C(\Gamma, \widehat{C}, \beta)$ *such that, for any* $(\mathbf{u}_h, p_h) \in V_h \times Q_h$*, we have*

$$\sup_{\substack{(\mathbf{v}_h,q_h)\in V_h\times Q_h\\ (\mathbf{v}_h,q_h)\neq(\mathbf{0},0)}} \frac{B_h(\mathbf{u}_h, p_h; \mathbf{v}_h, q_h)}{\|(\mathbf{v}_h, q_h)\|_{V\times Q}} \geq C\|(\mathbf{u}_h, p_h)\|_{V\times Q}.$$

Proof. It suffices to prove that there exist two constants $C_1 = C_1(\widehat{K}, \beta)$ and $C_2 = C_2(\widehat{K}, \beta)$ such that, for any $(\mathbf{u}_h, p_h) \in V_h \times Q_h$, there is a $(\mathbf{v}_h, q_h) \in V_h \times Q_h$ satisfying

$$\text{(A.1)}\quad B_h(\mathbf{u}_h, p_h; \mathbf{v}_h, q_h) \geq C_1\|(\mathbf{u}_h, p_h)\|^2_{V\times Q}, \quad \|(\mathbf{v}_h, q_h)\|_{V\times Q} \leq C_2\|(\mathbf{u}_h, p_h)\|_{V\times Q}.$$

We proceed as in [20] but in the frame of anisotropic meshes. The analogue of the so-called Verfürth's trick [41] is used in the presence of anisotropic meshes.

From Lemma A.1 there exists a function $\mathbf{v} \in V$ such that

$$\text{(A.2)}\qquad \|\nabla\mathbf{v}\|_{L^2(\Omega)} = \frac{1}{\mu}\|p_h\|_{L^2(\Omega)}$$

and

$$\frac{\beta}{\mu}\|p_h\|^2_{L^2(\Omega)} \leq -(p_h, \nabla\cdot\mathbf{v}). \tag{A.3}$$

Let $R_h(\mathbf{v})$ be the Clément interpolant of $\mathbf{v}$. Using the definition $(4.4)_1$ of the stabilized bilinear form $B_h(\cdot;\cdot)$ and integrating by parts we have

$$\begin{aligned} B_h(\mathbf{u}_h, p_h; R_h(\mathbf{v}), 0) &= \mu(\nabla\mathbf{u}_h, \nabla R_h(\mathbf{v})) - (p_h, \nabla\cdot R_h(\mathbf{v})) \\ &= \mu(\nabla\mathbf{u}_h, \nabla\mathbf{v}) - (p_h, \nabla\cdot\mathbf{v}) \\ &- \mu(\nabla\mathbf{u}_h, \nabla(\mathbf{v} - R_h(\mathbf{v}))) + (p_h, \nabla\cdot(\mathbf{v} - R_h(\mathbf{v}))) \\ &= \mu(\nabla\mathbf{u}_h, \nabla\mathbf{v}) - (p_h, \nabla\cdot\mathbf{v}) \\ &- \mu\sum_{K\in\mathcal{T}_h}(\nabla\mathbf{u}_h, \nabla(\mathbf{v} - R_K(\mathbf{v})))_K - \sum_{K\in\mathcal{T}_h}(\nabla p_h, \mathbf{v} - R_K(\mathbf{v}))_K. \end{aligned}$$

Notice that the stabilization term in the definition of $B_h(\cdot,\cdot)$ vanishes in this case. We now use the Cauchy–Schwarz inequality with (A.2) and (A.3) to obtain

$$\begin{aligned} B_h(\mathbf{u}_h, p_h; R_h(\mathbf{v}), 0) &\geq -\|\nabla\mathbf{u}_h\|_{L^2(\Omega)}\|p_h\|_{L^2(\Omega)} + \frac{\beta}{\mu}\|p_h\|^2_{L^2(\Omega)} \\ - \mu\sum_{K\in\mathcal{T}_h}\|\nabla\mathbf{u}_h\|_{L^2(K)}&\|\nabla(\mathbf{v} - R_K(\mathbf{v}))\|_{L^2(K)} - \sum_{K\in\mathcal{T}_h}\|\nabla p_h\|_{L^2(K)}\|\mathbf{v} - R_K(\mathbf{v})\|_{L^2(K)}. \end{aligned}$$

From the vectorial extensions of Lemma 2.3 and Proposition 2.5 (see Remark 2), there is a constant $C = C(\Gamma, \widehat{C})$ such that

$$\begin{aligned} B_h(\mathbf{u}_h, p_h; R_h(\mathbf{v}), 0) \geq &- \|\nabla\mathbf{u}_h\|_{L^2(\Omega)}\|p_h\|_{L^2(\Omega)} + \frac{\beta}{\mu}\|p_h\|^2_{L^2(\Omega)} \\ &- C\sum_{K\in\mathcal{T}_h}\left\{\left(\frac{\mu}{\lambda_{2,K}}\|\nabla\mathbf{u}_h\|_{L^2(K)} + \|\nabla p_h\|_{L^2(K)}\right)\right. \\ &\left.[\lambda^2_{1,K}(\mathbf{r}^T_{1,K}G_K(\mathbf{v})\mathbf{r}_{1,K}) + \lambda^2_{2,K}(\mathbf{r}^T_{2,K}G_K(\mathbf{v})\mathbf{r}_{2,K})]^{1/2}\right\}. \end{aligned}$$

Using Young inequality $ab \leq \frac{\gamma a^2}{2} + \frac{b^2}{2\gamma}$, with a, b and $\gamma > 0$, we then obtain

$$\begin{aligned} B_h(\mathbf{u}_h, p_h; R_h(\mathbf{v}), 0) \geq &-\frac{\mu}{2\beta}\|\nabla\mathbf{u}_h\|^2_{L^2(\Omega)} + \frac{\beta}{2\mu}\|p_h\|^2_{L^2(\Omega)} \\ &- C\,\frac{\gamma}{2}\sum_{K\in\mathcal{T}_h}\left(\mu\|\nabla\mathbf{u}_h\|^2_{L^2(K)} + \frac{\lambda^2_{2,K}}{\mu}\|\nabla p_h\|^2_{L^2(K)}\right) \\ &- C\,\frac{\mu}{\gamma}\sum_{K\in\mathcal{T}_h}\left\{\frac{1}{\lambda^2_{2,K}}[\lambda^2_{1,K}(\mathbf{r}^T_{1,K}G_K(\mathbf{v})\mathbf{r}_{1,K})\right. \\ &\left.+ \lambda^2_{2,K}(\mathbf{r}^T_{2,K}G_K(\mathbf{v})\mathbf{r}_{2,K})]\right\}, \end{aligned} \tag{A.4}$$

where γ has still to be chosen. Recalling the relation (2.11) and the hypothesis (2.1) on Δ_K and using (A.2) we have

$$\begin{aligned}\sum_{K\in\mathcal{T}_h}\left\{\frac{1}{\lambda_{2,K}^2}[\lambda_{1,K}^2(\mathbf{r}_{1,K}^T G_K(\mathbf{v})\mathbf{r}_{1,K})+\lambda_{2,K}^2(\mathbf{r}_{2,K}^T G_K(\mathbf{v})\mathbf{r}_{2,K})]\right\}&\\ \le C\left(1+\max_{K\in\mathcal{T}_h} s_K^2\right)\|\nabla\mathbf{v}\|_{L^2(\Omega)}^2 = C\left(1+\max_{K\in\mathcal{T}_h} s_K^2\right)\frac{1}{\mu^2}\|p_h\|_{L^2(\Omega)}^2,&\end{aligned}\tag{A.5}$$

with $C=C(\Gamma)$. Replacing the above estimate in (A.4) we can then choose γ so that

$$\begin{aligned}B_h(\mathbf{u}_h,p_h;R_h(\mathbf{v}),0)\ge\frac{\beta}{4\mu}\|p_h\|_{L^2(\Omega)}^2&\\ -\frac{C}{\beta}\left(1+\max_{K\in\mathcal{T}_h}s_K^2\right)\left(\mu\|\nabla\mathbf{u}_h\|_{L^2(\Omega)}^2+\sum_{K\in\mathcal{T}_h}\frac{\lambda_{2,K}^2}{\mu}\|\nabla p_h\|_{L^2(K)}^2\right),&\end{aligned}\tag{A.6}$$

where $C=C(\Gamma,\widehat{C})$. On the other hand, from Lemma 4.2 and (4.7) we know that

$$B_h(\mathbf{u}_h,p_h;\mathbf{u}_h,-p_h)=\mu\|\nabla\mathbf{u}_h\|_{L^2(\Omega)}^2+\sum_{K\in\mathcal{T}_h}\frac{\lambda_{2,K}^2}{\mu}\|\nabla p_h\|_{L^2(K)}^2.\tag{A.7}$$

Therefore, combining relations (A.6) and (A.7) we have

$$\begin{aligned}B_h(\mathbf{u}_h,p_h;\mathbf{u}_h+\delta R_h(\mathbf{v}),-p_h)\ge\delta\frac{\beta}{4\mu}\|p_h\|_{L^2(\Omega)}^2&\\ +\left(1-\delta\frac{C}{\beta}\left(1+\max_{K\in\mathcal{T}_h}s_K^2\right)\right)\left(\mu\|\nabla\mathbf{u}_h\|_{L^2(\Omega)}^2+\sum_{K\in\mathcal{T}_h}\frac{\lambda_{2,K}^2}{\mu}\|\nabla p_h\|_{L^2(K)}^2\right)&\end{aligned}$$

for any $\delta>0$. Choosing, for instance, δ such that

$$\frac{1}{2}=1-\delta\frac{C}{\beta}\left(1+\max_{K\in\mathcal{T}_h}s_K^2\right),$$

i.e.,

$$\delta=\frac{\beta}{2C\left(1+\max\limits_{K\in\mathcal{T}_h}s_K^2\right)},\tag{A.8}$$

we finally obtain

$$B_h(\mathbf{u}_h,p_h;\mathbf{u}_h+\delta R_h(\mathbf{v}),-p_h)\ge C\|(\mathbf{u}_h,p_h)\|_{V\times Q}^2,$$

with $C=C(\Gamma,\widehat{C},\beta)$. With reference to Lemma A.2, for each pair $(\mathbf{u}_h,p_h)\in V_h\times Q_h$, we can thus identify the test functions $(\mathbf{v}_h,q_h)\in V_h\times Q_h$ satisfying $(\text{A.1})_1$ as $(\mathbf{v}_h,q_h)=(\mathbf{u}_h+\delta R_h(\mathbf{v}),-p_h)$, where $\mathbf{v}\in V$ satisfies Lemma A.1 for $p=p_h$, and δ is given, e.g., by (A.8). In order to prove (A.1), it then remains to bound $\|(\mathbf{u}_h+\delta R_h(\mathbf{v}),-p_h)\|_{V\times Q}$ in terms of the norm $\|(\mathbf{u}_h,p_h)\|_{V\times Q}$. Young inequality yields

$$\begin{aligned}\|(\mathbf{u}_h+\delta R_h(\mathbf{v}),-p_h)\|_{V\times Q}^2\le 2\mu\Big(\|\nabla\mathbf{u}_h\|_{L^2(\Omega)}^2+\delta^2\|\nabla R_h(\mathbf{v})\|_{L^2(\Omega)}^2\Big)&\\ +\frac{1}{\mu}\frac{1}{\left(1+\max\limits_{K\in\mathcal{T}_h}s_K^2\right)}\|p_h\|_{L^2(\Omega)}^2.&\end{aligned}$$

Now, since

$$\|\nabla R_h(\mathbf{v})\|_{L^2(\Omega)} \leq \|\nabla \mathbf{v}\|_{L^2(\Omega)} + \|\nabla(\mathbf{v} - R_h(\mathbf{v}))\|_{L^2(\Omega)},$$

using relation (A.2), the vectorial extension of Proposition 2.5 (see Remark 2), and estimate (A.5), we obtain

$$\|\nabla R_h(\mathbf{v})\|_{L^2(\Omega)} \leq \frac{C}{\mu}\left(1 + \max_{K\in\mathcal{T}_h} s_K^2\right)^{1/2} \|p_h\|_{L^2(\Omega)},$$

where $C = C(\Gamma, \widehat{C})$. Finally, using the definition (A.8) of δ we obtain

$$\|(\mathbf{u}_h + \delta R_h(\mathbf{v}), -p_h)\|^2_{V\times Q} \leq C\|(\mathbf{u}_h, p_h)\|^2_{V\times Q},$$

with $C = C(\Gamma, \widehat{C}, \beta)$. $\square$

Acknowledgment. We thank Luca Formaggia for helpful suggestions and discussions.

REFERENCES

[1] T. Apel, *Anisotropic Finite Elements: Local Estimates and Applications*, Adv. Numer. Math., Teubner, Stuttgart, 1999.

[2] T. Apel and G. Lube, *Anisotropic mesh refinement in stabilized Galerkin methods*, Numer. Math., 74 (1996), pp. 261–282.

[3] F.P.T. Baaijens, *Mixed finite element methods for viscoelastic flow analysis: A review*, J. Non-Newtonian Fluid Mech., 79 (1998), pp. 361–385.

[4] R. Becker, *An Adaptive Finite Element Method for the Incompressible Navier-Stokes Equations on Time-Dependent Domains*, Ph.D. thesis, Institute of Applied Mathematics, University of Heidelberg, Heidelberg, Germany, 1995.

[5] R. Becker and R. Rannacher, *Finite element solution of the incompressible Navier-Stokes equations on anisotropically refined meshes*, Notes Numer. Fluid Mech., 49 (1995), pp. 52–62.

[6] M. Behr, L. Franca, and T. Tezduyar, *Stabilized finite element methods for the velocity-pressure-stress formulation of incompressible flows*, Comput. Methods Appl. Mech. Engrg., 104 (1993), pp. 31–48.

[7] J. Bonvin, M. Picasso, and R. Stenberg, *GLS and EVSS methods for a three-field Stokes problem arising from viscoelastic flows*, Comput. Methods Appl. Mech. Engrg., 190 (2001), pp. 3893–3914.

[8] H. Borouchaki and P. Laug, *The BL2D Mesh Generator: Beginner's Guide, User's and Programmer's Manual*, Technical report RT-0194, Institut National de Recherche en Informatique et Automatique (INRIA), Rocquencourt, Le Chesnay, France, 1996.

[9] F. Brezzi, L.P. Franca, T.J.R. Hughes, and A. Russo, $b = \int g$, Comput. Methods Appl. Mech. Engrg., 145 (1997), pp. 329–339.

[10] F. Brezzi and A. Russo, *Choosing bubbles for advection-diffusion problems*, Math. Models Methods Appl. Sci., 4 (1994), pp. 571–587.

[11] A.N. Brooks and T.J.R. Hughes, *Streamline upwind/Petrov-Galerkin formulations for convective dominated flows with particular emphasis on the incompressible Navier-Stokes equations*, Comput. Methods Appl. Mech. Engrg., 32 (1982), pp. 199–259.

[12] D. Chapelle and R. Stenberg, *Stabilized finite element formulations for shells in a bending dominated state*, SIAM J. Numer. Anal., 36 (1998), pp. 32–73.

[13] Ph. Ciarlet, *The Finite Element Method for Elliptic Problems*, North-Holland, Amsterdam, 1978.

[14] Ph. Clément, *Approximation by finite element functions using local regularization*, Rev. Française Automat. Informat. Recherche Opérationnelle Sér. Rouge Anal. Numér., 9 (1975), pp. 77–84.

[15] J. Douglas and J. Wang, *An absolutely stabilized finite element method for the Stokes problem*, Math. Comp., 52 (1989), pp. 495–508.

[16] K. Eriksson and C. Johnson, *Adaptive streamline diffusion finite element methods for stationary convection-diffusion problems*, Math. Comp., 60 (1993), pp. 167–188.

[17] L. Formaggia, S. Micheletti, and S. Perotto, *Anisotropic mesh adaption with application to CFD problems*, in Proceedings of WCCM V, Fifth World Congress on Computational Mechanics, H.A. Mang, F.G. Rammerstorfer, and J. Eberhardsteiner, eds., Vienna, Austria, 2002, available online from http://wccm.tuwien.ac.at.

[18] L. Formaggia and S. Perotto, *New anisotropic a priori error estimates*, Numer. Math., 89 (2001), pp. 641–667.

[19] L. Formaggia and S. Perotto, *Anisotropic error estimates for elliptic problems*, Numer. Math., 94 (2003), pp. 67–92.

[20] L.P. Franca and R. Stenberg, *Error analysis of some Galerkin least squares methods for the elasticity equations*, SIAM J. Numer. Anal., 28 (1991), pp. 1680–1697.

[21] T.J.R. Hughes, L. Franca, and M. Balestra, *A new finite element formulation for computational fluid dynamics.* V. *Circumventing the Babuška-Brezzi condition: A stable Petrov-Galerkin formulation of the Stokes problem accommodating equal-order interpolations*, Comput. Methods Appl. Mech. Engrg., 59 (1986), pp. 85–99.

[22] L.P. Franca, S.L. Frey, and T.J.R Hughes, *Stabilized finite element methods.* I. *Application to the advective-diffusive model*, Comput. Methods Appl. Mech. Engrg., 95 (1992), pp. 253–276.

[23] L.P. Franca and S.L. Frey, *Stabilized finite element methods.* II. *The incompressible Navier-Stokes equations*, Comput. Methods Appl. Mech. Engrg., 99 (1992), pp. 209–233.

[24] L.P. Franca and T.J.R Hughes, *Convergence analyses of Galerkin least-squares methods for symmetric advective-diffusive forms of the Stokes and incompressible Navier-Stokes equations*, Comput. Methods Appl. Mech. Engrg., 105 (1993), pp. 285–298.

[25] L.P. Franca and A. Madureira, *Element diameter free stability parameters for stabilized methods applied to fluids*, Comput. Methods Appl. Mech. Engrg., 105 (1993), pp. 395–403.

[26] J.-F. Gerbeau, *A stabilized finite element method for the incompressible magnetohydrodynamic equations*, Numer. Math., 87 (2000), pp. 83–111.

[27] G.H. Golub and C.F. Van Loan, *Matrix Computations*, 2nd ed., The Johns Hopkins University Press, Baltimore, MD, 1989.

[28] I. Harari and T.J.R Hughes, *What are C and h?: Inequalities for the analysis and design of finite element methods*, Comput. Methods Appl. Mech. Engrg., 97 (1992), pp. 157–192.

[29] T.J.R. Hughes, L.P. Franca, and G.M. Hulbert, *A new finite element formulation for computational fluid dynamics:* VIII. *The Galerkin/least-squares method for advective-diffusive equations*, Comput. Methods Appl. Mech. Engrg., 73 (1989), pp. 173–189.

[30] G. Kunert, *A Posteriori Error Estimation for Anisotropic Tetrahedral and Triangular Finite Element Meshes*, Ph.D. thesis, Fakultät für Mathematik der Technischen Universität Chemnitz, Chemnitz, Germany, 1999.

[31] G. Kunert, *Robust a posteriori error estimation for a singularly perturbed reaction-diffusion equation on anisotropic tetrahedral meshes*, Adv. Comput. Math., 15 (2001), pp. 237–259.

[32] J.L. Lions and E. Magenes, *Non-Homogeneous Boundary Value Problem and Application*, Vol. I, Springer-Verlag, Berlin, 1972.

[33] S. Micheletti, *Stabilized finite elements for semiconductor device simulation*, Comput. Vis. Sci., 3 (2001), pp. 177–183.

[34] S. Micheletti, S. Perotto, and M. Picasso, *Some Remarks on the Stability Coefficients and Bubble Stabilization of FEM on Anisotropic Grids*, MOX Report 6, MOX - Modeling and Scientific Computing, Dipartimento di Matematica "F. Brioschi," Politecnico di Milano, Milano, Italy, 2002.

[35] J.J.H. Miller, E. O'Riordan, and G.I. Shishkin, *Fitted Numerical Methods for Singular Perturbation Problems. Error Estimates in the Maximum Norm for Linear Problems in One and Two Dimensions*, World Scientific, River Edge, NJ, 1996.

[36] S. Mittal, *On the performance of high aspect ratio elements for incompressible flows*, Comput. Methods Appl. Mech. Engrg., 188 (2000), pp. 269–287.

[37] M. Picasso, *An anisotropic error indicator based on Zienkiewicz–Zhu error estimator : Application to elliptic and parabolic problems*, SIAM J. Sci. Comput., 24 (2003), pp. 1328–1355.

[38] H.G. Roos, M. Stynes, and L. Tobiska, *Numerical Methods for Singularly Perturbed Differential Equations. Convection-Diffusion and Flow Problems*, Springer Ser. Comput. Math. 24, Springer-Verlag, Berlin, 1996.

[39] A. Russo, *Bubble stabilization of finite element methods for the linearized incompressible Navier-Stokes equations*, Comput. Methods Appl. Mech. Engrg., 132 (1996), pp. 335–343.

[40] K.G. Siebert, *An a posteriori error estimator for anisotropic refinement*, Numer. Math., 73 (1996), pp. 373–398.

[41] R. Verfürth, *Error estimates for a mixed finite element approximation of the Stokes equations*, RAIRO Anal. Numér., 18 (1984), pp. 175–182.

SIAM J. NUMER. ANAL.
Vol. 41, No. 3, pp. 1163–1194

SURFACE PRESSURE POISSON EQUATION FORMULATION OF THE PRIMITIVE EQUATIONS: NUMERICAL SCHEMES*

ROGER SAMELSON[†], ROGER TEMAM[‡], CHENG WANG[‡], AND SHOUHONG WANG[‡]

Abstract. Numerical methods for the primitive equations (PEs) of oceanic flow are presented in this paper. First, a two-dimensional Poisson equation with a suitable boundary condition is derived to solve the surface pressure. Consequently, we derive a new formulation of the PEs in which the surface pressure Poisson equation replaces the nonlocal incompressibility constraint, which is known to be inconvenient to implement. Based on this new formulation, backward Euler and Crank–Nicolson schemes are presented. The marker and cell scheme, which gives values of physical variables on staggered mesh grid points, are chosen as spatial discretization. The convergence analysis of the fully discretized scheme is established in detail. The accuracy check for the scheme is also shown.

Key words. the primitive equations, surface pressure, staggered grid, convergence analysis

AMS subject classifications. 35Q35, 65M06, 86A10

DOI. 10.1137/S0036142901396284

Introduction. The primary purpose of this paper is to propose and analyze numerical methods for the three-dimensional (3-D) primitive equations (PEs) of large scale oceanic flow using the surface pressure Poisson equation with a suitable boundary condition.

The hydrostatic balance results in the decomposition of the total pressure field into two parts: the integral of the density variable in the vertical direction, and the pressure field at surface level $z = 0$, i.e., the surface pressure. It was shown by Lions, Temam, and Wang [13] that the surface pressure is the Lagrange multiplier of an incompressibility constraint (namely, the vertically averaged horizontal velocity is divergence-free). Based on this remark, they introduced a new mathematical formulation of the PEs in which the surface pressure disappears by projecting the PEs onto the function space of the divergence-free averaged horizontal velocity field.

In this paper, the preoccupations are different: we want to develop numerical algorithms for the solution of the PEs. Contrary to the approach in [13], the surface pressure will play a central role in the algorithm; it is dynamically updated in the momentum equation, instead of being treated as a Lagrange multiplier. In particular, we will display a Poisson equation for the surface pressure and derive an approximate boundary condition for this Poisson equation. As a result, the surface pressure Poisson equation replaces the nonlocal constraint for the horizontal velocity field. The vertical velocity is calculated by integrating the horizontal divergence of the horizontal velocity field, due to the 3-D incompressibility.

Numerical methods are then proposed for the PEs formulated in the surface pres-

[*] Received by the editors October 10, 2001; accepted for publication (in revised form) January 6, 2003; published electronically July 30, 2003.
http://www.siam.org/journals/sinum/41-3/39628.html

[†] College of Oceanic and Atmospheric Sciences, Oregon State University, 104 Ocean Admin. Bldg, Corvallis, OR 97331-5503 (rsamelson@oce.orst.edu). The research of this author was supported in part by ONR grant N00014-98-1-0813 and by the NSF under grant OCE-9907854.

[‡] Department of Mathematics and Institute for Scientific Computing and Applied Mathematics, Indiana University, Rawles Hall, Bloomington, IN 47405-5701 (temam@indiana.edu, cwang@indiana.edu, showang@indiana.edu). The research of the second author was supported by NSF grant DMS-0074334 and the Research Fund of Indiana University. The research of the fourth author was supported in part by ONR grant N00014-96-1-0425 and by NSF grant DMS-0072612.

sure Poisson equation. At each time step, the surface pressure field is determined by a two-dimensional (2-D) Poisson solver after the data of the horizontal velocity field and the density field are updated by the momentum equations and the density equations. In turn, the gradient of the surface pressure is updated at the next time step. The temporal discretization is implemented by either the backward Euler or the Crank–Nicolson method. For the spatial discretization, we adopt the idea of the 3-D marker and cell (MAC) grid. Different variables in the PEs are evaluated on different staggered grids. The derivatives are replaced by second order centered-difference operators, while the integration in the vertical direction is implemented by the trapezoidal rule. Following the approach related to the development of a local vorticity boundary condition, we derive a consistent and second order accurate boundary condition for the surface pressure at the discrete level. The main advantage of the MAC scheme can be seen in the fact that the computed horizontal velocity field has exactly zero mean-divergence in a discrete level. Because of such a property, the 3-D calculated velocity field is orthogonal to the horizontal and vertical gradients of the total pressure field in a discrete L^2 space, which plays an important role in the convergence analysis. The idea is similar to that of the finite element approach, yet it dramatically simplifies the computation. To our knowledge, this is the first theoretical analysis of the PEs on the MAC grid (which is usually referred to as a "C grid" in the geophysical fluid dynamics (GFD) literature). It should be possible to use similar methods to analyze other related GFD models.

The paper is organized as follows. In section 1 we recall the formulation of the PEs and introduce the alternate formulation using the surface pressure Poisson equation. Backward Euler and Crank–Nicolson schemes (in temporal discretization) are presented in section 2. The description of the 3-D MAC scheme is given in section 3, and the detailed convergence analysis of the backward Euler method combined with the MAC staggered grid is provided in section 4. Finally, a numerical accuracy check is given in section 5, using a set of exact solutions to compare with the profiles computed by our scheme.

1. The PEs and the surface pressure Poisson equation. We start with the nondimensional PEs with proper scaling:

$$\tag{1.1} \begin{cases} \boldsymbol{v}_t + (\boldsymbol{v}\cdot\nabla)\boldsymbol{v} + w\dfrac{\partial \boldsymbol{v}}{\partial z} + \dfrac{f}{Ro}k\times\boldsymbol{v} + \dfrac{1}{Ro}\left(\displaystyle\int_z^0 \nabla\rho(x,y,s)\,ds + \nabla p_s\right) = L_1\boldsymbol{v}, \\ \rho_t + (\boldsymbol{v}\cdot\nabla)\rho + w\dfrac{\partial\rho}{\partial z} = L_2\rho, \\ \nabla\cdot\displaystyle\int_{-H_0}^0 \boldsymbol{v}\,dz = 0. \end{cases}$$

See, e.g., Pedlosky [19] and Lions, Temam, and Wang [12, 13] for a detailed derivation. In the above system, $\boldsymbol{u} = (\boldsymbol{v}, w) = (u, v, w)$ is the 3-D velocity vector field, $\boldsymbol{v} = (u, v)$ the horizontal velocity, $w = \mathcal{W}(\boldsymbol{v})$ the vertical velocity in which the operator $\mathcal{W}$ will be introduced in (1.7), ρ the density field, p_s the surface pressure, and Ro a Rossby number. The term $fk \times \boldsymbol{v}$ corresponds to the Coriolis force in its β-plane approximation, with the parameter $f = f_0 + \beta y$. The operators ∇, $\nabla\cdot$, $\triangle$ represent the gradient, divergence, and Laplacian in the (x, y)-plane, respectively. The diffusion operators are given by $L_1 = (\frac{1}{Re_1}\triangle + \frac{1}{Re_2}\partial_z^2)$ and $L_2 = (\frac{1}{Rt_1}\triangle + \frac{1}{Rt_2}\partial_z^2)$. For simplicity, we denote $\nu_1 = \frac{1}{Re_1}$, $\nu_2 = \frac{1}{Re_1}$, $\kappa_1 = \frac{1}{Rt_1}$, $\kappa_2 = \frac{1}{Rt_1}$. The computational domain is taken as $\mathcal{M} = \mathcal{M}_0 \times [-H_0, 0]$, where $\mathcal{M}_0$ is the surface part of the ocean. The boundary condition for (1.1) is given by

$$\begin{aligned} \nu_2 \frac{\partial \boldsymbol{v}}{\partial z} &= \tau_0, \quad \kappa_2 \frac{\partial \rho}{\partial z} = \rho_f \quad \text{at } z = 0, \\ \nu_2 \frac{\partial \boldsymbol{v}}{\partial z} &= 0, \quad \kappa_2 \frac{\partial \rho}{\partial z} = 0 \quad \text{at } z = -H_0, \end{aligned} \tag{1.2}$$

$$\boldsymbol{v} = 0 \quad \text{and} \quad \frac{\partial \rho}{\partial \boldsymbol{n}} = 0 \quad \text{on } \partial\mathcal{M}_0 \times [-H_0, 0], \tag{1.3}$$

in which the term τ_0 represents the wind stress force and ρ_f represents the heat flux at the surface of the ocean.

Furthermore, the PEs (1.1) are supplemented with the following initial data:

$$\boldsymbol{v}(x, y, z, 0) = \boldsymbol{v}_0(x, y, z), \quad \rho(x, y, z, 0) = \rho_0(x, y, z), \tag{1.4}$$

in which $\boldsymbol{v}_0$ satisfies the mean divergence-free property as will be stated below.

The momentum equation for the horizontal velocity $\boldsymbol{v}$ comes from its original form

$$\boldsymbol{v}_t + (\boldsymbol{v}\cdot\nabla)\boldsymbol{v} + w\frac{\partial \boldsymbol{v}}{\partial z} + \frac{1}{Ro}\Big(fk \times \boldsymbol{v} + \nabla p\Big) = \Big(\nu_1 \triangle + \nu_2 \partial_z^2\Big)\boldsymbol{v}, \tag{1.5}$$

combined with the hydrostatic balance

$$\frac{\partial p}{\partial z} = -\rho, \qquad \text{which implies} \quad p(x, y, z) = \int_z^0 \rho(x, y, s)\, ds + p_s(x, y). \tag{1.6}$$

We note that p denotes the total pressure field; the surface pressure $p_s(x, y) = p(x, y, 0)$ is a 2-D field in the horizontal plane. We refer the readers to [3, 20, 24] for other physical and numerical considerations related to surface pressure.

The representation formula for the vertical velocity w comes from the vertical integration of the continuity equation $\nabla\cdot\boldsymbol{v} + \partial_z w = 0$, using the vanishing boundary condition for w at the top ($z = 0$) and at the bottom ($z = -H_0$):

$$w(x, y, z) = -\nabla\cdot \int_{-H_0}^{z} \boldsymbol{v}(x, y, s)\, ds \equiv \mathcal{W}(\boldsymbol{v}). \tag{1.7}$$

It was first observed by Lions, Temam, and Wang in [13] that the surface pressure appears to be the Lagrange multiplier of the nonlocal constraint $\nabla\cdot\int_{-H_0}^{0} \boldsymbol{v}\, dz = 0$. For instance, in view of studying the balance of energy of the system, we multiply the first equation in (1.1) by $\boldsymbol{v}$; then the integral $\int_{\mathcal{M}} \boldsymbol{v}\cdot\nabla p_s\, d\boldsymbol{x}$ vanishes.

1.1. Determination of the surface pressure variable. A major difficulty in the numerical approximation of the PEs is the absence of a time evolution equation for the surface pressure field. The main objective in this section is to derive an alternate formulation equivalent to the usual formulation (1.1) such that the surface pressure variable p_s can be determined by the horizontal velocity field $\boldsymbol{v}$ and the density field ρ, which can be updated by the momentum equations and the density equations. Some earlier work on this issue can be found in [1, 2, 4, 5, 7, 8, 11, 14, 15, 16, 18, 21, 22, 23, 25, 26].

1.1.1. Surface pressure Poisson equation. We now derive an equation for the surface pressure function $p_s(x,y)$. The starting point is the nonlocal constraint $\nabla\cdot\int_{-H_0}^0 \boldsymbol{v}\,dz=0$. By taking the horizontal divergence of the momentum equation in (1.1), integrating over $(-H_0,\,0)$ in the z-direction, dividing by H_0, and keeping in mind that p_s is a variable in the (x,y) plane, we arrive at the following equation:

$$\begin{aligned}(\partial_t-\nu_1\triangle)\Big(\overline{\nabla\cdot\boldsymbol{v}}\Big)-\nu_2\overline{\Big(\partial_z^2(\nabla\cdot\boldsymbol{v})\Big)}+\frac{1}{H_0}\int_{-H_0}^0\nabla\cdot\Big((\boldsymbol{v}\cdot\nabla)\boldsymbol{v}+w\frac{\partial\boldsymbol{v}}{\partial z}\Big)\,dz\\ +\frac{1}{Ro}\overline{\nabla\cdot\Big(fk\times\boldsymbol{v}\Big)}+\frac{1}{H_0}\frac{1}{Ro}\int_{-H_0}^0\int_z^0\triangle\rho(x,y,s)\,ds\,dz+\frac{1}{Ro}\triangle p_s=0,\end{aligned}\tag{1.8}$$

where $\overline{f}$ represents the average of the variable f in the z-direction. The first term in (1.8) vanishes, since $\overline{\nabla\cdot\boldsymbol{v}}$ is identically 0 in the horizontal domain. The second term turns out to be

$$\nu_2\overline{\Big(\partial_z^2(\nabla\cdot\boldsymbol{v})\Big)}=\frac{\nu_2}{H_0}\int_{-H_0}^0\partial_z^2(u_x+v_y)\,dz=\frac{\nu_2}{H_0}\Big(u_{zx}+v_{zy}\Big)\Big|_{-H_0}^0.\tag{1.9}$$

The boundary condition in (1.2) indicates that $(u_{zx}+v_{zy})=\frac{1}{\nu_2}(\partial_x(\tau_0)_1+\partial_y(\tau_0)_2)$ at $z=0$ and $(u_{zx}+v_{zy})=0$ at $z=-H_0$. Inserting (1.9) into (1.8) and setting $\tau_d=\nabla\cdot\tau_0)$ at $z=0$, which is a known function, we conclude that the surface pressure p_s solves the following Poisson equation:

$$\begin{aligned}\triangle p_s=\frac{Ro}{H_0}\nabla\cdot\tau_0-\frac{Ro}{H_0}\int_{-H_0}^0\nabla\cdot\Big((\boldsymbol{v}\cdot\nabla)\boldsymbol{v}+w\frac{\partial\boldsymbol{v}}{\partial z}+\frac{1}{Ro}(fk\times\boldsymbol{v})\Big)\,dz\\ -\frac{1}{H_0}\int_{-H_0}^0\int_z^0\triangle\rho(x,y,s)\,ds\,dz.\end{aligned}\tag{1.10}$$

The Poisson equation (1.10), together with the boundary condition described below, determines the surface pressure field by the velocity field and the density field without involving time derivative profiles.

1.1.2. Boundary condition for the surface pressure. Another point we have to emphasize is that there should be a boundary condition imposed for the surface pressure Poisson equation (1.10) if the Dirichlet boundary condition (1.3) for horizontal velocity field $\boldsymbol{v}$ is imposed on the lateral boundary section $\partial\mathcal{M}_0\times[-H_0,0]$.

Integrating the momentum equation in (1.1) over $(-H_0,\,0)$ in the z-direction and dividing by H_0 gives

$$\begin{aligned}\overline{\boldsymbol{v}}_t+\overline{\Big((\boldsymbol{v}\cdot\nabla)\boldsymbol{v}+w\frac{\partial\boldsymbol{v}}{\partial z}\Big)}+\frac{1}{Ro}\overline{fk\times\boldsymbol{v}}+\frac{1}{H_0}\frac{1}{Ro}\int_{-H_0}^0\int_z^0\nabla\rho(\cdot,s)\,ds\,dz\\ +\frac{1}{Ro}\nabla p_s=\nu_1\triangle\overline{\boldsymbol{v}}+\nu_2\overline{\partial_z^2\boldsymbol{v}},\end{aligned}\tag{1.11}$$

assuming that p_s is independent of the z-variable. On the lateral boundary $\partial\mathcal{M}_0\times[-H_0,0]$, the time marching term $\overline{\boldsymbol{v}}_t$ and all the convection terms vanish because of the no-penetration, no-slip boundary condition (the term $w\frac{\partial\boldsymbol{v}}{\partial z}$ disappears since $\frac{\partial\boldsymbol{v}}{\partial z}$ is zero on the boundary). The term $\overline{\partial_z^2\boldsymbol{v}}$ also vanishes, since $\partial_z^2\boldsymbol{v}$ is also 0 on $\partial\mathcal{M}_0\times[-H_0,0]$. Therefore, by taking the inner product of (1.11) with the unit normal vector field on the boundary $\partial\mathcal{M}_0$ (of the 2-D domain $\mathcal{M}_0$), we arrive at the following boundary condition for the surface pressure:

$$\frac{\partial p_s}{\partial \boldsymbol{n}} = \nu_1\, Ro\, \triangle \overline{\boldsymbol{v}} \cdot \boldsymbol{n} \quad \text{on} \quad \partial\mathcal{M}_0. \tag{1.12}$$

1.2. Alternate formulation of the PEs. We then have the following formulation, in which the nonlocal constraint $\nabla\cdot\int_{-H_0}^{0} \boldsymbol{v}\, dz = 0$ is replaced by the surface pressure Poisson equation and a mean divergence-free boundary condition for the horizontal velocity:

$$\boldsymbol{v}_t + (\boldsymbol{v}\cdot\nabla)\boldsymbol{v} + \mathcal{W}(\boldsymbol{v})\frac{\partial \boldsymbol{v}}{\partial z} + \frac{f}{Ro}k\times\boldsymbol{v} + \frac{1}{Ro}\left(\int_z^0 \nabla\rho(x,y,s)\,ds + \nabla p_s\right) = L_1\boldsymbol{v}, \tag{1.13a}$$

$$\rho_t + (\boldsymbol{v}\cdot\nabla)\rho + \mathcal{W}(\boldsymbol{v})\frac{\partial \rho}{\partial z} = L_2\rho, \tag{1.13b}$$

$$\begin{aligned} \triangle p_s = {} & \frac{Ro}{H_0}\tau_d - \frac{Ro}{H_0}\int_{-H_0}^{0} \nabla\cdot\Big((\boldsymbol{v}\cdot\nabla)\boldsymbol{v} + \mathcal{W}(\boldsymbol{v})\frac{\partial \boldsymbol{v}}{\partial z} + \frac{1}{Ro}(fk\times\boldsymbol{v})\Big)\,dz \\ & - \frac{1}{H_0}\int_{-H_0}^{0}\int_z^0 \triangle\rho(x,y,s)\,ds\,dz, \end{aligned} \tag{1.13c}$$

$$\begin{aligned} & \frac{\partial \boldsymbol{v}}{\partial z} = \frac{\tau_0}{\nu_2} \quad \text{at} \quad z=0, \qquad \frac{\partial \boldsymbol{v}}{\partial z} = 0 \quad \text{at} \quad z=-H_0, \\ & \frac{\partial \rho}{\partial z} = \frac{\rho_f}{\kappa_2} \quad \text{at} \quad z=0, \qquad \frac{\partial \rho}{\partial z} = 0 \quad \text{at} \quad z=-H_0, \\ & \boldsymbol{v} = 0 \quad \text{and} \quad \frac{\partial \rho}{\partial \boldsymbol{n}} = 0 \quad \text{on} \quad \partial\mathcal{M}_0\times[-H_0,0], \end{aligned} \tag{1.13d}$$

$$(\overline{\nabla\cdot\boldsymbol{v}}) = 0 \quad \text{on} \quad \partial\mathcal{M}_0. \tag{1.13e}$$

PROPOSITION 1.1. *For $\boldsymbol{v},\rho \in L^\infty([0,T],H^3)$, $\partial_t\boldsymbol{v},\partial_t\rho \in L^\infty([0,T],H^1)$, the original formulation* (1.1)–(1.3) *of the PEs is equivalent to the alternate formulation* (1.13a)–(1.13e).

Proof. Assume $(\boldsymbol{v},\rho,p_s)$ is a solution of (1.1)–(1.3). We observe that p_s satisfies the Poisson equation (1.13b), which can be obtained by taking the horizontal divergence of the momentum equation in (1.1) and averaging in the vertical direction as shown in the above derivation. In addition, taking the vertical derivative of the representation formula for the vertical velocity in (1.1) indicates that the horizontal velocity $\boldsymbol{v}$ satisfies the constraint $\overline{\nabla\cdot\boldsymbol{v}} = 0$. The usage of the regularity for $\boldsymbol{v}\in L^\infty([0,T],H^3)$ shows that $\boldsymbol{v}$ satisfies the additional boundary condition (1.13e). This concludes that $(\boldsymbol{v},\rho,p_s)$ is also a solution of (1.13).

Conversely, assume $(\boldsymbol{v},\rho,p_s)$ is a solution of (1.13). We need to show that $\overline{\nabla\cdot\boldsymbol{v}} = 0$. Taking the divergence of (1.13a) and integrating in the vertical direction leads to

$$\partial_t(\overline{\nabla\cdot\boldsymbol{v}}) - \nu_1\triangle(\overline{\nabla\cdot\boldsymbol{v}}) = 0, \tag{1.14a}$$

since *all the other terms are canceled by the surface pressure Poisson equation.* Hence the heat equation (1.14a) for the scalar quantity $\overline{\nabla\cdot\boldsymbol{v}}$, together with the homogeneous initial data

$$(\overline{\nabla\cdot\boldsymbol{v}})(x,y,t=0)=0, \tag{1.14b}$$

and the additional mean divergence-free boundary condition for the horizontal velocity imposed by (1.13e),

$$(\overline{\nabla\cdot\boldsymbol{v}})=0 \quad \text{on } \partial\mathcal{M}_0, \tag{1.14c}$$

show that $(\overline{\nabla\cdot\boldsymbol{v}})=0$; namely, the third equation in (1.1) is satisfied for all $t>0$. Therefore, $(\boldsymbol{v},\rho,p_s)$ is also a solution of (1.1)–(1.3). That completes the proof of Proposition 1.1. □

Remark 1.2. The above arguments show that the system (1.13a)–(1.13e) implies the original system of PEs (1.1)–(1.3), and therefore it implies (1.12), a Neumann-type boundary condition for the surface pressure p_s, since (1.12) is derived from the momentum equation in (1.1). In other words, the boundary condition (1.12) must be satisfied by any solution of the system (1.13a)–(1.13e). For the computations, the additional boundary condition (1.13e), a mean divergence-free boundary condition for the horizontal velocity, is not convenient to use. Instead, we will replace it by (1.12), a boundary condition for the surface pressure, to solve for the Poisson equation (1.13b); note that we are not claiming that the systems are equivalent if we replace (1.13e) by (1.12), leaving (1.13a)–(1.13d) unchanged. However, as we show below, such an equivalence occurs in the case of the MAC scheme, the spatially discrete scheme that will be studied in section 3.

1.3. Analogy with the 2-D Navier–Stokes equations. It could be observed that the boundary condition (1.13e) is coupled with the surface pressure Poisson equation (1.13b) and the momentum equation (1.13a), (1.13d). In more detail, in the derived formulation (1.13), four boundary conditions are prescribed for the horizontal velocity field: $\frac{\partial\boldsymbol{v}}{\partial z}=\frac{\tau_0}{\nu_2}$ on $z=0$, $\frac{\partial\boldsymbol{v}}{\partial z}=0$ on $z=-H_0$, $\boldsymbol{v}=0$ on $\partial\mathcal{M}_0\times[-H_0,0]$, and $(\overline{\nabla\cdot\boldsymbol{v}})=0$ on $\partial\mathcal{M}_0$, while there is no boundary condition for the surface pressure p_s. This subtle fact appears in a similar way for the formulations of incompressible fluid equations, such as the Navier–Stokes equation (NSE). For example, the vorticity-stream function formulation of the 2-D NSE in a simply connected domain, which is also a derived formulation, can be written as

$$\begin{cases} \partial_t\omega+(\boldsymbol{v}\cdot\nabla)\omega=\nu\triangle\omega, \\ \triangle\psi=\omega, \\ u=-\partial_y\psi, \qquad v=\partial_x\psi, \end{cases} \tag{1.15}$$

where $\boldsymbol{v}=(u,v)$ denotes the 2-D velocity field, $\omega=\nabla\times\boldsymbol{u}=-\partial_y u+\partial_x v$ denotes the vorticity, and the no-penetration, no-slip boundary condition can be written in terms of the stream function ψ:

$$\psi=0, \quad \frac{\partial\psi}{\partial\mathbf{n}}=0 \qquad \text{on} \qquad \partial\mathcal{M}_0. \tag{1.16}$$

Similar to our derived formulation (1.13), in the coupled system (1.15) and (1.16), there are two boundary conditions for the stream function ψ (both Dirichlet and Neumann) and no explicit boundary condition for the vorticity ω. On the other hand, updating the dynamic equation in (1.15) requires the vorticity boundary values; see [6, 10, 17, 18, 27, 28] for detailed description, derivation, and analysis of vorticity boundary conditions. A similar difficulty arises in the formulation (1.13): what boundary condition should be imposed to solve surface pressure p_s? Of course, the Neumann boundary condition (1.12) is a good choice to replace (1.13e); their equivalence is not claimed at the PDE level, as noted in Remark 1.2. However, in the MAC spatial discretization with a staggered grid described in section 3, the boundary condition $(\overline{\nabla\cdot\boldsymbol{v}})\ |_{\partial\mathcal{M}_0}= 0$ is converted by a second order accurate realization into the surface pressure boundary condition. Furthermore, in such a staggered grid, the equivalence between the derived boundary condition and the nonlocal constraint on the boundary as in (1.13e) can be fully proven.

Remark 1.3. The precise approximation of the pressure field via the pressure Poisson equation is a well-known difficulty in the incompressible flow calculation if the physical boundary condition is presented. The approach for solving the 2-D and 3-D NSEs by utilizing a local pressure boundary condition was recently introduced by Johnston and Liu in [11]. Some ideas in their paper can be adapted in our work.

Remark 1.4. The PEs with general boundary conditions or noncylindric domains were investigated in earlier literatures by Lions, Temam, and Wang [12, 13, 14, 15, 16] in a PDE level. The corresponding numerical methods can be accordingly derived using finite element approaches. We hope to report that issue in a future paper.

2. Temporal discretization.

Two computational methods for the PEs in surface pressure Poisson equation formulation (1.13) are proposed in this section. The horizontal velocity field and the density field are updated by the momentum equation (1.13a) and the density equation (1.13b). The surface pressure field, which is essentially a Lagrange multiplier in a horizontal plane, is determined by a 2-D Poisson solver, using the information of the velocity field and the density field at the same time step (stage). Henceforth, the surface pressure gradient is treated as a force term in the dynamic evolution of the momentum equation in the next time step (the stage). As a result, the surface pressure term is decoupled from the diffusion term; thus the Stokes solver is avoided. That dramatically simplifies the computation.

For simplicity, we use implicit treatment of the diffusion terms in the momentum equation and the density equation, which makes the stability and convergence analysis of the numerical scheme easier to handle. The backward Euler scheme is chosen as the example of the first order method (in temporal discretization) and the Crank–Nicolson scheme as the second order version.

2.1. Backward Euler method.

Given the velocity field $\boldsymbol{u}^n$, surface pressure field p_s^n, and density field ρ^n at time t^n, we update all the profiles at the time step t^{n+1} through the following procedure.

Step 1. The semi-implicit scheme for the momentum equation and the density equation is given, leaving the convection term and the surface pressure gradient explicit:

$$\text{(2.1a)}\quad \left\{\begin{aligned} &\frac{\boldsymbol{v}^{n+1}-\boldsymbol{v}^n}{\triangle t} + (\boldsymbol{v}^n\cdot\nabla)\boldsymbol{v}^n + \mathcal{W}(\boldsymbol{v}^n)\frac{\partial \boldsymbol{v}^n}{\partial z} + \frac{f}{Ro}k\times\boldsymbol{v}^n \\ &\qquad + \frac{1}{Ro}\int_z^0 \nabla\rho^n(x,y,s)\,d + \frac{1}{Ro}\nabla p_s^n(x,y) = \Big(\nu_1\triangle + \nu_2\partial_z^2\Big)\boldsymbol{v}^{n+1}, \\ &\frac{\rho^{n+1}-\rho^n}{\triangle t} + (\boldsymbol{v}^n\cdot\nabla)\rho^n + w^n\frac{\partial\rho^n}{\partial z} = \Big(\kappa_1\triangle + \kappa_2\partial_z^2\Big)\rho^{n+1}, \end{aligned}\right.$$

which are three standard Poisson-like equations, with the boundary condition

$$\begin{aligned}
&\frac{\partial \boldsymbol{v}^{n+1}}{\partial z} = \frac{\tau_0}{\nu_2}, \quad \frac{\partial \rho^{n+1}}{\partial z} = \frac{\rho_f}{\kappa_2} \quad \text{at } z=0,\\
&\frac{\partial \boldsymbol{v}^{n+1}}{\partial z} = 0, \quad \frac{\partial \rho^{n+1}}{\partial z} = 0 \quad \text{at } z=-H_0,\\
&\boldsymbol{v}^{n+1} = 0 \quad \text{and} \quad \frac{\partial \rho^{n+1}}{\partial \boldsymbol{n}} = 0 \quad \text{on } \partial\mathcal{M}_0 \times [-H_0, 0].
\end{aligned} \tag{2.1b}$$

Step 2. With all the velocity field $\boldsymbol{v}^{n+1}$, w^{n+1} at hand, we can solve for the surface pressure field at the time step t^{n+1} by the 2-D Poisson equation

$$\begin{aligned}
\triangle p_s^{n+1} = &\frac{Ro}{H_0}\tau_d^{n+1} - \frac{1}{H_0}\int_{-H_0}^0 \int_z^0 \triangle \rho^{n+1}(x,y,s)\,ds\,dz\\
&- \frac{Ro}{H_0}\int_{-H_0}^0 \nabla\cdot\Big((\boldsymbol{v}^{n+1}\cdot\nabla)\boldsymbol{v}^{n+1} + \mathcal{W}(\boldsymbol{v}^{n+1})\frac{\partial \boldsymbol{v}^{n+1}}{\partial z} + \frac{1}{Ro}(fk\times\boldsymbol{v}^{n+1})\Big)\,dz,
\end{aligned} \tag{2.2a}$$

supplemented with the derived boundary condition (1.12), as argued in Remark 1.2 and section 1.3:

$$\frac{\partial p_s^{n+1}}{\partial \boldsymbol{n}} = \nu_1\, Ro\, \triangle \overline{\boldsymbol{v}}^{n+1}\cdot\boldsymbol{n} \quad \text{on} \quad \partial\mathcal{M}_0. \tag{2.2b}$$

2.2. Crank–Nicolson method. The updating from time step t^n to t^{n+1} is carried out by the following steps.

Step 1. Solve for the momentum equations and the density equations

$$\begin{cases}
\dfrac{\boldsymbol{v}^{n+1}-\boldsymbol{v}^n}{\triangle t} + RHS1^{n+\frac12} + \dfrac{1}{Ro}\nabla p_s^{n+\frac12} = \dfrac12\Big(\nu_1\triangle + \nu_2\partial_z^2\Big)(\boldsymbol{v}^n + \boldsymbol{v}^{n+1}),\\
\dfrac{\rho^{n+1}-\rho^n}{\triangle t} + RHS2^{n+\frac12} = \dfrac12\Big(\kappa_1\triangle + \kappa_2\partial_z^2\Big)(\rho^n + \rho^{n+1}),
\end{cases} \tag{2.3}$$

using the boundary condition (2.1b), where

$$\begin{aligned}
RHS1^{n+\frac12} &= (\boldsymbol{v}^{n+\frac12}\cdot\nabla)\boldsymbol{v}^{n+\frac12} + \mathcal{W}(\boldsymbol{v}^{n+\frac12})\frac{\partial \boldsymbol{v}^{n+\frac12}}{\partial z} + \frac{f}{Ro}k\times\boldsymbol{v}^{n+\frac12}\\
&\quad + \frac{1}{Ro}\int_z^0 \nabla\rho^{n+\frac12}(x,y,s)\,ds,\\
RHS2^{n+\frac12} &= (\boldsymbol{v}^{n+\frac12}\cdot\nabla)\rho^{n+\frac12} + \mathcal{W}(\boldsymbol{v}^{n+\frac12})\frac{\partial \rho^{n+\frac12}}{\partial z}.
\end{aligned} \tag{2.4}$$

The velocity and the density profiles $(\boldsymbol{u}, \rho) = (\boldsymbol{v}, w, \rho)$, along with the surface pressure p_s, at the time step $t^{n+\frac12}$ are evaluated by second order explicit extrapolation in time

$$\boldsymbol{u}^{n+\frac12} = \frac32\boldsymbol{u}^n - \frac12\boldsymbol{u}^{n-1}, \quad \rho^{n+\frac12} = \frac32\rho^n - \frac12\rho^{n-1}, \quad p_s^{n+\frac12} = \frac32 p_s^n - \frac12 p_s^{n-1}. \tag{2.5}$$

Note that the system (2.3) is also composed of three standard Poisson-like equations.

Step 2. The surface pressure field at the time step t^{n+1} is solved by the 2-D Poisson equation (2.2a) with the derived boundary condition (2.2b), as in the second step of the backward Euler scheme.

3. Spatial discretization: MAC scheme. We consider hereafter the oceanic basin given by $\mathcal{M}_0 = [0,1]^2$ and assume for simplicity that $\triangle x = \triangle y = \triangle z = h$. The analysis of the spatial discretization with regular grids is quite difficult. In this paper, we consider the MAC staggered grid as spatial discretization. Some well-known difficulties in the numerical simulation of NSE, such as enforcement of the incompressibility condition and lack of proper evolutionary equation for the pressure and associate boundary condition, were elegantly resolved in the celebrated MAC scheme, which was first proposed by Harlow and Welch in [9]. For the system of the PEs, the 3-D MAC staggered grid is used in the computational method.

An illustration of the MAC mesh on the section $z_k = (k+1/2)\triangle z$ is given in Figure 3.1. The surface pressure variable p_s is evaluated at the square points $(i \pm 1/2, j \pm 1/2)$, the velocity u is evaluated at the triangle points $(i, j \pm 1/2, k \pm 1/2)$, the velocity v is evaluated at the circle points $(i \pm 1/2, j, k \pm 1/2)$, and the velocity w and the density function ρ are evaluated at the dot points $(i \pm 1/2, j \pm 1/2, k)$. The advantage of such a staggered grid is the convenience to assure the divergence-free property of the numerical velocity field, which can be observed later.

The following centered differences using different stencils at different grid points is introduced to facilitate the description below:

$$\begin{aligned} &D_x g(x) = \frac{g(x+\frac{1}{2}h) - g(x-\frac{1}{2}h)}{h}, \quad \tilde{D}_x g(x) = \frac{g(x+h) - g(x-h)}{2h}, \\ &D_x^2 g(x) = \frac{g(x-h) - 2g(x) + g(x+h)}{h^2}, \end{aligned} \tag{3.1}$$

which are second order approximations to ∂_x, ∂_x^2, respectively. The corresponding operators in y- and z-directions, such as D_y, $\tilde{D}_y$, D_y^2, D_z, $\tilde{D}_z$, D_z^2, can be defined in the same fashion.

The discrete divergence of the total velocity field $\boldsymbol{u}$ is evaluated at the square points:

$$(\nabla_h \cdot \boldsymbol{u})_{i+1/2,j+1/2,k+1/2} = \Big(D_x u + D_y v + D_z w \Big)_{i+1/2,j+1/2,k+1/2}. \tag{3.2}$$

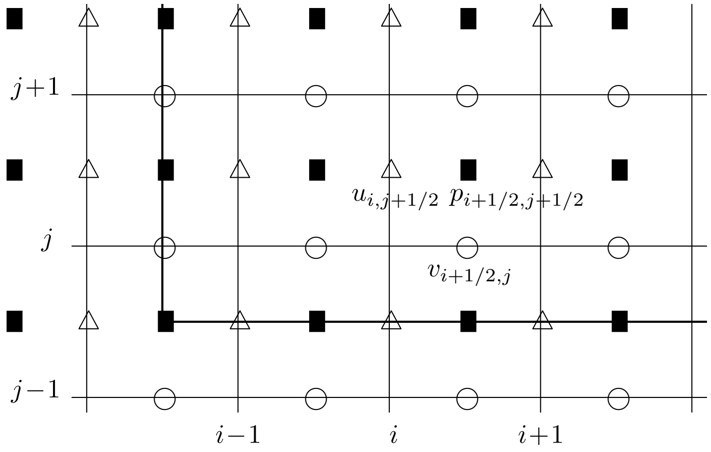


FIG. 3.1. *MAC mesh at $z_k = (k+1/2)\triangle z$; Harlow and Welch* [9].

The diffusion term for the velocity u is approximated by

$$\Big(\nu_1 \triangle + \nu_2 \partial_z^2\Big) u = \Big(\nu_1 \triangle_h + \nu_2 D_z^2\Big) u = \Big(\nu_1 (D_x^2 + D_y^2) + \nu_2 D_z^2\Big) u \tag{3.3}$$

at $(i, j+1/2, k+1/2)$. The other diffusion terms, $\triangle_h v$, $D_z^2 v$, at the mesh point $(i+1/2, j, k+1/2)$, and $\triangle_h \rho$, $D_z^2 \rho$, at the mesh point $(i+1/2, j+1/2, k)$, can be given in the same way. The gradient of density and the surface pressure appearing in the momentum equation is discretized by $(D_x\rho)_{i,j+1/2,k+1/2}$, $(D_y\rho)_{i,j+1/2,k+1/2}$ (and $(D_x p_s)_{i,j+1/2}$, $(D_y p_s)_{i+1/2,j}$), respectively.

The approximation to the nonlinear convection term $(\boldsymbol{v}\cdot\nabla)\boldsymbol{v} + w\frac{\partial \boldsymbol{v}}{\partial z}$, $(\boldsymbol{v}\cdot\nabla)\rho + w\frac{\partial \rho}{\partial z}$ at the corresponding mesh points for u, v, ρ relies on the introduction of average value of u, v, w at the staggered grid. For example, at the mesh point $(i, j+1/2, k+1/2)$ where u is located, the average v, w can be introduced as

$$\begin{aligned} \overline{\overline{v}}_{i,j+1/2,k+1/2} = \; & \frac{1}{4}(v_{i-1/2,j,k+1/2} + v_{i+1/2,j,k+1/2} \\ & + v_{i-1/2,j+1,k+1/2} + v_{i+1/2,j+1,k+1/2}), \end{aligned} \tag{3.4a}$$

$$\begin{aligned} \overline{\overline{w}}_{i,j+1/2,k+1/2} = \; & \frac{1}{4}(w_{i-1/2,j+1/2,k} + w_{i+1/2,j+1/2,k} \\ & + w_{i-1/2,j+1/2,k+1} + w_{i+1/2,j+1/2,k+1}), \end{aligned} \tag{3.4b}$$

and the corresponding convection term for u: $uu_x + vu_y + wu_z$ can be defined as

$$\mathcal{N}_h(\boldsymbol{u}, u) = u\tilde{D}_x u + \overline{\overline{v}}\tilde{D}_y u + \overline{\overline{w}}\tilde{D}_z u \qquad \text{at} \qquad (i, j+1/2, k+1/2). \tag{3.5}$$

The two other convection terms, $\mathcal{N}_h(\boldsymbol{u}, v)_{i+1/2,j,k+1/2}$, $\mathcal{N}_h(\boldsymbol{u}, \rho)_{i+1/2,j+1/2,k}$, which are approximations to $uv_x + vv_y + wv_z$, $u\rho_x + v\rho_y + w\rho_z$ at the corresponding mesh points, can be similarly defined. In addition, the Coriolis force term $fk \times \boldsymbol{v} = (-fv, fu)$ is evaluated at the mesh points for u, v, respectively, by taking the average of v and u at the required grid points as in (5.4):

$$\begin{aligned} (-fv)_{i,j+1/2,k+1/2} &= -f_{i,j+1/2}\overline{\overline{v}}_{i,j+1/2,k+1/2}, \\ (fu)_{i+1/2,j,k+1/2} &= f_{i+1/2,j}\overline{\overline{u}}_{i+1/2,j,k+1/2}. \end{aligned} \tag{3.6}$$

Clearly, the truncation errors of these approximations are of second order. The momentum equation for u is implemented at triangle points, the second momentum equation is implemented at circle points, and the density equation is implemented at the mesh points $(i+1/2, j+1/2, k)$. The discrete version of the term $\int_z^0 \nabla\rho(x,y,s)\,ds$ appearing in the momentum equation is a discrete integral of $\tilde{D}_x\rho$, $\tilde{D}_y\rho$ (which are defined at mesh points $(i, j+1/2, k)$, $(i+1/2, j, k)$, respectively, as given in (3.3)), in the z-direction. More accurately, $\mathcal{P}NRX$ is defined as the discrete version of $\int_z^0 \rho_x(x,y,s)\,ds$:

(3.7a)

$$\begin{aligned} &\mathcal{P}NRX_{i,j+1/2,N-1/2} = \frac{1}{2}\triangle z\,(D_x\rho)_{i,j+1/2,N}, \\ &\mathcal{P}NRX_{i,j+1/2,k-1/2} = \mathcal{P}NRX_{i,j+1/2,k+1/2} + \triangle z\,(D_x\rho)_{i,j+1/2,k}, \quad k \le N_z - 1, \end{aligned}$$

and $\mathcal{P}NRY$ can be given in a similar way:

(3.7b)

$$\begin{aligned} &\mathcal{P}NRY_{i+1/2,j,N-1/2} = \frac{1}{2}\triangle z\,(D_y\rho)_{i+1/2,j,N}, \\ &\mathcal{P}NRY_{i+1/2,j,k-1/2} = \mathcal{P}NRY_{i+1/2,j,k+1/2} + \triangle z\,(D_y\rho)_{i+1/2,j,k}, \quad k \le N_z - 1. \end{aligned}$$

Both formulae are second order approximation to the integral of the density gradient from z_k to 0.

The 2-D discrete Poisson equation for surface pressure p_s is implemented at square points. In more detail, we denote

$$\begin{aligned} \mathcal{F}U &= -\mathcal{N}_h(\boldsymbol{u}, u) + \frac{1}{Ro}(fv) - \frac{1}{Ro}\mathcal{P}NRX \quad \text{at } (i, j+1/2, k+1/2), \\ \mathcal{F}V &= -\mathcal{N}_h(\boldsymbol{u}, v) - \frac{1}{Ro}(fu) - \frac{1}{Ro}\mathcal{P}NRY \quad \text{at } (i+1/2, j, k+1/2) \end{aligned} \tag{3.8}$$

as the convection terms (including the Coriolis force term) for the momentum equation; therefore, the Poisson equation for surface pressure p_s can be written as

(3.9)

$$\begin{aligned} &(\triangle_h p_s)_{i+1/2,j+1/2} = \frac{Ro}{H_0}\Big(D_x\tau_{0,1} + D_y\tau_{0,2}\Big)_{i+1/2,j+1/2} + Ro\,\overline{\mathcal{F}P}_{i+1/2,j+1/2}, \text{ where} \\ &\mathcal{F}P_{i+1/2,j+1/2,k+1/2} = (D_x\mathcal{F}U)_{i+1/2,j+1/2,k+1/2} + (D_y\mathcal{F}V)_{i+1/2,j+1/2,k+1/2}, \end{aligned}$$

on the mesh points $(i+1/2, j+1/2)$ in the 2-D region $\mathcal{M}_0 = [0,1]^2$. The average of $\mathcal{F}P$, which is evaluated at the same numerical mesh grid as p, is defined as

$$\overline{\mathcal{F}P}_{i+1/2,j+1/2} = \frac{1}{H_0}\sum_{k=0}^{nz-1}(\triangle z\,\mathcal{F}P_{i+1/2,j+1/2,k+1/2}), \tag{3.10}$$

which is a second order approximation to the integral of $\mathcal{F}P$ in the z-direction. Such an evaluation of the discrete integral in the z-direction can be applied to any variable whose z-direction grid is indexed as $k \pm 1/2$.

It should be remarked that some suitable boundary condition is needed to solve the 2-D Poisson equation (3.9). Such a choice of the boundary condition assures the discrete divergence $(\nabla_h \cdot \boldsymbol{v})$ has mean zero (in the z-direction) on the boundary $\partial\mathcal{M}_0$. Details will be discussed in a later section.

On the physical boundary section $i = 0$, the no-penetration, no-slip boundary condition $\boldsymbol{v} = 0$ is translated by the reflection rule, whose application in the case of the 2-D NSE can be found in earlier work [4, 6, 7, 9],

$$u_{0,j+1/2,k+1/2} = 0, \quad v_{-1/2,j,k+1/2} + v_{1/2,j,k+1/2} = 0, \tag{3.11}$$

and the no-flux boundary condition for the density function is imposed by

$$(D_x\rho)_{0,j+1/2,k} = 0, \quad \text{which implies } \rho_{-1/2,j+1/2,k} = \rho_{1/2,j+1/2,k}. \tag{3.12}$$

Similarly, on the boundary section $j = 0$, the boundary condition $\boldsymbol{v} = 0$ is imposed by $v_{i+1/2,0,k+1/2} = 0$, $u_{i,-1/2,k+1/2} + u_{i,1/2,k+1/2} = 0$, and the boundary condition $\frac{\partial\rho}{\partial\boldsymbol{n}} = 0$ is imposed by $\rho_{i+1/2,-1/2,k} = \rho_{i+1/2,1/2,k}$.

On the bottom boundary $z = -H_0$, i.e., $k = 0$, the boundary condition $\frac{\partial \boldsymbol{v}}{\partial z} = 0$, $\frac{\partial \rho}{\partial z} = 0$ can be written as

$$\begin{aligned} &u_{i,j+1/2,-1/2} = u_{i,j+1/2,1/2}, \quad v_{i+1/2,j,-1/2} = v_{i+1/2,j,1/2}, \\ &\rho_{i+1/2,j+1/2,-1} = \rho_{i+1/2,j+1/2,1}, \end{aligned} \tag{3.13}$$

using a similar argument as in (3.12).

3.1. Boundary condition for surface pressure $\boldsymbol{p_s}$. The derived boundary condition (1.12) is needed to solve the surface pressure Poisson equation (3.9). As assumed earlier, in the case that $\mathcal{M}_0 = [0,1]^2$, we concentrate on the left boundary $x = 0$ for simplicity of presentation. The other three boundary sections $x = 1$, $y = 0, 1$ can be dealt with in the same manner. In PDE formulation, on the left boundary section $x = 0$, (1.12) indicates that

$$\frac{\partial p_s}{\partial x} = \nu_1 \, Ro \, \triangle \overline{u} = \nu_1 \, Ro \, \partial_x^2 \overline{u}, \tag{3.14}$$

where the second step is based on the fact that the velocity u vanishes; henceforth $\overline{u}$ vanishes on the boundary, too. The MAC mesh grid near the left boundary is shown in Figure 3.1.

Our methodology for approximating $\partial_x^2 \overline{u}$ as in (3.14) follows the approach taken by numerical methods for the 2-D NSE formulated in the vorticity-stream function, as given in (1.15), (1.16), based on local vorticity boundary conditions. The earliest work in this direction is due to Thom [27]. The more recent works [6], [10], [28] revived interest in the use of local formulae for vorticity on the boundary and analyzed the stability and convergence of a class of such formulae. The key point in these approaches is to convert the Neumann boundary condition for the stream function ψ, which states the no-slip velocity boundary condition, into a local vorticity boundary condition, such as Thom's formula.

A similar idea can be used in the approximation to $\partial_x^2 \overline{u}$ as in (3.14). In our scheme, the mean divergence-free boundary condition for the horizontal velocity, $(\nabla \cdot \boldsymbol{v}) \mid_{\partial \mathcal{M}_0} = 0$, can be converted into an approximation of the Neumann boundary condition for the surface pressure as derived in (1.12). In more detail, the following finite-difference method is applied on the boundary grid point $(0, j \pm 1/2)$:

$$\begin{aligned} \partial_x^2 \overline{u}_{0,j+1/2} &= \frac{\overline{u}_{-1,j+1/2} - 2\overline{u}_{0,j+1/2} + \overline{u}_{1,j+1/2}}{\triangle x^2} + O(h^2) \\ &= \frac{\overline{u}_{-1,j+1/2} + \overline{u}_{1,j+1/2}}{\triangle x^2} + O(h^2), \end{aligned} \tag{3.15}$$

where the second step is based on the fact that the velocity field $\boldsymbol{v}$ vanishes on the boundary. The second-order approximation (5.15) requires a value for $\overline{u}$ at grid point $(-1, j+1/2)$, which is a "ghost" point outside the computational domain. A consistent prescription of the value for $\overline{u}_{-1,j+1/2}$ relies on a second order centered difference of the mean divergence-free boundary condition $\overline{\nabla \cdot \boldsymbol{v}} \mid_{x=0} = 0$,

$$0 = (\partial_x \overline{u} + \partial_y \overline{v}) \mid_{x=0} = 0 + \partial_y \overline{v} \mid_{x=0} = \frac{\overline{u}_{1,j+1/2} - \overline{u}_{-1,j+1/2}}{2\triangle x} + O(h^2), \tag{3.16}$$

where the second step is due to the boundary condition $\overline{v} = 0$ on $\partial\mathcal{M}_0$. The finite-difference identity (3.16) directs us to take

$$\overline{u}_{-1,j+1/2} = \overline{u}_{1,j+1/2}, \tag{3.17}$$

whose substitution into (3.15), (3.14) gives a second order approximation of the derived Neumann boundary condition (1.12),

$$\frac{\partial p_s}{\partial \boldsymbol{n}} \mid_{0,y_{j+1/2}} = \frac{\partial p_s}{\partial x} \mid_{0,y_{j+1/2}} = Ro\,\frac{2\nu_1}{\triangle x^2}\overline{u}_{1,j+1/2}. \qquad \text{(ASPBC)} \tag{3.18}$$

The evaluation of $\frac{\partial p_s}{\partial \boldsymbol{n}}$ at three other boundary sections can be derived in the same fashion. We refer to the above formula as the accurate surface pressure boundary condition (ASPBC). A similar derivation for the local pressure boundary condition in the spatially discrete level of the incompressible NSE can be found in a recent paper of Johnston and Liu [11]. Therefore, we have the following set of boundary conditions for p_s in the discrete version:

$$\begin{aligned} (p_s)_{-1/2,j+1/2} &= (p_s)_{1/2,j+1/2} - Ro\,\frac{2\nu_1}{\triangle x}\overline{u}_{1,j+1/2}, \\ (p_s)_{i+1/2,-1/2} &= (p_s)_{i+1/2,1/2} - Ro\,\frac{2\nu_1}{\triangle y}\overline{v}_{i+1/2,1}. \end{aligned} \tag{3.19}$$

3.2. The MAC scheme for the PEs. Thus the system of MAC spatial discretization of the PEs can be written as

$$\left\{ \begin{aligned} & u_t + \mathcal{N}_h(\boldsymbol{u}, u) + \frac{1}{Ro}\Big(-f\overline{\overline{v}} + \mathcal{P}NRX + D_x p_s\Big) = L_{1,h}u \quad \text{at } \triangle, \\ & v_t + \mathcal{N}_h(\boldsymbol{u}, v) + \frac{1}{Ro}\Big(f\overline{\overline{u}} + \mathcal{P}NRY + D_y p_s\Big) = L_{1,h}v \quad \text{at } \bigcirc, \\ & D_z\boldsymbol{v} \mid_{z=-H_0} = 0, \quad D_z\boldsymbol{v} \mid_{z=0} = \frac{\tau_0}{\nu_2}, \\ & \boldsymbol{v}\cdot\boldsymbol{n} = 0, \quad \boldsymbol{v}\cdot\tau = 0 \quad \text{on } \partial\mathcal{M}_0 \times [-H_0, 0], \end{aligned} \right. \tag{3.20a}$$

$$\left\{ \begin{aligned} & (\triangle_h p_s)_{i+1/2,j+1/2} = Ro\,\overline{\mathcal{F}P}_{i+1,j+1/2}, \\ & \frac{\partial p_s}{\partial \boldsymbol{n}} = Ro\,\nu_1\,(\triangle_h \boldsymbol{v})\cdot\boldsymbol{n}, \end{aligned} \right. \tag{3.20b}$$

$$w_{i+1/2,j+1/2,k} = -\triangle z \sum_{l=0}^{k-1}\Big((D_x u)_{i+1/2,j+1/2,l+1/2} + (D_y v)_{i+1/2,j+1/2,l+1/2}\Big), \tag{3.20c}$$

$$\left\{ \begin{aligned} & \rho_t + \mathcal{N}_h(\boldsymbol{u}, \rho) = L_{2,h}\rho \quad \text{at } (i+1/2, j+1/2, k), \\ & \tilde{D}_z\rho \mid_{z=0} = \frac{\rho_f}{\kappa_2}, \quad \tilde{D}_z\rho \mid_{z=-H_0} = 0, \\ & \frac{\partial \rho}{\partial \boldsymbol{n}} = 0 \quad \text{on } \partial\mathcal{M}_0 \times [-H_0, 0]. \end{aligned} \right. \tag{3.20d}$$

Hereafter we denote $L_{1,h} = (\nu_1\triangle_h + \nu_2 D_z^2)$, $L_{2,h} = (\kappa_1\triangle_h + \kappa_2 D_z^2)$ for simplicity of presentation.

3.3. Mean divergence-free property. In this section, we argue that the numerical velocity field $\boldsymbol{v}_h$, the solution of the system (3.20), has free mean-divergence in a discrete level; i.e.,

$$D_x\overline{u} + D_y\overline{v} = 0 \quad \text{on mesh point} \quad (i+1/2, j+1/2), \tag{3.21}$$

where $\overline{u}$, $\overline{v}$ are defined in the same way as in (3.10). To see this, we use a similar argument as in (1.14a), (1.14b), and (1.14c). Taking the discrete divergence of the two momentum equations in (3.20a) at mesh points $(i+1/2, j+1/2, k+1/2)$, summing in the z-direction, and keeping in mind the discrete Poisson equation for p_s as in (3.8), (3.9), we have

$$(\overline{\nabla_h\cdot\boldsymbol{v}})_t = \nu_1\triangle_h\overline{\nabla_h\cdot\boldsymbol{v}} \quad \text{at} \quad (i+1/2, j+1/2). \tag{3.22}$$

In the derivation of (3.22), we used the fact that the composition of discrete divergence and discrete gradient (D_x, D_y) gives exactly the five-point Laplacian in the context of the MAC spatial discretization. Another important fact we used in the derivation of (3.22) is that, on MAC grids, the Laplacian operator $\triangle_h$ and the divergence operator are commutative. These two points represents a crucial advantage of the MAC grid.

The homogeneous initial data for $\overline{\nabla_h\cdot\boldsymbol{v}}$ is obvious:

$$\Big((\overline{\nabla_h\cdot\boldsymbol{v}})(\cdot, t=0)\Big)_{i+1/2,j+1/2} = 0. \tag{3.23}$$

It remains to make sure it vanishes on the lateral boundary $\partial\mathcal{M}_0$. We concentrate on the boundary section $x=0$. The discrete divergence of $\overline{\boldsymbol{v}}$ on $x=0$ can be evaluated as

$$\begin{aligned}(\overline{\nabla_h\cdot\boldsymbol{v}})_{0,j+1/2} &= \frac{1}{2}\Big((\overline{\nabla_h\cdot\boldsymbol{v}})_{-1/2,j+1/2} + (\overline{\nabla_h\cdot\boldsymbol{v}})_{1/2,j+1/2}\Big)\\ &= \frac{1}{2}\Big(\frac{\overline{u}_{0,j+1/2} - \overline{u}_{-1,j+1/2}}{\triangle x} + \frac{\overline{v}_{-1/2,j+1} - \overline{v}_{-1/2,j}}{\triangle y}\\ &\quad + \frac{\overline{u}_{1,j+1/2} - \overline{u}_{0,j+1/2}}{\triangle x} + \frac{\overline{v}_{1/2,j+1} - \overline{v}_{1/2,j}}{\triangle y}\Big),\\ &= \frac{\overline{u}_{1,j+1/2} - \overline{u}_{-1,j+1/2}}{2\triangle x} + \frac{1}{2}\Big(\frac{\overline{v}_{-1/2,j+1} - \overline{v}_{-1/2,j}}{\triangle y} + \frac{\overline{v}_{1/2,j+1} - \overline{v}_{1/2,j}}{\triangle y}\Big),\end{aligned} \tag{3.24}$$

where $\overline{u}_{-1,j+1/2}$, $\overline{v}_{-1/2,j}$ are "ghost" point computational values for $\overline{u}$, $\overline{v}$. Meanwhile, the reflection rule (3.11) (due to the no-slip boundary condition for v) gives that the last two terms in (3.12) vanish, i.e.,

$$(\overline{\nabla_h\cdot\boldsymbol{v}})_{0,j+1/2} = \frac{\overline{u}_{1,j+1/2} - \overline{u}_{-1,j+1/2}}{2\triangle x}. \tag{3.25}$$

By the identity (3.17) that $\overline{u}_{-1,j+1/2} = \overline{u}_{1,j+1/2}$, which is used for the derivation of the Neumann boundary condition for the surface pressure p_s, we conclude that the mean discrete divergence of $\boldsymbol{v}$ vanishes on the boundary $x=0$. In other words, the ASPBC (3.18) conversely indicates the choice for $\overline{u}_{-1,j+1/2}$ as in (3.17). The substitution of (3.17) into (3.25) gives

$$(\overline{\nabla_h\cdot\boldsymbol{v}})_{0,j+1/2} = 0. \tag{3.26}$$

The combination of (3.26), (3.22), (3.23) indicates (3.21), which states that the numerical solution $\boldsymbol{v}_h$ of the system (3.20) has exactly zero discrete mean-divergence. Henceforth, the formula (3.20c) for the determination of vertical velocity is consistent with the combination of divergence-free property of the numerical velocity $\boldsymbol{u}_h$:

$$\nabla_h \cdot \boldsymbol{v}_h + D_z w = 0, \tag{3.27}$$

and the boundary condition for the vertical velocity w at $z = 0$ and $z = -H_0$:

$$w_{i+1/2,j+1/2,0} = w_{i+1/2,j+1/2,N} = 0. \tag{3.28}$$

4. Convergence analysis of the fully discretized scheme using the backward Euler method combined with the MAC grid.

The MAC spatial discretization can be easily implemented in practical computations, combined with either backward Euler or Crank–Nicolson schemes as outlined in section 2. For technical simplicity, the periodic boundary condition is assumed in the horizontal (x, y)-plane so that only the top and bottom boundary sections need to be taken into consideration in the convergence analysis below. The scheme with physical lateral boundary conditions can be dealt with in a similar fashion, with more technical details involved in the consistency analysis. We skip it for the sake of brevity.

The fully discretized scheme using backward Euler temporal discretization is formulated as below. The corresponding Crank–Nicolson method can be similarly proposed and analyzed. We omit it in this article.

$$\begin{cases} \dfrac{u^{n+1} - u^n}{\triangle t} + \mathcal{N}_h(\boldsymbol{u}^n, u^n) + \dfrac{1}{Ro}\Big(-f\overline{\overline{v}}^n + \mathcal{P}NRX^n + D_x p_s^n\Big) \\ \quad = L_{1,h} u^{n+1} \quad \text{at} \ \ \triangle, \\ \dfrac{v^{n+1} - v^n}{\triangle t} + \mathcal{N}_h(\boldsymbol{u}^n, v^n) + \dfrac{1}{Ro}\Big(f\overline{\overline{u}}^n + \mathcal{P}NRY^n + D_y p_s^n\Big) \\ \quad = L_{1,h} v^{n+1} \quad \text{at} \ \ \bigcirc, \\ D_z \boldsymbol{v} \mid_{z=-H_0} = 0, \quad D_z \boldsymbol{v} \mid_{z=0} = 0, \end{cases} \tag{4.1a}$$

$$(\triangle_h p_s)^{n+1}_{i+1/2,j+1/2} = Ro\, \overline{\mathcal{F}P}^{n+1}_{i+1,j+1/2}, \tag{4.1b}$$

$$w^{n+1}_{i+1/2,j+1/2,k} = -\triangle z \sum_{l=0}^{k-1} \Big((D_x u)^{n+1}_{i+1/2,j+1/2,l+1/2} + (D_y v)^{n+1}_{i+1/2,j+1/2,l+1/2} \Big), \tag{4.1c}$$

$$\begin{cases} \dfrac{\rho^{n+1} - \rho^n}{\triangle t} + \mathcal{N}_h(\boldsymbol{u}^n, \rho^n) = L_{2,h} \rho^{n+1} \quad \text{at} \ \ \bullet, \\ \tilde{D}_z \rho \mid_{z=0} = 0, \quad \tilde{D}_z \rho \mid_{z=-H_0} = 0. \end{cases} \tag{4.1d}$$

4.1. Main theorem and some notations.

The following notations of L^2 norms in a discrete level need to be introduced.

Notation 4.1. For any pair of variables u^a, u^b which are defined at the mesh points $(i, j+1/2, k+1/2)$ (such as u, $\tilde{D}_x u$, $\tilde{D}_y u$, $\tilde{D}_z u$, etc.), the discrete L^2-inner product is given by

$$\langle u^a,\, u^b\rangle_1 = \sum_{k=0}^{N-1}\sum_{j=0}^{N-1}\sum_{i=0}^{N-1} u^a_{i,j+1/2,k+1/2}\, u^b_{i,j+1/2,k+1/2}\, h^3. \tag{4.2a}$$

For any pair of variables v^a, v^b which are defined at the mesh points $(i+1/2, j, k+1/2)$ (such as v, $\tilde{D}_x v$, $\tilde{D}_y v$, $\tilde{D}_z v$, etc.), the discrete L^2-inner product is given by

$$\langle v^a,\, v^b\rangle_2 = \sum_{k=0}^{N-1}\sum_{j=0}^{N-1}\sum_{i=0}^{N-1} v^a_{i+1/2,j,k+1/2}\, v^b_{i+1/2,j,k+1/2}\, h^3. \tag{4.2b}$$

For any pair of variables ρ^a, ρ^b defined at the mesh points $(i+1/2, j+1/2, k)$ (such as ρ, w, $\tilde{D}_x \rho$, $\tilde{D}_y \rho$, $\tilde{D}_z \rho$, etc.), the discrete L^2-inner product is given by

(4.2c)

$$\begin{aligned}\langle \rho^a,\, \rho^b\rangle_3 = \sum_{j=0}^{N-1}\sum_{i=0}^{N-1}\Big(&\sum_{k=1}^{N-1} \rho^a_{i+1/2,j+1/2,k}\, \rho^b_{i+1/2,j+1/2,k} \\ &+ \frac{1}{2}\rho^a_{i+1/2,j+1/2,0}\, \rho^b_{i+1/2,j+1/2,0} + \frac{1}{2}\rho^a_{i+1/2,j+1/2,N}\, \rho^b_{i+1/2,j+1/2,N}\Big)\, h^3.\end{aligned}$$

Finally, for any pair of variables p^a, p^b defined at the mesh points $(i+1/2, j+1/2, k+1/2)$ (such as p, $D_x u$, $D_y v$, $D_z w$), the discrete L^2-inner product is defined by

$$\langle p^a,\, p^b\rangle_4 = \sum_{k=0}^{N-1}\sum_{j=0}^{N-1}\sum_{i=0}^{N-1} p^a_{i+1/2,j+1/2,k+1/2}\, p^b_{i+1/2,j+1/2,k+1/2}\, h^3. \tag{4.2d}$$

Clearly, all the discrete L^2-inner products defined above are second order accurate. The corresponding L^2_h norms can be defined accordingly. In addition, we set a vector norm for the horizontal velocity as $\|\boldsymbol{v}\|^2_{\boldsymbol{L}^2_h} = \|u\|_1^2 + \|v\|_2^2$, where $\|u\|_1^2 = \langle u,\, u\rangle_1$, $\|v\|_2^2 = \langle v,\, v\rangle_2$.

The following is the main theorem of this paper.

THEOREM 4.1. *Let $\boldsymbol{u}_e = (\boldsymbol{v}_e, w_e)$, p_e, ρ_e be the exact solution of the PEs* (1.1), (1.2), *with periodic boundary condition in the horizontal (x, y)-plane, and let $(\boldsymbol{v}_{\triangle t,h}, w_{\triangle t,h}, \rho_{\triangle t,h})$ be the numerical solution of the backward Euler coupled with the MAC grid in* (4.1). *We assume that $\triangle t \le Ch$, in which C is an arbitrary fixed constant. Then the following convergence result holds as $\triangle t$ and h go to zero:*

$$\|\boldsymbol{v}_e - \boldsymbol{v}_{\triangle t,h}\|_{L^\infty(0,T;\boldsymbol{L}^2_h)} + \|\rho_e - \rho_{\triangle t,h}\|_{L^\infty(0,T;L^2_h)} \le C(\triangle t + h^2), \tag{4.3a}$$

where the constant C depends only on the regularity of the exact solution

(4.3b)

$$C = C\Big(\|\boldsymbol{u}_e\|_{L^\infty(0,T;C^{7,\alpha})}, \|\rho_e\|_{L^\infty(0,T;C^{7,\alpha})}, \|\boldsymbol{u}_e\|_{C^4(0,T;C^{2,\alpha})}, \|\rho_e\|_{C^4(0,T;C^{2,\alpha})}\Big).$$

The rest of the paper is devoted to the proof of Theorem 4.1. The main steps include the following: 1. The numerical horizontal velocity is shown to have vanishing averaged divergence. 2. The leading order consistency analysis, which gives a construction of approximate velocity and density profiles satisfying the numerical scheme up to an $O(\triangle t + h^2)$ error. Moreover, the constructed horizontal velocity satisfies zero mean-divergence property at the discrete level. 3. Higher order expansion, up to $O(\triangle t^3 + h^4)$ expansion, of the numerical scheme. That makes the recovery of the L^∞ a priori assumption possible, for both the horizontal and the vertical velocity fields in the full nonlinear system of the PEs, by the usage of inverse inequalities. 4. The energy estimate for the error functions. The four steps will be presented in sections 4.2–4.5 below, respectively.

4.2. Evolution for the mean divergence of $\boldsymbol{v}$. To facilitate the proof of Theorem 4.1, we show that the calculated horizontal velocity at each time step has free mean-divergence at the discrete level; i.e., (3.21) is satisfied for $\boldsymbol{v}$ at each time step. The argument is similar to the one in section 3.3. Taking the discrete divergence of the momentum equations in (4.1a) at mesh points $(i+1/2, j+1/2, k+1/2)$ and summing in the z-direction gives

$$\frac{\overline{\nabla_h \cdot \boldsymbol{v}^{n+1}} - \overline{\nabla_h \cdot \boldsymbol{v}^{n}}}{\triangle t} = \nu_1 \triangle_h (\overline{\nabla_h \cdot \boldsymbol{v}^{n+1}}) \quad \text{at} \quad (i+1/2, j+1/2), \tag{4.4a}$$

since all other terms at the time step t^n were canceled by the surface pressure Poisson equation (4.1b) at the same time step. The combination of the evolution equation in (4.4a) and the homogeneous initial data,

$$\overline{\nabla_h \cdot \boldsymbol{v}^{0}} = 0, \tag{4.4b}$$

shows that the numerical solution $\boldsymbol{v}_{h,\triangle t}$ of the scheme (4.1) has exactly zero discrete mean-divergence. As a result, the combined system (3.27), (3.28) is valid for $\boldsymbol{v}^n$, w^n at any time step t^n.

Furthermore, the numerical scheme (4.1a) for the momentum equation and the discrete Poisson equation (4.1b) can also be reformulated in a form similar to that of (1.5), (1.6) in the PDE level, for the sake of simplicity of the convergence analysis given below. We denote the total pressure variable p at mesh points $(i+1/2, j+1/2, k+1/2)$ as

$$p_{i+1/2,j+1/2,k+1/2} = \mathcal{P}R_{i+1/2,j+1/2,k+1/2} + (p_s)_{i+1/2,j+1/2}, \tag{4.5a}$$

where $\mathcal{P}R$, a discrete realization of $\int_z^0 \rho(x,y,s)\,ds$, is defined in a similar way as in (3.7):

$$\begin{aligned} &\mathcal{P}R_{i+1/2,j+1/2,N-1/2} = \frac{1}{2} \triangle z \, \rho_{i+1/2,j+1/2,N}, \\ &\mathcal{P}R_{i+1/2,j+1/2,k-1/2} = \mathcal{P}R_{i+1/2,j,k+1/2} + \triangle z \, \rho_{i+1/2,j+1/2,k}. \end{aligned} \tag{4.5b}$$

Clearly, (4.5) is a discrete version of the hydrostatic equation. One obvious fact is that

$$D_z p = \rho \qquad \text{at the mesh point } (i+1/2, j+1/2, k). \tag{4.6}$$

Therefore the scheme (4.1) can be rewritten as the following system:

$$\text{(4.7a)}\quad \begin{cases} \dfrac{u^{n+1}-u^n}{\triangle t} + \mathcal{N}_h(\boldsymbol{u}^n, u^n) + \dfrac{1}{Ro}\Big(-f\overline{\overline{v}}^n + D_x p^n\Big) = L_{1,h} u^{n+1} \quad \text{at} \;\; \triangle, \\ \dfrac{v^{n+1}-v^n}{\triangle t} + \mathcal{N}_h(\boldsymbol{u}^n, v^n) + \dfrac{1}{Ro}\Big(f\overline{\overline{u}}^n + D_y p^n\Big) = L_{1,h} v^{n+1} \quad \text{at} \;\; \bigcirc, \\ D_z \boldsymbol{v}^{n+1} \mid_{z=-H_0} = 0, \quad D_z \boldsymbol{v}^{n+1} \mid_{z=0} = 0, \end{cases}$$

$$\text{(4.7b)}\qquad D_z p^{n+1} = \rho^{n+1} \quad \text{at} \;\; (i+1/2, j+1/2, k),$$

$$\text{(4.7c)}\qquad \begin{cases} \nabla_h \cdot \boldsymbol{v}^{n+1} + D_z w^{n+1} = 0, \\ w^{n+1}_{i+1/2,j+1/2,0} = w^{n+1}_{i+1/2,j+1/2,N} = 0, \end{cases}$$

$$\text{(4.7d)}\qquad \begin{cases} \dfrac{\rho^{n+1}-\rho^n}{\triangle t} + \mathcal{N}_h(\boldsymbol{u}^n, \rho^n) = L_{2,h}\rho^{n+1} \quad \text{at} \;\; \bullet, \\ \tilde{D}_z \rho \mid_{z=0} = 0, \quad \tilde{D}_z \rho \mid_{z=-H_0} = 0. \end{cases}$$

We remark that the mean divergence-free property for the numerical horizontal velocity field and the corresponding identities (3.27), (3.28) assure that the 3-D velocity field is orthogonal to the horizontal and vertical gradients of the total pressure field p in the staggered L^2 space introduced in (4.2), i.e.,

$$\text{(4.8)}\qquad \langle u, D_x p\rangle_1 + \langle v, D_y p\rangle_2 + \langle w, D_z p\rangle_3 = -\langle (\nabla_h \cdot \boldsymbol{v} + D_z w), p\rangle_4 = 0,$$

by usage of summing by parts in the MAC grid and of the boundary condition for the velocity field. This crucial point makes possible the convergence analysis of the whole numerical scheme using the MAC spatial discretization.

4.3. Leading order consistency analysis. Our goal is to construct approximate velocity profiles $\boldsymbol{V}^0 = (U^0, V^0)$, $\boldsymbol{W}^0$ and approximate density profile Θ^0, and to show that their combination with exact pressure profile p_e satisfies the numerical scheme (4.7) up to an $O(\triangle t + h^2)$ error. Furthermore, the constructed $\boldsymbol{V}^0$ has to be assured to have zero mean-divergence in the discrete sense, i.e.,

$$\text{(4.9)}\qquad \overline{\nabla_h \cdot \boldsymbol{V}^0} = 0 \quad \text{at} \quad (i+1/2, j+1/2),$$

so that the vertical velocity $\boldsymbol{W}^0$ can be determined by the formula in the same way as in (4.7c) consistent with its boundary condition:

$$\text{(4.10)}\qquad \boldsymbol{W}^0_{i+1/2,j+1/2,k} = -\triangle z \sum_{l=0}^{k-1} \Big((D_x U^0)_{i+1/2,j+1/2,l} + (D_y V^0)_{i+1/2,j+1/2,l} \Big).$$

In other words, the combination of (4.9) and (4.10) gives

$$\text{(4.11)}\qquad \begin{cases} \nabla_h \cdot \boldsymbol{V}^0 + D_z W^0 = 0, \\ \boldsymbol{W}^0_{i+1/2,j+1/2,0} = \boldsymbol{W}^0_{i+1/2,j+1/2,N} = 0, \end{cases}$$

which is analogous to (4.7c).

The construction of the leading term for the horizontal velocity field $\boldsymbol{V}^0$ relies on the fact that any C^1 function g in $\mathcal{M}$ can be uniquely recovered from its average in the z-direction and its derivative with respect to z by

$$g(x,y,z) = \int_{-H_0}^{z} g_z(x,y,s)\,ds + \overline{g}(x,y) - \frac{1}{H_0}\int_{-H_0}^{0}\int_{-H_0}^{z'} g_z(x,y,s)\,ds\,dz'. \tag{4.12}$$

As a result, the exact horizontal velocity field $\boldsymbol{v}_e$ can be represented as

(4.13)

$$\begin{aligned} u_e(x,y,z) &= \int_{-H_0}^{z} \partial_z u_e(x,y,s)\,ds + \overline{u_e}(x,y) - \frac{1}{H_0}\int_{-H_0}^{0}\int_{-H_0}^{z} \partial_z u_e(x,y,s)\,ds\,dz,\\ v_e(x,y,z) &= \int_{-H_0}^{z} \partial_z v_e(x,y,s)\,ds + \overline{v_e}(x,y) - \frac{1}{H_0}\int_{-H_0}^{0}\int_{-H_0}^{z} \partial_z v_e(x,y,s)\,ds\,dz. \end{aligned}$$

The discrete form of the recovery formula (4.13) applied to U^0, V^0 can be written as follows:

$$\begin{aligned} U^0_{i,j+1/2,k+1/2} &= -\mathcal{P}UZ^0_{i,j+1/2,k+1/2} + \overline{U^0}_{i,j+1/2} + \overline{PUZ}^0_{i,j+1/2},\\ V^0_{i+1/2,j,k+1/2} &= -\mathcal{P}VZ^0_{i+1/2,j,k+1/2} + \overline{V^0}_{i+1/2,j} + \overline{PVZ}^0_{i+1/2,j}, \end{aligned} \tag{4.14}$$

the construction of the mean velocity field $\overline{U}^0$, $\overline{V}^0$ will be given later, and $\mathcal{P}UZ^0$, $\mathcal{P}VZ^0$ represents the discrete integral of $\partial_z u_e$, $\partial_z v_e$ from $-H_0$ up to $z_k = (k+\frac{1}{2})\triangle z$, respectively. Keeping in mind that $\partial_z u_e$, $\partial_z v_e$ are defined on the numerical grids $(i, j \pm 1/2, k)$, $(i \pm 1/2, j, k)$, respectively, we express such integrals as

$$\begin{aligned} \mathcal{P}UZ^0_{i,j+1/2,-1/2} &= -\frac{1}{2}\triangle z\,(\partial_z u_e)_{i,j+1/2,0},\\ \mathcal{P}UZ^0_{i,j+1/2,k+1/2} &= \mathcal{P}UZ^0_{i,j+1/2,k-1/2} + \triangle z\,(\partial_z u_e)_{i,j+1/2,k}, \end{aligned} \tag{4.15a}$$

$$\begin{aligned} \mathcal{P}VZ^0_{i+1/2,j,-1/2} &= -\frac{1}{2}\triangle z\,(\partial_z v_e)_{i+1/2,j,0},\\ \mathcal{P}VZ^0_{i+1/2,j,k+1/2} &= \mathcal{P}VZ^0_{i+1/2,j,k-1/2} + \triangle z\,(\partial_z v_e)_{i+1/2,j,k}. \end{aligned} \tag{4.15b}$$

Obviously, the combination of (4.14) and (4.15) gives

$$\begin{aligned} &\sum_{k=0}^{N_z-1} (\triangle z\, U^0_{i,j+1/2,k+1/2}) = \overline{U^0}_{i,j+1/2}, \quad (D_z U^0)_{i,j+1/2,k} = (\partial_z u_e)_{i,j+1/2,k},\\ &\sum_{k=0}^{N_z-1} (\triangle z\, V^0_{i+1/2,j,k+1/2}) = \overline{V^0}_{i+1/2,j}, \quad (D_z V^0)_{i+1/2,j,k} = (\partial_z v_e)_{i+1/2,j,k}. \end{aligned} \tag{4.16}$$

We use the "mean stream function" corresponding to the exact velocity solution $\boldsymbol{v}_e$ to construct the mean velocity field $\overline{\boldsymbol{V}^0}$ appearing in the construction formula (4.14). Since the average of the exact velocity field $\overline{\boldsymbol{v}_e}$ is divergence-free in the 2-D domain $\mathcal{M}_0$, as shown in (1.4), there exists a mean stream function $\overline{\psi_e}$ such that

$$\overline{\boldsymbol{v}_e} = \nabla^{\perp}\overline{\psi_e} = (-\partial_y\overline{\psi_e}, \partial_x\overline{\psi_e}). \tag{4.17}$$

Subsequently, the average of $\boldsymbol{V}^0$ (in the z-direction) can be determined via second order finite-difference of the exact "mean stream function"

$$\begin{aligned} \overline{U^0} &= -D_y\overline{\psi_e} = -\frac{\overline{\psi_e}_{i,j+1} - \overline{\psi_e}_{i,j}}{\triangle y} \quad \text{at} \quad (i, j+1/2), \\ \overline{V^0} &= D_x\overline{\psi_e} = \frac{\overline{\psi_e}_{i+1,j} - \overline{\psi_e}_{i,j}}{\triangle x} \quad \text{at} \quad (i+1/2, j). \end{aligned} \tag{4.18}$$

It should be remarked that the mean stream function is evaluated at regular mesh points (i, j), $0 \le i, j \le N$. Obviously, (4.18) gives

$$D_x\overline{U^0} + D_y\overline{V^0} = -D_x(D_y\overline{\psi_e}) + D_y(D_x\overline{\psi_e}) = 0, \tag{4.19}$$

which along with the identity (4.16) assures that the mean divergence-free property is automatically satisfied for the constructed leading velocity field

$$\overline{\nabla_h \cdot \boldsymbol{V}^0} = 0 \quad \text{at} \quad (i+1/2, j+1/2). \tag{4.20}$$

Accordingly, the recovery formula analogous to (4.2) is used to construct the leading vertical velocity

$$\boldsymbol{W}^0_{i+1/2,j+1/2,k} = -\triangle z \sum_{l=0}^{k-1}\Big((D_xU^0)_{i+1/2,j+1/2,l+1/2} + (D_yV^0)_{i+1/2,j+1/2,l+1/2}\Big), \tag{4.21}$$

which is compatible with the boundary condition $\boldsymbol{W}^0_{i+1/2,j+1/2,0} = \boldsymbol{W}^0_{i+1/2,j+1/2,N} = 0$.

The proposition below states that the constructed leading velocity profile, together with its temporal derivative, is within $O(h^2)$ difference with the exact velocity $\boldsymbol{u}_e = (\boldsymbol{v}_e, w_e)$. Its verification is omitted in this paper for brevity and will appear elsewhere.

PROPOSITION 4.2. *The following estimates for $\boldsymbol{V}^0$, $\boldsymbol{W}^0$ hold:*

$$\|\boldsymbol{V}^0 - \boldsymbol{v}_e\|_{W^{m,\infty}(\mathcal{M})} \le Ch^2\|\boldsymbol{v}_e\|_{C^{m+3}} \quad \textit{for} \;\; m = 0, 1, 2\ldots, \tag{4.22a}$$

$$\|W^0 - w_e\|_{W^{m,\infty}(\mathcal{M})} \le Ch^2\|\boldsymbol{v}_e\|_{C^{m+4}}. \tag{4.22b}$$

Here $\|\cdot\|_{W^{m,\infty}(\mathcal{M})}$ represents the maximum value at the corresponding mesh points of the given function up to mth order finite-difference over the 3-D domain $\mathcal{M}$. Furthermore, the difference between the time derivatives of $\boldsymbol{V}^0$ and $\boldsymbol{v}_e$ can be controlled by

$$\partial_t^m\boldsymbol{V}^0 = \partial_t^m\boldsymbol{v}_e + O(h^2)\|\partial_t^m\boldsymbol{v}_e\|_{C^3} \quad \textit{for} \;\; m \ge 1. \tag{4.23}$$

In addition, we observe that $\boldsymbol{V}^0$ exactly satisfies the boundary condition in the discrete form as given in (4.7) at the top $z = 0$ and at the bottom $z = -H_0$:

$$\begin{aligned} &(D_zU^0)_{i,j+1/2,0} = 0, \quad (D_zV^0)_{i+1/2,j,0} = 0, \\ &(D_zU^0)_{i,j+1/2,N} = 0, \quad (D_zV^0)_{i+1/2,j,N} = 0, \end{aligned} \tag{4.24}$$

due to its construction in (4.15) and the fact that $\partial_z \boldsymbol{v}_e = 0$ at the two boundary sections.

The leading order density profile is composed of the exact density and a correction term

$$\Theta^0 = \rho_e + h^2 \Theta^1. \tag{4.25}$$

The addition of the $O(h^2)$ correction terms $h^2\Theta^1$ is due to the higher order consistency of the approximate profile Θ with the boundary condition given in the numerical scheme (4.7d), which is required in the error analysis presented later. The correction function Θ^1 is constructed as the solution of the Poisson equation with Neumann boundary condition

(4.26)
$$\begin{cases} \triangle \Theta^1 = C^1 \equiv \dfrac{1}{|\mathcal{M}|} \left(\displaystyle\int_{\mathcal{M}_0} \frac{1}{6} \partial_z^3 \rho_e(x,y,-H_0)\, d\boldsymbol{x} - \int_{\mathcal{M}_0} \frac{1}{6} \partial_z^3 \rho_e(x,y,0)\, d\boldsymbol{x} \right), \\ \partial_z \Theta^1(x,y,-H_0) = -\dfrac{1}{6}\, \partial_z^3 \rho_e(x,y,-H_0), \quad \partial_z \Theta^1(x,y,0) = -\dfrac{1}{6}\, \partial_z^3 \rho_e(x,y,0). \end{cases}$$

Note that the number C^1 (a function of time t) is chosen so that $\int_{\mathcal{M}} C^1 \, d\boldsymbol{x}\, dz = \int_{\partial\mathcal{M}} \frac{\partial \Theta^1}{\partial \boldsymbol{n}}\, d\boldsymbol{n}$ to maintain the consistency. The Schauder's estimate applied to the Poisson equation (4.26) gives that

$$\|\Theta^1\|_{C^{m,\alpha}} \le \|\rho_e\|_{C^{m+2,\alpha}}, \quad \|\partial_t^k \Theta^1\|_{C^{m,\alpha}} \le \|\partial_t^k \rho_e\|_{C^{m+2,\alpha}} \qquad \text{for } \ m \ge 2. \tag{4.27}$$

The choice of the boundary condition for Θ^1 in (4.26) implies that the approximated density Θ as given in the expansion (4.25) satisfies the discrete boundary condition in (4.7d) to an $O(h^5)$ order. It can be seen by local Taylor expansion for the exact density field ρ_e around the bottom boundary that

(4.28)
$$(\rho_e)_{i+1/2,j+1/2,-1} = (\rho_e)_{i+1/2,j+1/2,1} - \frac{\triangle z^3}{3}\, \partial_z^3 \rho_e(x_{i+1/2}, y_{j+1/2}, -H_0) + O(h^5) \|\rho_e\|_{C^5},$$

in which the no-flux boundary condition is used. The insertion of the boundary condition given by (4.26) into the Taylor expansion of Θ^1, along with Schauder's estimate $\|\Theta_1\|_{C^2} \le C \|\rho_e\|_{C^{5,\alpha}}$ given by (4.27), leads to

(4.29)
$$\Theta^1_{i+1/2,j+1/2,-1} = \Theta^1_{i+1/2,j+1/2,1} + \frac{\triangle z}{3}\, \partial_z^3 \rho_e(x_{i+1/2}, y_{j+1/2}, -H_0) + O(h^3) \|\rho_e\|_{C^{5,\alpha}}.$$

The combination of (4.28) and (4.29) results in the following estimate for $\Theta^0 = \rho_e + h^2\Theta^1$:

$$\Theta^0_{i+1/2,j+1/2,-1} = \Theta^0_{i+1/2,j+1/2,1} + O(h^5) \|\rho_e\|_{C^{5,\alpha}}, \tag{4.30}$$

which proves our claim. The top boundary $z = 0$ can be dealt with in the same manner.

It is straightforward to verify the following local truncation estimates:

$$
\text{(4.31a)}\quad \begin{cases} \dfrac{(U^0)^{n+1}-(U^0)^n}{\triangle t} + \mathcal{N}_h((\boldsymbol{U}^0)^n,(U^0)^n) + \dfrac{1}{Ro}\Big(-f(\overline{\overline{V^0}})^n + D_x p_e^n\Big) \\ \qquad = L_{1,h}(U^0)^{n+1} + \triangle t E_{\triangle t}^{u(0),n} + h^2 E_h^{u(0),n} \quad \text{at } \triangle, \\ \dfrac{(V^0)^{n+1}-(V^0)^n}{\triangle t} + \mathcal{N}_h((\boldsymbol{U}^0)^n,(V^0)^n) + \dfrac{1}{Ro}\Big(f(\overline{\overline{U^0}})^n + D_y p_e^n\Big) \\ \qquad = L_{1,h}(V^0)^{n+1} + \triangle t E_{\triangle t}^{v(0),n} + h^2 E_h^{v(0),n} \quad \text{at } \bigcirc, \\ (U^0)^{n+1}_{i,j+1/2,-1/2} = (U^0)^{n+1}_{i,j+1/2,1/2}, \quad (V^0)^{n+1}_{i+1/2,j,-1/2} = (V^0)^{n+1}_{i+1/2,j,1/2}, \end{cases}
$$

$$
\text{(4.31b)}\qquad D_z p_e^{n+1} = (\Theta^0)^{n+1} + h^2 E_h^{p(0),n} \quad \text{at } (i+1/2, j+1/2, k),
$$

$$
\text{(4.31c)}\qquad \begin{cases} \nabla_h \cdot (\boldsymbol{V}^0)^{n+1} + D_z(\boldsymbol{W}^0)^{n+1} = 0, \\ (\boldsymbol{W}^0)^{n+1}_{i+1/2,j+1/2,0} = (\boldsymbol{W}^0)^{n+1}_{i+1/2,j+1/2,N} = 0, \end{cases}
$$

$$
\text{(4.31d)}\qquad \begin{cases} \dfrac{(\Theta^0)^{n+1}-(\Theta^0)^n}{\triangle t} + \mathcal{N}_h((\boldsymbol{U}^0)^n,(\Theta^0)^n) \\ \qquad = L_{2,h}(\Theta^0)^{n+1} + (\triangle t + h^2) E^{\rho(0),n} \quad \text{at } \bullet, \\ (\Theta^0)^{n+1}_{i+1/2,j+1/2,-1} = (\Theta^0)^{n+1}_{i+1/2,j+1/2,1} + h^5 \boldsymbol{e}_{\rho b}, \end{cases}
$$

via high order Taylor expansion of the constructed solution $\boldsymbol{V}^0$, W^0, Θ^0, along with the usage of Proposition 4.2. The local error terms satisfy

$$
\text{(4.32)}\quad \begin{aligned} &|E^{u(0)}|, |E^{v(0)}| \le C\Big(\|\partial_t \boldsymbol{v}_e\|_{C^2} + \|\partial_t^2 \boldsymbol{v}_e\|_{C^2} + \|\boldsymbol{u}_e\|_{C^6}(1+\|\boldsymbol{u}_e\|_{C^3}) + \|p_e\|_{C^4}\Big), \\ &|E^{p(0)}| \le C\|\rho_e\|_{C^2}, \quad |E^{\rho(0)}| \le C\Big(\|\partial_t \rho_e\|_{C^2} + \|\partial_t^2 \rho_e\|_{C^2} + \|\rho_e\|_{C^5}(1+\|\boldsymbol{u}_e\|_{C^4})\Big). \end{aligned}
$$

4.4. Higher order expansion of the numerical scheme.

The consistency analysis (4.31) is not enough to recover the L^∞ a priori estimates for the approximate velocity field in the full nonlinear system of the PEs. We need to construct further fields, $(\boldsymbol{V}_h^1, \boldsymbol{W}_h^1, \Theta_h^1, P_h^1)$, $(\boldsymbol{V}_{\triangle t}^1, \boldsymbol{W}_{\triangle t}^1, \Theta_{\triangle t}^1, P_{\triangle t}^1)$, $(\boldsymbol{V}_{\triangle t}^2, \boldsymbol{W}_{\triangle t}^2, \Theta_{\triangle t}^2, P_{\triangle t}^2)$, and to introduce, for the error analysis, the fields $\boldsymbol{V}$, $\boldsymbol{W}$, Θ, P defined by

$$
\text{(4.33)}\quad \begin{aligned} &\boldsymbol{V} = \boldsymbol{V}^0 + h^2 \boldsymbol{V}_h^1 + \triangle t \boldsymbol{V}_{\triangle t}^1 + \triangle t^2 \boldsymbol{V}_{\triangle t}^2, \\ &\boldsymbol{W} = \boldsymbol{W}^0 + h^2 \boldsymbol{W}_h^1 + \triangle t \boldsymbol{W}_{\triangle t}^1 + \triangle t^2 \boldsymbol{W}_{\triangle t}^2, \\ &\Theta = \Theta^0 + h^2 \Theta_h^1 + \triangle t \Theta_{\triangle t}^1 + \triangle t^2 \Theta_{\triangle t}^2, \quad P = p_e + h^2 P_h^1 + \triangle t P_{\triangle t}^1 + \triangle t^2 P_{\triangle t}^2. \end{aligned}
$$

These new fields depend solely on $(\boldsymbol{V}^0, \boldsymbol{W}^0, \Theta^0, p_e)$, namely, on the exact solution. Their construction is straightforward but lengthy; we omit the details. The expanded profiles satisfy the backward Euler scheme combined with the MAC grid up to order $O(\triangle t^3 + h^4)$:

(4.34a)
$$\begin{cases} \dfrac{U^{n+1} - U^n}{\triangle t} + \mathcal{N}_h(\boldsymbol{U}^n, U^n) + \dfrac{1}{Ro}\Big(-f\overline{\overline{V^n}} + D_x P^n\Big) \\ \qquad = L_{1,h} U^{n+1} + (\triangle t^3 + h^4) E^{u,n}, \\ \dfrac{V^{n+1} - V^n}{\triangle t} + \mathcal{N}_h(\boldsymbol{U}^n, V^n) + \dfrac{1}{Ro}\Big(f\overline{\overline{U^n}} + D_y P^n\Big) \\ \qquad = L_{1,h} V^{n+1} + (\triangle t^3 + h^4) E^{v,n}, \\ U^{n+1}_{i,j+1/2,-1/2} = U^{n+1}_{i,j+1/2,1/2} + h^5 \boldsymbol{e}_{ub}, \quad V^{n+1}_{i+1/2,j,-1/2} = V^{n+1}_{i+1/2,j,1/2} + h^5 \boldsymbol{e}_{ub}, \end{cases}$$

$$D_z P^{n+1} = \Theta^{n+1} + h^4 E^{p,n}, \tag{4.34b}$$

$$\begin{cases} \nabla_h \cdot \boldsymbol{V}^{n+1} + D_z \boldsymbol{W}^{n+1} = 0, \\ \boldsymbol{W}^{n+1}_{i+1/2,j+1/2,0} = \boldsymbol{W}^{n+1}_{i+1/2,j+1/2,N} = 0, \end{cases} \tag{4.34c}$$

$$\begin{cases} \dfrac{\Theta^{n+1} - \Theta^n}{\triangle t} + \mathcal{N}_h(\boldsymbol{U}^n, \Theta^n) = L_{2,h} \Theta^{n+1} + (\triangle t^3 + h^4) E^{\rho,n}, \\ \Theta^{n+1}_{i+1/2,j+1/2,-1} = \Theta^{n+1}_{i+1/2,j+1/2,1} + h^5 \boldsymbol{e}_{\rho b}, \end{cases} \tag{4.34d}$$

in which the local truncation error and the boundary error terms are bounded in the L^∞ norm

$$|E^u|, |E^v|, |E^p|, |E^\rho||\boldsymbol{e}_{ub}|, |\boldsymbol{e}_{vb}|, |\boldsymbol{e}_{\rho b}| \le \mathcal{C}^*, \tag{4.34e}$$

with the constant $\mathcal{C}^*$ depending on the exact solution. This completes the consistency analysis.

Remark 4.3. As stated earlier, the purpose of the higher order expansion (4.33) is to obtain the L^∞ estimate of the error function via its L^2 norm in higher order accuracy by utilizing an inverse inequality in spatial discretization, which will be shown below. Such expansion is always possible under suitable regularity assumption of the exact solution. A detailed analysis shows that

$$|\boldsymbol{v}_e - \boldsymbol{V}| + |\boldsymbol{w}_e - \boldsymbol{W}| + |\rho_e - \Theta| \le C(\triangle t + h^2), \tag{4.35}$$

with C introduced in Theorem 4.1. This estimate will be used later.

Remark 4.4. We note that there is no $O(h^3)$ term in the higher order expansion (4.33). This is due to the centered difference we used in the spatial discretization, which gives local truncation errors with only even order, etc., $O(h^2)$, $O(h^4)$.

4.5. Proof of Theorem 4.1. We consider the following error functions:

$$\tilde{\boldsymbol{v}} = (\tilde{u}, \tilde{v}) = \boldsymbol{V} - \boldsymbol{v} = (U - u, V - v), \quad \tilde{w} = \boldsymbol{W} - w, \quad \tilde{p} = P - p, \quad \tilde{\rho} = \Theta - \rho. \tag{4.36}$$

Subtracting (4.7) from (4.34), we obtain the following system for the error functions:

(4.37a)
$$\begin{cases} \dfrac{\tilde{u}^{n+1} - \tilde{u}^n}{\triangle t} + \mathcal{E}NLU^n + \dfrac{1}{Ro}\Big(-f \overline{\overline{\tilde{v}}}^n + D_x \tilde{p}^n \Big) = L_{1,h} \tilde{u}^{n+1} + (\triangle t^3 + h^4) E^{u,n}, \\ \dfrac{\tilde{v}^{n+1} - \tilde{v}^n}{\triangle t} + \mathcal{E}NLV^n + \dfrac{1}{Ro}\Big(f \overline{\overline{\tilde{u}}}^n + D_y \tilde{p}^n \Big) = L_{1,h} \tilde{v}^{n+1} + (\triangle t^3 + h^4) E^{v,n}, \\ \tilde{u}^{n+1}_{i,j+1/2,-1/2} = \tilde{u}^{n+1}_{i,j+1/2,1/2} + h^5 \boldsymbol{e}_{ub}, \quad \tilde{v}^{n+1}_{i+1/2,j,-1/2} = \tilde{v}^{n+1}_{i+1/2,j,1/2} + h^5 \boldsymbol{e}_{ub}, \end{cases}$$

(4.37b)
$$D_z \tilde{p}^{n+1} = \tilde{\rho}^{n+1} + h^4 E^{p,n},$$

(4.37c)
$$\begin{cases} \nabla_h \cdot \tilde{\boldsymbol{v}}^{n+1} + \tilde{D}_z \tilde{w}^{n+1} = 0, \\ \tilde{w}^{n+1}_{i+1/2,j+1/2,0} = \tilde{w}^{n+1}_{i+1/2,j+1/2,N} = 0, \end{cases}$$

(4.37d)
$$\begin{cases} \dfrac{\tilde{\rho}^{n+1} - \tilde{\rho}^n}{\triangle t} + \mathcal{E}NLR^n = L_{2,h} \tilde{\rho}^{n+1} + (\triangle t^3 + h^4) E^{\rho,n}, \\ \tilde{\rho}^{n+1}_{i+1/2,j+1/2,-1} = \tilde{\rho}^{n+1}_{i+1/2,j+1/2,1} + h^5 \boldsymbol{e}_{\rho b}\,; \end{cases}$$

the nonlinear error terms corresponding to the convection have the following decomposition:

(4.37e)
$$\begin{aligned} \mathcal{E}NLU &= \mathcal{N}_h(\boldsymbol{U}, U) - \mathcal{N}_h(\boldsymbol{u}, u) = \mathcal{N}_h(\tilde{\boldsymbol{u}}, U) + \mathcal{N}_h(\boldsymbol{u}, \tilde{u}), \\ \mathcal{E}NLV &= \mathcal{N}_h(\boldsymbol{U}, V) - \mathcal{N}_h(\boldsymbol{u}, v) = \mathcal{N}_h(\tilde{\boldsymbol{u}}, V) + \mathcal{N}_h(\boldsymbol{u}, \tilde{v}), \\ \mathcal{E}NLR &= \mathcal{N}_h(\boldsymbol{U}, \Theta) - \mathcal{N}_h(\boldsymbol{u}, \rho) = \mathcal{N}_h(\tilde{\boldsymbol{u}}, \Theta) + \mathcal{N}_h(\boldsymbol{u}, \tilde{\rho}). \end{aligned}$$

4.5.1. Preliminary results.

The following preliminary results will be needed in the energy estimate of the system (4.37). The proofs are straightforward so that we omit the detail.

LEMMA 4.5. *We have the following:*

(i) *Inverse inequality in* 3*-D:*

(4.38a)
$$\|f\|_{L^\infty} \le \frac{C}{h^{\frac{3}{2}}} \|f\|_{L^2}.$$

(ii) *Suppose* $w_{i+1/2,j+1/2,0} = w_{i+1/2,j+1/2,N}$*; then*

(4.38b)
$$\|\overline{\overline{u}}\|_2 \le \|u\|_1, \quad \|\overline{\overline{v}}\|_1 \le \|v\|_2, \quad \|\overline{\overline{w}}\|_1 \le \|w\|_3, \quad \|\overline{\overline{w}}\|_2 \le \|w\|_3.$$

(iii) *For* $\tilde{w}$ *determined by* (4.37c), *we have*

(4.38c)
$$\|\tilde{w}\|_3 \le \|D_x^+ \tilde{u}\|_1 + \|D_y^+ \tilde{v}\|_2.$$

(iv) *Suppose* $\tilde{\boldsymbol{v}} = (\tilde{u}, \tilde{v})$ *and* $\tilde{\rho}$ *satisfy the boundary condition in* (4.37); *then*

(4.38d)
$$\begin{aligned} &\|\tilde{D}_x \tilde{u}\|_1 \le \|D_x^+ \tilde{u}\|_1, \quad \|\tilde{D}_y \tilde{u}\|_1 \le \|D_y^+ \tilde{u}\|_1, \quad \|\tilde{D}_z \tilde{u}\|_1 \le \|D_z^+ \tilde{u}\|_1 + h^4, \\ &\|\tilde{D}_y \tilde{v}\|_2 \le \|D_y^+ \tilde{v}\|_2, \quad \|\tilde{D}_x \tilde{v}\|_2 \le \|D_x^+ \tilde{v}\|_2, \quad \|\tilde{D}_z \tilde{v}\|_2 \le \|D_z^+ \tilde{v}\| + h^4, \\ &\|\tilde{D}_x \tilde{\rho}\|_3 \le \|D_x^+ \tilde{\rho}\|_3, \quad \|\tilde{D}_z \tilde{\rho}\|_3 \le \|D_z^+ \tilde{\rho}\|_3, \quad \|\tilde{D}_y \tilde{\rho}\|_3 \le \|D_y^+ \tilde{\rho}\| + h^4. \end{aligned}$$

(v) *Suppose* $w_{i+1/2,j+1/2,0} = w_{i+1/2,j+1/2,N} = 0$; *then*

$$\text{(4.38e)}\quad \langle u, \tilde{D}_x p\rangle_1 + \langle v, \tilde{D}_y p\rangle_2 + \langle w, \tilde{D}_z p\rangle_3 = -\langle(\nabla_h \cdot \boldsymbol{v} + \tilde{D}_z w), p\rangle_4.$$

4.5.2. Energy estimate for the error functions. Assume a priori that

$$\|\tilde{\boldsymbol{v}}\|_{L^\infty} + \|\tilde{w}\|_{L^\infty} \le \frac{1}{2}. \tag{4.39}$$

Such a priori assumption will be verified later using the inverse inequality (4.38a).

Taking the inner product of the first momentum error equation in (4.37a) with $\tilde{u}^{n+1}$ at the mesh point $(i, j+1/2, k+1/2)$, the second momentum error equation with $\tilde{v}^{n+1}$ at the mesh point $(i+1/2, j, k+1/2)$, the equation in (4.37b) with $\frac{\tilde{w}^{n+1}}{Ro}$ at the mesh point $(i+1/2, j+1/2, k)$, and summing up gives

(4.40)

$$\begin{aligned}
&\frac{1}{2}\cdot\frac{1}{\triangle t}\Big(\|\tilde{u}^{n+1}\|_1^2 - \|\tilde{u}^n\|_1^2 + \|\tilde{u}^{n+1} - \tilde{u}^n\|_1^2 + \|\tilde{v}^{n+1}\|_2^2 - \|\tilde{v}^n\|_2^2 + \|\tilde{v}^{n+1} - \tilde{v}^n\|_2^2\Big)\\
&+ \Big\langle \tilde{u}^{n+1}, \mathcal{E}NLU^n \Big\rangle_1 + \Big\langle \tilde{v}^{n+1}, \mathcal{E}NLV^n \Big\rangle_2 + \frac{1}{Ro}\Big(\langle \tilde{u}^{n+1}, -f\overline{\overline{\tilde{v}}}^n\rangle_1 + \langle \tilde{v}^{n+1}, f\overline{\overline{\tilde{u}}}^n\rangle_2\Big)\\
&+ \frac{1}{Ro}\Big(\langle \tilde{u}^{n+1}, D_x\tilde{p}^n\rangle_1 + \langle \tilde{v}^{n+1}, D_y\tilde{p}^n\rangle_2 + \langle \tilde{w}^{n+1}, D_z\tilde{p}^n\rangle_3\Big)\\
&- \Big\langle \tilde{u}^{n+1}, (\nu_1\triangle_h + \nu_2 D_z^2)\tilde{u}^{n+1}\Big\rangle_1 - \Big\langle \tilde{v}^{n+1}, (\nu_1\triangle_h + \nu_2 D_z^2)\tilde{v}^{n+1}\Big\rangle_2\\
&= \frac{1}{2}\cdot\frac{1}{\triangle t}\Big(\|\tilde{u}^{n+1}\|_1^2 - \|\tilde{u}^n\|_1^2 + \|\tilde{u}^{n+1} - \tilde{u}^n\|_1^2 + \|\tilde{v}^{n+1}\|_2^2 - \|\tilde{v}^n\|_2^2\\
&\qquad\qquad + \|\tilde{v}^{n+1} - \tilde{v}^n\|_2^2\Big) + I_{cu}^n + I_{cv}^n + \frac{1}{Ro}I_{cg}^n + \frac{1}{Ro}I_p^n + I_{du}^{n+1} + I_{dv}^{n+1}\\
&= \langle \tilde{u}^{n+1}, (\triangle t^3 + h^4)E^{u,n}\rangle_1 + \langle \tilde{v}^{n+1}, (\triangle t^3 + h^4)E^{v,n}\rangle_2\\
&\quad - \frac{1}{Ro}\langle \tilde{w}^{n+1}, \tilde{\rho}^n\rangle_3 + \frac{h^4}{Ro}\langle \tilde{w}^{n+1}, E^{p,n}\rangle_3.
\end{aligned}$$

A direct application of part (v) in Lemma 4.5 gives that I_p^n appearing in (4.40) vanishes indeed:

$$\begin{aligned}
I_p^n &= \langle \tilde{u}^{n+1}, D_x\tilde{p}^n\rangle_1 + \langle \tilde{v}^{n+1}, D_y\tilde{p}^n\rangle_2 + \langle \tilde{w}^{n+1}, D_z\tilde{p}^n\rangle_3\\
&= -\langle(\nabla_h \cdot \tilde{\boldsymbol{v}}^{n+1} + D_z\tilde{w}^{n+1}), \tilde{p}^n\rangle_4 = 0,
\end{aligned} \tag{4.41}$$

due to the fact that $\tilde{\boldsymbol{u}} = (\tilde{u}, \tilde{v}, \tilde{w})$ is identically divergence-free at the discrete level and the vertical velocity vanishes on the top and bottom boundaries. The identity (4.41) is analogous to (4.8), which shows that the 3-D velocity field is orthogonal to the pressure gradient in the staggered L^2 space. This represents the main advantage of the MAC grid.

The term I_{cg}^n, which corresponds to the Coriolis force, can be controlled directly by the Cauchy inequality and the application of part (ii) in Lemma 4.5:

(4.42)

$$|I_{cg}^n| = \Big|\langle \tilde{u}^{n+1}, -f\overline{\overline{\tilde{v}}}^n\rangle_1 + \langle \tilde{v}^{n+1}, f\overline{\overline{\tilde{u}}}^n\rangle_2\Big| \le \frac{f_0 + \beta}{2}\Big(\|\tilde{u}^{n+1}\|_1^2 + \|\tilde{u}^n\|_1^2 + \|\tilde{v}^{n+1}\|_2^2 + \|\tilde{v}^n\|_2^2\Big).$$

Next we consider the terms I_{du}^{n+1}, I_{dv}^{n+1} corresponding to the diffusion of $\tilde{u}$ and $\tilde{v}$. A direct calculation shows that

$$-\langle \tilde{u}^{n+1}, D_x^2 \tilde{u}^{n+1}\rangle_1 = \|D_x^+ \tilde{u}^{n+1}\|_1^2, \quad -\langle \tilde{u}^{n+1}, D_y^2 \tilde{u}^{n+1}\rangle_1 = \|D_y^+ \tilde{u}^{n+1}\|_1^2, \tag{4.43a}$$

$$\begin{aligned} &-\langle \tilde{u}^{n+1}, D_z^2 \tilde{u}^{n+1}\rangle_1 = \|D_z \tilde{u}^{n+1}\|_1^2 + \mathcal{B}_{uz}^{n+1}, \\ &\mathcal{B}_{uz}^{n+1} = h^3 \sum_{j=0}^{N-1} \sum_{i=1}^{N-1} \Big(h^3 \tilde{u}_{i,j+1/2,1/2}^{n+1} \boldsymbol{e}_{ub} + h^3 \tilde{u}_{i,j+1/2,N-1/2}^{n+1} \boldsymbol{e}_{ut} \Big), \end{aligned} \tag{4.43b}$$

by utilizing the boundary condition for $\tilde{u}^{n+1}$ given in (4.37a). The boundary error term $\mathcal{B}_{uz}^{n+1}$ can be bounded from below as follows:

(4.44)
$$\begin{aligned} \mathcal{B}_{uz}^{n+1} &\geq h^3 \sum_{j=0}^{N-1} \sum_{i=1}^{N-1} \Big(-\frac{1}{2} (\tilde{u}_{i,j+1/2,1/2}^{n+1})^2 - \frac{1}{2} h^6 \boldsymbol{e}_{ub}^2 - \frac{1}{2} (\tilde{u}_{i,j+1/2,N-1/2}^{n+1})^2 - \frac{1}{2} h^6 \boldsymbol{e}_{ut}^2 \Big) \\ &\geq -\frac{1}{2} \|\tilde{u}^{n+1}\|_1^2 - \frac{1}{2} h^9 \sum_{j=0}^{N-1} \sum_{i=1}^{N-1} (\boldsymbol{e}_{ub}^2 + \boldsymbol{e}_{ut}^2) \geq -\frac{1}{2} \|\tilde{u}^{n+1}\|_1^2 - \frac{1}{2} h^6 \,; \end{aligned}$$

in the second step we absorbed the terms $\tilde{u}_{i,j+1/2,1/2}^2$ and $\tilde{u}_{i,j+1/2,N-1/2}^2$ into $\|\tilde{u}\|_1^2$ by its definition. Then we obtain

(4.45)
$$I_{du}^{n+1} \geq \nu_0 (\|D_x^+ \tilde{u}^{n+1}\|_1^2 + \|D_y^+ \tilde{u}^{n+1}\|_1^2 + \|D_z^+ \tilde{u}^{n+1}\|_1^2) - \frac{1}{2} \nu_2 \|\tilde{u}^{n+1}\|_1^2 - \frac{1}{2} \nu_2 h^6,$$

in which $\nu_0 = \min(\nu_1, \nu_2, \kappa_1, \kappa_2)$. Similar estimates can be obtained for I_{dv}^{n+1}:

$$I_{dv}^{n+1} \geq \nu_0 (\|D_x^+ \tilde{v}^{n+1}\|_2^2 + \|D_y^+ \tilde{v}^{n+1}\|_2^2 + \|D_z^+ \tilde{v}^{n+1}\|_2^2) - \frac{1}{2} \nu_2 \|\tilde{v}^{n+1}\|_2^2 - \frac{1}{2} \nu_2 h^6. \tag{4.46}$$

It remains to estimate I_{cu}^n and I_{cv}^n corresponding to the nonlinear convection terms. Using the decomposition for $\mathcal{E}NLU$ as shown in (4.37e) yields

$$I_{cu}^n = \Big\langle \tilde{u}^{n+1}, \mathcal{E}NLU^n \Big\rangle_1 = \Big\langle \tilde{u}^{n+1}, \mathcal{N}_h(\tilde{\boldsymbol{u}}^n, U^n) \Big\rangle_1 + \Big\langle \tilde{u}^{n+1}, \mathcal{N}_h(\boldsymbol{u}^n, \tilde{u}^n) \Big\rangle_1 . \tag{4.47}$$

The application of the Cauchy inequality to the first integral appearing on the right-hand side of (4.47) indicates

(4.48)
$$-\Big\langle \tilde{u}^{n+1}, \mathcal{N}_h(\tilde{\boldsymbol{u}}^n, U^n) \Big\rangle_1 \leq \tilde{C}_1 \Big(\|\tilde{u}^{n+1}\|_1^2 + \|\tilde{u}^n\|_1^2 + \|\overline{\overline{\tilde{v}}}^n\|_1^2 \Big) + \frac{2\tilde{C}_1^2}{\nu_0} \|\tilde{u}^{n+1}\|_1^2 + \frac{1}{8} \nu_0 \|\overline{\overline{\tilde{w}}}^n\|_1^2,$$

where $\tilde{C}_1 = \|U\|_{W^{1,\infty}}$. The consistency analysis (4.35) shows that $\tilde{C}_1 \leq \|\boldsymbol{v}_e\|_{C^1} + 1$. Meanwhile, the combination of parts (ii) and (iii) in Lemma 4.5 gives

$$\|\overline{\overline{\tilde{w}}}^n\|_1^2 \leq \|\tilde{w}^n\|_3^2 \leq 2(\|D_x^+ \tilde{u}^n\|^2 + \|D_y^+ \tilde{v}^n\|^2), \quad \|\overline{\overline{\tilde{v}}}^n\|_1^2 \leq \|\tilde{v}^n\|_2^2, \tag{4.49}$$

whose insertion into (4.48) leads to

$$\begin{aligned}-\Big\langle \tilde u^{n+1}, \mathcal{N}_h(\tilde{\boldsymbol u}^n, U^n)\Big\rangle_1 &\le \frac{3\tilde C_1^2}{\nu_0}\|\tilde u^{n+1}\|_1^2 + \tilde C_1(\|\tilde u^n\|_1^2 + \|\tilde v^n\|_2^2)\\ &\quad + \frac{1}{4}\nu_0(\|D_x^+\tilde u^n\|_1^2 + \|D_y^+\tilde v^n\|_2^2).\end{aligned} \tag{4.50}$$

The second inner product appearing on the right-hand side of (4.47) can be controlled in a similar way:

$$-\Big\langle \tilde u^{n+1}, \mathcal{N}_h(\boldsymbol u^n, \tilde u^n)\Big\rangle_1 \le 2\cdot\frac{\tilde C_2^2}{\nu_0}\|\tilde u^{n+1}\|_1^2 + \frac{1}{4}\nu_0\Big(\|D_x^+\tilde u^n\|_1^2 + \|D_y^+\tilde u^n\|_1^2 + \|D_z^+\tilde u^n\|_1^2\Big), \tag{4.51}$$

where $\tilde C_2 = \|\boldsymbol u\|_{L^\infty}$. The a priori assumption (4.39) and the consistency analysis (4.35) assures that

$$\tilde C_2 \le \|\boldsymbol V\|_{L^\infty} + \|\boldsymbol W\|_{L^\infty} + \frac{1}{2} \le \|\boldsymbol u_e\|_{C^0} + C(\triangle t + h^2) + \frac{1}{2} \le \|\boldsymbol u_e\|_{C^0} + 1, \tag{4.52}$$

provided that $\triangle t$ and h are small enough. Applying part (iv) of Lemma 4.5 into (4.52) results in

$$\begin{aligned}&-\Big\langle \tilde u^{n+1}, \mathcal{N}_h(\boldsymbol u^n, \tilde u^n)\Big\rangle_1\\ &\le \frac{2\tilde C_2^2}{\nu_0}\|\tilde u^{n+1}\|_1^2 + \frac{1}{4}\nu_0\Big(\|D_x^+\tilde u^n\|_1^2 + \|D_y^+\tilde u^n\|^2 + \|D_z^+\tilde u^n\|_1^2 + 2h^6\Big).\end{aligned} \tag{4.53}$$

Thus the combination of (4.51) and (4.53) gives

$$\begin{aligned}I_{cu}^n &\ge -\Big(\frac{3\tilde C_1^2}{\nu_0} + \frac{2\tilde C_2^2}{\nu_0}\Big)\|\tilde u^{n+1}\|_1^2 - \tilde C_1(\|\tilde u^n\|_1^2 + \|\tilde v^n\|_2^2)\\ &\quad - \frac{1}{2}\nu_0\Big(\|D_x^+\tilde u^n\|_1^2 + \|D_y^+\tilde u^n\|_1^2 + \|D_z^+\tilde u^n\|_1^2\Big) - h^6,\end{aligned} \tag{4.54}$$

where $\tilde C_1 \le \|\boldsymbol v_e\|_{C^1}+1$, $\tilde C_2 \le \|\boldsymbol u_e\|_{C^0}+1$. The bound for I_{cv}^n can be similarly obtained:

$$\begin{aligned}I_{cv}^n &\ge -\Big(\frac{3\tilde C_1^2}{\nu_0} + \frac{2\tilde C_2^2}{\nu_0}\Big)\|\tilde v^{n+1}\|_2^2 - \tilde C_1(\|\tilde u^n\|_1^2 + \|\tilde v^n\|_2^2)\\ &\quad - \frac{1}{2}\nu_0\Big(\|D_x^+\tilde v^n\|_2^2 + \|D_y^+\tilde v^n\|_2^2 + \|D_z^+\tilde v^n\|_2^2\Big) - h^6.\end{aligned} \tag{4.55}$$

The four terms appearing on the right-hand side of (4.40) can be controlled by the Cauchy inequality, together with the application of part (iii) of Lemma 4.5:

$$\begin{aligned}&\langle \tilde u^{n+1}, (\triangle t^3 + h^4)E^{u,n}\rangle_1 \le \frac{1}{2}\|\tilde u^{n+1}\|_1^2 + (\triangle t^6 + h^8)\|E^{u,n}\|_1^2,\\ &\langle \tilde v^{n+1}, (\triangle t^3 + h^4)E^{v,n}\rangle_2 \le \frac{1}{2}\|\tilde v^{n+1}\|_2^2 + (\triangle t^6 + h^8)\|E^{v,n}\|_2^2,\\ &-\frac{1}{Ro}\langle \tilde w^{n+1}, \tilde\rho^n\rangle_3 \le \frac{1}{8}\nu_0(\|D_x^+\tilde u^{n+1}\|^2 + \|D_y^+\tilde v^{n+1}\|^2) + \frac{C}{\nu_0}\|\tilde\rho^n\|_3^2,\\ &\frac{h^4}{Ro}\langle \tilde w^{n+1}, E^{p,n}\rangle_3 \le \frac{1}{8}\nu_0(\|D_x^+\tilde u^{n+1}\|_1^2 + \|D_y^+\tilde v^{n+1}\|_2^2) + \frac{C}{\nu_0}h^8\,\|E^{p,n}\|_3^2.\end{aligned} \tag{4.56}$$

Substituting (4.56), (4.55), (4.54), (4.46), (4.45), (4.42), and (4.41) into (4.40), and denoting

$$\begin{aligned} \mathcal{IEV} &= \|D_x^+\tilde{u}\|_1^2 + \|D_y^+\tilde{u}\|_1^2 + \|D_z^+\tilde{u}\|_1^2 + \|D_x^+\tilde{v}\|_2^2 + \|D_y^+\tilde{v}\|_2^2 + \|D_z^+\tilde{v}\|_2^2, \\ \mathcal{IER} &= \|D_x^+\tilde{\rho}\|_3^2 + \|D_y^+\tilde{\rho}\|_3^2 + \|D_z^+\tilde{\rho}\|_3^2, \end{aligned} \tag{4.57}$$

we have the energy estimate for $\tilde{\boldsymbol{v}}$

$$\begin{aligned} &\frac{1}{2}\cdot\frac{1}{\triangle t}\Big(\|\tilde{u}^{n+1}\|_1^2 - \|\tilde{u}^n\|_1^2 + \|\tilde{v}^{n+1}\|_2^2 - \|\tilde{v}^n\|_2^2\Big) + \frac{3}{4}\nu_0\mathcal{IEV}^{n+1} \\ &\le \Big(\frac{4\tilde{C}_1^2}{\nu_0} + \frac{4\tilde{C}_2^2}{\nu_0} + C\Big)\|\tilde{\boldsymbol{v}}\|^2 + \frac{C}{\nu_0}\|\tilde{\rho}\|_3^2 + \frac{1}{2}\nu_0\mathcal{IEV}^n + \tilde{E}_u^n, \end{aligned} \tag{4.58a}$$

in which $\tilde{C}_1 \le \|\boldsymbol{v}_e\|_{C^1} + 1$, $\tilde{C}_2 \le \|\boldsymbol{u}_e\|_{C^0} + 1$, and the error term satisfies

$$\tilde{E}_u^n \le \frac{1}{2}(\triangle t^6 + h^8)\Big(\|E^{u,n}\|_1^2 + \|E^{v,n}\|_2^2) + \frac{\|E^{p,n}\|_3^2}{\nu_0}\Big) + Ch^6. \tag{4.58b}$$

The energy estimate for the density error function can be carried out in a similar way (we omit the details):

$$\begin{aligned} &\frac{1}{2}\cdot\frac{1}{\triangle t}\Big(\|\tilde{\rho}^{n+1}\|_3^2 - \|\tilde{\rho}^n\|_3^2\Big) + \frac{3}{4}\nu_0\mathcal{IER}^{n+1} \le \Big(\frac{\tilde{C}_3^2}{\nu_0} + \frac{4\tilde{C}_2^2}{\nu_0} + C\Big)\|\tilde{\rho}\|_3^2 + \frac{1}{2}\|\tilde{\boldsymbol{v}}\|^2 \\ &\quad + \frac{1}{8}\nu_0(\|D_x^+\tilde{u}^n\|_1^2 + \|D_y^+\tilde{v}^n\|_2^2) + \frac{1}{2}\nu_0\mathcal{IER}^n + C(\triangle t^6 + h^8)\,\|E^{\rho,n}\|_3^2, \end{aligned} \tag{4.59}$$

in which $\tilde{C}_3 = \|\Theta\|_{W^{1,\infty}} \le \|\rho_e\|_{C^1} + 1$. By setting $\|\tilde{\boldsymbol{v}}\|^2 = \|\tilde{u}\|_1^2 + \|\tilde{v}\|_2^2$, we arrive at

$$\begin{aligned} &\frac{1}{2}\cdot\frac{1}{\triangle t}\Big(\|\tilde{\boldsymbol{v}}^{n+1}\|^2 - \|\tilde{\boldsymbol{v}}^n\|^2 + \|\tilde{\rho}^{n+1}\|_3^2 - \|\tilde{\rho}^n\|_3^2\Big) + \frac{3}{4}\nu_0\,\mathcal{IEV}^{n+1} + \frac{3}{4}\nu_0\,\mathcal{IER}^{n+1} \\ &\le \Big(\frac{4\tilde{C}_1^2}{\nu_0} + \frac{4\tilde{C}_2^2}{\nu_0} + \frac{\tilde{C}_3^2}{\nu_0} + C\Big)(\|\tilde{\boldsymbol{v}}\|^2 + \|\tilde{\rho}\|_3^2) + \frac{5}{8}\nu_0\,\mathcal{IEV}^n + \frac{1}{2}\nu_0\,\mathcal{IER}^n + \tilde{E}^n, \text{ with} \\ &\qquad \tilde{E}^n \le C(\triangle t^6 + h^8)\Big(\|E^{u,n}\|_1^2 + \|E^{v,n}\|_2^2 + \|E^{\rho,n}\|_3^2) + \frac{\|E^{p,n}\|_3^2}{\nu_0}\Big) + Ch^6, \end{aligned} \tag{4.60}$$

since the term $(\|D_x^+\tilde{u}^n\|_1^2+\|D_y^+\tilde{v}^n\|_2^2)$ appearing on the right-hand side of (4.59) can be absorbed into $\mathcal{IEV}^n$. Summing (4.60) in time and applying the Gronwall inequality yield

$$\|\tilde{\boldsymbol{v}}^n\|^2 + \|\tilde{\rho}^n\|_3^2 \le \quad C\cdot\exp\Big(\frac{Ct}{\nu_0}\Big)\Big(\triangle t^6 + h^8\Big)(\mathcal{C}^*)^2 + CTh^6, \tag{4.61}$$

where C was given in Theorem 4.1 and $\mathcal{C}^*$ depends only on the exact solution. In the derivation of (4.61), we drop the gradient terms since the coefficients of $\mathcal{IEV}$, $\mathcal{IER}$ on the right-hand side of (4.60) are less than those on the left-hand side. The inequality (4.61) amounts to saying

$$\begin{aligned} &\|\boldsymbol{v}_{\triangle t,h} - \boldsymbol{V}\|_{L^\infty(0,T;\boldsymbol{L}_h^2)} + \|\rho_{\triangle t,h} - \Theta\|_{L^\infty(0,T;L_h^2)} \\ &\qquad \le C\mathcal{C}^*\Big(\exp\Big\{\frac{CT}{\nu_0}\Big\} + T\Big)\Big(\triangle t^3 + h^3\Big), \end{aligned} \tag{4.62}$$

whose combination with the estimate (4.35) gives the convergence result (4.3a). The inverse inequality in three dimensions as given in Lemma 4.5 shows that

$$\|\tilde{\boldsymbol{v}}\|_{L^\infty} \leq C\frac{\triangle t^3 + h^3}{h^{\frac{3}{2}}}, \tag{4.63}$$

and this is bounded by $Ch^{3/2}$, since we impose to $\triangle t$ a CFL-like condition $\triangle t \leq Ch$. Moreover, we have

$$\|\tilde{w}\|_{L^\infty} \leq \frac{C}{h}\|\tilde{\boldsymbol{v}}\|_{L^\infty} \leq Ch^{1/2}, \tag{4.64}$$

which comes from the determination identity for $\tilde{w}$ in (4.37). As a result, the a priori assumption (4.39) is satisfied if h is small enough. Thus Theorem 4.1 is proven.

Remark 4.6. The inverse inequality (4.38a) recovers the L^∞ a priori assumption (4.39) for the velocity field. This is the main advantage in the analysis of the fully discretized system. Since the vertical velocity is formulated as the integration of the divergence for the horizontal velocity, the $O(h^{\frac{5}{2}})$ estimate for the L^2 norm of $\tilde{\boldsymbol{v}}$ is required. This is the reason for the higher order consistency analysis in section 4.4.

Remark 4.7. The stability constraint in Theorem 4.1 is $\triangle t \leq Ch$. We infer from (4.62) that the backward Euler scheme is unconditionally stable for the $L^2(0, T; L^2)$ norm, as expected from a scheme with implicit treatment of the diffusion term. The stability constraint $\triangle t \leq Ch$ is introduced after (4.63), (4.64) to recover the $L^\infty([0, T] \times \mathcal{M})$ stability, and C is an arbitrary fixed constant; note that the usual CFL constraint has the same form with $C = |\boldsymbol{u}|^{-1}_{L^\infty}$, but in the present case C is arbitrary, since the CFL condition is needed only to ensure additional stability.

5. Numerical accuracy check. In this section we check the numerical accuracy of the computational scheme. The exact velocity and density are chosen to be

$$\begin{aligned} u_e(x, y, z, t) &= \frac{1}{\pi^2}\text{sin}(\pi x)\text{sin}(\pi y)\text{cos}(\pi z)\text{cos}t, \\ v_e(x, y, z, t) &= \frac{1}{\pi^2}\text{sin}(\pi x)\text{sin}(\pi y)\text{cos}(\pi z)\text{cos}t, \\ \rho_e(x, y, z, t) &= \frac{1}{\pi^2}\text{cos}(\pi x)\text{cos}(\pi y)\text{cos}(\pi z)\text{cos}t. \end{aligned} \tag{5.1}$$

The corresponding exact vertical velocity w_e and exact pressure variable p_e are determined by the incompressibility $\nabla\cdot\boldsymbol{v}_e + \partial_z w_e$ and hydrostatic balance $\frac{\partial p_e}{\partial z} = -\rho_e$, respectively:

$$\begin{aligned} w_e(x, y, z, t) &= -\frac{1}{\pi^2}\Big(\text{cos}(\pi x)\text{sin}(\pi y) + \text{sin}(\pi x)\text{cos}(\pi y)\Big)\text{sin}(\pi z)\text{cos}t, \\ p_e(x, y, z, t) &= \frac{1}{\pi^3}\text{cos}(\pi x)\text{cos}(\pi y)\Big(1 - \text{sin}(\pi z)\Big)\text{cos}t, \end{aligned} \tag{5.2}$$

in which we set the exact surface pressure as

$$p_{se}(x, y, t) = p_e(x, y, 0, t) = \frac{1}{\pi^3}\text{cos}(\pi x)\text{cos}(\pi y)\text{cos}t. \tag{5.3}$$

TABLE 5.1

Error and order of accuracy for velocity and density at $t = 1$ when the Crank–Nicolson *scheme using* MAC *spatial discretization is used.* $\triangle t = \frac{1}{4}\triangle x$. *The physical parameters: Rossby number* $Ro = 0.5$, *Coriolis force* $f = 0.5 + y$.

	N	L^1 error	L^1 order	L^2 error	L^2 order	L^∞ error	L^∞ order
u	16	6.67e-05		9.15e-05		3.50e-04	
	32	1.66e-05	2.00	2.29e-05	1.99	9.06e-05	1.95
	64	4.15e-06	2.00	5.71e-06	2.00	2.29e-05	1.98
	128	1.03e-07	2.01	1.43e-06	2.00	5.73e-06	2.00
v	16	2.56e-04		3.78e-04		1.28e-03	
	32	6.40e-05	2.00	9.46e-05	1.99	3.23e-04	1.99
	64	1.60e-06	2.00	2.37e-05	2.00	8.10e-05	2.00
	128	4.01e-06	2.00	5.93e-06	2.00	2.03e-05	2.00
ρ	16	4.78e-05		6.17e-05		2.01e-04	
	32	1.19e-05	2.00	1.54e-05	2.00	5.22e-05	1.95
	64	2.98e-06	2.00	3.68e-06	2.00	1.32e-05	1.98
	128	7.48e-07	1.99	9.68e-07	2.00	3.30e-06	2.00

Then we arrive at the following system of PEs with force terms $\boldsymbol{f}$, $\boldsymbol{g}$ in the momentum equation and the density equation

(5.4a)

$$\begin{cases} \partial_t \boldsymbol{v}_e + (\boldsymbol{v}_e \cdot \nabla)\boldsymbol{v}_e + w_e \dfrac{\partial \boldsymbol{v}_e}{\partial z} + \dfrac{1}{Ro}\Big(f k \times \boldsymbol{v}_e + \nabla p_e \Big) = \Big(\nu_1 \triangle + \nu_1 \partial_z^2 \Big) \boldsymbol{v}_e + \boldsymbol{f}, \\ \dfrac{\partial p_e}{\partial z} = -\rho_e, \\ \nabla \cdot \boldsymbol{v}_e + \partial_z w_e = 0, \\ \partial_t \rho_e + (\boldsymbol{v}_e \cdot \nabla)\rho_e + w_e \dfrac{\partial \rho_e}{\partial z} = \Big(\kappa_1 \triangle + \kappa_2 \partial_z^2 \Big) \rho_e + \boldsymbol{g}, \end{cases}$$

with the boundary condition

$$\begin{aligned} &\frac{\partial \boldsymbol{v}_e}{\partial z} = 0, \quad w_e = 0, \quad \frac{\partial \rho_e}{\partial z} = 0 \quad \text{at}\, z = 0, -H_0, \\ &\boldsymbol{v}_e = 0, \quad \frac{\partial \rho_e}{\partial \boldsymbol{n}} = 0 \quad \text{on} \quad \partial \mathcal{M}_0 \times [-H_0, 0]. \end{aligned} \tag{5.4b}$$

The computational domain is chosen as $\mathcal{M} = \mathcal{M}_0 \times [-H_0, 0]$, where $\mathcal{M}_0 = [0,1]^2$, $H_0 = 1$. The viscosity parameters are given by $\nu_1 = \nu_2 = 0.005$, $\kappa_1 = \kappa_2 = 0.005$. In a usual GFD model, the Rossby number ranges from $O(1)$ to $O(10^{-3})$. We choose $Ro = 0.5$ in the numerical experiment. The Coriolis force parameter is set to be $f_0 = 0.5$, $\beta = 1$.

The system (5.4) can be formulated in the same fashion as (1.13) such that the surface pressure Poisson equation replaces the nonlocal incompressibility constraint for the horizontal velocity. Note that a force term $\nabla \cdot \boldsymbol{f}$ appears in the Poisson equation. Based on such formulation, we apply the Crank–Nicolson method, a second order numerical scheme with implicit diffusion terms, using the MAC spatial grid, to solve the PEs (5.4). The force terms $\boldsymbol{f}$, $\boldsymbol{g}$ and $\nabla \cdot \boldsymbol{f}$ are added when we update the momentum equation and the density equation and solve the surface pressure Poisson equation. The final time is taken to be $t = 1.0$. Table 5.1 lists the absolute errors between the numerical and exact solutions for velocity and density. All the error functions are measured in L^1, L^2, and L^∞ norms in a discrete level similar to that

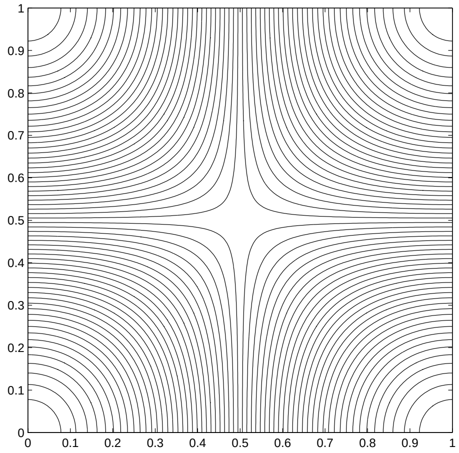


Fig. 5.1. *The contour plot of the surface pressure at $t = 1$ with $N = 128$.*

in the notation (4.2). As can be seen, exactly second order accuracy for the velocity field $\boldsymbol{v} = (u, v)$ and the density field ρ, in both L^1, L^2, and L^∞ norms, is obtained.

The contour plot of the surface pressure at the final time $t = 1.0$ (calculated by the resolution $N = 128$) is also presented in Figure 5.1, which shows a smooth numerical profile and verifies the robustness of the computational method. Such a plot gives an accurate approximation to the exact surface pressure given by (5.3).

Acknowledgment. The authors are grateful to the referees for their insightful comments.

REFERENCES

[1] A. S. Almgren, J. B. Bell, and W. G. Szymczak, *A numerical method for the incompressible Navier–Stokes equations based on an approximate projection*, SIAM J. Sci. Comput., 17 (1996), pp. 358–369.

[2] A. J. Chorin, *Numerical solution of the Navier-Stokes equations*, Math. Comp., 22 (1968), pp. 745–762.

[3] J. K. Dukowicz and R. D. Smith, *A reformulation and implementation of Bryan-Cox-Semtner ocean model on the connection machine*, J. Atmos. Oceanic Tech., 10 (1993), pp. 195–208.

[4] W. E and J.-G. Liu, *Projection method* I: *Convergence and numerical boundary layers*, SIAM J. Numer. Anal., 32 (1995), pp. 1017–1057.

[5] W. E and J.-G. Liu, *Projection method* II: *Godunov–Ryabenski analysis*, SIAM J. Numer. Anal., 33 (1996), pp. 1597–1621.

[6] W. E and J.-G. Liu, *Vorticity boundary condition and related issues for finite difference schemes*, J. Comput. Phys., 124 (1996), pp. 368–382.

[7] W. E and J.-G. Liu, *Projection method* III: *Spatial discretization on the staggered grid*, Math. Comp., 71 (2002), pp. 27–47.

[8] P. Gresho and R. Sani, *On pressure boundary conditions for the incompressible Navier-Stokes equations*, Internat. J. Numer. Methods Fluids, 7 (1987), pp. 1111–1145.

[9] F. Harlow and J. Welch, *Numerical calculation of time-dependent viscous incompressible flow of fluid with free surface*, Phys. Fluids, 8 (1965), pp. 2182–2189.

[10] T. Y. Hou and B. T. R. Wetton, *Convergence of a finite difference scheme for the Navier–Stokes equations using vorticity boundary conditions*, SIAM J. Numer. Anal., 29 (1992), pp. 615–639.

[11] H. E. Johnston and J.-G. Liu, *Finite difference schemes for incompressible flow based on local pressure boundary conditions*, J. Comput. Phys., 180 (2002), pp. 120–154.

[12] J. L. Lions, R. Temam, and S. Wang, *New formulations of the primitive equations of the atmosphere and applications*, Nonlinearity, 5 (1992), pp. 237–288.

[13] J. L. Lions, R. Temam, and S. Wang, *On the equations of large-scale ocean*, Nonlinearity, 5 (1992), pp. 1007–1053.

[14] J. L. Lions, R. Temam, and S. Wang, *Models for the coupled atmosphere and ocean (CAO* I*)*, Comput. Mech. Adv., 1 (1993), pp. 3–54.

[15] J. L. Lions, R. Temam, and S. Wang, *Numerical analysis of the coupled atmosphere and ocean models (CAO*II*)*, Comput. Mech. Adv., 1 (1993), pp. 55–120.

[16] J. L. Lions, R. Temam, and S. Wang, *Mathematical problems of the coupled models of atmosphere and ocean (CAO*III*)*, J. Math. Pures Appl. (9), 74 (1995), pp. 105–163.

[17] S. Orszag and M. Israeli, *Numerical simulation of viscous incompressible flow*, Annu. Rev. Fluid Mech., 6 (1974), pp. 281–318.

[18] S. Orszag and M. Israeli, *Boundary conditions for incompressible flows*, J. Sci. Comput., 1 (1986), pp. 75–111.

[19] J. Pedlosky, *Geophysical Fluid Dynamics*, 2nd ed., Springer-Verlag, New York, 1987.

[20] N. Pinardi, A. Rosati, and R. C. Pacanowski, *The sea surface pressure formulation of frigid lid models. Implications for altimetric data assimilation studies*, J. Marine Systems, 6 (1995), pp. 109–119.

[21] J. Shen, *On error estimates of projection methods for Navier–Stokes equations: First-order schemes*, SIAM J. Numer. Anal., 29 (1992), pp. 57–77.

[22] J. Shen, *On error estimates of some higher order projection and penalty-projection methods for Navier-Stokes equations*, Numer. Math., 62 (1992), pp. 49–73.

[23] J. Shen and S. Wang, *A fast and accurate numerical scheme for the primitive equations of the atmosphere*, SIAM J. Numer. Anal., 36 (1999), pp. 719–737.

[24] R. D. Smith, J. K. Dukowicz, and R. C. Malone, *Parallel ocean general circulation modeling*, Phys. D, 60 (1992), pp. 38–61.

[25] R. Temam, *Sur l'approximation de la solution des equations Navier-Stokes par la méthode des fractionnarires ii*, Arch. Rational Mech. Anal., 33 (1969), pp. 377–385.

[26] R. Temam, *Navier Stokes Equations: Theory and Numerical Analysis*, North-Holland, Amsterdam, 1984.

[27] A. Thom, *The flow past circular cylinders at low speeds*, Proc. Roy. Soc. London Ser. A, 141 (1933), pp. 651–669.

[28] C. Wang and J.-G. Liu, *Analysis of finite difference schemes for unsteady Navier-Stokes equations in vorticity formulation*, Numer. Math., 91 (2002), pp. 543–576.

SIAM J. NUMER. ANAL.
Vol. 41, No. 4, pp. 1195–1208

p INTERPOLATION ERROR ESTIMATES FOR EDGE FINITE ELEMENTS OF VARIABLE ORDER IN TWO DIMENSIONS*

L. DEMKOWICZ† AND I. BABUŠKA†

Abstract. We derive optimal p interpolation error estimates for triangular edge elements of variable order.

Key words. Maxwell's equations, hp finite elements

AMS subject classifications. 65N30, 35L15

DOI. 10.1137/S0036142901387932

1. Introduction.

This paper is concerned with the finite element discretization of Maxwell's equations in two dimensions. Critical to the theory of such discretizations is the so-called de Rham diagram relating two exact sequences of spaces, on both continuous and discrete levels, and corresponding interpolation operators,

$$\begin{tikzcd}
\mathbb{R} \arrow[r] \arrow[d, "id"] & H^{1+\epsilon} \arrow[r, "\boldsymbol{\nabla}"] \arrow[d, "\Pi"] & \boldsymbol{H}^{\epsilon} \cap \boldsymbol{H}(\mathrm{curl}) \arrow[r, "\boldsymbol{\nabla}\times"] \arrow[d, "\Pi^{\mathrm{curl}}"] & L^2 \arrow[r] \arrow[d, "{P,}"] & \mathbf{0}, \\
\mathbb{R} \arrow[r] & \mathcal{P}^{p+1}_{p_e+1} \arrow[r, "\boldsymbol{\nabla}"] & \mathbf{P}^{p}_{p_e} \arrow[r, "\boldsymbol{\nabla}\times"] & \mathcal{P}^{p-1} \arrow[r] & \mathbf{0}\,.
\end{tikzcd} \tag{1.1}$$

All functional spaces are defined on the equilateral, master triangular element[1]

$$T = \left\{ (x_1, x_2) \,:\, x_2 > 0,\ x_2 < \sqrt{3}\left(x_1 + \frac{1}{2}\right),\ x_2 < -\sqrt{3}\left(x_1 - \frac{1}{2}\right) \right\} .$$

In (1.1) and throughout this paper, $\epsilon > 0$ denotes a small positive constant (always smaller than exponent r representing regularity of functions being interpolated), and all constants that appear in presented estimates depend, in general, on ϵ.

The polynomial spaces present in the diagram are defined as follows.

- $\mathcal{P}^p_{p_e}$—space of polynomials of order p, defined on the triangle, whose traces to edges e reduce to (possibly lower) order $p_e, e = 1, 2, 3$.
- $\boldsymbol{P}^p_{p_e}$—space of vector-valued polynomials of order p, defined on triangle T, with traces of their tangential components on edges e of (possibly lower) order p_e.
- $\mathcal{P}^p$—space of polynomials of order p, defined on triangle T.

In particular, $\mathcal{P}^p_{-1}$ denotes the space of polynomials of order p, vanishing on the boundary of the triangle, and $\boldsymbol{P}^p_{-1}$ stands for the space of vector-valued polynomials of order p, with the trace of the tangential component on the boundary equal

*Received by the editors April 11, 2001; accepted for publication (in revised form) November 20, 2002; published electronically August 22, 2003. The work of the first author was supported by the Air Force under contract F49620-98-1-0255, and the work of the second author was supported by NSF grant DMS 9802357.

http://www.siam.org/journals/sinum/41-4/38793.html

†Texas Institute for Computational and Applied Mathematics, University of Texas at Austin, Austin, TX 78712 (leszek@ticam.utexas.edu, babuska@ticam.utexas.edu).

[1] We shall restrict our presentation to triangular elements only.

to zero. The assumption that edge orders p_e should not exceed polynomial order p, $p_e \leq p, e = 1, 2, 3$, is realized in practice by implementing the *minimum rule* that sets an edge order p_e to the minimum of orders p corresponding to the adjacent elements.

The three interpolation operators present in the diagram map are H^1-, $H(\text{curl})$-, and L^2-conforming operators. The last operator is simply the L^2-projection, and the first two will be discussed in detail in the next sections. The operators are constructed in such a way that, when applied elementwise to a function defined on a whole mesh, they yield a discretization that is globally conforming. Simply speaking, if u is continuous, then the union of element H^1-interpolants is continuous as well. Similarly, if field $\boldsymbol{E}$ has a continuous tangential component across every interelement boundary, the corresponding $H(\text{curl})$-conforming interpolant will have the same property.

This paper is concerned with p interpolation error estimates, with respect to order of approximation p, with minimum regularity assumptions that would yield optimal convergence rates. Such estimates are critical in proving discrete compactness property, and consequently convergence of Maxwell eigenvalues [6], which in turn implies (and it is necessary for) the stability of finite element approximations to the time-harmonic Maxwell equations. The p interpolation error estimates are also the first step towards proving exponential convergence of hp discretizations; see, e.g., [19], analysis of two-grid and multigrid algorithms, etc.

The presented H^1 interpolation error estimate derives directly from the works of Babuška and his collaborators. To the best of our knowledge, the corresponding result for the $H(\text{curl})$-conforming interpolation is new.

There are two related known results for quad edge elements. Monk [13] proved (suboptimal) p interpolation error estimates for square and hexahedral elements using the Nédélec interpolation [14, 15]. Stenberg and Suri [21] studied Brezzi–Douglas–Fortin–Marini (BDFM) elements (in two dimensions $H(\text{curl})$-conforming elements can be obtained by "rotating" $H(\text{div})$-conforming elements), and proved ϵ-optimal L^2 and $H(\text{curl})$ estimates for the original BDFM interpolation operator, generalizing and improving the earlier work of Suri [22] for Brezzi–Douglas–Marini spaces.

In both cases, the proofs were based on expansions in terms of Legendre polynomials and do not seem to be generalizable to the triangular elements.

For details on implementation of the variable order edge elements, see [17, 18].

2. Preliminaries.

Fundamental spaces and norms. We shall use a number of standard inner products and the corresponding norms.

The L^2 product on triangle T will be denoted (u, v), and the corresponding L^2 norm will be denoted $\|u\|$. The notions extend in a standard way to vector-valued functions.

The standard H^1-norm will be denoted $\|u\|_{H^1(T)}$. We shall also need the corresponding seminorm,

$$|u|_{H^1(T)} = \|\boldsymbol{\nabla} u\| \,.$$

Space $H^{\frac{1}{2}}(\partial T)$ will be defined as the space of traces of functions from $H^1(T)$ to boundary ∂T. The corresponding seminorm can be defined as

$$|u|^2_{H^{\frac{1}{2}}(\partial T)} = \inf_{U|_{\partial T} = u} |U|^2_{H^1(T)} = \|\boldsymbol{\nabla} \tilde{u}\|^2 = \left\langle \frac{\partial \tilde{u}}{\partial n}, \tilde{u} \right\rangle \,. \tag{2.1}$$

Here $\langle\cdot,\cdot\rangle$ stands for the duality pairing between $H^{-\frac{1}{2}}(\partial T)$ and $H^{\frac{1}{2}}(\partial T)$, $\tilde{u}$ is the harmonic lift of function $u \in H^{\frac{1}{2}}(\partial T)$, and the normal derivative is in the subspace of functionals with vanishing average,

$$H_0^{-\frac{1}{2}}(\partial T) = \{\phi \in H^{-\frac{1}{2}}(\partial T) \,:\, \langle \phi, 1\rangle = 0\}\,.$$

The parallelogram law allows us to learn the corresponding inner product,

$$\begin{aligned}(u,v)_{H^{\frac{1}{2}}(\partial T)} &= \frac{1}{4}\left\{|u+v|^2_{H^{\frac{1}{2}}(\partial T)} - |u-v|^2_{H^{\frac{1}{2}}(\partial T)}\right\}\\ &= \frac{1}{4}\{\|\boldsymbol{\nabla}(\tilde{u}+\tilde{v})\|^2 - \|\boldsymbol{\nabla}(\tilde{u}-\tilde{v}\|^2\}\\ &= (\boldsymbol{\nabla}\tilde{u}, \boldsymbol{\nabla}\tilde{v})\\ &= \left\langle \frac{\partial \tilde{u}}{\partial n}, \tilde{v}\right\rangle = \left\langle \frac{\partial \tilde{v}}{\partial n}, \tilde{u}\right\rangle.\end{aligned}$$

In an analogous way we can use the norm and inner product in $H^1(T)$ to introduce the corresponding norm and inner products in $H^{\frac{1}{2}}(\partial T)$. On the quotient space $H^{\frac{1}{2}}(\partial T)/\mathbb{R}$, the seminorm turns into a norm, equivalent to the standard norm for quotient spaces. Subspace $H_0^{-\frac{1}{2}}(\partial T)$ can be identified with the dual of $H^{\frac{1}{2}}(\partial T)/\mathbb{R}$.

By Riesz theorem, $\phi \in H_0^{-\frac{1}{2}}(\partial T)$ can be identified with the equivalence class of a function $u_\phi \in H^{\frac{1}{2}}(\partial T)$ such that

$$\langle \phi, v\rangle = (\boldsymbol{\nabla}\tilde{u}_\phi, \boldsymbol{\nabla}\tilde{v}) = \left\langle \frac{\partial \tilde{u}_\phi}{\partial n}, \tilde{v}\right\rangle\,.$$

Consequently,

$$\|\phi\|_{H_0^{-\frac{1}{2}}(\partial T)} = \sup_{v\in H^{\frac{1}{2}}(\partial T)} \frac{|\langle \phi, v\rangle|}{|v|_{H^{\frac{1}{2}}(\partial T)}} = \sup_{v\in H^{\frac{1}{2}}(\partial T)} \frac{(\boldsymbol{\nabla}\tilde{u}_\phi, \boldsymbol{\nabla}\tilde{v})}{|\boldsymbol{\nabla}\tilde{v}|} = |\boldsymbol{\nabla}\tilde{u}_\phi|\,, \tag{2.2}$$

and the parallelogram law again can be used to recover the scalar product,

$$(\phi,\psi)_{H_0^{-\frac{1}{2}}(\partial T)} = (\boldsymbol{\nabla}\tilde{u}_\phi, \boldsymbol{\nabla}\tilde{v}_\psi)\,.$$

Let e be one of the triangle edges. Space $H_{00}^{\frac{1}{2}}(e)$ is defined as the collection of restrictions of functions from $H^{\frac{1}{2}}(\partial T)$ vanishing along the two remaining edges,

$$H_{00}^{\frac{1}{2}}(e) = \{u|_e \,:\, u \in H^{\frac{1}{2}}(\partial T),\ u = 0 \text{ on } \partial T - e\}\,.$$

Equivalently, u is in $H_{00}^{\frac{1}{2}}(e)$ if its zero extension $\tilde{u}$ is in $H^{\frac{1}{2}}(\partial T)$. The seminorm and the corresponding product in $H^{\frac{1}{2}}(\partial T)$ define the norm and the scalar product in $H_{00}^{\frac{1}{2}}(e)$,

$$\begin{aligned}\|u\|_{H_{00}^{\frac{1}{2}}(e)} &= |\tilde{u}|_{H^{\frac{1}{2}}(\partial T)},\\ (u,v)_{H_{00}^{\frac{1}{2}}(e)} &= (\tilde{u},\tilde{v})_{H^{\frac{1}{2}}(\partial T)}.\end{aligned}$$

Sobolev spaces of fractional order. We shall need a number of technical but standard facts about fractional order spaces; see, e.g., [12, 1, 20, 2, 19, 5].

We may use Fourier transform to introduce spaces $H^r(\mathbb{R}^2)$, $r \geq 0$. Functions from $H^r(T)$, $r \geq 0$, can be identified with restrictions of functions from $H^r(\mathbb{R}^2)$. Elements of $H^r(\partial T)$, $r > 0$, can be identified with traces of functions from $H^{r+\frac{1}{2}}(T)$. Their duals define fractional boundary spaces with negative exponents $H^r(\partial T)$, $r < 0$. For $-1 < r < 1$, the boundary spaces can be defined directly using local maps, and spaces $H^r(\mathbb{R})$ can be defined through Fourier transform, and the two definitions are equivalent.

For an edge e, spaces $H^r(e)$, $r \geq 0$, can be constructed by considering restrictions from $H^r(\partial T)$. They are naturally isomorphic with space $H^r(I)$ for unit interval $I = (0,1)$, which can be obtained via restrictions of functions from $H^r(\mathbb{R})$. For $\frac{1}{2} < r < 1$, boundary spaces $H^r(\partial T)$ can be conveniently characterized:

$$H^r(\partial T) = \{u \,:\, u|_e \in H^r(e), \text{ for each edge } e, \text{ and } u \text{ is continuous at vertices}\}$$

(Notice that, for $r > \frac{1}{2}$, $H^r(e)$ is embedded in $C(\bar{e})$; see Lemma 1.)

Finally, space $H^{\frac{1}{2}}_{00}(I)$, isomorphic with $H^{\frac{1}{2}}_{00}(e)$, can be equipped with an explicit, equivalent norm,

$$\|u\|^2_{H^{\frac{1}{2}}_{00}(I)} = \|u\|^2_{H^{\frac{1}{2}}(I)} + \int_0^1 \frac{u^2}{x}\,dx + \int_0^1 \frac{u^2}{1-x}\,dx. \tag{2.3}$$

We shall need a couple of nonstandard results involving the fractional spaces.

LEMMA 1. *Let $r > \frac{1}{2}$. Subspace of $H^r(I)$, consisting of functions vanishing at the endpoints,*

$$\{u \in H^r(0,1) \,:\, u(0) = u(1) = 0\}\,,$$

is continuously embedded in $H^{\frac{1}{2}}_{00}(I)$.

Proof. Due to the existence of a continuous extension operator from $H^r(I)$ into $H^r(\mathbb{R})$, $r \geq 0$ [20], it is sufficient to consider functions from $H^r(\mathbb{R})$.

Recall first that space $H^{\frac{1}{2}+\epsilon}(\mathbb{R})$ is continuously embedded in the space of Hölder continuous functions $C^\epsilon(\mathbb{R})$; see, e.g., [1, Thm. 7.57].

We now use the explicit norm (2.3). It is sufficient to show that the last two terms can be bounded by the norm in $H^{\frac{1}{2}+\epsilon}(I)$. But this follows immediately from the embedding into the Hölder continuous functions; e.g., for the first term we have

$$\int_0^1 \frac{u^2}{x}\,dx \leq C\|u\|^2_{H^{\frac{1}{2}+\epsilon}(I)} \int_0^1 \frac{x^{2\epsilon}}{x}\,dx \leq C\|u\|^2_{H^{\frac{1}{2}+\epsilon}(I)}\,. \qquad \square$$

LEMMA 2. *With norms* (2.1) *and* (2.2), *the tangential derivative defines an isometry from $H^{\frac{1}{2}}(\partial T)/\mathbb{R}$ onto $H_0^{-\frac{1}{2}}(\partial T)$,*

$$\frac{\partial}{\partial s} \,:\, H^{\frac{1}{2}}(\partial T)/\mathbb{R} \ni u \to \frac{\partial u}{\partial s} \in H_0^{-\frac{1}{2}}(\partial T)\,.$$

The construction extends to an isomorphism between fractional spaces,

$$\frac{\partial}{\partial s} : H^{\frac{1}{2}+r}(\partial T)/\mathbb{R} \ni u \to \frac{\partial u}{\partial s} \in H^{-\frac{1}{2}+r}(\partial T)$$

for $0 < r < \frac{1}{2}$.

Proof. Let u_1 be the harmonic lift of u, and let u_2 be the harmonic function such that $\frac{\partial u_2}{\partial n} = \frac{\partial u}{\partial s}$. The first result follows from the fact that u_1 and u_2 are conjugate harmonic functions and, consequently,

$$\|\boldsymbol{\nabla} u_1\| = \|\boldsymbol{\nabla} u_2\| .$$

The second result follows by the standard interpolation argument applied to the first case and an obvious case with integer order spaces,

$$H^1(\partial T)/\mathbb{R} \ni u \to \frac{\partial u}{\partial s} \in L^2_0(T) . \qquad \square$$

3. Discrete Friedrichs inequality. The classical Friedrichs inequalities (see, e.g., [2] and the literature therein) formulated for Dirichlet and Neumann boundary conditions state that there exist positive constants C_1, C_2 such that

(3.1)
$$\|\boldsymbol{E}\| \le C_1 \|\mathrm{curl}\boldsymbol{E}\| \quad \text{for every } \boldsymbol{E} \in \boldsymbol{H}_0(\mathrm{curl}, T) \quad \text{such that } (\boldsymbol{E}, \boldsymbol{\nabla}\phi) = 0\, \forall \phi \in H^1_0(T),$$
$$\|\boldsymbol{E}\| \le C_2 \|\mathrm{curl}\boldsymbol{E}\| \quad \text{for every } \boldsymbol{E} \in \boldsymbol{H}(\mathrm{curl}, T) \quad \text{such that } (\boldsymbol{E}, \boldsymbol{\nabla}\phi) = 0\, \forall \phi \in H^1(T).$$

Here $H^1_0(T)$ stands for the subspace of all functions in $H^1(T)$, vanishing on boundary ∂T, and $\boldsymbol{H}_0(\mathrm{curl}, T)$ is the subspace of all fields in $\boldsymbol{H}(\mathrm{curl}, T)$, with the tangential component vanishing on boundary ∂T.

Both inequalities trivially extend to polynomial spaces,

(3.2)
$$\|\boldsymbol{E}\| \le C_1 \|\mathrm{curl}\boldsymbol{E}\| \quad \text{for every } \boldsymbol{E} \in \boldsymbol{P}^p_{-1} \quad \text{such that } (\boldsymbol{E}, \boldsymbol{\nabla}\phi) = 0\, \forall \phi \in \mathcal{P}^{p+1}_{-1},$$
$$\|\boldsymbol{E}\| \le C_2 \|\mathrm{curl}\boldsymbol{E}\| \quad \text{for every } \boldsymbol{E} \in \boldsymbol{P}^p \quad \text{such that } (\boldsymbol{E}, \boldsymbol{\nabla}\phi) = 0\, \forall \phi \in \mathcal{P}^{p+1}.$$

Indeed, the right-hand sides in (3.2) define equivalent norms on the involved subspace of discrete divergence-free polynomials. The question is, How do the constants C_1, C_2 depend upon order p ?

THEOREM 1 (discrete Friedrichs inequalities). *There exist* C_1, C_2 *in* (3.2) *that are independent of* p.

Proof. The first two steps are identical for both cases. We shall present the Neumann case with an obvious generalization to the Dirichlet case.

Step 1. In order to prove $(3.2)_2$, it is sufficient to show that

$$\min_{\phi \in \mathcal{P}^{p+1}} \|\boldsymbol{E} - \boldsymbol{\nabla}\phi\| \le C \|\mathrm{curl}\boldsymbol{E}\| \quad \forall \boldsymbol{E} \in \boldsymbol{P}^p . \tag{3.3}$$

Indeed, if $(\boldsymbol{E}, \boldsymbol{\nabla}\phi) = 0$, for all $\phi \in \mathcal{P}^{p+1}$, then

$$\min_{\phi \in \mathcal{P}^{p+1}} \|\boldsymbol{E} - \boldsymbol{\nabla}\phi\| = \|\boldsymbol{E}\| .$$

Step 2. In order to prove (3.3), it is sufficient to construct a *continuous* left inverse of the curl operator,

$$A \,:\, \mathcal{P}^{p-1} \ni \psi \to A\psi = \boldsymbol{E} \in \mathcal{P}^p, \quad \mathrm{curl}(A\psi) = \psi \,,$$

with the continuity constant $\|A\|$ *independent* of p. Indeed,

$$\min_{\phi\in\mathcal{P}^{p+1}} \|\boldsymbol{E}-\boldsymbol{\nabla}\phi\| \le \|\boldsymbol{E}-(\boldsymbol{E}-A(\mathrm{curl}\boldsymbol{E})\| = \|A(\mathrm{curl}\boldsymbol{E})\| \le C\|\mathrm{curl}\boldsymbol{E}\|\,.$$

Notice that $\mathrm{curl}(\boldsymbol{E}-A(\mathrm{curl}\boldsymbol{E})) = \mathrm{curl}\boldsymbol{E}-\mathrm{curl}\boldsymbol{E}=0$ and, therefore, $\boldsymbol{E}-A(\mathrm{curl}\boldsymbol{E})$ is the gradient of some polynomial of order $p+1$.

Step 3. The first two steps are identical for both the Dirichlet $(3.2)_1$ and Neumann $(3.2)_2$ versions of the discrete Friedrichs inequality. In the final step, we have to consider the two cases separately.

Case of "Neumann boundary conditions." Use Poincaré's map for operator A,

$$E_1(\boldsymbol{x}) = -\quad x_2\int_0^1 t\psi(t\boldsymbol{x})\,dt,$$

$$E_2(\boldsymbol{x}) = \quad x_1\int_0^1 t\psi(t\boldsymbol{x})\,dt.$$

The map is well defined as both E_1, E_2 are polynomials of order p, provided ψ is a polynomial of order $p-1$. *Right inverse:*

$$\begin{aligned}
\mathrm{curl}\boldsymbol{E} &= \frac{\partial E_2}{\partial x_1}-\frac{\partial E_1}{\partial x_2}\\
&= \int_0^1 t\psi(t\boldsymbol{x})\,dt + x_1\int_0^1 t\frac{\partial\psi}{\partial y_1}t\,dt + \int_0^1 t\psi(t\boldsymbol{x})\,dt + x_2\int_0^1 t\frac{\partial\psi}{\partial y_2}t\,dt\\
&= 2\int_0^1 t\psi(t\boldsymbol{x})\,dt + \int_0^1 t^2\frac{d}{dt}(\psi(t\boldsymbol{x}))\,dt\\
&= 2\int_0^1 t\psi(t\boldsymbol{x})\,dt - \int_0^1 2t\psi(t\boldsymbol{x})\,dt + t^2\psi(t\boldsymbol{x})|_0^1\\
&= \psi(\boldsymbol{x}).
\end{aligned}$$

Continuity of the operator follows from the continuity of the Poincaré map at the continuous level. It is sufficient to demonstrate it, e.g., for the right triangle,

$$T = \{(x_1,x_2)\,:\,0<x_1<1,\ 0<x_2<1-x_1\}\,.$$

We have

$$\begin{aligned}
\int_T (E_1^2+E_2^2)\,dT &= \int_T \underbrace{(x_1^2+x_2^2)}_{\le 1}\left(\int_0^1 t\psi(t\boldsymbol{x})\,dt\right)^2 dT\\
&\le \int_T\left(\int_0^1 t^2\psi^2(t\boldsymbol{x})\,dt\right)\,dT\\
&= \int_0^1 t^2\left(\int_T \psi^2(t\boldsymbol{x})\,dT\right)\,dt \qquad (\text{ use substitution } t\boldsymbol{x}=\boldsymbol{x}')\\
&= \int_0^1 \left(\int_{T'}\psi^2(\boldsymbol{x}')\right)dT'\\
&\le \int_T \psi^2(\boldsymbol{x})\,dT \qquad (\text{ image } T' \text{ is contained in } T\ ).
\end{aligned}$$

Case of "Dirichlet boundary conditions." Assume $p > 2$. Let $\psi \in \mathcal{P}^{p-1}$. We decompose ψ into a constant c and ψ_0 with zero average, $\int_T \psi_0 \, dT = 0$. Let $\boldsymbol{E}_0$ be a unique polynomial $\boldsymbol{E}_0 \in \boldsymbol{P}^2_{-1}$ such that $\operatorname{curl}\boldsymbol{E}_0 = 1$, and $(\boldsymbol{E}_0, \boldsymbol{\nabla}\phi) = 0$ for all $\phi \in \mathcal{P}^3_{-1}$. We define the map A as follows:

$$A\psi = A(c + \psi_0) = c\boldsymbol{E}_0 + A\psi_0 \, .$$

In order to define $A\psi_0$, we first evaluate the Poincaré map at ψ_0 to obtain a field $\boldsymbol{E}$ and learn that tangential component $\boldsymbol{n} \times \boldsymbol{E}$ vanishes along the horizontal and vertical edges and has a zero average along the inclined edge e_2. Consequently, there exists a polynomial ϕ of order p in space $\mathcal{P}^p_{-1}(e_2)$ such that $\frac{\partial \phi}{\partial s} = \boldsymbol{n} \times \boldsymbol{E}$. By the polynomial extension theorem of Babuška and Suri [3] and Babuška et al. [4], ψ admits a polynomial extension $\tilde{\phi}$, defined on T such that

$$\begin{aligned}(3.4)\quad \|\boldsymbol{\nabla}\tilde{\phi}\| \le C\|\phi\|_{H^{\frac{1}{2}}_{00}(e_2)} = C\left\|\frac{\partial \phi}{\partial s}\right\|_{H^{-\frac{1}{2}}(\partial T)} = C\|\boldsymbol{n} \times \boldsymbol{E}\|_{H^{-\frac{1}{2}}(\partial T)} &\le C\|\boldsymbol{E}\|_{H(\mathrm{curl},T)} \\ &\le C\|\psi_0\|.\end{aligned}$$

In the estimate above we have used Lemma 2 and the continuity of the Poincaré map. Finally, we define

$$A\psi_0 = \boldsymbol{E} - \boldsymbol{\nabla}\tilde{\phi} \, .$$

Adding the gradient does not change the curl of $A\psi_0$, and the L^2 boundedness follows from (3.4). □

Remark 1. The relevance of the Poincaré map in the construction of Nédélec's edge elements of the first type was first noticed by Hiptmair [11].

4. $\boldsymbol{H^1}$-conforming interpolation. Given a function $u \in H^{1+\epsilon}(T)$, we define the corresponding interpolant $\Pi u := u^p \in \mathcal{P}^p_{p_e}(T)$ as the sum of three components,

$$(4.1)\qquad u^p = u_1 + \underbrace{\sum_{e=1}^{3} u^p_{2,e}}_{u^p_2} + u^p_3 \, .$$

Step 1: *Linear interpolant.* We construct $u_1 \in \mathcal{P}^1(T)$ by the standard, linear interpolation at the vertices,

$$u_1 \in \mathcal{P}^1(T), \quad u_1(v) = u(v) \quad \text{for every vertex } v \, .$$

Step 2: *Edge interpolants.* For each edge e, we project trace of difference $u - u_1$ onto space $\mathcal{P}^{p_e}_{-1}(e)$ of polynomials of order p_e, vanishing at the edge endpoints, using the $H^{\frac{1}{2}}_{00}(e)$-norm:

$$\left\{\begin{array}{l} u_{2,e} \in \mathcal{P}^{p_e}_{-1}(e), \\ \|u_{2,e} - (u - u_1)|_e\|_{H^{\frac{1}{2}}_{00}(e)} \to \min \, . \end{array}\right.$$

Next, we extend $u_{2,e}$ to a polynomial $u^p_{2,e}$ from $\mathcal{P}^p_{p_e}(T)$, vanishing along the two remaining edges. The sum of edge interpolants $u^p_{2,e}$ will be denoted u^p_2.

Step 3: *Interior interpolant.* We project difference $u - u_1 - u_2^p$ onto space $\mathcal{P}^p_{-1}(T)$ of polynomials of order p, vanishing on ∂T, in the H^1-seminorm:

$$\begin{cases} u_3^p \in \mathcal{P}^p_{-1}(T), \\ |u_3^p - (u - u_1 - u_2^p)|_{H^1(T)} \to \min . \end{cases}$$

Remark 2. Notice that even though extensions $u^p_{2,e}$ are not uniquely defined, the final interpolant is unique. To ensure the uniqueness of the extensions, we could use, e.g., discrete harmonic extensions, i.e., request for the orthogonality

$$(\boldsymbol{\nabla} u^p_{2,e}, \boldsymbol{\nabla} v) = 0 \quad \forall v \in \mathcal{P}^p_{-1}(T) .$$

PROPOSITION 1. *Operator* $\Pi : H^{1+\epsilon}(T) \to H^1(T)$ *is well defined and bounded, with the norm independent of orders* p, p_e.

Proof. Step 1. It follows from the continuous embedding of $H^{1+\epsilon}(T)$ into $C(\bar{T})$ and equivalence of norms on a finite dimensional space that

$$\|u_1\|_{H^1(T)} \leq \|u_1\|_{H^{1+\epsilon}(T)} \leq C \left(\sum_{\boldsymbol{a}} |u(\boldsymbol{a})|^2 \right)^{\frac{1}{2}} \leq C \|u\|_{H^{1+\epsilon}(T)} ,$$

where $\boldsymbol{a}$ denote vertices of the triangle.

Step 2. By Lemma 1, restriction $(u - u_1)|_e$ to edge e is in $H^{\frac{1}{2}}_{00}(e)$, so the edge minimization problem is well defined and

$$\begin{aligned} \|u_{2,e}\|_{H^{\frac{1}{2}}_{00}(e)} \leq \|(u-u_1)|_e\|_{H^{\frac{1}{2}}_{00}(e)} \leq C \|(u-u_1)|_e\|_{H^{\frac{1}{2}+\epsilon}(e)} &\leq C \|u - u_1\|_{H^{1+\epsilon}(T)} \\ &\leq C \|u\|_{H^{1+\epsilon}(T)} . \end{aligned}$$

By the polynomial extension theorem of Babuška and Suri [3] and Babuška et al. [4], there exists an extension $u^{p_e}_{2,e} \in \mathcal{P}^{p_e}(T)$, vanishing along the two remaining edges, such that

$$\|u^{p_e}_{2,e}\|_{H^1(T)} \leq C \|u^p_{2,e}|_e\|_{H^{\frac{1}{2}}_{00}(e)} ,$$

with constant C *independent* of p_e.

Step 3. By Poincaré's inequality and the results of the first two steps,

$$\begin{aligned} \|u_3^p\|_{H^1(T)} &\leq C |u_3^p|_{H^1(T)} \leq C |u - (u_1 + u_2^p)|_{H^1(T)} \\ &\leq C (\|u\|_{H^1(T)} + \|u_1\|_{H^1(T)} + \|u_2^p\|_{H^1(T)}) \leq C \|u\|_{H^1(T)} . \end{aligned}$$

Applying the triangle inequality to the definition of the interpolant finishes the proof. □

THEOREM 2 (H^1-conforming interpolation error estimate). *There exists a constant* C, *dependent upon* ϵ *but independent of* p, p_e, *such that*

$$\|u - \Pi u\|_{H^1(T)} \leq C \inf_{v \in \mathcal{P}^p_{p_e}} \|u - v\|_{H^{1+\epsilon}(T)} \leq C p_{min}^{-(r-\epsilon)} \|u\|_{H^{1+r}(T)}$$

for every $r > 1$ *and* $0 < \epsilon < r$. *Here* $p_{min} = \min_e p_e$.

Proof. It is sufficient to consider "small" $\epsilon, 0 < \epsilon < \min\{r, \frac{1}{2}\}$. As the interpolation preserves polynomials ψ from $\mathcal{P}^p_{p_e}(T)$, we have

$$\|u - \Pi u\|_{H^1(T)} = \|u - \psi - \Pi(u - \psi)\|_{H^1(T)} \leq (1 + \|\Pi\|) \|u - \psi\|_{H^{1+\epsilon}(T)} ,$$

where $\|\Pi\|$ is the norm of the interpolation operator in space $\mathcal{L}(H^{1+\epsilon}(T), H^1(T))$ and ψ is an arbitrary polynomial from $\mathcal{P}^p_{p_e}(T)$. Consequently, it follows from the best approximation results for polynomial spaces (see, e.g., [4]) that

$$\|u - \Pi u\|_{H^1(T)} \leq C \inf_{\psi \in \mathcal{P}^{p_{min}}(T)} \|u - \psi\|_{H^{1+\epsilon}(T)} \leq C p_{min}^{-(r-\epsilon)} \|u\|_{H^r(T)} \,. \qquad \square$$

Remark 3.

1. The described interpolation procedure is a generalization of the *hp*-interpolation proposed in [16] and used in [9]. The original *hp*-interpolation uses stronger $H^1_0(e)$-norms along the edges and requires more regularity ($u \in H^{3/2+\epsilon}(T)$). Consequently, it does not yield optimal (up to ϵ) convergence rates.
2. According to Lemma 2, projection in the $H^{\frac{1}{2}}_{00}(e)$-norm over the edges can be reinterpreted as the projection in the $H^{-\frac{1}{2}}(\partial T)$-norm of the (tangential) derivatives,

$$\|u_{2,e} - w\|_{H^{\frac{1}{2}}_{00}(e)} = \|\tilde{u}_{2,e} - \tilde{w}\|_{H^{\frac{1}{2}}(\partial T)} = \left\| \frac{\partial}{\partial s}(\tilde{u}_{2,e} - \tilde{w}) \right\|_{H^{-\frac{1}{2}}(\partial T)} \,.$$

   Here $w = (u - u_1)|_e$, and $\tilde{u}_{2,e}, \tilde{w}$ denote zero extensions of $u_{2,e}, w$ to the whole boundary ∂T.

5. $\boldsymbol{H}$(curl)-conforming interpolation.

Given a function $\boldsymbol{E} \in \boldsymbol{H}^\epsilon \cap \boldsymbol{H}(\text{curl}, T)$, we construct interpolant $\Pi^{\text{curl}}\boldsymbol{E} := \boldsymbol{E}^p \in \boldsymbol{P}^p_{p_e}(T)$ again in three steps,

$$\boldsymbol{E}^p = \boldsymbol{E}_1 + \underbrace{\sum_{e=1}^{3} \boldsymbol{E}^p_{2,e}}_{\boldsymbol{E}^p_2} + \boldsymbol{E}^p_3 \,. \tag{5.1}$$

Step 1: *Whitney's (lowest order) interpolant.* For each edge e, let $\boldsymbol{\phi}^e \in \boldsymbol{P}^1(T)$ denote the vector-valued, linear polynomial such that

$$\phi^e_t = \boldsymbol{n} \times \boldsymbol{\phi}^e = \begin{cases} 1 \text{ along edge } e, \\ 0 \text{ along the remaining edges.} \end{cases}$$

Here $\boldsymbol{n}$ is the outward normal unit vector to ∂T, and $\phi_t = \boldsymbol{n} \times \boldsymbol{\phi} = (-n_2)\phi_1 + n_1\phi_2$ denotes the trace of the tangential component of vector-valued function $\boldsymbol{\phi}$ to boundary ∂T. The Whitney interpolant is then defined as

$$\boldsymbol{E}_1 = \sum_e \left(\int_e E_t \right) \boldsymbol{\phi}^e \,.$$

Step 2: *Edge interpolants.* It follows from the construction of the Whitney interpolant that the trace of tangential component $\boldsymbol{n} \times (\boldsymbol{E} - \boldsymbol{E}_1)$ has zero average over each edge e. Thus, we can introduce a scalar-valued function ψ, defined on boundary ∂T, such that

$$\frac{\partial \psi}{\partial s} = \boldsymbol{n} \times (\boldsymbol{E} - \boldsymbol{E}_1), \quad \psi = 0 \text{ at vertices.}$$

For each edge e then, we project restriction $\psi|_e$ in the $H^{\frac{1}{2}}_{00}$-norm onto polynomials $\mathcal{P}^{p_e+1}_{-1}(e)$,

$$\begin{cases} \psi_{2,e} \in \mathcal{P}^{p_e+1}_{-1}(e), \\ \|\psi_{2,e} - \psi|_e\|_{H^{\frac{1}{2}}_{00}(e)} \to \min \, . \end{cases}$$

We take then any polynomial extension $\psi^{p+1}_{2,e} \in \mathcal{P}^{p+1}_{p_e+1}(T)$ that vanishes along the two remaining edges and define the edge interpolant by the gradient of the extension,

$$\boldsymbol{E}^p_{2,e} = \boldsymbol{\nabla}\psi^{p+1}_{2,e} \in \boldsymbol{P}^p_{p_e}(T) \, .$$

Step 3: *Interior interpolant.* We solve the constrained minimization problem,

$$\begin{cases} \boldsymbol{E}^p_3 \in \boldsymbol{P}^p_{-1}(T), \\ \|\mathrm{curl}(\boldsymbol{E}^p_3 - (\boldsymbol{E} - \boldsymbol{E}_1 - \boldsymbol{E}^p_2))\| = \|\mathrm{curl}(\boldsymbol{E}^p_3 - (\boldsymbol{E} - \boldsymbol{E}_1))\| \to \min, \\ (\boldsymbol{E}^p_3 - (\boldsymbol{E} - \boldsymbol{E}_1 - \boldsymbol{E}^p_2), \boldsymbol{\nabla}\phi) = 0 \quad \forall\phi \in \mathcal{P}^{p+1}_{-1}(T) \, . \end{cases}$$

Remark 4.

1. Notice again that the final interpolant is uniquely defined despite the possibility of many extensions $\psi^{p+1}_{2,e}$. In fact, we can even use extensions $\boldsymbol{E}^p_{2,e}$ of tangential trace $\frac{\partial\psi 2,e}{\partial s}$ with nonzero curl.
2. The edge interpolants can again be defined directly using the projection in $H^{-\frac{1}{2}}(\partial T)$-norm; compare Remark 3.

Proposition 2. *Operator* $\Pi^{\mathrm{curl}} : \boldsymbol{H}^{\epsilon}(T) \cap \boldsymbol{H}(\mathrm{curl}, T) \to \boldsymbol{H}(\mathrm{curl}, T)$ *is well defined and bounded, with a norm independent of orders* p, p_e*.*

Proof. Step 1. Consider a test function $\phi \in H^{1-\epsilon}(T)$, $\phi = 1$ on edge e, $\phi = 0$ on the two remaining edges. It follows from the integration by parts formula

$$\int_T \mathrm{curl}\boldsymbol{E}\,\phi = \int_T \boldsymbol{E}(\boldsymbol{\nabla}\times\phi) \;+ \int_{\partial T} (\boldsymbol{n}\times\boldsymbol{E})\phi$$

$(\mathrm{curl}\boldsymbol{E} = \frac{\partial E_2}{\partial x_1} - \frac{\partial E_1}{\partial x_2},\ \boldsymbol{\nabla}\times\phi = (\frac{\partial\phi}{\partial x_2}, -\frac{\partial\phi}{\partial x_1})\ )$ that functional

$$\boldsymbol{H}^{\epsilon}(T) \cap \boldsymbol{H}(\mathrm{curl}, T) \ni \boldsymbol{E} \to \int_e E_t = \int_e \boldsymbol{n}\times\boldsymbol{E}$$

is well defined and continuous. By the finite dimensionality argument, we get

$$\|\boldsymbol{E}_1\|_{H(\mathrm{curl})} \le (\|\boldsymbol{E}_1\|^2_{H^\epsilon} + \|\mathrm{curl}\boldsymbol{E}_1\|^2)^{\frac{1}{2}} \le C(\|\boldsymbol{E}\|^2_{H^\epsilon} + \|\mathrm{curl}\boldsymbol{E}\|^2)^{\frac{1}{2}} \, .$$

In fact, one can show directly [8] that $\|\mathrm{curl}\boldsymbol{E}_1\| \le \|\mathrm{curl}\boldsymbol{E}\|$.

Step 2. From the result of Step 1 and the integration by parts,

$$\int_T \mathrm{curl}(\boldsymbol{E} - \boldsymbol{E}_1)\phi = \int_T (\boldsymbol{E} - \boldsymbol{E}_1)(\boldsymbol{\nabla}\times\phi) + \int_{\partial T} \boldsymbol{n}\times(\boldsymbol{E} - \boldsymbol{E}_1)\phi \, ,$$

it follows that $\boldsymbol{n}\times(\boldsymbol{E} - \boldsymbol{E}_1) \in H^{-\frac{1}{2}+\epsilon}(\partial T)$. From the construction of $\boldsymbol{E}_1$ it follows that $\boldsymbol{n}\times(\boldsymbol{E} - \boldsymbol{E}_1)$ has zero average,

$$\int_{\partial T} \boldsymbol{n}\times(\boldsymbol{E} - \boldsymbol{E}_1) = 0 \, .$$

Consequently, by Lemma 2, potential ψ is well defined and it lives in $H^{\frac{1}{2}+\epsilon}(\partial K)$. Repeating arguments from the proof of Proposition 1, we show that

$$\|\psi^p_{2,e}\|_{H^1(T)} \leq C\|\boldsymbol{n} \times (\boldsymbol{E} - \boldsymbol{E}_1)\|_{H^{-\frac{1}{2}+\epsilon}(\partial T)} \leq C(\|\boldsymbol{E}\|^2_{H^\epsilon(T)} + \|\mathrm{curl}\boldsymbol{E}\|^2)^{\frac{1}{2}} \ .$$

Consequently,

$$\|\boldsymbol{E}^p_2\| = \|\boldsymbol{E}^p_2\|_{H(\mathrm{curl})} \leq C(\|\boldsymbol{E}\|^2_{H^\epsilon(T)} + \|\mathrm{curl}\boldsymbol{E}\|^2)^{\frac{1}{2}} \ .$$

Step 3. We use the discrete Helmholtz decomposition,

$$\boldsymbol{E}^p_3 = \boldsymbol{E}^p_{3,0} + \boldsymbol{\nabla}\psi^{p+1} \ , \tag{5.2}$$

where $\psi \in \mathcal{P}^{p+1}_{-1}(T)$, $\boldsymbol{E}^p_{3,0} \in \boldsymbol{P}^p_{-1}$, $(\boldsymbol{E}^p_{3,0}, \boldsymbol{\nabla}\phi) = 0$ for all $\phi \in \mathcal{P}^{p+1}_{-1}(T)$.

We have

$$\|\mathrm{curl}\boldsymbol{E}^p_{3,0}\| = \|\mathrm{curl}\boldsymbol{E}^p_3\| \leq \|\mathrm{curl}(\boldsymbol{E} - \boldsymbol{E}_1)\| \leq 2\|\mathrm{curl}\boldsymbol{E}\| \ ,$$

and by the discrete Friedrichs inequality (3.2),

$$\|\boldsymbol{E}^p_{3,0}\| \leq C\|\mathrm{curl}\boldsymbol{E}^p_{3,0}\| \leq 2C\|\mathrm{curl}\boldsymbol{E}\| \ .$$

Finally, by the results of the first two steps,

$$\|\boldsymbol{\nabla}\psi^{p+1}\| \leq \|\boldsymbol{E} - \boldsymbol{E}_1 - \boldsymbol{E}^p_2\| \leq C(\|\boldsymbol{E}\|^2_{H^\epsilon} + \|\mathrm{curl}\boldsymbol{E}\|^2)^{\frac{1}{2}} \ .$$

Applying the triangle inequality to the definition of the interpolant finishes the proof. □

PROPOSITION 3. *de Rham diagram* (1.1) *commutes.*

Proof. Both horizontal sequences in the diagram are exactl; compare [9, 10].

Commutativity of the first part of the diagram is obvious; operator Π preserves constants.

Let now $\boldsymbol{E} = \boldsymbol{\nabla}\phi, \phi \in H^{1+\epsilon}(T)$. It follows from the definitions of the interpolation operators that the Whitney interpolant of $\boldsymbol{\nabla}\phi$ coincides with the gradient of linear interpolant ϕ_1 of ϕ, $\boldsymbol{E}_1 = \boldsymbol{\nabla}\phi_1$. Consequently, $\phi - \phi_1$ coincides with function ψ used to define $\boldsymbol{E}^p_2$, and extensions $\psi^{p+1}_{2,e}$ and $\phi^{p+1}_{2,e}$ may be taken to be the same. Finally, $\boldsymbol{E}^p_{3,0}$ in the Helmholtz decomposition of $\boldsymbol{E}^p_3$ is zero, and ψ^{p+1} in (5.2) coincides with ϕ^{p+1}_3.

Finally, we need to show that

$$(\mathrm{curl}(\boldsymbol{E}^p - \boldsymbol{E}), \mathrm{curl}\boldsymbol{F}) = (\mathrm{curl}(\boldsymbol{E}^p_1 + \boldsymbol{E}^p_3 - \boldsymbol{E}), \mathrm{curl}\boldsymbol{F}) = 0 \quad \forall \boldsymbol{F} \in \boldsymbol{P}^p_{p_e} \ .$$

It follows from our discussion (see also [8]) that any $\boldsymbol{F} \in \boldsymbol{P}^p_{p_e}$ can be decomposed as

$$\boldsymbol{F} = \boldsymbol{F}_1 + \boldsymbol{F}^p_2 + \boldsymbol{F}^p_3,$$

where $\boldsymbol{F}^p_2$ is a gradient. Orthogonality with $\boldsymbol{F}_1$ follows from the integration by parts and the definition of the Whitney interpolant,

$$(\mathrm{curl}(\boldsymbol{E}^p_1+\boldsymbol{E}^p_3-\boldsymbol{E}), \mathrm{curl}\boldsymbol{F}_1) = \int_{\partial T}\boldsymbol{n}\times(\boldsymbol{E}_1+\boldsymbol{E}^p_3-\boldsymbol{E})\mathrm{curl}\boldsymbol{F}_1 = \int_{\partial T}\boldsymbol{n}\times(\boldsymbol{E}_1-\boldsymbol{E})\mathrm{curl}\boldsymbol{F}_1 = 0.$$

Orthogonality with $\boldsymbol{F}^p_3$ is a consequence of the definition of $\boldsymbol{E}^p_3$. □

Best approximation result. Let $\boldsymbol{E} \in \boldsymbol{H}^r(T)$ and curl$\boldsymbol{E} \in H^r(T), 0 < r \leq 1$. Consider the Helmholtz decomposition,

$$\boldsymbol{E} = \boldsymbol{E}_0 + \boldsymbol{\nabla}\psi \quad (\boldsymbol{E}_0, \boldsymbol{\nabla}\phi) = 0 \quad \forall\phi \in H^1(T) .$$

Potential ψ is the solution of the Poisson equation,

$$\begin{cases} \psi \in H^1(T)/\mathbb{R}, \\ (\boldsymbol{\nabla}\psi, \boldsymbol{\nabla}\phi) = (\boldsymbol{E}, \boldsymbol{\nabla}\phi) \quad \forall\phi \in H^1(T) . \end{cases}$$

It follows from the regularity results of Costabel and Dauge [7] that

- $\boldsymbol{E}_0 \in \boldsymbol{H}^{1+r}(T)$ and

$$\|\boldsymbol{E}_0\|_{H^{1+r}} \leq C\|\text{curl}\boldsymbol{E}_0\|_{H^r} = \|\text{curl}\boldsymbol{E}\|_{H^r};$$

- $\psi \in H^{1+r}(T)$ and

$$\|\psi\|_{H^{1+r}} \leq C\|\boldsymbol{E}\|_{H^r} .$$

Consequently, there exists a vector-valued polynomial $\boldsymbol{E}_0^p \in \boldsymbol{P}^p(T)$ such that

$$\|\boldsymbol{E}_0 - \boldsymbol{E}_0^p\|_{H^1} \leq Cp^{-r}\|\text{curl}\boldsymbol{E}\|_{H^r}$$

and, consequently,

$$(\|\boldsymbol{E}_0 - \boldsymbol{E}_0^p\|_{H^\epsilon}^2 + \|\text{curl}(\boldsymbol{E}_0 - \boldsymbol{E}_0^p)\|^2)^{\frac{1}{2}} \leq Cp^{-r}\|\text{curl}\boldsymbol{E}\|_{H^r} .$$

Also, there exists a polynomial $\psi^{p+1} \in \mathcal{P}^{p+1}$ such that

$$\|\boldsymbol{\nabla}\psi^{p+1} - \boldsymbol{\nabla}\psi\|_{H^\epsilon} \leq Cp^{-(r-\epsilon)}\|\psi\|_{H^{1+r}(T)} .$$

Summing up $\boldsymbol{E}_0^p$ and $\boldsymbol{\nabla}\psi^{p+1}$,

$$\boldsymbol{E}^p = \boldsymbol{E}_0^p + \boldsymbol{\nabla}\psi^{p+1} ,$$

and using the triangle inequality we get

$$(\|\boldsymbol{E} - \boldsymbol{E}^p\|_{H^\epsilon}^2 + \|\text{curl}(\boldsymbol{E} - \boldsymbol{E}^p)\|^2)^{\frac{1}{2}} \leq Cp^{-(r-\epsilon)}(\|\boldsymbol{E}\|_{H^r}^2 + \|\text{curl}\boldsymbol{E}\|_{H^r}^2)^{\frac{1}{2}} .$$

Theorem 3 ($\boldsymbol{H}$(curl)-conforming interpolation error estimate). *There exists $C > 0$, dependent upon ϵ but independent of p, p_e, such that*

$$\begin{aligned} \|\boldsymbol{E} - \Pi^{\text{curl}}\boldsymbol{E}\|_{H(\text{curl},T)} &\leq C \inf_{\boldsymbol{F}\in\boldsymbol{P}_{p_e}^p} (\|\boldsymbol{E} - \boldsymbol{F}\|_{H^\epsilon}^2 + \|\text{curl}(\boldsymbol{E} - \boldsymbol{F})\|^2)^{\frac{1}{2}} \\ &\leq Cp_{min}^{-(r-\epsilon)}(\|\boldsymbol{E}\|_{H^r}^2 + \|\text{curl}\boldsymbol{E}\|_{H^r}^2)^{\frac{1}{2}} \end{aligned} \tag{5.3}$$

for every $0 < r < 1$ and $0 < \epsilon < r$. Here $p_{min} = \min_e p_e$.

Proof. Combining Proposition 2 with the best approximation error estimate, we have

$$\begin{aligned} \|\boldsymbol{E} - \Pi^{\text{curl}}\boldsymbol{E}\|_{H(\text{curl},T)} &= \|\boldsymbol{E} - \boldsymbol{F} - \Pi^{\text{curl}}(\boldsymbol{E} - \boldsymbol{F})\|_{H(\text{curl},T)} \qquad (\forall\boldsymbol{F} \in \boldsymbol{P}_{p_e}^p) \\ &\leq (1 + \|\Pi^{\text{curl}}\|) \inf_{\boldsymbol{F}\in\boldsymbol{P}_{p_e}^p} (\|\boldsymbol{E} - \boldsymbol{F}\|_{H^\epsilon}^2 + \|\text{curl}(\boldsymbol{E} - \boldsymbol{F})\|^2)^{\frac{1}{2}} \\ &C \leq p^{-(r-\epsilon)}(\|\boldsymbol{E}\|_{H^r}^2 + \|\text{curl}\boldsymbol{E}\|_{H^r}^2)^{\frac{1}{2}} . \qquad \square \end{aligned}$$

6. Final remarks.

First family of Nédélec's elements. Results concerning the $H(\text{curl})$-conforming interpolation extend automatically to the diagram corresponding to the first family of Nédélec's family of triangular elements [14]:

$$\begin{tikzcd}
H^{1+\epsilon} \arrow[r, "\boldsymbol{\nabla}"] \arrow[d, "\Pi"] & \boldsymbol{H}^{\epsilon} \cap \boldsymbol{H}(\text{curl}) \arrow[r, "\boldsymbol{\nabla}\times"] \arrow[d, "\Pi^{\text{curl}}"] & L^2, \arrow[d, "P,"] \\
\mathcal{P}^{p+1}_{p_e+1} \arrow[r, "\boldsymbol{\nabla}"] & \mathbf{P}^{p}_{p_e} \oplus \tilde{\mathbf{P}}^{p+1}_{p_e} \arrow[r, "\boldsymbol{\nabla}\times"] & \mathcal{P}^{p}\,.
\end{tikzcd} \tag{6.1}$$

Here $\tilde{\boldsymbol{P}}^{p+1}_{p_e+1}$ corresponds to the direct sum decompositions,

$$\begin{aligned}
\mathcal{P}^{p+2}_{p_e+1} &= \mathcal{P}^{p+1}_{p_e+1} \oplus \tilde{\mathcal{P}}^{p+1}_{p_e}, \\
\boldsymbol{P}^{p+1}_{p_e} &= \boldsymbol{P}^{p}_{p_e} \oplus \boldsymbol{\nabla}\tilde{\mathcal{P}}^{p+2}_{p_e+1} \oplus \tilde{\boldsymbol{P}}^{p+1}_{p_e}\,;
\end{aligned}$$

see [10] for a more detailed discussion. Contrary to the original diagram, order p may now be equal to zero, except for $p = 0$ (when $\tilde{\mathcal{P}}^2_1$ is trivial, space $\boldsymbol{P}^{p}_{p_e} \oplus \tilde{\boldsymbol{P}}^{p+1}_{p_e}$ is not uniquely defined and depends upon the choice of algebraic component $\tilde{\boldsymbol{P}}^{p+1}_{p_e}$, unless one requests additionally for orthogonality of the components [10]).

The advantage of using the first Nédélec family is that both $\boldsymbol{E}$ and curl $\boldsymbol{E}$ are now interpolated with polynomials of the same order.

Extensions. An extension to the case of square elements in two dimensions seems to be straightforward. As the interpolation takes place on the master element, the results extend automatically to the case of curvilinear, parametric elements [9]. The results extend also to three-dimensional tetrahedral elements with a minimum regularity of $H^{\frac{3}{2}+\epsilon}(T)$ for the H^1-conforming interpolation. The limitation comes from the definition of the linear, vertex interpolant. It looks like the analysis for three-dimensional elements (nD differential forms in general [11]) will require nonlocal interpolation operators. The question of how to define them so that the de Rham diagram will commute is open.

REFERENCES

[1] R. A. ADAMS, *Sobolev Spaces*, Academic Press, New York, 1978.

[2] C. AMROUCHE, C. BERNARDI, M. DAUGE, AND V. GIRAULT, *Vector potentials in three-dimensional non-smooth domains*, Math. Methods Appl. Sci., 21 (1998), pp. 823–864.

[3] I. BABUŠKA AND M. SURI, *The optimal convergence rate of the p-version of the finite element method*, SIAM J. Numer. Anal., 24 (1987), pp. 750–776.

[4] I. BABUŠKA, A. CRAIG, J. MANDEL, AND J. PITKÄRANTA, *Efficient preconditioning for the p-version finite element method in two dimensions*, SIAM J. Numer. Anal., 28 (1991), pp. 624–661.

[5] J. BERGH AND J. LÖFSTÖM, *Interpolation Spaces*, Springer-Verlag, Berlin, 1976.

[6] D. BOFFI, *A note on the discrete compactness property and the de Rham diagram*, Appl. Math. Lett., 14 (2001), pp. 33–38.

[7] M. COSTABEL AND M. DAUGE, *Singularities of electromagnetic fields in polyhedral domains*, Arch. Ration. Mech. Anal., 151 (2000), pp. 221–276.

[8] L. DEMKOWICZ, P. MONK, CH. SCHWAB, AND L. VARDAPETYAN, *Maxwell eigenvalues and discrete compactness in two dimensions*, Comput. Math. Appl., 40 (2000), pp. 598–605.

[9] L. Demkowicz, P. Monk, L. Vardapetyan, and W. Rachowicz, *de Rham diagram for hp finite element spaces*, Comput. Math. Appl., 39 (2000), pp. 29–38.

[10] L. Demkowicz, *Edge finite elements of variable order for Maxwell's equations*, in Scientific Computing in Electrical Engineering, Lecture Notes in Comput. Sci. Eng. 18, Springer-Verlag, Berlin, 2000.

[11] R. Hiptmair, *Higher Order Whitney Forms*, Report 156, Sonderforschungsbereich 382, Universität Tübingen, Tübingen, Germany, 2000.

[12] J. L. Lions and E. Magenes, *Non Homogeneous Boundary Value Problems and Applications*, Vol. 1, Springer-Verlag, Berlin, 1972.

[13] P. Monk, *On the p− and hp−extension of Nédélec's curl-conforming elements*, J. Comput. Appl. Math., 53 (1994), pp. 117–137.

[14] J. C. Nédélec, *Mixed finite elements in* $\mathbb{R}^3$, Numer. Math., 35 (1980), pp. 315–341.

[15] J. C. Nédélec, *A new family of mixed finite elements in* $\mathbb{R}^3$, Numer. Math., 50 (1986), pp. 57–81.

[16] J. T. Oden, L. Demkowicz, W. Rachowicz, and T. A. Westermann, *Toward a universal hp adaptive finite element strategy, Part* 2. *A posteriori error estimation*, Comput. Methods Appl. Mech. Engrg., 77 (1989), pp. 113–180.

[17] W. Rachowicz and L. Demkowicz, *An hp-adaptive finite element method for electromagnetics. Part* 1: *Data structure and constrained approximation*, Comput. Methods Appl. Mech. Engrg., 187 (2000), pp. 307–337.

[18] W. Rachowicz and L. Demkowicz, *A three-dimensional hp-adaptive finite element package for electromagnetics*, Internat. J. Numer. Methods Engrg., 53 (2002), pp. 147–180.

[19] Ch. Schwab, *p and hp-Finite Element Methods*, Clarendon Press, Oxford, UK, 1998.

[20] E. M. Stein, *Singular Integrals and Differentiability Properties of Functions*, Princeton University Press, Princeton, NJ, 1970.

[21] R. Stenberg and M. Suri, *An hp error analysis of MITC plate elements*, SIAM J. Numer. Anal., 34 (1997), pp. 544–568.

[22] M. Suri, *On the stability and convergence of higher-order mixed finite element methods for second-order elliptic problems*, Math. Comp., 54 (1990), pp. 1–19.

SIAM J. Numer. Anal.
Vol. 41, No. 4, pp. 1209–1231

RESTRICTED ADDITIVE SCHWARZ PRECONDITIONERS WITH HARMONIC OVERLAP FOR SYMMETRIC POSITIVE DEFINITE LINEAR SYSTEMS*

XIAO-CHUAN CAI[†], MAKSYMILIAN DRYJA[‡], AND MARCUS SARKIS[§]

Abstract. A restricted additive Schwarz (RAS) preconditioning technique was introduced recently for solving general nonsymmetric sparse linear systems. In this paper, we provide one-level and two-level extensions of RAS for symmetric positive definite problems using the so-called harmonic overlaps (RASHO). Both RAS and RASHO outperform their counterparts of the classical additive Schwarz variants (AS). The design of RASHO is based on a much deeper understanding of the behavior of Schwarz-type methods in overlapping subregions and in the construction of the overlap. In RASHO, the overlap is obtained by extending the nonoverlapping subdomains only in the directions that do not cut the boundaries of other subdomains, and all functions are made harmonic in the overlapping regions. As a result, the subdomain problems in RASHO are smaller than those of AS, and the communication cost is also smaller when implemented on distributed memory computers, since the right-hand sides of discrete harmonic systems are always zero and therefore do not need to be communicated. We also show numerically that RASHO-preconditioned CG takes fewer iterations than the corresponding AS-preconditioned CG. A nearly optimal theory is included for the convergence of RASHO-preconditioned CG for solving elliptic problems discretized with a finite element method.

Key words. restricted additive Schwarz preconditioner, two-level domain decomposition, harmonic overlap, elliptic equations, finite elements

AMS subject classifications. 65N30, 65F10

DOI. 10.1137/S0036142901389621

1. Introduction.

A restricted additive Schwarz (RAS) preconditioning technique was introduced recently for solving general nonsymmetric sparse linear systems [1, 5, 7, 14, 16, 17, 20]. RAS outperforms the classical additive Schwarz (AS) preconditioner [8, 24] in the sense that it requires fewer iterations, as well as lower communication and CPU time costs when implemented on distributed memory computers [1]. Unfortunately, RAS in its original form is nonsymmetric, and therefore the CG method cannot be used [15]. Although a symmetrized version was constructed in [7], our numerical experiments show that it often takes more iterations than the corresponding AS/CG. In this paper we propose another modification of RAS and show in both theory and numerical experiments that this new variant works well for symmetric positive definite sparse linear systems and is superior to AS. Recall that the basic building blocks of classical Schwarz-type algorithms are realized by solving

*Received by the editors May 21, 2001; accepted for publication (in revised form) October 24, 2002; published electronically August 22, 2003.
http://www.siam.org/journals/sinum/41-4/38962.html

[†]Department of Computer Science, University of Colorado, Boulder, CO 80309 (cai@cs.colorado.edu). The work of this author was supported in part by the NSF grants ACI-0072089, ACI-0083004, and ACI-0112930.

[‡]Faculty of Mathematics, Informatics and Mechanics, Warsaw University, Warsaw 00-097, Poland (dryja@mimuw.edu.pl). The work of this author was supported in part by the NSF grant CCR-9732208 and in part by the Polish Science Foundation grant 2 P03A 021 16.

[§]Mathematical Sciences Department, Worcester Polytechnic Institute, Worcester, MA 01609 (msarkis@wpi.edu) and Instituto de Matemática Pura e Aplicada, Estrada Dona Castorina 110, Jardim Botânico, Rio de Janeiro, RJ 22460-320, Brazil. The work of this author was supported in part by the NSF grant CCR-9984404.

the linear systems of the form

$$A_i^\delta w = R_i^\delta v \tag{1.1}$$

on each extended subdomain, where A_i^δ is the extended subdomain stiffness matrix and R_i^δ is the restriction operator for the extended subdomain. (Formal definitions will be given later in the paper.) The key idea of RAS is that (1.1) is replaced by

$$A_i^\delta w = \begin{cases} v & \text{inside the unextended subdomain,} \\ 0 & \text{in the overlapping part of the subdomain.} \end{cases} \tag{1.2}$$

Note that the solution of (1.2) is discrete harmonic in the overlapping part of the subdomain and therefore carries minimum energy in some sense. Setting part of the right-hand-side vector to zero reduces the energy of the solution and also destroys the symmetry of the additive Schwarz operator. In this paper, we further explore the idea of "harmonic overlap" and at the same time keep the symmetry of the Schwarz preconditioner. We mention that other approaches can also be taken to modifying the Schwarz algorithm in the overlapping regions, such as allowing the functions to be discontinuous [4].

The algorithm to be discussed below is applicable for general symmetric positive definite problems. However, in order to provide a complete mathematical analysis, we restrict our discussion to a finite element problem [3]. We consider a simple variational problem: Find $u \in H_0^1(\Omega)$ such that

$$a(u, v) = f(v) \qquad \forall\ v \in H_0^1(\Omega), \tag{1.3}$$

where

$$a(u, v) = \int_\Omega \nabla u \cdot \nabla v \, dx \quad \text{and} \quad f(v) = \int_\Omega f v \, dx \qquad \text{for } f \in L^2(\Omega).$$

For simplicity, let Ω be a bounded polygonal region in $\Re^2$ with a diameter of size $O(1)$. The extension of the results to $\Re^3$ can be carried out easily by using the theory developed here in this paper and the well-known three-dimensional AS techniques; see [9, 10, 12]. Let $\mathcal{T}^h(\Omega)$ be a shape-regular quasi-uniform triangulation of size $O(h)$ of Ω, and $\mathcal{V} \subset H_0^1(\Omega)$ the finite element space consisting of continuous piecewise linear functions associated with the triangulation. We are interested in solving the following discrete problem associated with (1.3): Find $u^* \in \mathcal{V}$ such that

$$a(u^*, v) = f(v) \qquad \forall\ v \in \mathcal{V}. \tag{1.4}$$

Using the standard basis functions, (1.4) can be rewritten as a linear system of equations

$$Au^* = f. \tag{1.5}$$

For simplicity, we understand u^* and f both as functions and vectors, depending on the situation.

The paper is organized as follows. In section 2, we introduce notation. The new algorithm is described in section 3. Section 4 is devoted to the mathematical analysis of the new algorithm. We conclude the paper in section 5 by providing some numerical results and final remarks. Throughout this paper, C is a positive generic constant that is independent of any of the mesh parameters and the number of subdomains. All the domains and subdomains are assumed to be open; i.e., boundaries are not included in their definitions.

2. Notation. Let n be the total number of interior nodes of $\mathcal{T}^h(\Omega)$, and W the set containing all the interior nodes. We assume that a node-based partitioning has been applied and has resulted in N nonoverlapping subsets $W_i^0, i = 1, \ldots, N$, whose union is W. For each W_i^0, we define a subregion Ω_i^R to be the union of all elements of $\mathcal{T}^h(\Omega)$ that have all three vertices in $W_i^0 \cup \partial\Omega$. Note that $\cup\bar{\Omega}_i^R$ is not equal to $\bar{\Omega}$; see Figure 2.1(b). We denote by H the representative size (diameter) of the subregion Ω_i^R.

We define the overlapping partition of W as follows. Let $\{W_i^1\}$ be the one-overlap partition of W, where $W_i^1 \supset W_i^0$ is obtained by including all the immediate neighboring vertices of all vertices in W_i^0; see Figure 2.1(c). Using the idea recursively, we can define a δ-overlap partition of W,

$$W = \bigcup_{i=1}^{N} W_i^\delta.$$

Here the integer δ indicates the level of overlap with its neighboring subdomains, and δh is approximately the length of the extension. The definition of W_i^δ, as well as many other subsets, can be found in an illustrative picture, Figure 2.1.

We next define a subregion of Ω induced by a subset of nodes of $\mathcal{T}^h(\Omega)$ as follows. Let Z be a subset of W. The induced subregion, denoted by $\Omega(Z)$, is defined as the union of (1) the set Z itself, (2) the union of all the open elements (triangles) of $\mathcal{T}^h(\Omega)$ that have at least one vertex in Z, and (3) the union of the open edges of these triangles that have at least one endpoint as a vertex of Z. Note that $\Omega(Z)$ is always an open region. The extended subregion Ω_i^δ is defined as $\Omega(W_i^\delta)$, and the corresponding subspace as

$$\mathcal{V}_i^\delta \equiv \mathcal{V} \cap H_0^1(\Omega_i^\delta) \quad \text{extended by zero to } \Omega\backslash\Omega_i^\delta.$$

It is easy to verify that

$$\mathcal{V} = \mathcal{V}_1^\delta + \mathcal{V}_2^\delta + \cdots + \mathcal{V}_N^\delta.$$

This decomposition is used in defining the classical one-level AS algorithm [8]. Note that for $\delta = 0$ this decomposition is a direct sum. Let us define $P_i^\delta : \mathcal{V} \to \mathcal{V}_i^\delta$ by the following: For any $u \in \mathcal{V}$,

$$a(P_i^\delta u, v) = a(u, v) \qquad \forall v \in \mathcal{V}_i^\delta. \tag{2.1}$$

Then, the classical one-level AS operator has the form

$$P^\delta = P_1^\delta + \cdots + P_N^\delta.$$

In the classical AS as defined above, all the nodes of W_i^δ are treated equally even through some subsets of the nodes play different roles in determining the convergence rate of the AS-preconditioned CG. To further understand the issue, we classify the nodes as follows. Let $\Gamma_i^\delta = \partial\Omega_i^\delta \backslash \partial\Omega$, i.e., the part of the boundary of Ω_i^δ that does not belong to the Dirichlet part of the physical boundary $\partial\Omega$. We define the interface-overlapping boundary Γ^δ as the union of all Γ_i^δ; i.e., $\Gamma^\delta = \cup_{i=1}^N \Gamma_i^\delta$. We also need to define the following subsets of W (see, for example, Figure 2.1, where $\delta = 1$):

- $W^{\Gamma^\delta} \equiv W \bigcap \Gamma^\delta$ (interface nodes),
- $W_i^{\Gamma^\delta} \equiv W^{\Gamma^\delta} \bigcap W_i^\delta$ (local interface nodes),

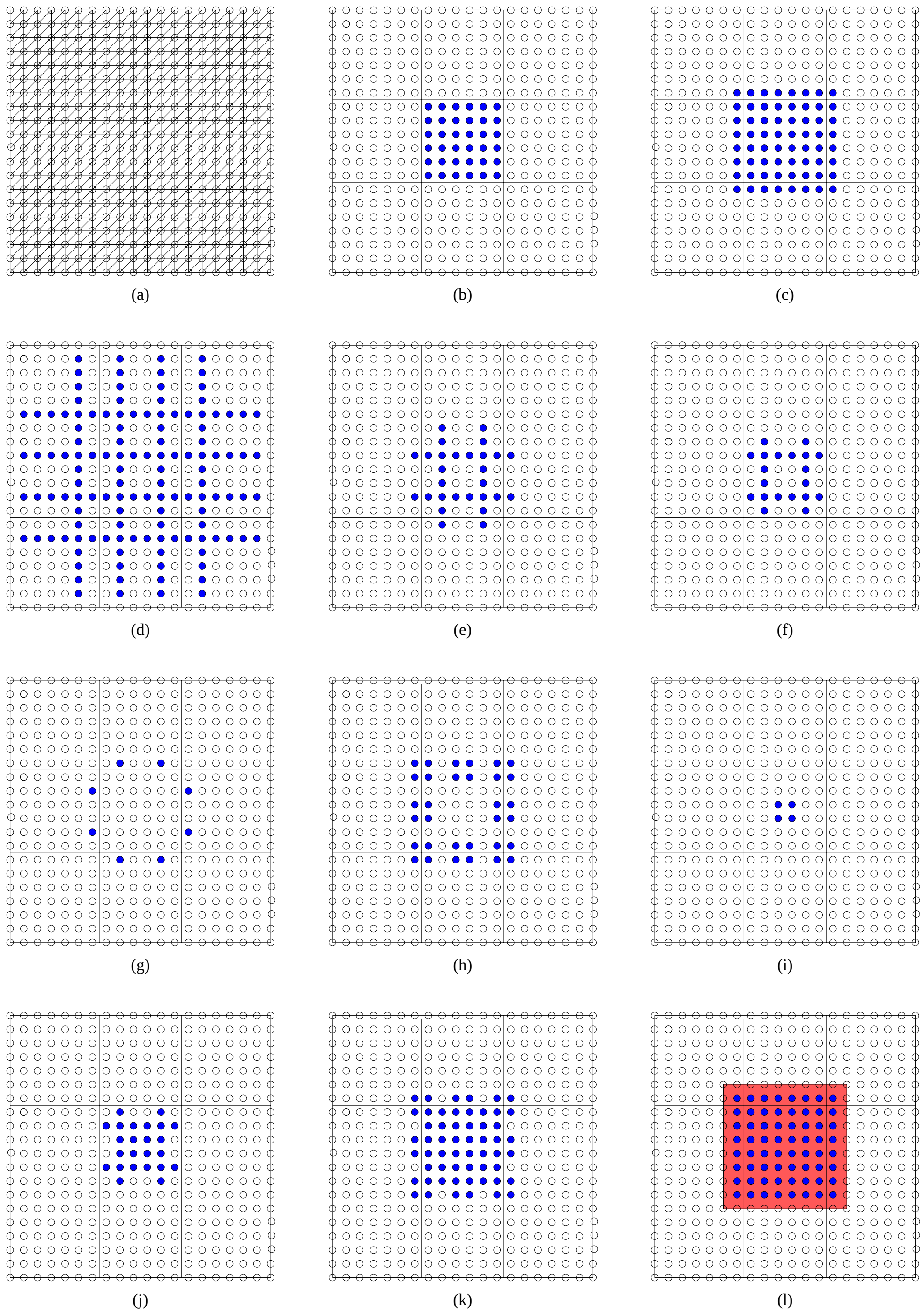


Fig. 2.1. *The partition of a finite element mesh into nine subdomains with the overlapping factor* $\delta = 1$. (a) *The finite element mesh and nodal points;* (b) *a node-based partition of the mesh into nine nonoverlapping subsets, and the collection of "●" forms the set* W_i^0*;* (c) W_i^δ*;* (d) W^{Γ^δ}*;* (e) $W_i^{\Gamma^\delta}$*;* (f) $W_{i,in}^{\Gamma^\delta}$*;* (g) $W_{i,cut}^{\Gamma^\delta}$*;* (h) $W_{i,ovl}^\delta$*;* (i) $W_{i,non}^\delta$*;* (j) $W_{i,in}^\delta$*;* (k) $\widetilde{W}_i^\delta$*;* (l) *the shadowed area is* Ω_i^δ.

- $W_{i,in}^{\Gamma^\delta} \equiv W^{\Gamma^\delta} \bigcap W_i^0$ (local internal interface nodes),
- $W_{i,cut}^{\Gamma^\delta} \equiv W_i^{\Gamma^\delta} \backslash W_{i,in}^{\Gamma^\delta}$ (local cut interface nodes),
- $W_{i,ovl}^{\delta} \equiv (W_i^\delta \backslash W_i^{\Gamma^\delta}) \bigcap (\bigcup_{j \neq i} W_j^\delta)$ (local overlapping nodes),
- $W_{i,non}^{\delta} \equiv W_i^\delta \backslash (W_i^{\Gamma^\delta} \bigcup W_{i,ovl}^\delta)$ (local nonoverlapping nodes),
- $W_{i,in}^{\delta} \equiv W_{i,non}^\delta \bigcup W_{i,in}^{\Gamma^\delta}$ (internal nodes).

We note that the most northwest and the southeast nodes in Figure 2.1(c) were added to Γ_i^δ in order to make Ω_i^δ a rectangle. This is just to simplify the presentation, and it is not required in the implementation of the algorithms.

We frequently use functions that are discrete harmonic at certain nodes. Let $x_k \in W$ be a mesh point and $\phi_{x_k}(x) \in \mathcal{V}$ the finite element basis function associated with x_k; i.e., $\phi_{x_k}(x_k) = 1$, and $\phi_{x_k}(x_j) = 0, j \neq k$. We say that $u \in \mathcal{V}$ is discrete harmonic at x_k if

$$a(u, \phi_{x_k}) = 0.$$

If u is discrete harmonic at a set of nodal points Z, we say that u is discrete harmonic in $\Omega(Z)$.

Our new algorithm will be built on the subspace $\widetilde{\mathcal{V}}_i^\delta$ defined as a subspace of $\mathcal{V}_i^\delta$. $\widetilde{\mathcal{V}}_i^\delta$ consists of all functions that vanish on the cutting nodes $W_{i,cut}^{\Gamma^\delta}$ and are discrete harmonic at the nodes of $W_{i,ovl}^\delta$. Note that the degrees of freedom associated with the subspace $\widetilde{\mathcal{V}}_i^\delta$ are

$$\widetilde{W_i^\delta} \equiv W_i^\delta \backslash W_{i,cut}^{\Gamma^\delta},$$

and, since the values at the harmonic nodes are not independent, they cannot be counted toward the degrees of freedom. The dimension of $\widetilde{\mathcal{V}}_i^\delta$ is

$$\dim(\widetilde{\mathcal{V}}_i^\delta) = |W_{i,in}^\delta|.$$

Let $\Omega(\widetilde{W_i^\delta})$ be the induced domain. It is easy to see that $\Omega(\widetilde{W_i^\delta})$ is the same as Ω_i^δ but with cuts. We denote $\Omega(\widetilde{W_i^\delta})$ by $\widetilde{\Omega}_i^\delta$. We then have $\widetilde{\mathcal{V}}_i^\delta = \mathcal{V} \cap H_0^1(\widetilde{\Omega}_i^\delta)$, and hence the functions in $\widetilde{\mathcal{V}}_i^\delta$ are discrete harmonic on $\Omega(W_{i,ovl}^\delta)$. We denote $\Omega(W_{i,ovl}^\delta)$ by $\Omega_{i,ovl}^\delta$.

We define $\widetilde{\mathcal{V}}^\delta \subset \mathcal{V}^\delta$ as

$$\widetilde{\mathcal{V}}^\delta = \widetilde{\mathcal{V}}_1^\delta \oplus \cdots \oplus \widetilde{\mathcal{V}}_N^\delta,$$

which is a direct sum. We remark that functions in $\widetilde{\mathcal{V}}^\delta$ are, by definition, the sum of functions $u_i \in \widetilde{\mathcal{V}}_i^\delta$, $i = 1, \ldots, N$. Functions in $\widetilde{\mathcal{V}}^\delta$ can, in fact, be characterized easily as in the following lemma.

Lemma 2.1. *If $u \in \mathcal{V}$ and u is discrete harmonic at all the overlapping nodes, i.e., on $\cup_{i=1}^N W_{i,ovl}^\delta$, then $u \in \widetilde{\mathcal{V}}^\delta$.*

Proof. To prove that $u \in \widetilde{\mathcal{V}}^\delta$, all we need is to find a decomposition

$$u = \sum_{i=1}^N u_i, \quad \text{with } u_i \in \widetilde{\mathcal{V}}_i^\delta, \quad i = 1, \ldots, N.$$

For the given u, we define u_i piece by piece as follows. On the nodes in $W_{i,in}^\delta$ we let $u_i = u$. On the nodes in $W_{i,cut}^\delta$ we let u_i be zero. On the nodes outside W_i^δ we set u_i to zero. We now need only to define u_i on the nodes belonging to $W_{i,ovl}^\delta$. There, we extend u_i as a discrete harmonic function with boundary data given by u_i just defined. ◻

3. One-level RASHO method. Using notation introduced in the previous section, we now describe a new method, namely a RASHO.

We first define $\widetilde{P}_i^\delta : \widetilde{\mathcal{V}}^\delta \rightarrow \widetilde{\mathcal{V}}_i^\delta$ as a projection operator such that, for any $u \in \widetilde{\mathcal{V}}^\delta$,

$$a(\widetilde{P}_i^\delta u, v) = a(u, v) \qquad \forall v \in \widetilde{\mathcal{V}}_i^\delta. \tag{3.1}$$

The RASHO operator can then be defined as

$$\widetilde{P}^\delta = \widetilde{P}_1^\delta + \cdots + \widetilde{P}_N^\delta. \tag{3.2}$$

Note, however, that the solution u^* of (1.4) (see also (1.5)), is not, generally speaking, in the subspace $\widetilde{\mathcal{V}}^\delta$; therefore, the operator $\widetilde{P}^\delta$ cannot be used to solve the linear system (1.5) directly. We will need to modify the right-hand side of system (1.5). A reformulated (1.5) will be presented in Lemma 3.1 below. We will show that the elimination of the variables associated with the overlapping nodes is not needed in order to apply $\widetilde{P}^\delta$ to any given vector $v \in \widetilde{P}^\delta$.

We now introduce a matrix form of (3.2). We define the restriction operator, or a matrix, $\widetilde{R}_i^\delta$ as follows. Let $v = (v_1, \ldots, v_n)^T$ be a vector corresponding to the nodal values of a function $u \in \mathcal{V}$; namely, for any node $x_k \in W$, $v_k = u(x_k)$. For convenience, we say "v is defined on W." Its restriction on $\widetilde{W}_i^\delta$, $\widetilde{R}_i^\delta v$ is defined as

$$(\widetilde{R}_i^\delta v)(x_k) = \begin{cases} v_k & \text{if } x_k \in \widetilde{W}_i^\delta, \\ 0 & \text{otherwise.} \end{cases} \tag{3.3}$$

The matrix representation of $\widetilde{R}_i^\delta$ is given by a diagonal matrix, with 1 for nodal points in $\widetilde{W}_i^\delta$ and 0 for the remaining nodal points. We remark that, by way of definition, the operator $\widetilde{R}_i^\delta$ is symmetric; i.e., $(\widetilde{R}_i^\delta)^T = \widetilde{R}_i^\delta$. Using this restriction operator, we define the subdomain stiffness matrix as

$$\widetilde{A}_i^\delta = \widetilde{R}_i^\delta \; A \; (\widetilde{R}_i^\delta)^T,$$

which can also be obtained by the discretization of the original finite element problem on $\widetilde{W}_i^\delta$ with zero Dirichlet data on nodes $W \setminus \widetilde{W}_i^\delta$. The matrix $\widetilde{A}_i^\delta$ is block diagonal with blocks corresponding to the structure of $\widetilde{R}_i^\delta$, and its inverse is understood as an inverse of the nonzero block. A matrix representation of $\widetilde{P}_i^\delta$, denoted also by $\widetilde{P}_i^\delta$, is equal to

$$\widetilde{P}_i^\delta = (\widetilde{A}_i^\delta)^{-1} \; A$$

and

$$\widetilde{P}^\delta = ((\widetilde{A}_1^\delta)^{-1} + \cdots + (\widetilde{A}_N^\delta)^{-1}) \; A. \tag{3.4}$$

Using the matrix notations, the next lemma shows how to modify system (1.5) so that its solution belongs to $\widetilde{\mathcal{V}}^\delta$.

LEMMA 3.1. *Let u^* and f be the exact solution and the right-hand side of* (1.5)*, and*

$$w = \sum_{i=1}^{N} (\widetilde{A}_i^\delta)^{-1} \widetilde{R}_i^0 f; \tag{3.5}$$

then we have $\widetilde{u}^* = u^* - w \in \widetilde{\mathcal{V}}^\delta$*, which is the solution of the modified linear system of equations*

$$A\widetilde{u}^* = f - Aw = \widetilde{f}.$$

Proof. If we can show that

$$a(w, \phi_k) = f(\phi_k)$$

for a regular basis function associated with an arbitrary overlapping node $x_k \in W^\delta_{i,ovl}$, for some i, then we will have

$$a(u^* - w, \phi_k) = f(\phi_k) - f(\phi_k) = 0, \tag{3.6}$$

which says that $\widetilde{u}^* = u^* - w$ is discrete harmonic at the overlapping node x_k. We can then use Lemma 2.1 to conclude the proof. Let us now consider

$$w_i = (\widetilde{A}^\delta_i)^{-1}\widetilde{R}^0_i f,$$

which, by definition, is the same as

$$a(w_i, \phi_j) = (\phi_j, \widetilde{R}^0_i f) \qquad \forall x_j \in \widetilde{W}^\delta_i.$$

Here and in the rest of the proof, ϕ_j is the basis function associated with the node $x_j \in \widetilde{W}^\delta_i$. Using that $\widetilde{R}^0_i$ is symmetric and

$$(\phi_j, \widetilde{R}^0_i f) = (f, \widetilde{R}^0_i \phi_j) = a(u^*, \widetilde{R}^0_i \phi_j),$$

we get

$$a(w_i, \phi_j) = a(u^*, \widetilde{R}^0_i \phi_j). \tag{3.7}$$

Let us compute $a(w_i, \phi_k)$. Since x_k is an overlapping node, it cannot be on the boundary of $\widetilde{\Omega}^\delta_i$. This leaves us with the following two cases.

Case 1. The support of $\phi_k(x)$ belongs to the exterior of $\widetilde{\Omega}^\delta_i$. Since the supports of w_i and ϕ_k do not overlap, we have

$$a(w_i, \phi_k) = 0.$$

Case 2. The support of $\phi_k(x)$ belongs to the interior of $\widetilde{\Omega}^\delta_i$. In this case, we have

$$a(w_i, \phi_k) = a(u^*, \widetilde{R}^0_i \phi_k).$$

Taking the sum of the above equality for $i = 1, \ldots, N$, we get

$$a(w, \phi_k) = a\left(\sum_{i=1}^N w_i, \phi_k\right) = a\left(u^*, \sum_{i=1}^N \widetilde{R}^0_i \phi_k\right) = a(u^*, \phi_k),$$

which proves (3.6). Here the fact that $\sum_{i=1}^N \widetilde{R}^0_i = I$ has been used. □

There are basically two ways to compute w in practice. Suppose that subdomain problems are solved using some LU factorization–based method. One can use the same factorization of $\widetilde{A}^\delta_i$ to modify the right-hand side of the system and to solve subdomain problems in the preconditioning steps as that suggested in Lemma 3.1.

Alternatively, one can obtain w by solving several small Dirichlet problems on each subdomain with zero Dirichlet boundary conditions in the overlapping regions $\Omega^\delta_{i,ovl}$. In both strategies, the computation can be done in parallel, and no communication is needed in a distributed memory implementation. In the first approach, $\widetilde{u}^*$ is discrete harmonic in $W^\delta_{i,ovl} \cup W^\delta_{i,non}$, and in the second approach, $\widetilde{u}^*$ is discrete harmonic only in $W^\delta_{i,ovl}$. We note that the discrete harmonicity of $\widetilde{u}^*$ on $W^\delta_{i,non}$ is not required for the algorithms and for the corresponding theory developed in this paper.

Let $\widetilde{f} = f - Aw$; then $\widetilde{u}^*$ is the solution of the following linear system of equations:

$$A\widetilde{u}^* = \widetilde{f}. \tag{3.8}$$

Since $\widetilde{u}^* \in \widetilde{\mathcal{V}}^\delta$,

$$g \equiv \widetilde{P}^\delta \widetilde{u}^*$$

is well defined and can be computed without knowing $\widetilde{u}^*$ by using the following relations:

$$a(\widetilde{P}^\delta_i \widetilde{u}^*, v) = a(\widetilde{u}^*, v) = (\widetilde{f}, v) \qquad \forall v \in \widetilde{\mathcal{V}}^\delta_i \text{ and } i = 1, \dots, N.$$

More precisely, we can obtain g by solving the subdomain problems

$$a(g_i, v) = (\widetilde{f}, v) \qquad \forall v \in \widetilde{\mathcal{V}}^\delta_i$$

for $i = 1, \dots, N$ and taking $g = g_1 + \cdots + g_N$. With such a right-hand side, we introduce a new linear system

$$\widetilde{P}^\delta \widetilde{u}^* = g, \tag{3.9}$$

which is equivalent to the linear system (3.8); see Theorem 5.1. The system (3.9) is a symmetric positive definite system under the usual energy inner product and therefore can be solved using the CG method. RASHO has a few advantages over the classical AS preconditioner. Let us recall AS briefly. Let

$$\left(R^\delta_i v\right)(x_k) = \begin{cases} v_k & \text{if } x_k \in W^\delta_i, \\ 0 & \text{otherwise.} \end{cases} \tag{3.10}$$

Then the AS operator takes the following matrix form:

$$P^\delta = \left((A^\delta_1)^{-1} + \cdots + (A^\delta_N)^{-1}\right) A, \tag{3.11}$$

where $A^\delta_i = R^\delta_i A (R^\delta_i)^T$. Because of the inclusion of the cut interface nodes, the size of the matrix A^δ_i is $|W^\delta_i|$, which is slightly larger than the size of the matrix $\widetilde{A}^\delta_i$, which is $|\widetilde{W}^\delta_i|$. In a distributed memory implementation, the operation $R^\delta_i v$ involves moving data from one processor to another, but the operation $\widetilde{R}^\delta_i v$ does not involve any communication. More precisely, in RASHO, if $u \in \widetilde{\mathcal{V}}^\delta$, then it is easy to see that

$$\widetilde{R}^\delta_i A u = \widetilde{R}^\delta_{i,in} A u, \tag{3.12}$$

where $\widetilde{R}^\delta_{i,in}$ is defined as

$$(\widetilde{R}^\delta_{i,in} v)(x_k) = \begin{cases} v_k & \text{if } x_k \in W^\delta_{i,in}, \\ 0 & \text{otherwise.} \end{cases} \tag{3.13}$$

Therefore, for functions in $\widetilde{\mathcal{V}}^\delta$ we can rewrite $\widetilde{P}^\delta$, as in (3.4), in the following form:

$$\widetilde{P}^\delta = ((\widetilde{A}_1^\delta)^{-1}\widetilde{R}_{1,in}^\delta + \cdots + (\widetilde{A}_N^\delta)^{-1}\widetilde{R}_{N,in}^\delta)\, A. \tag{3.14}$$

Although the operator (3.14) does not look like a symmetric operator, it is indeed symmetric when applied to functions in the subspace $\widetilde{\mathcal{V}}^\delta$. The form (3.12) takes advantage of the fact that the operator $\widetilde{R}_{i,in}^\delta$ is communication-free in the sense that it needs only the residual associated with nodes in $W_{i,in}^\delta \subset \Omega_i^0$.

We make some further comments on how the residual Au can be calculated in a distributed memory environment for a given vector $u \in \widetilde{\mathcal{V}}^\delta$. In a typical implementation, the matrix A is constructed and stored in the form $\{\widetilde{A}_i^\delta\}$; each processor has one or several of the subdomain matrix $\widetilde{A}_i^\delta$. Similarly, u is stored in the form $\{u_i\}$, where $u_i \in \widetilde{\mathcal{V}}_i^\delta$. We note, however, that to compute the residual at nodes $W_{i,in}^{\Gamma^\delta}$ some communications are required. The processor associated with subdomain Ω_i^δ needs to obtain the local solution from the neighboring subdomains at nodes connected to $W_{i,in}^{\Gamma^\delta}$. It is important to note that the amount of communication does not depend on the size of the overlap, since only one layer of nodes is required. This shows that in terms of communication, the RASHO is superior to AS and RAS.

4. Some two-level versions. As with other domain decomposition methods, the convergence rate of the single-level method depends on the number of subdomains. To make the algorithm more scalable with respect to the number of subdomains, we next introduce two two-level versions of RASHO in this section. This includes an additive version and a hybrid version using the same coarse space.

Standard coarse spaces cannot be used since they are usually not discrete harmonic in the overlapping regions. To construct a coarse subspace $\widetilde{\mathcal{V}}_0$ of $\widetilde{\mathcal{V}}$, we introduce the coarse basis functions $\phi^i(x), i = 1, \ldots, N$, based on a partition of unity [21] on the interface nodes W^{Γ^δ}. For each subdomain, we define the nodal values of $\phi^i(x) \in \widetilde{\mathcal{V}}_i$ as follows:

$$\phi^i(x_k) = \begin{cases} 1 & \text{if } x_k \in W_{i,in}^{\Gamma^\delta}, \\ \text{discrete harmonic} & \text{if } x_k \in W_{i,ovl}^\delta \cup W_{i,non}^\delta, \\ 0 & \text{if } x_k \in W \backslash \widetilde{W}_i^\delta. \end{cases} \tag{4.1}$$

Let us denote $\Omega(W_{i,non}^\delta)$ by $\Omega_{i,non}^\delta$. Then $\phi^i(x_k) = 1$ at $x_k \in W_{i,non}^\delta$ for the case $\Omega_{i,non}^\delta \cap \partial\Omega = \emptyset$ since all the boundary nodal values of $\Omega_{i,non}^\delta$ belong to $W_{i,in}^{\Gamma^\delta}$ and therefore have nodal values equal to one. For the case $\Omega_{i,non}^\delta \cap \partial\Omega \neq \emptyset$, we have chosen to define $\phi^i(\Omega_{i,non}^\delta)$ as the discrete harmonic extension with boundary nodal values equal to one on $W_{i,in}^{\Gamma^\delta}$ and equal to zero at $\bar{\Omega}_{i,non}^\delta \cap \partial\Omega$; note, however, that we do not require that $\widetilde{\mathcal{V}}_0^\delta$ be discrete harmonic on $\Omega_{i,non}^\delta$. If we had chosen ϕ^i equal to one at all nodes of $\Omega_{i,non}^\delta$ also for the $\Omega_{i,non}^\delta \cap \partial\Omega \neq \emptyset$ case, ϕ^i would have a jump from one to zero on the neighboring elements of $\partial\Omega$. This jump would give lower bounds that depend on the factor h/H, and such bounds would be poor if the overlap were very small. Another possibility for avoiding the discrete harmonicity of ϕ^i on $\Omega_{i,non}^\delta$ in the $\Omega_{i,non}^\delta \cap \partial\Omega \neq \emptyset$ case would be the use of the boundary layer technique developed in [21]. We note, however, that the bounds of Theorem 5.1 would remain the same as well as the analysis, with some minor modifications.

The coarse space $\widetilde{\mathcal{V}}_0^\delta$ is simply the space spanned by all linear combinations of the coarse basis functions $\phi^i, i = 1, \ldots, N$. We define $\widetilde{P}_0^\delta : \mathcal{V} \to \widetilde{\mathcal{V}}_0^\delta$ as the operator such that, for any $u \in \mathcal{V}$,

$$a(\widetilde{P}_0^\delta u, v) = a(u, v) \qquad \forall v \in \widetilde{\mathcal{V}}_0^\delta.$$

A two-level additive version of RASHO can now be introduced with the operator

$$\widetilde{P}_C^\delta = \sum_{i=0}^{N} \widetilde{P}_i^\delta. \tag{4.2}$$

The convergence properties of this two-level algorithm will be studied in the next section. To describe the computational aspects of the coarse problem, we rewrite the above definitions in matrix notation. Recall that n is the total number of nodes in W, N is the total number of subdomains, and ϕ^i is the coarse basis function. We write the fine-to-coarse restriction operator as an $N \times n$ matrix

$$(\widetilde{R}_0)_{N\times n} = \left(\phi^i(x_k)\right)_{i=1,N;k=1,n}.$$

The matrix form of the coarse projection operator $\widetilde{P}_0^\delta$ is

$$\widetilde{P}_0^\delta = \widetilde{R}_0^T \widetilde{A}_0^{-1} \widetilde{R}_0 A, \tag{4.3}$$

where $\widetilde{A}_0 = \widetilde{R}_0 A \widetilde{R}_0^T$ is an $N \times N$ matrix.

We remark that $\widetilde{A}_0$ is more sparse than coarse space matrices that appear in other methods such as Neumann–Neumann or FETI-type algorithms [12, 13, 18, 23], since only connections with the neighboring subdomains appear in the stencils associated with a coarse basis function. Another feature of this coarse space problem is that the computation of the right-hand side, i.e., $\widetilde{R}_0 A u$ for some u, can be done inside each Ω_i^δ; this is a clear advantage over regular coarse spaces.

The two-level additive algorithm (4.2) is easy to code, but the performance isn't as good as expected. Some examples are given in the numerical experiments section of this paper. We next introduce another two-level algorithm—a hybrid Schwarz operator (see [19]) with the error propagation operator given by

$$\left(I - \widetilde{P}_0^\delta\right)\left(I - \sum_{i=1}^{N} \widetilde{P}_i^\delta\right)\left(I - \widetilde{P}_0^\delta\right). \tag{4.4}$$

This is a symmetric operator with which we can work essentially without any extra cost, since, when forming powers of the operator (4.4) on building the Krylov space on the PCG, we can use the fact that $I - \widetilde{P}_0^\delta$ is a projection, and therefore $(I - \widetilde{P}_0^\delta)^2 = I - \widetilde{P}_0^\delta$. Subtracting the operator (4.4) from the identity operator I, we obtain the operator

$$\widetilde{P}_{hyb}^\delta = \widetilde{P}_0^\delta + \left(I - \widetilde{P}_0^\delta\right)\left(\sum_{i=1}^{N} \widetilde{P}_i^\delta\right)\left(I - \widetilde{P}_0^\delta\right). \tag{4.5}$$

The spectral properties of $\widetilde{P}_{hyb}^\delta$ will be studied in the next section. Some numerical results obtained using the additive and the hybrid two-level methods will be presented in the numerical experiments section of the paper, and they will both be compared with the single-level method.

5. Theoretical analysis.

The algorithm presented in the previous section is applicable for general sparse, symmetric positive definite linear systems. The notions of subdomains, harmonic overlaps, the classification of the nodal points, etc. can all be defined in terms of the graph of the sparse matrix. In this section we provide a nearly optimal estimate for a Poisson equation discretized with a piecewise linear finite element method. We estimate the condition number of the RASHO operators $\widetilde{P}^\delta$ and $\widetilde{P}^\delta_C$ in terms of the fine mesh size h, the subdomain size H, and the overlapping factor δ. We shall follow the abstract AS theory [24] in what follows.

Lemma 5.1. *Suppose that the following assumptions hold:*

(i) *There exists a constant C_0 such that for any $u \in \widetilde{\mathcal{V}}^\delta$ there exists a decomposition*

$$u = \sum_{i=0}^{N} u_i,$$

where $u_i \in \widetilde{\mathcal{V}}^\delta_i$, and

$$\sum_{i=0}^{N} |u_i|^2_{H^1(\Omega)} \le C_0^2 |u|^2_{H^1(\Omega)}.$$

(ii) *There exist constants $\epsilon_{ij}, i, j = 1, \ldots, N$, such that*

$$a(u_i, u_j) \le \epsilon_{ij}\, a(u_i, u_i)^{1/2} a(u_j, u_j)^{1/2} \qquad \forall u_i \in \widetilde{\mathcal{V}}^\delta_i,\ \forall u_j \in \widetilde{\mathcal{V}}^\delta_j.$$

Then $\widetilde{P}^\delta_C$ is invertible, symmetric; i.e., $a(\widetilde{P}^\delta_C u, v) = a(u, \widetilde{P}^\delta_C v)$,

$$C_0^{-2} a(u,u) \le a(\widetilde{P}^\delta_C u, u) \le (\rho(\mathcal{E}) + 1) a(u,u) \qquad \forall u \in \widetilde{\mathcal{V}}^\delta. \tag{5.1}$$

Here $\rho(\mathcal{E})$ is the spectral radius of $\mathcal{E}$, which is an $(N) \times (N)$ matrix made of $\{\epsilon_{ij}\}$.

It is trivial to see that $\rho(\mathcal{E}) \le C$. Thus our focus in the rest of the section is on bounding C_0. For the case of the single-level RASHO, the lemma above can be modified by replacing $u = \sum_{i=0}^N u_i$, $\widetilde{P}^\delta_C$, and $(\rho(\mathcal{E}) + 1)$ with $u = \sum_{i=1}^N u_i$, $\widetilde{P}^\delta$, and $\rho(\mathcal{E})$, respectively.

To analyze the hybrid algorithm, we use a result due to Mandel [19, Lemma 3.2], which in our context is given by the following.

Lemma 5.2. *The extreme eigenvalues of $\widetilde{P}^\delta_{hyb}$, $\widetilde{P}^\delta_C$, and $\widetilde{P}^\delta$ satisfy*

$$\lambda_{min}(\widetilde{P}^\delta_{hyb}) \ge \lambda_{min}(\widetilde{P}^\delta_C) \qquad \text{and} \qquad \lambda_{max}(\widetilde{P}^\delta_{hyb}) \le \lambda_{max}(\widetilde{P}^\delta).$$

5.1. The partition of unity and a comparison function.

The construction of a partition of unity is one of the key steps in an AS analysis. Consider $\phi^i(x)$ defined in (4.1). It is easy to see that $\{\phi^i(x), i = 1, \ldots, N\}$ restricted to W^{Γ^δ} forms a partition of unity.

In addition to $\phi^i(x)$, we also need to construct a *comparison function* $\theta_i(x)$ for each subdomain Ω^δ_i. Comparison functions, or barrier functions, are very useful for many Schwarz algorithms, such as these on nonmatching grids [6]. We will show that, even though $\theta_i(x) \in \mathcal{V}^\delta_i$, and is not in $\widetilde{\mathcal{V}}^\delta_i$ as we wished, it can still be used to bound functions in $\widetilde{\mathcal{V}}^\delta_i$. Both $\theta_i(x)$ and $\phi^i(x)$ depend on the overlapping factor δ. Because $\phi^i(x)$ is discrete harmonic at $W^\delta_{i,ovl} \cup W^\delta_{i,non}$ and identical to θ_i at the remaining nodes, we have

$$a(\phi^i, \phi^i) \le a(\theta_i, \theta_i).$$

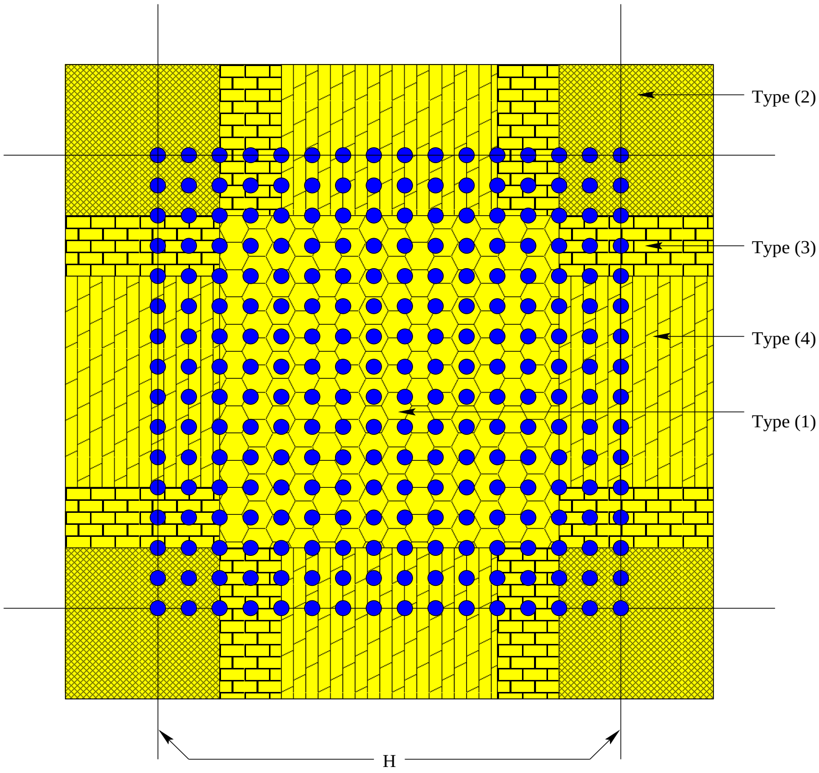


FIG. 5.1. *The partition of Ω_i^δ into the union of four types of subregions. This is a "floating" subdomain with $\delta = 2$. The collection of "$\bullet$" forms the set W_i^0.*

To construct the function $\theta_i(x)$, we first consider the case in which Ω_i^0 is a floating square subdomain. "Floating" refers to the fact that the subdomain doesn't touch the boundary $\partial\Omega$. The extension to cases in which Ω_i^δ touches the boundary is simple, and we will comment on it later. To further simplify our arguments, we assume that Ω_i^δ and its neighboring extended subdomains Ω_j^δ are squares of the same size, i.e., sides of length equal to $H + 2(\delta + 1)h$. This assumption is equivalent to claiming that Ω^R has size H and that δ levels of overlap are applied; see Figure 5.1. We also assume that the overlap is not too large; for the analysis given below, δh no larger than $H/4$ is enough. Our techniques can be modified to consider larger overlaps and more complex subdomains, although too large of an overlap has little practical value.

Roughly speaking, $\theta_i(x)$ is equal to $\phi^i(x)$ on $W \backslash W_{i,ovl}^\delta$. On the overlapping region $W_{i,ovl}^\delta$, we need to define $\theta_i(x)$ carefully so that we can control its energy in the H^1 seminorm. For this purpose, we decompose Ω_i^δ into subregions of four types (see Figure 5.1), $\Omega_{i,non}^\delta$ (Type (1)), $\Omega_i^{\delta\delta}$ (Type (2)), $\Omega_i^{\delta H}$ (Type (3)), and $\Omega_i^{\delta\tilde{\delta}}$ (Type (4)), and define $\theta_i(x)$ on each piece of the subregion separately.

Type (1). The first subregion is $\Omega_{i,non}^\delta$, which is a square with sides of size $H - 2\delta h$.

Type (2). The second subregion $\Omega_i^{\delta\delta}$ is the area in which Ω_i^δ overlaps simultaneously with three neighbors Ω_j^δ. $\Omega_i^{\delta\delta}$ therefore represents the union of the four corner pieces of Ω_i^δ, i.e., four squares with sides of size $(2\delta + 1)h$.

Types (3) and (4). The area in which Ω_i^δ overlaps only one neighbor is four rectangles of size $H-2\delta h\times(2\delta+1)h$. We further partition each of the four rectangles into three smaller rectangles; i.e., two of them are of $\Omega_i^{\delta\tilde{\delta}}$ type and one of them of $\Omega_i^{\delta H}$ type. For instance, without lost of generality, let us consider the intersection of Ω_i^{δ} with its right-hand neighbor Ω_j^{δ}, excluding the corner parts. In this case, the subregion to be partitioned is a rectangle of size $(2\delta+1)h$ in the x direction and $H-2\delta h$ in the y direction. The partition of this rectangles gives two smaller rectangles of $\Omega_i^{\delta\tilde{\delta}}$ type with dimensions $2(\delta+1)h\times\delta h$, and each one has an edge in common with a square of $\Omega_i^{\delta\delta}$ type. We define them as transition subregions because they are placed between a corner-type subregion $\Omega_i^{\delta\delta}$ and a face-type subregion $\Omega_i^{\delta H}$. The $\Omega_i^{\delta H}$ face-type subregions are the smaller rectangles that are placed between the two smaller rectangles of $\Omega_i^{\delta\tilde{\delta}}$ type. $\Omega_i^{\delta H}$ face-type regions are of size $(2\delta+1)h$ by $H-4\delta h$.

For any node x belonging to a Type (1) region $\Omega_{i,non}^{\delta}$, we define $\theta_i(x)$ to be equal to one, i.e., equal to $\phi^i(x)$. Therefore

$$|\phi^i(x)|^2_{H^1(\Omega^\delta_{i,non})}=|\theta_i(x)|^2_{H^1(\Omega^\delta_{i,non})}=0.$$

We next define $\theta_i(x)$, node by node, in $\Omega^\delta_{i,ovl}$, which is the union of corner-, transition-, and face-type regions defined above.

For a Type (2) region $\Omega_i^{\delta\delta}$, let Q be such a square with vertices $V_1=(a,b)$, $V_2=(a+(2\delta+1)h,b)$, $V_3=(a,b+(2\delta+1)h)$, and $V_4=(a+(2\delta+1)h,b+(2\delta+1)h)$. We assume that V_1,V_2, and V_4 belong to $\partial\Omega_i^\delta$. In other words, Q is located on the southeast corner of Ω_i^δ. Let us also introduce another square region $\widetilde{Q}$, with vertices $V_3=(a,b+(2\delta+1)h)$, $\widetilde{V}_1=(a,b+\delta h)$, $\widetilde{V}_2=(a+(\delta+1)h,b+\delta h)$, and $\widetilde{V}_4=(a+(\delta+1)h,b+(2\delta+1)h)$. Note that $\widetilde{Q}$ is contained in Q, with V_3 as the common vertex. To define $\theta_i(x)$ on Q, we set $\theta_i(V_3)=1$, $\theta_i(\widetilde{V}_1)=0$, $\theta_i(\widetilde{V}_2)=0$, $\theta_i(\widetilde{V}_4)=0$. At the remaining nodes x on the edges $\widetilde{V}_1\widetilde{V}_2$ and $\widetilde{V}_2\widetilde{V}_4$ we set $\theta_i(x)=0$, and on the edges $V_3\widetilde{V}_1$ and $V_3\widetilde{V}_4$ we set $\theta_i(x)=1$. For nodes on $Q\backslash\widetilde{Q}$ we set $\theta_i(x)=0$. It remains only to define $\theta_i(x)$ for nodes x in the interior of $\widetilde{Q}$. To define $\theta_i(x)$ there, we use a well-known cutoff function technique, such as the one introduced in Lemma 4.4 of [10], but for two-dimensional square regions. An illustrative picture of $\theta_i(x)$ in a typical region $\Omega_i^{\delta\delta}$ is shown in Figure 5.2. For the completeness of this paper, we include the construction below. Let C be the center of the square $\widetilde{Q}$. The construction of $\theta_i(x)$ is defined by the following steps:

(1) Define $\theta_i(V_3)=1$, $\theta_i(\widetilde{V}_2)=0$, $\theta_i(\widetilde{V}_1)=0$, and $\theta_i(\widetilde{V}_4)=0$.

(2) For a point P that belongs to the segments $V_3\widetilde{V}_4$ or $V_3\widetilde{V}_1$, define $\theta_i(P)=1$. For a point P that belongs to the segments $\widetilde{V}_4\widetilde{V}_2$ or $\widetilde{V}_1\widetilde{V}_2$, define $\theta_i(P)=0$.

(3) For a point Y that belongs to the line segment connecting C to V_3, define $\theta_i(Y)$ by linear interpolation between values $\theta_i(C)=1/2$ and $\theta_i(V_3)=1$. For a point Y that belongs to the line segment connecting C to $\widetilde{V}_2$, define $\theta_i(Y)$ by linear interpolation between values $\theta_i(C)=1/2$ and $\theta_i(\widetilde{V}_2)=0$.

(4) For a point S that belongs to a line segment connecting a point Y to a vertex $\widetilde{V}_1$ or $\widetilde{V}_4$, define $\theta_i(S)=\theta_i(Y)$.

(5) Note that the θ_i is defined everywhere on $\widetilde{Q}\cup\partial\widetilde{Q}$. θ_i is continuous everywhere except at the points $\widetilde{V}_1$ and $\widetilde{V}_4$. We redefine θ_i as the continuous piecewise linear finite element function given by the standard pointwise interpolation.

The most important observation of the construction of $\theta_i(x)$ inside $\widetilde{Q}$ is that $|\nabla\theta_i(x)|\le C/r$ near $\widetilde{V}_1$ or $\widetilde{V}_4$. Here r is the distance of x from $\widetilde{V}_1$ or $\widetilde{V}_4$. Therefore,

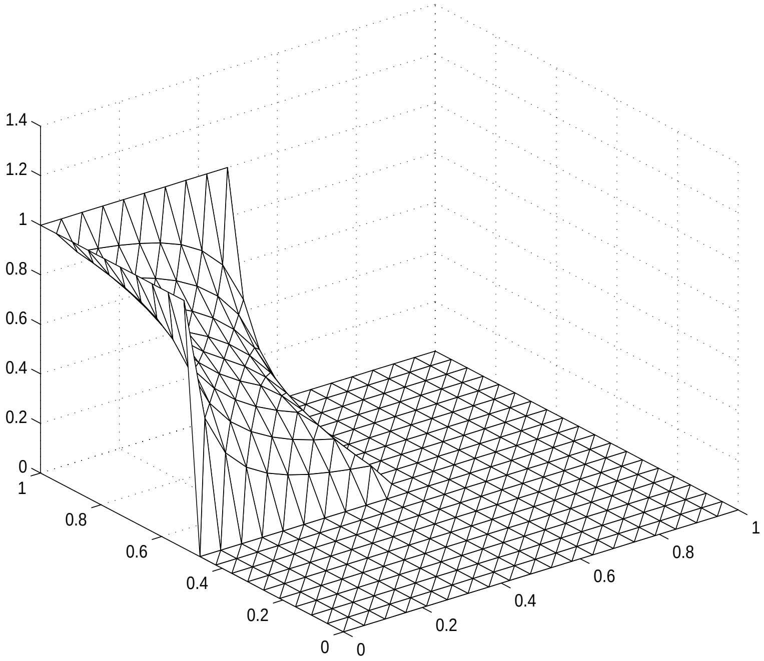


FIG. 5.2. *An illustrative picture of $\theta_i(x)$ in a typical region $\Omega_i^{\delta\delta}$.*

we obtain (see [10] and [23])

$$|\theta_i(x)|^2_{H^1(Q)} = |\theta_i(x)|^2_{H^1(\widetilde{Q})} \leq C\left(1 + \log\left(\frac{(\delta+1)h}{h}\right)\right) = C(1 + \log(\delta + 1)).$$

Since inside of Ω_i^δ there are four of those squares, we obtain

$$|\theta_i(x)|^2_{H^1(\Omega_i^{\delta\delta})} \leq C\left(1 + \log(\delta + 1)\right).$$

Type (3) regions consist of transition-type rectangles. Let us consider one of them and denote it by T, which we assume has vertices at $V_3 = (a, b + (2\delta + 1)h)$, $V_4 = (a+(2\delta+1)h, b+(2\delta+1)h)$, $V_5 = (a, b+(3\delta+1)h)$, and $V_6 = (a+(2\delta+1)h, b+(3\delta+1)h)$. Note that T stands on top of the square Q introduced above and has the common edge V_3V_4. We define $\theta_i(x)$ over the edge V_3V_4 to be equal to $\phi^i(x)$. Over the edge V_3V_5, we set $\theta_i(x) = 1$. Over the edge V_4V_6, we set $\theta_i(x) = 0$. And over the edge V_5V_6 we let $\theta_i(x)$ decrease linearly from the value 1 to 0. What remains is to define $\theta_i(x)$ inside T. Let us define the nodes $V_l = (a+\delta h, b+(2\delta+1)h)$ and $V_r = (a+(\delta+1)h, b+(2\delta+1)h)$, which is the same as the node $\widetilde{V}_4$ used in the description of Type (2) regions. The nodes V_l and V_r are exactly the places on the edge V_3V_4 where $\phi^i(x)$ jumps from 1 to 0. On the triangle $V_3V_lV_5$ we set $\theta_i(x) = 1$. On the triangle $V_rV_4V_6$ we set $\theta_i(x) = 0$. On the region $V_lV_rV_6V_5$, we let $\theta_i(x)$ decrease linearly in the x direction from the value 1 to 0. We note that next to the nodes V_lV_r, $\theta_i(x)$ has a singular behavior similar to $|\nabla\theta_i(x)| \leq C/r$, where r is the distance from x to the line $V_l\ V_r$. Similarly, we have

$$|\theta_i(x)|^2_{H^1(T)} \leq C\left(1 + \log(\delta + 1)\right).$$

Since there are eight rectangles of Type (3) inside $\Omega_i^{\delta\tilde{\delta}}$, we obtain

$$|\theta_i(x)|^2_{H^1(\Omega_i^{\delta\tilde{\delta}})} \leq C\left(1 + \log(\delta+1)\right).$$

Type (4) regions are rectangles of face type. Let R be one of them, and assume that the vertices are given by $V_5 = (a, b+(3\delta+1)h)$, $V_6 = (a+(2\delta+1)h, b+(3\delta+1)h)$, $V_7 = (a, b+H-(\delta-1)h)$, and $V_8 = (a+(2\delta+1)h, b+H-(\delta-1)h)$. Note that R is on the top of the rectangle T defined above, and its height is $H - 4\delta h$. The vertices V_6 and V_8 are the vertices that belong to $\partial\Omega_i^\delta$. We define $\theta_i(x) = 1$ if x is on the edge V_5V_7, and $\theta_i(x) = 0$ if x is on the edge V_6V_8, and $\theta_i(x)$ is linear in the horizontal direction for the remaining points. We then obtain

$$|\theta_i(x)|^2_{H^1(R)} \leq \frac{H-4\delta h}{(2\delta+1)h}.$$

Since there are four of those rectangles inside $\Omega_i^{\delta H}$, we obtain

$$|\theta_i(x)|_{H^1(\Omega_i^{\delta H})} \leq C\frac{H-4\delta h}{(2\delta+1)h} \leq C\frac{H}{(2\delta+1)h}.$$

For the cases in which Ω_i^0 touches the boundary $\partial\Omega$, the analysis needs to be modified slightly. The first modification is because the shape of the overlapping region changes slightly, i.e., the longer side is shorter; it is easy to see that we get similar bounds as before. The other modification is because ϕ^i on $\Omega_{i,non}^\delta$ is not identically equal to one and therefore the corresponding energy is not necessarily zero; for this case we can design θ_i similarly and obtain

$$|\theta_i(x)|^2_{H^1(\Omega_{i,non}^\delta)} \leq C\left(1 + \log\left(\frac{H}{h}\right)\right).$$

Putting all the pieces of $\theta_i(x)$ together, we see that $\theta_i(x) \in \mathcal{V}_i^\delta$, and it is equal to $\phi^i(x)$ on W^{Γ^δ}. Adding all the estimates on subregions of the four types, we arrive at the following lemma.

Lemma 5.3. *For $i = 1, \ldots, N$, $\theta_i(x) \in \mathcal{V}_i^\delta$, $\phi^i(x) \in \tilde{\mathcal{V}}_i^\delta$, and the following hold:*

(1) $|\phi^i|^2_{H^1(\Omega_i^\delta)} \leq |\theta_i|^2_{H^1(\Omega_i^\delta)}$.

(2)

$$|\theta_i|^2_{H^1(\Omega_i^\delta \backslash \Omega_{i,non}^\delta)} \leq C\left(1 + \log(\delta+1) + \frac{H}{(2\delta+1)h}\right).$$

(3) *If $\Omega_{i,non}^\delta \cap \partial\Omega = \emptyset$, then $|\theta_i|^2_{H^1(\Omega_{i,non}^\delta)} = 0$.*

(4) *If $\Omega_{i,non}^\delta \cap \partial\Omega \neq \emptyset$, then*

$$|\theta_i|^2_{H^1(\Omega_{i,non}^\delta)} \leq C\left(1 + \log\left(\frac{H}{h}\right)\right).$$

Here $C > 0$ is independent of the parameters h, H, and δ.

5.2. A bounded partition lemma. To obtain the parameter C_0 of assumption (i) of the abstract AS theory (see Lemma 5.1), we construct a decomposition of $\widetilde{\mathcal{V}}^\delta$ and prove its boundedness below.

Lemma 5.4. *There exists a constant $C > 0$, independent of h, H, and δ, such that for any $u \in \widetilde{\mathcal{V}}^\delta$ there exist $v_i \in \widetilde{\mathcal{V}}_i^\delta$ such that*

$$u = \sum_{i=0}^{N} v_i \tag{5.2}$$

and

$$\begin{aligned}\sum_{i=0}^{N} |v_i|_{H^1(\Omega)}^2 \le C\left(\left(\frac{H}{(2\delta+1)h}\right)\right)|u|_{H^1(\Omega)}^2 \\ + C(1+\log(\delta+1))\left(1+\log\left(\frac{H}{h}\right)\right)|u|_{H^1(\Omega)}^2.\end{aligned} \tag{5.3}$$

In addition, there exist $u_i \in \widetilde{\mathcal{V}}_i^\delta$ such that

$$u = \sum_{i=1}^{N} u_i \tag{5.4}$$

and

$$\begin{aligned}\sum_{i=1}^{N} |u_i|_{H^1(\Omega)}^2 \le C\,(1+\log(\delta+1))\left(1+\log\left(\frac{H}{h}\right)\right)|u|_{H^1(\Omega)}^2 \\ + C\frac{1}{H^2}\left(1+\log(\delta+1)+\frac{H}{(2\delta+1)h}\right)|u|_{H^1(\Omega)}^2.\end{aligned} \tag{5.5}$$

Proof. We first construct the decomposition (5.4). For any given $u \in \widetilde{\mathcal{V}}^\delta$ we define $u_i \in \widetilde{\mathcal{V}}_i^\delta$ as

$$u_i(x_k) = \begin{cases} u(x_k) & \text{if } x_k \in W_{i,in}^\delta, \\ \text{discrete harmonic} & \text{if } x_k \in W_{i,ovl}^\delta, \\ 0 & \text{if } x_k \in W \backslash \widetilde{W}_i^\delta. \end{cases}$$

It is easy to see that (5.4) holds. We next construct the decomposition (5.2). For $i = 1, \dots, N$, let us define $v_i \in \widetilde{\mathcal{V}}_i^\delta$ by

$$v_i = u_i - \bar{u}_i \phi^i \in \widetilde{\mathcal{V}}_i^\delta,$$

where

$$\bar{u}_i = \frac{1}{|\Omega_i^\delta|}\int_{\Omega_i^\delta} u dx$$

is the average of u on the extended region Ω_i^δ. Here $|\Omega_i^\delta|$ is the area of the region Ω_i^δ. We also define

$$v_0 = \sum_{i=1}^{N} \bar{u}_i \phi^i.$$

It is easy to see that (5.2) holds.

The next step is to bound $\sum_{i=1}^{N} |v_i|^2_{H^1(\Omega)}$. To bound each term $|v_i|^2_{H^1(\Omega)}, i = 1, \dots, N$, we use $\theta_i(x), i = 1, \dots, N$, introduced before. Consider $\tilde{v}_i \in \mathcal{V}_i^\delta$ defined as follows:

$$\tilde{v}_i(x) = I_h(\theta_i(x)(u(x) - \bar{u}_i)).$$

Note that $\tilde{v}_i(x)$ is equal to $v_i(x)$ on $W_i^{\Gamma^\delta}$ and on $\partial\Omega_i^\delta$. On $\Omega^\delta_{i,ovl}$, v_i is discrete harmonic. Therefore, we have

$$|v_i|^2_{H^1(\Omega^\delta_{i,ovl})} \leq |\tilde{v}_i|^2_{H^1(\Omega^\delta_{i,ovl})}.$$

In addition, $v_i(x)$ is identical to $\tilde{v}_i$ on $\Omega^\delta_{i,non}$ whenever $\Omega^\delta_{i,non}$ does not touch $\partial\Omega$. For such cases, we next devote the proof to the estimate of $|\tilde{v}_i|^2_{H^1(\Omega_i^\delta)}$ in terms of $|u|^2_{H^1(\Omega_i^\delta)}$. The estimate of $|v_i|^2_{H^1(\Omega^\delta_{i,non})}$ for the case in which $\Omega^\delta_{i,non}$ does not touch $\partial\Omega$ is done afterwards in (5.10).

Let K be an element of Ω_i^δ, and let us define $w_i = u - \bar{u}_i$; then

$$(5.6)\qquad |\tilde{v}_i|^2_{H^1(K)} = |I_h(\theta_i w_i)|^2_{H^1(K)} \leq 2|\bar{\theta}_i w_i|^2_{H^1(K)} + 2|I_h((\bar{\theta}_i - \theta_i)w_i)|^2_{H^1(K)}.$$

Here, $\bar{\theta}_i$ is the average of θ_i on K, and I_h is the standard pointwise interpolation. To estimate the first part of (5.6) we use the fact that $|\bar{\theta}_i| \leq 1$ to obtain

$$|\bar{\theta}_i w_i|^2_{H^1(K)} = |\bar{\theta}_i(u - \bar{u}_i)|^2_{H^1(K)} \leq |u - \bar{u}_i|^2_{H^1(K)} = |u|^2_{H^1(K)}.$$

The last equality comes from the fact that $\bar{u}_i$ is a constant. For the second part of (5.6), according to an inverse inequality, we have

$$(5.7)\qquad |I_h((\bar{\theta}_i - \theta_i)w_i)|^2_{H^1(K)} \leq C\frac{1}{h^2}\|I_h((\bar{\theta}_i - \theta_i)w_i)\|^2_{L^2(K)}.$$

To obtain the bound for the right-hand side of (5.7), we consider the element K in four different situations corresponding to the four types of subregions into which the the subregion Ω_i^δ is split, i.e., $\Omega^\delta_{i,non}$, $\Omega_i^{\delta H}$, $\Omega_i^{\delta\tilde{\delta}}$, and $\Omega_i^{\delta\delta}$.

The proof for the cases $K \subset \Omega_i^{\delta H}$ and $K \subset \Omega_i^{\delta\tilde{\delta}}$ are nearly the same, so we only consider one of them here. For $K \subset \Omega_i^{\delta H}$, since

$$\|\bar{\theta}_i - \theta_i\|^2_{L^\infty(K)} \leq C\left(\frac{h}{(2\delta+1)h}\right)^2,$$

we obtain

$$\frac{1}{h^2}\|I_h((\bar{\theta}_i - \theta_i)w_i)\|^2_{L^2(K)} \leq C\frac{1}{((2\delta+1)h)^2}\|w_i\|^2_{L^2(K)}.$$

Applying a technique developed in Dryja and Widlund [11], we obtain

$$(5.8)\ \frac{1}{((2\delta+1)h)^2}\|w_i\|^2_{L^2(\Omega_i^{\delta H})} \leq C\left(\frac{H}{(2\delta+1)h}|w_i|^2_{H^1(\Omega_i^\delta)} + \frac{1}{H((2\delta+1)h)}\|w_i\|^2_{L^2(\Omega_i^\delta)}\right).$$

Using the fact that $|w_i|^2_{H^1(\Omega_i^\delta)} = |u|^2_{H^1(\Omega_i^\delta)}$ and a Friedrichs inequality, we have

$$(5.9)\qquad \|w_i\|^2_{L^2(\Omega_i^\delta)} \leq CH^2|u|^2_{H^1(\Omega_i^\delta)}.$$

Combining the estimates (5.8) and (5.9), we obtain

$$\frac{1}{((2\delta+1)h)^2}\|w_i\|^2_{L^2(\Omega_i^{\delta H})} \le C\frac{H}{(2\delta+1)h}|u|^2_{H^1(\Omega_i^\delta)}.$$

For the case when $K \subset \Omega_i^{\delta\delta}$, we use similar arguments as in Dryja, Smith, and Widlund [10] to obtain

$$\sum_{K\in\Omega_i^{\delta\delta}} \frac{1}{h^2}\|I_h((\bar{\theta}_i - \theta_i)w_i)\|^2_{L^2(K)} \le \sum_{K\in\Omega_i^{\delta\delta}} C\frac{1}{r^2}\|w_i\|^2_{L^2(K)}, \tag{5.10}$$

where $ch \le r \le C((\delta+1)h)$ is the distance to those "cut pieces." We have used here that $\theta_i(x)$ has the singular behavior C/r on $\Omega_i^{\delta\delta}$. We then have

$$\sum_{K\in\Omega_i^{\delta\delta}} \frac{1}{r^2}\|w_i\|^2_{L^2(K)} \le C\int_{ch}^{C(\delta+1)h}\int_\alpha r^{-2}r\|w_i\|^2_{L^\infty(\Omega_i^{\delta\delta})}d\alpha dr \tag{5.11}$$

and

$$\|w_i\|^2_{L^\infty(\Omega_i^{\delta\delta})} \le C\left(1+\log\left(\frac{H}{h}\right)\right)|u|^2_{H^1(\Omega_i^\delta)}. \tag{5.12}$$

For the inequality (5.12), we have used a well-known result (see Bramble [2])

$$\|u-\bar{u}_i\|^2_{L^\infty(\Omega_i^{\delta\delta})} \le \|u-\bar{u}_i\|_{L^\infty(\Omega_i^{\delta})} \le C\left(1+\log\left(\frac{H}{h}\right)\right)\|u-\bar{u}_i\|^2_{H^1(\Omega_i^\delta)}$$

and that $\bar{u}_i$ is the average of u on Ω_i^δ, i.e., a Friedrichs inequality,

$$\|u-\bar{u}_i\|^2_{H^1(\Omega_i^\delta)} \le C|u|^2_{H^1(\Omega_i^\delta)}.$$

Putting (5.11) and (5.12) together, we obtain

$$\sum_{K\in\Omega_i^{\delta\delta}} \frac{1}{r^2}\|w_i\|^2_{L^2(K)} \le C\left((1+\log(\delta+1))\left(1+\log\left(\frac{H}{h}\right)\right)\right)|u|^2_{H^1(\Omega_i^\delta)}. \tag{5.13}$$

For the case $K \subset \Omega^\delta_{i,non}$, if Ω_i^0 is a floating subdomain, which is to say that $\Omega^\delta_{i,non}$ does not touch $\partial\Omega$, then $\bar{\theta}_i - \theta_i$ is zero. If $\Omega^\delta_{i,non}$ touches the boundary $\partial\Omega$, then the estimate becomes

$$\begin{aligned} |v_i|^2_{H^1(\Omega^\delta_{i,non})} &\le C\left(|u|^2_{H^1(\Omega^\delta_{i,non})} + |\bar{u}_i|^2|\phi^i|^2_{H^1(\Omega^\delta_{i,non})}\right) \\ &\le C\left(1+\log\left(\frac{H}{h}\right)\right)|u|^2_{H^1(\Omega_i^\delta)}. \end{aligned} \tag{5.14}$$

Here we have used Lemma 5.3 and that for the cases $i \in \partial\Omega$ we can use a Poincaré inequality to obtain

$$\sum_{i\in\partial\Omega}|\bar{u}_i|^2 \le C\sum_{i\in\partial\Omega}\frac{1}{H^2}\|u\|^2_{L^2(\Omega_i^\delta)} \le C\sum_{i\in\partial\Omega}|u|^2_{H^1(\Omega_i^\delta)} \le C|u|^2_{H^1(\Omega)}. \tag{5.15}$$

Here we have introduced the notation $i \in \partial\Omega$ to denote the subdomains Ω_i^0 that touch the boundary $\partial\Omega$ with a face.

Putting everything together, we have shown that

$$\sum_{i=1}^{N} |v_i|_{H^1(\Omega)}^2 \leq C\left(\left(\frac{H}{(2\delta+1)h}\right)\right)|u|_{H^1(\Omega)}^2 + C(1+\log(\delta+1))\left(1+\log\left(\frac{H}{h}\right)\right)|u|_{H^1(\Omega)}^2. \tag{5.16}$$

We remark that the bound (5.3) follows from (5.16). To see this, we use that $v_0 = u - \sum_i v_i$, the triangular inequalities, and (5.16) to obtain (5.3).

We now consider the bound for the one-level RASHO method, i.e., to bound $\sum_{i=1}^N u_i$. Note that

$$\sum_{i=1}^{N} u_i = \sum_{i=1}^{N} v_i + \sum_{i=1}^{N} \bar{u}_i \phi^i.$$

For the second sum above, we first use Lemma 5.3 to obtain

$$\sum_{i=1}^{N} |\bar{u}_i \phi^i|_{H^1(\Omega)}^2 \leq C\left(1+\log\left(\frac{H}{h}\right)\right)\sum_{i\in\partial\Omega} |\bar{u}_i|^2 + C\left(1+\log(\delta+1)+\frac{H}{(2\delta+1)h}\right)\sum_{i=1}^{N}|\bar{u}_i|^2.$$

We then use the Cauchy–Schwarz and Friedrichs inequalities to obtain

$$\begin{aligned}\sum_{i=1}^{N}|\bar{u}_i|^2 = \sum_{i=1}^{N}\left(\frac{1}{|\Omega_i^\delta|}\int_{\Omega_i^\delta} u dx\right)^2 &\leq C\sum_{i=1}^{N}\frac{1}{H^2}\|u\|_{L^2(\Omega_i^\delta)}^2 \\ &\leq C\frac{1}{H^2}\|u\|_{L^2(\Omega)}^2 \leq C\frac{1}{H^2}|u|_{H^1(\Omega)}^2.\end{aligned}$$

For the cases $i \in \partial\Omega$, we use (5.15). The inequality (5.5) then follows. □

5.3. The main theorem. We state the main theorem of this paper here. The proof follows directly from all the abstract Schwarz theory given by Lemmas 5.1, 5.2, and 5.4.

THEOREM 5.1. *The RASHO operators $\widetilde{P}^\delta$, $\widetilde{P}_C^\delta$, and $\widetilde{P}_{hyb}^\delta$ are symmetric in the inner product $a(\cdot,\cdot)$, nonsingular, and bounded from below and above:*

$$C_0^{-2}a(u,u) \leq a(\widetilde{P}_C^\delta u, u) \leq C_1 a(u,u) \qquad \forall u \in \widetilde{\mathcal{V}}^\delta,$$

$$\hat{C}_0^{-2}a(u,u) \leq a(\widetilde{P}^\delta u, u) \leq \hat{C}_1 a(u,u) \qquad \forall u \in \widetilde{\mathcal{V}}^\delta,$$

and

$$\kappa(\widetilde{P}_{hyb}^\delta) \leq \kappa(\widetilde{P}_C^\delta).$$

Here

$$C_0^2 = C\left(\frac{H}{(2\delta+1)h} + (1+\log(\delta+1))\left(1+\log\left(\frac{H}{h}\right)\right)\right)$$

and

$$\hat{C}_0^2 = C\left((1+\log(\delta+1))\left(1+\log\left(\frac{H}{h}\right)\right) + \frac{1}{H^2}\left(1+\log(\delta+1)+\frac{H}{(2\delta+1)h}\right)\right).$$

The constants $C, C_1, \hat{C}_1 > 0$ are independent of h, H, and δ.

We remark that the corresponding convergence rate estimate for the regular one-level AS methods [11], in terms of the constant $\hat{C}_0$, is

$$\hat{C}_0^2 = C\left(1+\frac{1}{H(2\delta+1)h}\right),$$

and that for the two-level additive Schwarz method is

$$C_0^2 = C\left(1+\frac{H}{\delta h}\right).$$

The lower bound $\hat{C}_0^2$ of the one-level RASHO algorithm is theoretically slightly worse than the lower bound of regular AS algorithm in the case of large overlap, but roughly the same for small overlap. For small overlap, the lower bounds of both algorithms behave like $O(H/h)$. When the overlap gets larger, the RASHO scheme starts to feel the factor $\log(H/h)$, and the performance gets worse than the additive version for large overlap. On the other hand, the upper bound C_1 of RASHO is smaller than the upper bounds of AS. We can see this since $\widetilde{\mathcal{V}}_k^\delta \subset \mathcal{V}_k^\delta \ \forall k$ implies that the positive numbers ϵ_{ij} defined in Lemma 5.1 are smaller for RASHO than the corresponding ϵ_{ij} for AS. Consequently, the spectral radius $\mathcal{E}$ of RASHO is smaller. Because C_1 of RASHO is smaller, the numerical performance of RASHO presented in the next section is better than that of AS for the practical cases. Similar considerations also apply to the two-level RASHO methods.

6. Numerical experiments. In this section, we present some numerical results for solving the Poisson equation on the unit square with zero Dirichlet boundary conditions. We compare the performance of RASHO- and AS-preconditioned CG methods in terms of the number of iterations and the condition numbers. We pay particular attention to the dependence of the performance on the number of subdomains and the size of overlap.

We first discuss a few implementation issues related to the new preconditioner. In order to apply the RASHO/CG method, it is necessary to force the solution to belong to $\widetilde{\mathcal{V}}^\delta$. To do this, a pre-CG-computation is needed, and it is done through the formula (3.5). We note that $u = u^* - w \in \widetilde{\mathcal{V}}^\delta$ (see Lemma 3.1), and therefore we can apply the regular preconditioned CG to the RASHO-preconditioned system (3.9). The AS/CG is the classical AS preconditioned CG as described in [8]. We note that in the case $\delta = 0$, i.e., $ovlp = h$, RASHO and AS are the same.

The stopping condition for the CG method is to reduce the initial residual by a factor of 10^{-6}. The exact solution of the equation is $u(x,y) = e^{5(x+y)}\sin(\pi x)\sin(\pi y)$. All subdomain problems are solved exactly. The iteration counts (iter), condition numbers (cond), maximum (max) and minimum (min) eigenvalues of the preconditioned matrix are summarized in Tables 6.1–6.5.

From Tables 6.1, 6.2, and 6.3, it is clear that for overlap not too large and for mesh not too small, which is the case of practical interest, the one-level RASHO/CG outperforms the classical one-level AS/CG in terms of the iteration counts and condition numbers. In this case of small overlap, the condition number of RASHO is

Table 6.1

One-level RASHO- and AS-preconditioned CG for solving the Poisson equation on a 128×128 *mesh decomposed into* $2 \times 2 = 4$ *subdomains with overlap* = *ovlp. The AS/CG results are shown in ( ). The "+1" is for the preprocessing step needed for RASHO.*

ovlp	iter	cond	max	min
h	42 (42)	129.(129.)	1.98 (1.98)	0.0154 (0.0154)
$3h$	24+1 (28)	48.4 (86.3)	1.94 (4.00)	0.0402 (0.0464)
$5h$	20+1 (23)	33.3 (51.8)	1.91 (4.00)	0.0574 (0.0773)
$7h$	18+1 (20)	27.2 (37.0)	1.89 (4.00)	0.0694 (0.1081)

Table 6.2

One-level RASHO- and AS-preconditioned CG for solving the Poisson equation on a $32 * DOM \times 32 * DOM$ *mesh decomposed into* $DOM \times DOM$ *subdomains with overlap* = $3h$*, i.e.,* $\delta = 1$.

$DOM \times DOM$	iter	cond	max	min
2×2	19+1 (20)	26.8 (43.7)	1.89 (4.00)	0.0708 (0.0916)
4×4	39+1 (42)	86.9 (145.)	1.95 (4.00)	0.0225 (0.0276)
8×8	75+1 (78)	328. (550.)	1.97 (4.00)	0.0060 (0.0073)
16×16	147+1 (156)	1295 (2168.)	1.98 (4.00)	0.0015 (0.0018)

Table 6.3

One-level RASHO-and AS-preconditioned CG for solving the Poisson equation on an $n \times n$ *mesh decomposed into* 4×4 *subdomains with overlap* = $3h$*, i.e.,* $\delta = 1$.

$DOM \times DOM$	iter	cond	max	min
64×64	30+1 (29)	50.1 (72.2)	1.91 (4.00)	0.0382 (0.0554)
128×128	39+1 (40)	86.9 (145.)	1.95 (4.00)	0.0225 (0.0276)
256×256	53+1 (56)	159.9 (290.7)	1.98 (4.00)	0.0124 (0.0138)
512×512	74+1 (77)	305.6 (582.1)	1.99 (4.00)	0.0065 (0.00069)

Table 6.4

Two-level hybrid and additive RASHO for solving the Poisson equation on a $32 * DOM \times 32 * DOM$ *mesh decomposed into* $DOM \times DOM$ *subdomains with overlap* = $3h$*, i.e.,* $\delta = 1$*; the two-level additive RASHO results are shown in ( ).*

$DOM \times DOM$	iter	cond	max	min
2×2	27+1 (30+1)	24.2 (45.9)	1.82 (2.90)	0.0751 (0.0634)
4×4	32+1 (46+1)	27.2 (53.3)	1.80 (2.93)	0.0662 (0.0551)
8×8	33+1 (52+1)	28.4 (55.3)	1.80 (2.94)	0.0634 (0.0533)
16×16	33+1 (52+1)	28.8 (55.8)	1.80 (2.94)	0.0625 (0.0528)

Table 6.5

Two-level hybrid and additive RASHO CG for solving the Poisson equation on a 512×512 *mesh decomposed into* $16 \times 16 = 256$ *subdomains with overlap* = *ovlp. The two-level additive RASHO results are shown in ( ).*

ovlp	iter	cond	max	min
h	86 +1 (109+1)	307 (275.7)	1.96 (3.74)	0.0064 (0.0136)
$3h$	44 +1 (68+1)	48.0 (95.7)	1.87 (2.98)	0.0391 (0.0312)
$5h$	36 +1 (58+1)	32.8 (70.1)	1.83 (2.95)	0.0558 (0.0421)
$7h$	31 +1 (53+1)	27.3 (59.8)	1.80 (2.93)	0.0662 (0.0491)

almost twofold smaller than AS. This is an important result since it is easy to modify a (parallel) one-level AS/CG code to obtain a one-level RASHO/CG implementation. Although we do not have any parallel results to report here, we confidently predict that RASHO/CG would be even better than AS/CG on a parallel computer with distributed memory, since many less communications are required. Also the local solvers in RASHO are slightly cheaper, since the local solvers have slightly smaller numbers of unknowns than for the regular AS. From Table 6.4 we see that both the two-level hybrid and additive versions of RASHO attain scalability in terms of number of iterations when the number of subdomains becomes large; the hybrid version reaches the asymptotic behavior sooner than the additive version. The hybrid version is superior to the additive version since the number of iterations is much smaller. Finally, from Table 6.5 we see that larger overlap reduces dramatically the number of iterations.

REFERENCES

[1] S. Balay, K. Buschelman, W. D. Gropp, D. Kaushik, M. Knepley, L. C. McInnes, B. Smith, and H. Zhang, *The Portable Extensible Toolkit for Scientific Computing (PETSc)*, available online at http://www.mcs.anl.gov/petsc, 2003.

[2] J. Bramble, *A second order finite difference analogue of the first biharmonic boundary value problem*, Numer. Math., 9 (1966), pp. 236–249.

[3] S. Brenner and R. Scott, *The Mathematical Theory of Finite Element Methods*, Springer, New York, 1994.

[4] X.-C. Cai, M. Casarin, F. Elliot, and O. Widlund, *Overlapping Schwarz algorithms for solving Helmholtz's equation*, in Contemp. Math. 218, AMS, Providence, RI, 1998, pp. 391–399.

[5] X.-C. Cai, C. Farhat, and M. Sarkis, *A minimum overlap restricted additive Schwarz preconditioner and applications in 3D flow simulations*, in Contemp. Math. 218, AMS, Providence, RI, 1998, pp. 479–485.

[6] X.-C. Cai, T. P. Mathew, and M. V. Sarkis, *Maximum norm analysis of overlapping nonmatching grid discretizations of elliptic equations*, SIAM J. Numer. Anal., 37 (2000), pp. 1709–1728.

[7] X.-C. Cai and M. Sarkis, *A restricted additive Schwarz preconditioner for general sparse linear systems*, SIAM J. Sci. Comput., 21 (1999), pp. 792–797.

[8] M. Dryja and O. Widlund, *An Additive Variant of the Schwarz Alternating Method for the Case of Many Subregions*, Technical Report 339, also Ultracomputer Note 131, Department of Computer Science, Courant Institute, New York, 1987.

[9] M. Dryja, M. Sarkis, and O. Widlund, *Multilevel Schwarz methods for elliptic problems with discontinuous coefficients in three dimensions*, Numer. Math., 72 (1996), pp. 313–348.

[10] M. Dryja, B. F. Smith, and O. B. Widlund, *Schwarz analysis of iterative substructuring algorithms for elliptic problems in three dimensions*, SIAM J. Numer. Anal., 31 (1994), pp. 1662–1694.

[11] M. Dryja and O. B. Widlund, *Domain decomposition algorithms with small overlap*, SIAM J. Sci. Comput., 15 (1994), pp. 604–620.

[12] M. Dryja and O. Widlund, *Schwarz methods of Neumann–Neumann type for three-dimensional elliptic finite element problems*, Comm. Pure Appl. Math., 48 (1995), pp. 121–155.

[13] C. Farhat and F. Roux, *A method of finite element tearing and interconnecting and its parallel solution algorithm*, Internat. J. Numer. Methods Engrg., 32 (1991), pp. 1205–1227.

[14] A. Frommer and D. B. Szyld, *An algebraic convergence theory for restricted additive Schwarz methods using weighted max norms*, SIAM J. Numer. Anal., 39 (2001), pp. 463–479.

[15] G. Golub and C. Van Loan, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, MD, 1983.

[16] W. Gropp, D. Kaushik, D. Keyes, and B. Smith, *Performance modeling and tuning of an unstructured mesh CFD application*, in Proceedings of the SC2000 High Performance Networking and Computing Conference, Dallas, 2000, IEEE Computer Society, 2000.

[17] M. Lesoinne, M. Sarkis, U. Hetmaniuk, and C. Farhat, *A linearized method for the frequency analysis of three-dimensional fluid/structure interaction problems in all flow regimes*, Comput. Methods Appl. Mech. Engrg., 190 (2001), pp. 3121–3146.

[18] J. Mandel, *Balancing domain decomposition*, Comm. Numer. Methods Engrg., 9 (1993), pp. 233–241.

[19] J. Mandel, *Hybrid domain decomposition with unstructured subdomains*, in Contemp. Math. 157, AMS, Providence, RI, 1994, pp. 103–112.

[20] A. Quarteroni and A. Valli, *Domain Decomposition Methods for Partial Differential Equations*, Oxford Science Publications, Oxford, UK, 1999.

[21] M. Sarkis, *Partition of unity coarse spaces and Schwarz methods with harmonic overlap*, in the Proceedings of the Workshop in Domain Decomposition, ETH Zurich, 2001, Springer-Verlag, pp. 75–92.

[22] Y. Saad and M. H. Schultz, *GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 856–869.

[23] M. Sarkis, *Nonstandard coarse spaces and Schwarz methods for elliptic problems with discontinuous coefficients using non-conforming elements*, Numer. Math., 77 (1997), pp. 383–406.

[24] B. Smith, P. Bjørstad, and W. Gropp, *Domain Decomposition: Parallel Multilevel Methods for Elliptic Partial Differential Equations*, Cambridge University Press, Cambridge, UK, 1996.

SIAM J. Numer. Anal.
Vol. 41, No. 4, pp. 1232–1241

A NUMERICAL METHOD FOR AN INTEGRO-DIFFERENTIAL EQUATION WITH MEMORY IN BANACH SPACES: QUALITATIVE PROPERTIES[*]

E. CUESTA[†] and C. PALENCIA[†]

Abstract. A first order method is considered for the discretization in time of an integro-differential equation, which can be written as $D^\alpha u(t) = Au(t) + f(t)$, $1 < \alpha < 2$, where $A : D(A) \subset X \to X$ is a sectorial operator in a Banach space X. Qualitative properties of the numerical solution, such as contractivity and positivity, are studied. A numerical illustration is provided.

Key words. integro-differential equations, memory, parabolic, sectorial operator, Banach space, Euler method, fractional quadrature, positivity, contractivity

AMS subject classifications. 45K05, 65J10, 65D32

DOI. 10.1137/S0036142902402481

1. Introduction.

In the present paper we consider a numerical method for the time discretization of the problem

$$\left\{\begin{array}{rcl} u'(t) &=& \dfrac{1}{\Gamma(\alpha-1)}\displaystyle\int_0^t (t-s)^{\alpha-2} Au(s)\,ds + f(t), \qquad 0 \le t \le T, \\ u(0) &=& u_0 \in D(A), \end{array}\right. \tag{1}$$

where $A : D(A) \subset X \to X$ is a linear, closed operator in a complex Banach space X and $f : [0,T] \to X$. The parameter α lies in the open interval $(1,2)$. The operator A is assumed to be sectorial, and the precise hypothesis concerning A is given in section 2.

These equations with memory are of interest in connection with several applications (see [1, 12]). Formally, the equation in (1) can be viewed as (see [5, 6])

$$D^\alpha u(t) = Au(t) + D^{\alpha-1} f(t),$$

where D^α and $D^{\alpha-1}$ stand for the fractional time derivatives, defined in terms of the Riemann–Liouville operator (see [5, 7, 11, 13]). Thus, the equation in (1) is intermediate between the one with $\alpha = 1$,

$$\left\{\begin{array}{rcl} u'(t) &=& Au(t) + f(t), \\ u(0) &=& u_0, \end{array}\right.$$

and the one with $\alpha = 2$,

$$\left\{\begin{array}{rcl} u''(t) &=& Au(t) + f'(t), \\ u(0) &=& u_0, \\ u_t(0) &=& 0. \end{array}\right.$$

[*] Received by the editors February 11, 2002; accepted for publication (in revised form) November 19, 2002; published electronically August 22, 2003. The research of the authors was supported by project MCYT BFM2001-2013.

http://www.siam.org/journals/sinum/41-4/40248.html

[†] Departamento Matemática Aplicada y Computación, Universidad de Valladolid, 47005 Valladolid, Spain (eduardo@mat.uva.es, palencia@mac.cie.uva.es).

Numerical methods for (1) have been considered in [10, 14] in the context of self-adjoint operators in Hilbert spaces and in [3] in the context of sectorial operators in Banach spaces. Here we consider the method introduced in [10] and [14]. This scheme combines the backward Euler method with an appropriate quadrature rule for the integral (see [8]). The analysis in [10] can be extended to the present framework of sectorial operators and Banach spaces. Moreover, as remarked at the end of section 3, the convergence can also be proved from the results in [8, 9]. Because of this, we focus only on some qualitative properties of the method.

A brief review of the continuous problem (1) is given in section 2, where some new qualitative properties are also deduced. The numerical method is presented in section 3. It is shown, in section 4, that the values provided by the method can be expressed as certain averages of the continuous solution under discretization. From this representation we deduce that the numerical method inherits qualitative properties such as contractivity and positivity from the continuous solution. A numerical illustration is given in section 5.

2. Continuous problem.

In this section, for the convenience of the reader, we recall some basic facts about the solutions of (1). We also derive some new results concerning their qualitative behavior.

Notice that problem (1) is well posed (see [2, 12]) when A is sectorial and its spectral angle is small enough. To be precise, when there exist $\omega \in \mathbb{R}$, $0 < \theta < \pi(2-\alpha)/2$, and $M \geq 1$ such that, for $z \in \mathbb{C}$ outside the sector

$$\omega + S_\theta = \omega + \{\mu \in \mathbb{C} \;/\; |\arg(-\mu)| \leq \theta\},$$

the resolvent $(z-A)^{-1} : X \to X$ exists and

$$\|(z-A)^{-1}\| \leq \frac{M}{|z-\omega|}, \qquad z \notin \omega + S_\theta. \tag{2}$$

In the rest of the paper we assume that A is sectorial in this sense.

Suppose first that $f = 0$, and let $u : [0,T] \to X$ be a solution of (1). Assume that u is of exponential growth. Then taking the Laplace transform on both sides of (1) leads to (see [2, 6])

$$U(z) = z^{\alpha-1}(z^\alpha - A)^{-1}u_0, \qquad |\arg(z-\omega)| < (\pi-\theta)/\alpha, \tag{3}$$

where U stands for the Laplace transform of u. Thus, by the inversion formula, it turns out that

$$u(t) = E_\alpha(t)u_0, \qquad t \geq 0, \tag{4}$$

where $E_\alpha(t) : X \to X$, $t > 0$, is the bounded operator

$$E_\alpha(t) := \frac{1}{2\pi i}\int_\Gamma e^{tz} z^{\alpha-1}(z^\alpha - A)^{-1}\,dz,$$

Γ being a suitable path connecting $-i\infty$ with $+i\infty$.

For arbitrary $u_0 \in X$ we define the generalized solution of the corresponding Cauchy problem (1) by the expression in (4). The above discussion shows that a solution of exponential growth is necessarily the generalized solution corresponding to its initial value.

If $f \neq 0$, then it is well known that the solution is represented by means of the variation-of-constant formula

$$u(t) = E_\alpha(t)u_0 + \int_0^t E_\alpha(t-s)f(s)\,ds, \qquad t \geq 0. \tag{5}$$

For $\omega = 0$ and $t > 0$ we can take $\Gamma = \Gamma_t$ as the positive boundary of the union of S_θ and the circle $|z| = 1/t^\alpha$. Then, by (2), it is straightforward to prove that

$$\|E_\alpha(t)\| \leq CM, \tag{6}$$

where $C > 0$ stands for a constant depending only on α and θ. This estimate can also be proved by other means (see [12, Corollary 6.4]).

For $\omega < 0$ we can choose $\Gamma = \tilde{\Gamma}_t$ to be the boundary of the sector S_θ. Using again (2), this time we obtain

$$\|E_\alpha(t)\| \leq \frac{CM}{(2-\alpha)|\omega|t^\alpha}, \qquad t > 0,$$

which combined with (6) results finally in

$$\|E_\alpha(t)\| \leq \frac{CM}{1 + (2-\alpha)|\omega|t^\alpha}. \tag{7}$$

The case $\omega > 0$ is a bit more cumbersome. By arguments similar to the ones used in [2] we can establish now that

$$\|E_\alpha(t)\| \leq CM(1 + \ln^+(t\omega^{1/\alpha}))e^{t\omega^{1/\alpha}}. \tag{8}$$

In the remainder of the paper, for simplicity, we will assume that $\omega = 0$. When $\omega \neq 0$ the estimates we will derive are affected by similar factors to those appearing in (7) or (8).

It is also true that the family $E_\alpha(t)$, $t > 0$, is strongly continuous. Continuity at 0^+ holds in the case where A is densely defined.

To end this section we present some new estimates relating the qualitative behavior of $E_\beta(t)$ to the ones of $E_\alpha(t)$, $1 < \beta < \alpha < 2$.

Lemma 2.1. *For $0 < \mu < 1$, let $K_\mu : (0, +\infty) \times (0, +\infty) \to \mathbb{C}$ be the mapping defined by*

$$K_\mu(\sigma, t) := \frac{1}{2\pi i}\int_{-i\infty}^{+i\infty} e^{st - \sigma s^\mu}\,ds, \qquad \sigma > 0, \quad t > 0.$$

Then there exist constants $c = c(\mu) > 0$ and $C = C(\mu) > 0$ such that, for any $a > 0$,

$$K_\mu(\sigma, t) \leq C \min\left\{\frac{1}{t}, \frac{e^{at - ca^\mu\sigma}}{\sigma^{1/\mu}}\right\}, \qquad \sigma > 0, \quad t > 0.$$

Moreover,

$$K_\mu(\sigma, t) \geq 0, \qquad \sigma > 0, \quad t > 0.$$

Proof. The positivity of K_μ is provided by Proposition 2 in Chapter IX of [16].

On the other hand, for $a > 0$ fixed, let Γ_1 be the positive boundary of a sector S_φ with $0 < \varphi < \pi/2$ and $(\pi - \varphi)\mu \geq \pi/2$. Notice that there exists $C = C(\mu) > 0$ such that

$$|e^{st-\sigma s^\mu}| \leq Ce^{-|s|t\cos\varphi}, \qquad s \in \Gamma_1.$$

Therefore,

$$|K_\mu(\sigma, t)| \leq \frac{C}{2\pi}\int_{\Gamma_1} e^{-|s|t\cos\varphi}\,|ds| \leq \frac{C}{\pi t\cos\varphi}.$$

Besides, let us consider the path $\Gamma_2(\xi) = a + i\xi$, $-\infty < \xi < +\infty$. Now, for $s = a + i\xi \in \Gamma_2$, we have

$$|s^\mu| = (a^2 + \xi^2)^{\mu/2} \geq 2^{\mu/2-1}(a^\mu + |\xi|^\mu)$$

and $|\arg(s^\mu)| \leq \pi\mu/2$ so that

$$|e^{st-\sigma s^\mu}| \leq e^{at}e^{-ca^\mu\sigma}e^{-c|\xi|^\mu\sigma},$$

where $c = \cos(\pi\mu/2)/2^{\mu/2-1}$. Therefore,

$$|K_\mu(\sigma, t)| \leq \frac{1}{\pi}e^{at}\,e^{-ca^\mu\sigma}\int_0^{+\infty} e^{-c\sigma\xi^\mu}\,d\xi = \frac{1}{\pi}e^{at}\,e^{-ca^\mu\sigma}\frac{\Gamma(1/\mu)}{\mu(c\sigma)^{1/\mu}}. \qquad \square$$

THEOREM 2.2. *For $1/\alpha < \mu < 1$ and $t > 0$ there holds*

$$E_{\alpha\mu}(t) = \frac{1}{\Gamma(1-\mu)}\int_0^t \frac{1}{(t-\tau)^\mu}\left[\int_0^{+\infty} K_\mu(\sigma, \tau)E_\alpha(\sigma)\,d\sigma\right]d\tau. \tag{9}$$

Proof. Set $h_\mu(t) = t^{-\mu}/\Gamma(1-\mu)$ and

$$E_\alpha^\mu(t) = \int_0^{+\infty} K_\mu(\sigma, t)E_\alpha(\sigma)\,d\sigma, \qquad t > 0.$$

Since the integrand is strongly continuous and, because of the previous lemma, absolutely convergent (this is true even for $\omega \neq 0$), it turns out that the integral exists in the strong sense and that the operators $E_\alpha^\mu(t)$ are uniformly bounded in $t > 0$.

Notice that for fixed $\sigma > 0$ the Laplace transform of $t \to K_\mu(\sigma, t)$ is the mapping $s \to e^{-\sigma s^\mu}$. Therefore, the Laplace transform of $t \to E_\alpha^\mu(t)$ is

$$\mathcal{L}E_\alpha^\mu(s) = \int_0^{+\infty} e^{-\sigma s^\mu}E_\alpha(\sigma)\,d\sigma, \qquad \Re(s) > 0.$$

Recalling (3) we deduce that

$$\mathcal{L}E_\alpha^\mu(s) = s^{\mu(\alpha-1)}(s^{\alpha\mu} - A)^{-1}, \qquad \Re(s) > 0,$$

so that the Laplace transform of $h_\mu * E_\alpha^\mu$ is

$$\mathcal{L}(h_\mu * E_\alpha^\mu)(s) = s^{\mu-1}s^{\mu(\alpha-1)}(s^{\alpha\mu} - A)^{-1} = s^{\mu\alpha-1}(s^{\alpha\mu} - A)^{-1}, \qquad \Re(s) > 0,$$

which, by (3) applied to $\alpha\mu$, is the Laplace transform of $E_{\alpha\mu}$. $\square$

Corollary 2.3. *Let $u_0 \in X$, and assume that $\|E_\alpha(t)u_0\| \leq H$ for $t > 0$. Then, for $1/\alpha < \mu < 1$, there holds*

$$\|E_{\alpha\mu}(t)u_0\| \leq H, \qquad t > 0.$$

Proof. Notice that the equation in (2.2) with $A = 0$ reduces to

$$\frac{1}{\Gamma(1-\mu)} \int_0^t \frac{1}{(t-\tau)^\mu} \left[\int_0^{+\infty} K_\mu(\sigma,\tau)\, d\sigma \right] d\tau = 1.$$

Then, since $K_\mu \geq 0$, taking norms in (9) leads to

$$\|E_{\mu\alpha}(t)u_0\| \leq \frac{H}{\Gamma(1-\mu)} \int_0^t \frac{1}{(t-\tau)^\mu} \left[\int_0^{+\infty} K_\mu(\sigma,\tau)\, d\sigma \right] d\tau \leq H. \qquad \square$$

In the same way we can prove the next corollary.

Corollary 2.4. *Assume that X is an ordered Banach lattice (see* [15]*). If $u_0 \in X$ is such that*

$$E_\alpha(t)u_0 \geq 0, \qquad t > 0,$$

then, for $1/\alpha < \mu < 1$, we also have

$$E_{\alpha\mu}(t)u_0 \geq 0, \qquad t > 0.$$

It is possible to prove that Theorem 2.2 and its corollaries are valid even for $\alpha = 2$, under the hypothesis that now A generates a cosine family (see [4]). Thus, for instance, since the one-dimensional wave equation with Dirichlet conditions

$$\left\{ \begin{array}{rcll} u_{tt}(t,x) & = & u_{xx}(t,x), & 0 \leq x \leq L, \quad t \geq 0, \\ u(0,x) & = & u_0(x), & 0 \leq x \leq L, \\ u_t(0,x) & = & 0, & 0 \leq x \leq L, \\ u(t,0) & = & 0, & t \geq 0, \\ u(t,L) & = & 0, & t \geq 0, \end{array} \right.$$

is contractive with respect to the maximum norm, i.e.,

$$\|u(t,\cdot)\|_\infty \leq \|u_0\|_\infty, \qquad t \geq 0,$$

Corollary 2.3 shows that $\|E_\beta(t)\| \leq 1$, $t \geq 0$, $1 < \beta < 2$, with respect to the maximum norm, where $E_\beta(t)$ is the evolution operator corresponding to (1) and the operator $A\phi := \phi''$ acting on the domain

$$D(A) = \{\phi \in \mathcal{C}[0,L] \;/\; \phi'' \in \mathcal{C}[0,L] \text{ and } \phi(0) = \phi(L) = 0\}.$$

In the case of Neumann boundary conditions, the one-dimensional wave equation preserves positivity. Therefore, by Corollary 2.4, this property remains valid for the corresponding $E_\beta(t)$, $t > 0$, for $1 < \beta < 2$. These qualitative results extend those in [5, 6].

3. The numerical method. The numerical method we propose for (1) has already been considered in [10] and [14] in the context of Hilbert spaces and self-adjoint operators. The basic idea is to combine the classical backward Euler method with a suitable quadrature rule for approximating the integral term, which is the fractional quadrature rule generated by the rectangle rule (see [2, 8]). The resulting numerical method reads

$$\frac{U_n - U_{n-1}}{\tau} = \sum_{j=1}^{n} q_{n-j}^{(\alpha-1)} A U_j + \tau f(t_n), \qquad 0 < n \leq [T/\tau], \tag{10}$$

where $\tau > 0$ stands for the step size, $t_n = n\tau$, and $U_n \in X$ are the approximations to $u(t_n)$ we are looking for. The weights $q_n^{(\alpha-1)}$ are given by

$$\sum_{n=0}^{+\infty} q_n^{(\alpha-1)} z^n := \left(\frac{\tau}{\delta(z)} \right)^{\alpha-1},$$

where

$$\delta(z) := 1 - z.$$

Therefore, the weights turn out to be

$$q_n^{(\alpha-1)} := \tau^{\alpha-1} (-1)^n \binom{1-\alpha}{n}.$$

As starting value U_0 we can take either u_0 or an available approximation to u_0.

Notice that (10) is an implicit scheme. To obtain U_n from $U_0, U_1, \ldots, U_{n-1}$ we must solve the linear equation

$$(I - \tau^\alpha A) U_n = U_{n-1} + \tau \sum_{j=1}^{n-1} q_{n-j}^{(\alpha-1)} A U_j + \tau f(t_n), \tag{11}$$

which, since $\omega = 0$, possesses a unique solution (in general, for $\omega \neq 0$, a sufficient condition for the uniqueness and solvability of (11) is $\max\{\omega, 0\} \cdot \tau^\alpha < 1$).

For the analysis of the method it is convenient to rewrite (10) in terms of generating functions. To this end, without loss of generality, we assume that f and u are defined on $[0, +\infty)$. Thus, set

$$U(z) := \sum_{n=1}^{+\infty} U_n z^n, \qquad F(z) := \sum_{n=1}^{+\infty} f(t_n) z^n, \qquad Q(z) := \frac{\tau}{\delta(z)}.$$

Multiplying (10) by z^n and summing up in n we obtain

$$U(z) - zU(z) - zU_0 = \tau Q(z)^{\alpha-1} A U(z) + \tau F(z)$$

or

$$(I - Q(z)^\alpha A) U(z) = \frac{z}{1-z} U_0 + \frac{\tau}{1-z} F(z).$$

Therefore, since

$$I - Q(z)^\alpha A = Q(z)^\alpha \left(\frac{1}{Q(z)^\alpha} - A \right),$$

and since, for $|z| = r < 1$, we have $\Re((1-z)/\tau) > 0$, it is clear that $I - Q(z)^\alpha A$ is invertible and that

$$(I - Q(z)^\alpha A)^{-1} = \frac{1}{Q(z)^\alpha}\left(\frac{1}{Q(z)^\alpha} - A\right)^{-1}$$

is a holomorphic operator-valued mapping. Cauchy's formula shows now that

$$\frac{z}{1-z}(I - Q(z)^\alpha A)^{-1} = \sum_{n=1}^{+\infty} D_n^{(\alpha)} z^n, \qquad |z| < 1,$$

where the bounded operators $D_n^{(\alpha)} : X \to X$, $n \geq 1$, are given by

$$D_n^{(\alpha)} = \frac{1}{2\pi i}\int_{|z|=r} \frac{1}{(1-z)z^n}(I - Q(z)^\alpha A)^{-1}\, dz, \qquad 0 < r < 1.$$

Going back to (11) we obtain the following expression for the numerical approximation:

$$U_n = D_n^{(\alpha)} U_0 + \tau \sum_{j=1}^{n} D_{n+1-j}^{(\alpha)} f(t_j), \qquad n \geq 1. \tag{12}$$

The above representation shows that the numerical scheme makes sense even for starting values $u_0 \in X$ which are not in $D(A)$.

It is also of interest to connect this method with the approach in [8, 9]. Assume that $U_0 = 0$. Now, in view of (5), we have

$$u(t) = \int_0^t E_\alpha(t-s) f(s)\, ds, \qquad t \geq 0.$$

Discretizing this convolution by the method in [8, 9] leads to approximations

$$U_n = \tau \sum_{j=0}^{n} L_{n-j}^{(\alpha)} f(t_j) \simeq u(t_n), \tag{13}$$

where the operators $L_n^{(\alpha)} : X \to X$ are defined by

$$\sum_{n=0}^{+\infty} L_n^{(\alpha)} z^n := \frac{1}{\tau} F_\alpha\left(\frac{\delta(z)}{\tau}\right), \tag{14}$$

and F_α stands for the Laplace transform of E_α, i.e.,

$$F_\alpha(\xi) := \int_0^{+\infty} e^{-\xi t} E_\alpha(t)\, dt = \xi^{\alpha-1}(\xi^\alpha - A)^{-1}.$$

Since

$$\frac{1}{\tau} F_\alpha\left(\frac{\delta(z)}{\tau}\right) = \frac{1}{1-z}(I - Q(z)^\alpha A)^{-1},$$

comparison of (13) with (12) shows readily that

$$D_n^{(\alpha)} = L_{n-1}^{(\alpha)}, \qquad n \geq 1, \tag{15}$$

a useful result we use later.

Finally, notice that (15), together with the stability of the method (see the next section), allows us to prove that the method is convergent of first order just by using Theorem 3.1 in [8]. Besides, Theorem 4.1 in [8] yields an optimal estimate in the case where $u_0 \notin D(A)$.

4. Qualitative behavior of the numerical solutions. The next theorem provides a representation of the discrete evolution operators $D_n^{(\alpha)}$ as an average of the continuous ones $E_\alpha(t)$, $t > 0$. This, combined with (12), allows us to obtain several interesting properties concerning the qualitative behavior of U_n.

For fixed $\tau > 0$, we set

$$\rho_n(t) := e^{-t/\tau} \left(\frac{t}{\tau}\right)^{n-1} \frac{1}{\tau(n-1)!}, \qquad t \geq 0, \quad n \geq 1.$$

Notice that $\rho_n(t) \geq 0$ and that

$$\int_0^{+\infty} \rho_n(t)\, dt = 1.$$

Theorem 4.1. *For $n \geq 1$ there holds*

$$D_n^{(\alpha)} = \int_0^{+\infty} E_\alpha(t) \rho_n(t)\, dt.$$

Proof. By definition (14), we have

$$\sum_{n=0}^{+\infty} L_n^{(\alpha)} z^n = \frac{1}{\tau} \int_0^{+\infty} E_\alpha(t) e^{-t\delta(z)/\tau}\, dt.$$

Then, since

$$\frac{1}{\tau} e^{-t\delta(z)/\tau} = \sum_{n=0}^{+\infty} \rho_{n+1}(t) z^n, \qquad t > 0,$$

we obtain

$$L_n^{(\alpha)} = \int_0^{+\infty} \rho_{n+1}(t) E_\alpha(t)\, dt, \qquad n \geq 0,$$

and the theorem is now clear because of (15). □

From this theorem we derive several corollaries whose proofs are obvious. For simplicity, we consider only $f = 0$. Related results for nonhomogeneous problems could be obtained by using (12).

Corollary 4.2. *Let $U_0 \in X$, and assume that $\|E_\alpha(t) U_0\| \leq H$, $t > 0$. Then*

$$\|D_n^{(\alpha)} U_0\| \leq H, \qquad n \geq 1.$$

In particular, the previous corollary shows that

$$\|D_n^{(\alpha)}\| \leq \sup_{t \geq 0} \|E_\alpha(t)\|, \qquad n \geq 1,$$

i.e., that the numerical scheme is stable. Using (6), we see that the stability constant is less than or equal to CM. Finally, notice that if $E_\alpha(t)$, $t > 0$, are contractions (recall Theorem 2.2), we also have that $D_n^{(\beta)}$, $n \geq 1$, are contractions for $1 < \beta \leq \alpha$.

Corollary 4.3. *Assume that X is an ordered Banach lattice (see* [15]*) and that for some $U_0 \in X$ we have $E_\alpha(t) U_0 \geq 0$, $t \geq 0$. Then*

$$D_n^{(\alpha)} U_0 \geq 0, \qquad n \geq 1.$$

Corollary 4.3 shows in particular that if $E_\alpha(t) \geq 0$, $t \geq 0$ (i.e. the operators $E_\alpha(t)$ are order-preserving), then $D_n^{(\alpha)} \geq 0$, $n \geq 1$. To illustrate the scope of Corollary 4.3 in situations where $E_\alpha(t)$ are not order-preserving let us consider the three-dimensional wave equation. For radial initial data

$$u_0(r) = \frac{f(r)}{r}$$

the solution of

$$\begin{cases} u_{tt}(t,r) &= c^2 \Delta u(t,r), \\ u(0,r) &= u_0(r), \\ u_t(0,r) &= 0 \end{cases}$$

turns out to be

$$u(t,r) = \frac{1}{2r}\left[f(r+ct) + f(r-ct)\right].$$

Therefore, if $u_0 \geq 0$, we have that $u(r,t) \geq 0$. Then, by Theorem 2.2, we deduce that $E_\alpha(t)u_0 \geq 0$, $1 < \alpha < 2$, where $E_\alpha(t)$ is the evolution operator that corresponds to $A = \Delta$ in (1). Now Corollary 4.3 shows that $D_n^{(\alpha)} u_0 \geq 0$, $n \geq 1$.

5. Numerical illustration. Let us consider the two-dimensional problem

$$\begin{cases} u_t(t,x,y) &= \dfrac{1}{\Gamma(\alpha-1)} \displaystyle\int_0^t (t-s)^{\alpha-2} \Delta u(s,x,y)\, ds, \qquad (x,y) \in \Omega, \\ u(0,x,y) &= u_0(x,y), \qquad (x,y) \in \Omega, \end{cases}$$

in the square $\Omega = [0,1] \times [0,1]$ with homogeneous Neumann boundary condition

$$D_n u(t,x,y), \qquad (x,y) \in \partial\Omega, \quad t \geq 0.$$

As an initial condition we take the indicator mapping of the subsquare $[1/3, 1/7] \times [1/3, 1/7]$. Notice that, by the maximum principle, for $\alpha = 1$ the solution is nonnegative. This problem is fully discretized by using first centered finite differences in space, with parameter h, and then applying method (10) to the resulting semidiscrete problem with step size τ.

Let A_h be the matrix corresponding to the discrete Laplacian, and denote by $E_{\alpha,h}(t)$ the corresponding evolution operators. Since A_h is symmetric, nonpositive, and diagonally dominant, it is well known that $E_{1,h}(t) = e^{tA_h}$ preserves the positivity. For $\alpha = 1.1, h = 1/50$, and $\tau = 1/200$ it turns out that $\min U_6 = -0.1164$. Because of Theorem 4.1 we deduce that $E_{\alpha,h}(t)$ cannot be positive for $1.1 \leq \alpha < 2$. Thus, in view of Corollary 4.3, this experiment suggests that $E_\alpha(t)$ is not positive either for $1.1 \leq \alpha < 2$. It is reasonable to conjecture that, in fact, $E_\alpha(t)$ does not preserve the positivity for $1 < \alpha < 2$.

Acknowledgment. The authors wish to express their gratitude to Prof. Ch. Lubich for several remarks which greatly improved the presentation of the paper.

REFERENCES

[1] C. Chen and T. Shi, *Finite Element Methods for Integro–Differential Equations*, World Scientific, Singapore, 1997.

[2] E. Cuesta, *Métodos Lineales Multipaso para Ecuaciones Integro–Diferenciales de Orden Fraccionario en Espacios de Banach*, Ph.D. Thesis, Universidad de Valladolid, Valladolid, Spain, 2001.

[3] E. Cuesta and C. Palencia, *A fractional trapezoidal rule for integro–differential equations of fractional order in Banach spaces*, Appl. Numer. Math., 45 (2003), pp. 139–159.

[4] H. O. Fattorini, *Second Order Linear Differential Equations in Banach Spaces*, North–Holland, Amsterdam, 1985.

[5] Y. Fujita, *Integro–differential equation which interpolates the heat equation and the wave equation*, Osaka J. Math., 27 (1990), pp. 319–327.

[6] Y. Fujita, *Integro–differential equation which interpolates the heat equation and the wave equation* II, Osaka J. Math., 27 (1990), pp. 797–804.

[7] J. L. Lavoie, T. J. Osler, and K. Tremblay, *Fractional derivatives and special functions*, SIAM Rev., 18 (1976), pp. 240–268.

[8] Ch. Lubich, *Convolution quadrature and discretized operational calculus* I, Numer. Math., 52 (1988), pp. 129–145.

[9] Ch. Lubich, *On convolution quadrature and Hille–Phillips operational calculus*, Appl. Numer. Math., 9 (1992), pp. 187–199.

[10] Ch. Lubich, I. H. Sloan, and V. Thomée, *Nonsmooth data error estimates for approximations of an evolution equation with a positive type memory term*, Math. Comp., 65 (1996), pp. 1–17.

[11] I. Podlubny, *Fractional Differential Equations*, Academic Press, San Diego, 1999.

[12] J. Prüss, *Evolutionary Integral Equations and Applications*, Birkhäuser, Basel, 1993.

[13] B. Ross, *Fractional calculus*, Math. Mag., 50 (1977), pp. 115–122.

[14] J. M. Sanz-Serna, *A numerical method for a partial integro-differential equation*, SIAM J. Numer. Anal., 25 (1988), pp. 319–327.

[15] H. H. Schaefer, *Topological Vector Spaces*, Springer-Verlag, New York, 1971.

[16] K. Yoshida, *Functional Analysis*, 6th ed., Springer-Verlag, Berlin, 1980.

SIAM J. NUMER. ANAL.
Vol. 41, No. 4, pp. 1242–1262

VARIABLE PRECONDITIONING VIA QUASI-NEWTON METHODS FOR NONLINEAR PROBLEMS IN HILBERT SPACE*

JÁNOS KARÁTSON† AND ISTVÁN FARAGÓ†

Abstract. The aim of this paper is to develop stepwise variable preconditioning for the iterative solution of monotone operator equations in Hilbert space and apply it to nonlinear elliptic problems. The paper is built up to reflect the common character of preconditioned simple iterations and quasi-Newton methods. The main feature of the results is that the preconditioners are chosen via spectral equivalence. The latter can be executed in the corresponding Sobolev space in the case of elliptic problems, which helps both the construction and convergence analysis of preconditioners. This is illustrated by an example of a preconditioner using suitable domain decomposition.

Key words. variable preconditioning, quasi-Newton methods, iterative methods in Hilbert space, nonlinear elliptic problems

AMS subject classifications. 35J65, 65J15

DOI. 10.1137/S0036142901384277

1. Introduction.

The aim of this paper is to develop stepwise variable preconditioning for the iterative solution of monotone operator equations

$$F(u) = 0 \tag{1}$$

in Hilbert space and apply it to nonlinear elliptic boundary value problems.

Nonlinear elliptic problems arise in many applications in physics and other fields, for instance in elastoplasticity, magnetic potential equations, and flow problems. The most frequently used numerical methods for nonlinear elliptic problems rely on some discretized form of the problem, whose solution is obtained by an iterative method. Simple iteration is often able to yield favorable speed of global convergence if supplied with suitable preconditioning, and in these cases its usage can be justified versus Newton's method owing to the extra work of forming the Jacobians (see, e.g., [2, 5]). Hence, similarly to linear problems, preconditioning is most times a crucial element of the construction of the iterative method. The choice of preconditioners is often helped by Hilbert space background, which helps both the construction of methods and the study of convergence. A typical example of this is the Sobolev gradient technique [25, 26]. (For the authors' related results see, e.g., [6, 15, 20].) In the case of monotone operators a natural kind of preconditioning is based on spectral equivalence, in an analogous way to symmetric linear equations. Namely, preconditioners are chosen to be globally spectrally equivalent to the derivatives $F'(u)$ of the operator in each point. A Hilbert space framework has been developed for this in [21], in which preconditioners for the discretized elliptic systems are found as projections of linear operators chosen as preconditioners for the original nonlinear differential operator.

The above described simple iterations are globally preconditioned in the sense that the preconditioners are the same in each step and rely on the global behavior of

*Received by the editors January 29, 2001; accepted for publication (in revised form) January 14, 2003; published electronically August 22, 2003. This research was supported by the Hungarian National Research Funds AMFK under Magyary Zoltán Scholarship and by OTKA under grants F022228 and T031807.

http://www.siam.org/journals/sinum/41-4/38427.html

†Department of Applied Analysis, ELTE University, H-1518 Budapest, Hungary (karatson@cs.elte.hu, faragois@cs.elte.hu).

$F'(u)$. However, this may be insufficient since the global convergence quotient may be very poor, as it is, e.g., for magnetic potential equations [23]. This insufficiency demands the stepwise improvement of contractivity, which necessarily involves the local properties of the Jacobians during the iteration. The stepwise comparison to $F'(u_n)$ leads to the framework of Newton-like or inexact Newton methods.

Inexact Newton methods, coupled with damping when global convergence is required, form a class that encompasses most iterative methods. A general description of these methods is found, e.g., in [10, 12], and with applications to BVPs in [1, 14]. The scope of these methods involves two main areas. In the first case the auxiliary equations contain exactly $F'(u_n)$ and inexactness comes from solving them approximately, often by an inner iteration. The other area (quasi-Newton methods) involves approximate Jacobians when they are not known exactly or in order to reduce work. (Several related methods are discussed in [13].) The generalization of Newton's method to Hilbert spaces has long been known [19], and many inexact versions can be put through as well.

The problem of preconditioning is connected to the second (quasi-Newton or approximate Jacobian) approach of inexact Newton methods. In both cases the iterative sequences are of the form

$$u_{n+1} = u_n - B_n^{-1} F(u_n) \qquad (n \in \mathbf{N}) \tag{2}$$

(which in fact contains all reasonable one-step iterative methods for (1)). The difference in approach is that preconditioning uses B_n to improve contractivity, whereas in quasi-Newton methods B_n has to approximate $F'(u_n)$. However, for elliptic problems, B_n required by the two methods may be similar, as is suggested by the preceding considerations and will be the case in our investigations. (Concerning the one-step sequence (2), we note that CG-type multistep methods in this context are in general unable to increase the order of convergence [9].)

Our investigation concerns variable preconditioning of iterations in Hilbert space, motivated by Sobolev space methods for nonlinear elliptic problems. The main feature of our results is that the preconditioners are chosen via spectral equivalence. This construction exploits the ellipticity properties of the equation. We note that our results also reflect more directly the way in which (2) shares equally the characters of simple and Newton iterations.

The main difficulty encountered in the convergence proof is that the variable preconditioners yield natural contractivity in stepwise different norms that are unable to produce a common estimate. Hence a transition is required to norms that can be compared to a common one. This will involve background investigations of suitable energy norms.

The approximate Jacobian approach, to which our investigation belongs, is relevant in applications to elliptic problems. Although the linear elliptic operators $F'(u)$ are then known exactly, it is worth looking for suitable approximations of $F'(u_n)$ in order to have simpler auxiliary equations. The use of Sobolev space background for such preconditioners seems an efficient approach for elliptic problems. This is the main scope of application for our method, the idea being analogous to the one suggested in [21] for simple iterations. Namely, preconditioning matrices can be obtained as projections of linear preconditioning operators chosen for the BVP itself on the continuous level, i.e., in the corresponding Sobolev space. In this way the original properties of the differential operator (mostly, those of its coefficients) can be exploited before discretization, and the obtained conditioning properties are mesh independent. As a main example, the operators B_n can be close to the Laplacian regarding their structure: for this purpose one may look for them as diagonal (scalar) coefficient operators with piecewise constant coefficients. (We note that the solution of the auxiliary linear systems in the steps of the iteration can rely on efficient standard methods, since these are highly developed; see, e.g., [3, 22]. In [3, 8] there is also discussed variable preconditioning in the context of linear equations.) We note that a summary on preconditioning operators is given in the book [16].

The paper is organized as follows. In section 2, first a result on simple iterations is quoted as a starting point; then some required properties of linear operators are given. The Hilbert space results on variable preconditioning are found in sections 3 and 4. First, section 3 provides local linear convergence using fixed spectral bounds for preconditioning. Although this theorem might be essentially derived from the one to come in section 4, it is worth formulating for two reasons. First of all, it illustrates more lucidly the preconditioning role of quasi-Newton methods, the result being an exact analogue of the quoted theorem on simple iterations in section 2. Besides, technical background is clearer when developed first for this simpler case. Section 4 contains the general method: global convergence up to second order is obtained using damped iteration and variable spectral bound preconditioning. Its proof is provided using the preceding technical background in the framework of damped quasi-Newton methods. Finally, in section 5 we apply the results to nonlinear elliptic boundary value problems. First we derive a general convergence result in Sobolev space and then give the construction of piecewise constant coefficient preconditioning operators using suitable domain decomposition. A numerical example illustrates the convergence results.

2. Preliminaries.

Let H be a Hilbert space with norm $\|\,.\,\|$. The notation

$$\langle u, v\rangle_A = \langle Au, v\rangle$$

will be used for the energy inner product of a self-adjoint positive operator. The corresponding norm has the obvious notation $\|\,.\,\|_A$.

2.1. Motivation: Simple preconditioning.

The following theorem gives a linear convergence result for preconditioned simple iterations.

THEOREM 2.1. *Let H be a real Hilbert space. Let the nonlinear operator $F : H \to H$ have a Gâteaux derivative satisfying the following properties:*

(i) *For any $u \in H$ the operator $F'(u)$ is self-adjoint.*

(ii) *(Ellipticity.) There exist constants $\Lambda \geq \lambda > 0$ satisfying*

$$\lambda\|h\|^2 \leq \langle F'(u)h, h\rangle \leq \Lambda\|h\|^2 \qquad (u, h \in H).$$

Denote u^* the unique solution of equation $F(u) = 0$. Assume that B is a self-adjoint linear operator satisfying

$$m\langle Bh, h\rangle \leq \langle F'(u)h, h\rangle \leq M\langle Bh, h\rangle \qquad (u, h \in H) \tag{3}$$

with some constants $M \geq m > 0$. Then for any $u_0 \in H$, the following sequence converges linearly to u^*:

$$u_{n+1} = u_n - \frac{2}{M+m} B^{-1}F(u_n) \qquad (n \in \mathbf{N}).$$

Namely,

$$\|u_n - u^*\| \leq C \cdot \left(\frac{M-m}{M+m}\right)^n \qquad (n \in \mathbf{N}) \tag{4}$$

with some constant $C > 0$.

Proof. The theorem is a special case of Theorem 3.2 in [21], where F itself is not assumed to be Gâteaux differentiable. Incidentally, in this form it follows easily from the well-known special case (see, e.g., [17]) when $B = I$ and (3) is no more than condition (ii). Namely, $B^{-1}F$ inherits the properties of F in the energy space H_B of the operator B (i.e., w.r.t. the inner product $\langle u, v\rangle_B = \langle Bu, v\rangle$).

We note that the basis of the convergence estimate (4) is the contractivity of the operator $I - \frac{2}{M+m} B^{-1}F$ with constant $\frac{M-m}{M+m}$ w.r.t. the energy norm $\| \, . \, \|_B$, which implies (4) in the $\| \, . \, \|_B$-norm. The equivalence of the two norms then yields (4).

Theorem 2.1 can be used for the preconditioning of the iterative solution of nonlinear elliptic BVPs. The straightforward application is to consider the discretized version of the problem and use the finite dimensional case of the theorem with a suitable preconditioning matrix B. The Hilbert space setting can be applied if F is the weak form of the differential operator and B is a weak linear elliptic operator defined in $H^1_0(\Omega)$ by

$$\langle Bh, v\rangle = \int_\Omega G(x)\, \nabla h \cdot \nabla v \qquad (h, v \in H^1_0(\Omega)),$$

where the coefficient matrix $G(x)$ can be chosen following the properties of $F'(u)$. Then for any discretization of the BVP, a suitable preconditioning matrix can be obtained as the projection of B under the same discretization, and this yields a mesh independent convergence estimate. Preconditioning strategies of this kind are summarized in [21]. (We note that often $B = I$ is already suitable in $H^1_0(\Omega)$; i.e., the original operator in strong form is preconditioned by the minus Laplacian. This works, e.g., for problems in plasticity [15]. The discrete Laplacian preconditioner is connected to the Sobolev gradient idea, developed for least-squares methods [25].) In both approaches the preconditioners so obtained are favorable, provided that M/m is not very large.

However, this kind of preconditioning is insufficient if M/m is large, and the latter may be unimprovable globally. This insufficiency demands the stepwise improvement of contractivity, i.e., varying B during the iteration to produce better spectral bounds. Since this modification involves the local properties of the Jacobians during the iteration (i.e., stepwise comparison to $F'(u_n)$), it leads to the framework of quasi-Newton methods.

As mentioned in the introduction, the following difficulty is encountered in this generalization. The variable preconditioners B_n yield contractivity in the above way in the stepwise different $\| \, . \, \|_{B_n}$ norms; hence we must verify contractivity in other norms that can be compared to a common one. Some properties required for this are given in the next subsection.

2.2. Properties of spectrally equivalent operators.

We formulate two lemmas for operators in a Hilbert space H satisfying the following spectral equivalence condition:

(C1) A and B be are self-adjoint linear operators in H with positive lower bound, and there exist constants $M \geq m > 0$ such that

$$m\langle Bh, h\rangle \leq \langle Ah, h\rangle \leq M\langle Bh, h\rangle \qquad (h \in H).$$

LEMMA 2.2. *Let the operators A and B satisfy condition* (C1). *Then*

$$m\langle A^{-1}h, h\rangle \leq \langle B^{-1}h, h\rangle \leq M\langle A^{-1}h, h\rangle \qquad (h \in H). \tag{5}$$

Proof. We prove only the right side of (5); the left one is similar. Let $v \in H$. Setting $h = B^{-1/2}v$, condition (C1) yields

$$\|A^{1/2}B^{-1/2}v\|^2 = \|A^{1/2}h\|^2 \le M\|B^{1/2}h\|^2 = M\|v\|^2.$$

Using also that for arbitrary bounded linear operator C there holds

$$\|C\| = \|C^*\|$$

and setting $C = B^{-1/2}A^{1/2}$, we obtain

$$\|B^{-1/2}A^{1/2}\|^2 = \|(B^{-1/2}A^{1/2})^*\|^2 = \|A^{1/2}B^{-1/2}\|^2 \le M,$$

which implies that

$$\langle B^{-1}h, h\rangle = \|B^{-1/2}A^{1/2}A^{-1/2}h\|^2 \le M\|A^{-1/2}h\|^2 = M\langle A^{-1}h, h\rangle \qquad (h \in H).$$

Lemma 2.3. *Let the operators A and B satisfy condition* (C1). *Then*

$$\left\| I - \frac{2}{M+m}AB^{-1} \right\|_{A^{-1}} \le \frac{M-m}{M+m}, \tag{6}$$

where I is the identity operator.

Proof. Let $C = \frac{M+m}{2}B$. The operator $I - AC^{-1}$ is self-adjoint w.r.t. the energy norm of A^{-1}, since

$$\langle AC^{-1}h, v\rangle_{A^{-1}} = \langle C^{-1}h, v\rangle = \langle h, C^{-1}v\rangle = \langle h, AC^{-1}v\rangle_{A^{-1}} \qquad (h, v \in H).$$

Hence

$$\|I - AC^{-1}\|_{A^{-1}} = \sup_{h\neq 0} \frac{|\langle (I - AC^{-1})h, h\rangle_{A^{-1}}|}{\|h\|^2_{A^{-1}}} = \sup_{h\neq 0} \frac{|\langle (A^{-1} - C^{-1})h, h\rangle|}{\langle A^{-1}h, h\rangle}. \tag{7}$$

Since $C^{-1} = \frac{2}{M+m}B^{-1}$, condition (C1) and Lemma 2.2 imply

$$\frac{2m}{M+m}\langle A^{-1}h, h\rangle \le \langle C^{-1}h, h\rangle \le \frac{2M}{M+m}\langle A^{-1}h, h\rangle.$$

Hence for any $h \in H$

$$-\frac{M-m}{M+m}\langle A^{-1}h, h\rangle \le \langle (A^{-1} - C^{-1})h, h\rangle \le \frac{M-m}{M+m}\langle A^{-1}h, h\rangle;$$

i.e., the supremum in (7) is indeed at most $\frac{M-m}{M+m}$.

3. Linear convergence by variable preconditioning.

3. Linear convergence by variable preconditioning. The following theorem provides local linear convergence using fixed spectral bounds for preconditioning. The result being locally an exact analogue of Theorem 2.1, it illustrates the preconditioning role of quasi-Newton methods.

Theorem 3.1. *Let H be a real Hilbert space. Assume that the nonlinear operator $F : H \to H$ has a Gâteaux derivative satisfying the following properties:*

(i) *For any $u \in H$ the operator $F'(u)$ is self-adjoint.*

(ii) *(Ellipticity.) There exist constants $\Lambda \ge \lambda > 0$ satisfying*

$$\lambda\|h\|^2 \le \langle F'(u)h, h\rangle \le \Lambda\|h\|^2 \qquad (u, h \in H).$$

(iii) *(Lipschitz continuity.) There exists* $L > 0$ *such that*

$$\|F'(u) - F'(v)\| \le L\|u - v\| \qquad (u, v \in H).$$

Denote u^* *the unique solution of equation* $F(u) = 0$. *We fix constants* $M \ge m > 0$. *Then there exists a neighborhood* $\mathcal{V}$ *of* u^* *such that for any* $u_0 \in \mathcal{V}$, *the sequence*

$$u_{n+1} = u_n - \frac{2}{M+m} B_n^{-1} F(u_n) \qquad (n \in \mathbf{N}),$$

with self-adjoint linear operators B_n *satisfying*

$$m\langle B_n h, h\rangle \le \langle F'(u_n)h, h\rangle \le M\langle B_n h, h\rangle \qquad (n \in \mathbf{N},\ h \in H), \tag{8}$$

converges linearly to u^*. *Namely,*

$$\|u_n - u^*\| \le C \cdot \left(\frac{M-m}{M+m}\right)^n \qquad (n \in \mathbf{N}) \tag{9}$$

with some constant $C > 0$.

(We note that constructive estimates are provided in the proof for the neighborhood $\mathcal{V}$; cf. (22) and Remark 1(b), and for the constant C, cf. (25).)

The proof of Theorem 3.1 is preceded by some required properties.

Lemma 3.2. *Let conditions* (i)–(iii) *of Theorem* 3.1 *hold. Then for any* $u, v, h \in H$,

$$\langle F'(u)h, h\rangle \le \langle F'(v)h, h\rangle \left(1 + L\lambda^{-2}\|F(u) - F(v)\|\right). \tag{10}$$

Proof. Condition (ii) of Theorem 3.1 implies that $\|F(u) - F(v)\| \ge \lambda\|u - v\|$. Hence

$$\langle F'(u)h, h\rangle \le \langle F'(v)h, h\rangle + L\|u-v\|\|h\|^2 \le \langle F'(v)h, h\rangle + L\lambda^{-2}\|F(u)-F(v)\|\langle F'(v)h, h\rangle.$$

Applying Lemma 3.2 to u and u^*, we obtain the following corollary.

Corollary 3.3. *If* $F(u^*) = 0$, *then for any* $u \in H$ *there holds*

$$\frac{1}{1+\mu(u)} \le \frac{\langle F'(u^*)h, h\rangle}{\langle F'(u)h, h\rangle} \le 1 + \mu(u),$$

where $\mu(u) = L\lambda^{-2}\|F(u)\| \le L\Lambda^{1/2}\lambda^{-2}\langle F'(u^*)F(u), F(u)\rangle^{1/2}$.

We introduce the norms

$$\|h\|_u = \langle F'(u)^{-1}h, h\rangle^{1/2} \quad (u, h \in H). \tag{11}$$

Then Corollary 3.3 and Lemma 2.2 imply directly the following corollary.

Corollary 3.4. *If* $F(u^*) = 0$, *then for any* $u \in H$ *there holds*

$$\frac{1}{1+\mu(u)} \le \frac{\|h\|_{u^*}^2}{\|h\|_u^2} \le 1 + \mu(u),$$

where $\mu(u)$ *is from Corollary* 3.3.

Proof of Theorem 3.1. Assumption (ii) and Lemma 2.2 imply that $\Lambda^{-1}\|h\|^2 \le \langle F'(u)^{-1}h, h\rangle \le \lambda^{-1}\|h\|^2$ for any $u, h \in H$. Hence the norms (11) satisfy

$$\lambda^{1/2}\|h\|_u \le \|h\| \le \Lambda^{1/2}\|h\|_u \qquad (u, h \in H), \tag{12}$$

and there also holds

$$\|F'(u)^{-1/2}\| \le \lambda^{-1/2} \qquad (u \in H). \tag{13}$$

Since the assumptions imply that $\lambda M^{-1}\|h\|^2 \le \langle B_n h, h\rangle$ for any $h \in H$, we obtain similarly to (13) that

$$\|B_n^{-1/2}\| \le \lambda^{-1/2} M^{1/2}. \tag{14}$$

The following norms (special cases of (11)) will be used throughout the proof:

$$\|\,.\,\|_n = \|\,.\,\|_{u_n} \quad (n \in \mathbf{N}), \qquad \|\,.\,\|_* = \|\,.\,\|_{u^*}. \tag{15}$$

The Lipschitz continuity of F implies that

$$F(u_{n+1}) = F(u_n) + F'(u_n)(u_{n+1} - u_n) + R(u_n), \tag{16}$$

where

$$\|R(u_n)\| \le \frac{L}{2}\|u_{n+1} - u_n\|^2. \tag{17}$$

Here

$$F(u_n) + F'(u_n)(u_{n+1} - u_n) = F(u_n) - \frac{2}{M+m} F'(u_n) B_n^{-1} F(u_n);$$

hence (8) and Lemma 2.3 imply that

$$\begin{aligned} \|F(u_n) + F'(u_n)(u_{n+1} - u_n)\|_n &\le \left\| I - \frac{2}{M+m} F'(u_n) B_n^{-1} \right\|_n \|F(u_n)\|_n \\ &\le \frac{M-m}{M+m}\|F(u_n)\|_n. \end{aligned} \tag{18}$$

Further, (12) and (17) yield

$$\|R(u_n)\|_n \le \frac{2L}{\lambda^{1/2}(M+m)^2}\|B_n^{-1}F(u_n)\|^2.$$

Here, using (14), (8), and Lemma 2.2, we have

$$\begin{aligned} \|B_n^{-1}F(u_n)\|^2 &\le \|B_n^{-1/2}\|^2 \|B_n^{-1/2}F(u_n)\|^2 \le M\lambda^{-1}\langle B_n^{-1}F(u_n), F(u_n)\rangle \\ &\le M^2\lambda^{-1}\langle F'(u_n)^{-1}F(u_n), F(u_n)\rangle = M^2\lambda^{-1}\|F(u_n)\|_n^2. \end{aligned}$$

Hence

$$\|R(u_n)\|_n \le \frac{2LM^2}{\lambda^{3/2}(M+m)^2}\|F(u_n)\|_n^2. \tag{19}$$

Altogether, (16), (18), and (19) yield

$$\|F(u_{n+1})\|_n \le \left(\frac{M-m}{M+m} + \frac{2LM^2}{\lambda^{3/2}(M+m)^2}\|F(u_n)\|_n \right) \|F(u_n)\|_n\,.$$

Finally, using Corollary 3.4, we obtain

$$\|F(u_{n+1})\|_* \le (1+\mu(u_n)) \left(\frac{M-m}{M+m} + \frac{2LM^2}{\lambda^{3/2}(M+m)^2} (1+\mu(u_n))^{1/2} \|F(u_n)\|_* \right) \|F(u_n)\|_* \,,$$

where $\mu(u_n) = L\Lambda^{1/2}\lambda^{-2}\|F(u_n)\|_*$. That is,

$$\|F(u_{n+1})\|_* \le \varphi(\|F(u_n)\|_*) \, \|F(u_n)\|_* \,, \tag{20}$$

where

$$\varphi(t) = (1+\beta\Lambda^{1/2}t) \left(Q + M^2\beta\alpha^{-2}\lambda^{1/2}(t/2) \left(1+\beta\Lambda^{1/2}t\right)^{1/2} \right) \tag{21}$$

and the notations

$$\alpha = \frac{M+m}{2}, \quad \beta = \frac{L}{\lambda^2}, \quad Q = \frac{M-m}{M+m}$$

are used. Then $\varphi : \mathbf{R}^+ \to \mathbf{R}^+$ is a strictly increasing continuous function and $\varphi(0) = Q$.

Estimate (20) puts us in the position to prove the required convergence estimate (9), provided that the assumption

$$r := \varphi(\|F(u_0)\|_*) \; < \; 1 \tag{22}$$

is satisfied for the initial guess.

First, we obtain by induction that

$$\|F(u_{n+1})\|_* \le r\|F(u_n)\|_* \qquad (n \in \mathbf{N}). \tag{23}$$

Namely, $\|F(u_1)\|_* = r\|F(u_0)\|_*$. Further, the assumption $\|F(u_{k+1})\|_* \le r\|F(u_k)\|_*$ $(k = 0, \dots, n-1)$ yields $\|F(u_n)\|_* < \|F(u_0)\|_*$; hence

$$\|F(u_{n+1})\|_* \le \varphi(\|F(u_n)\|_*) \, \|F(u_n)\|_* \le \varphi(\|F(u_0)\|_*) \, \|F(u_n)\|_* = r\|F(u_n)\|_* \,.$$

Inequality (23) implies that $\|F(u_n)\|_* \le r^n\|F(u_0)\|_* \to 0$, $\varphi(\|F(u_n)\|_*) \to Q$, and hence

$$\limsup \frac{\|F(u_{n+1})\|_*}{\|F(u_n)\|_*} \le \lim \varphi(\|F(u_n)\|_*) = Q.$$

From now on we use the notation

$$e_n = \|F(u_n)\|_* \,.$$

Then (20) implies that

$$e_n \le \left(\prod_{k=0}^{n-1} \varphi(e_k) \right) e_0 = \left(\prod_{k=0}^{n-1} \frac{\varphi(e_k)}{Q} \right) Q^n e_0 \qquad (n \in \mathbf{N}). \tag{24}$$

Using (21) and the notations $c = \beta\Lambda^{1/2}$, $d = (M^2\beta\alpha^{-2}\lambda^{1/2})/2$, we have

$$\varphi(t) = (1+ct) \left(Q + dt \, (1+ct)^{1/2} \right).$$

Here

$$\frac{\varphi(e_k)}{Q} = (1+ce_k)\left(1+\frac{d}{Q}e_k\,(1+ce_k)^{1/2}\right)$$

$$\leq (1+ce_k)\left(1+\frac{d}{Q}e_k\left(1+\frac{c}{2}e_k\right)\right) = 1+\left(c+\frac{d}{Q}\right)e_k+\frac{cd}{Q}e_k^2+\frac{c^2d}{2Q}e_k^3$$

$$\leq 1+\left(c+\frac{d}{Q}\right)e_0r^k+\frac{cd}{Q}e_0^2r^{2k}+\frac{c^2d}{2Q}e_0^3r^{3k}.$$

Since for any sequence $(a_k)\subset \mathbf{R}^+$ there holds $\prod_{k=0}^{n-1}(1+a_k)\leq \prod_{k=0}^{n-1}\exp(a_k)\leq \exp(\sum_{k=0}^{\infty}a_k)$, we obtain

$$\prod_{k=0}^{n-1}\frac{\varphi(e_k)}{Q}\leq \exp\left\{\left(c+\frac{d}{Q}\right)\frac{e_0}{1-r}+\frac{cd}{Q}\frac{e_0^2}{1-r^2}+\frac{c^2d}{2Q}\frac{e_0^3}{1-r^3}\right\} =: E\,.$$

Therefore (24) yields

$$e_n \;\leq\; e_0E\cdot Q^n \qquad (n\in\mathbf{N}).$$

Finally, using condition (ii) and (12), this implies

$$\|u_n-u^*\|\leq \lambda^{-1}\|F(u_n)\|\leq\; \lambda^{-1}\Lambda^{1/2}e_0E\cdot Q^n \qquad (n\in\mathbf{N}), \tag{25}$$

which coincides with the required convergence estimate with $C=\lambda^{-1}\Lambda^{1/2}e_0E$.

Remark 1. The convergence has been proved under the sufficient condition

$$\varphi(\|F(u_0)\|_*)\;<\;1 \tag{26}$$

for the initial guess, with φ defined in (21). In connection with this we note the following:

(a) The condition (26) is satisfied if

$$K\frac{L}{\lambda^2}\|F(u_0)\|_*\;<\;1\,,$$

where $K=\Lambda^{1/2}(M/m)\max\left\{1,\,2(M-m)^{-1}(\lambda/\Lambda)^{1/2}\right\}$. (This is proved in the appendix.) Relating this to the well-known sufficient condition $\frac{L}{2\lambda^2}\|F(u_0)\|<1$ of the exact Newton iteration, we observe that the order is similar (although K is obviously somewhat larger than 1/2).

(b) The sufficient condition of convergence can be given using the original norm as follows. Since the theoretical norm $\|\,.\,\|_*$ satisfies $\|F(u_0)\|_*\leq\lambda^{-1/2}\|F(u_0)\|$ by (12), and φ increases, therefore we obtain the condition

$$\varphi(\lambda^{-1/2}\|F(u_0)\|)\;<\;1$$

to be checked for u_0.

4. Damped quasi-Newton method as variable preconditioning. We recall the following definitions of norms (see (15)), where (u_n) is an iterative sequence and u^* is the solution of $F(u) = 0$:

$$\|h\|_n = \langle F'(u_n)^{-1}h, h\rangle^{1/2} \quad (n \in \mathbf{N}), \qquad \|h\|_* = \langle F'(u^*)^{-1}h, h\rangle^{1/2}. \tag{27}$$

The following theorem generalizes Theorem 3.1. Using damped iteration and variable spectral bound preconditioning, it provides global convergence up to second order.

Theorem 4.1. *Let H be a real Hilbert space. Let the operator $F : H \to H$ have a Gâteaux derivative satisfying the properties* (i)–(iii) *of Theorem* 3.1.

Denote u^* the unique solution of equation $F(u) = 0$. For arbitrary $u_0 \in H$ let (u_n) be the sequence defined by

$$u_{n+1} = u_n - \frac{2\tau_n}{M_n + m_n} B_n^{-1} F(u_n) \qquad (n \in \mathbf{N}), \tag{28}$$

where the following conditions hold:

(iv) *$M_n \geq m_n > 0$ and the self-adjoint linear operators B_n satisfy*

$$m_n\langle B_n h, h\rangle \leq \langle F'(u_n)h, h\rangle \leq M_n\langle B_n h, h\rangle \qquad (n \in \mathbf{N},\, h \in H);$$

further, using notation $\mu(u_n) = L\lambda^{-2}\|F(u_n)\|$, there exist constants $K > 1$ and $\varepsilon > 0$ such that $M_n/m_n \leq 1 + 2/(\varepsilon + K\mu(u_n))$.

(v) *We define*

$$\tau_n = \min\left\{1, \frac{1 - Q_n}{2\rho_n}\right\}, \tag{29}$$

where $Q_n = \frac{M_n - m_n}{M_n + m_n}(1 + \mu(u_n))$, $\rho_n = 2LM_n^2\lambda^{-3/2}(M_n + m_n)^{-2}\|F(u_n)\|_n(1 + \mu(u_n))^{1/2}$, $\mu(u_n)$ is as in condition (iv), *and $\|\,.\,\|_n$ is defined in* (27). *(This value of τ_n ensures optimal contractivity in the nth step in the $\|\,.\,\|_*$-norm.)*

Then there holds

$$\|u_n - u^*\| \leq \lambda^{-1}\|F(u_n)\| \to 0;$$

namely,

$$\limsup \frac{\|F(u_{n+1})\|_*}{\|F(u_n)\|_*} \leq \limsup \frac{M_n - m_n}{M_n + m_n} < 1\,. \tag{30}$$

Moreover, if in addition we assume $M_n/m_n \leq 1 + c_1\|F(u_n)\|^\gamma$ $(n \in \mathbf{N})$ with some constants $c_1 > 0$ and $0 < \gamma \leq 1$, then

$$\|F(u_{n+1})\|_* \leq d_1\|F(u_n)\|_*^{1+\gamma} \qquad (n \in \mathbf{N}) \tag{31}$$

with some constant $d_1 > 0$.

Owing to the equivalence of the norms $\|\,.\,\|$ and $\|\,.\,\|_*$, the orders of convergence corresponding to the estimates to (30) and (31) can be formulated with the original norm.

Corollary 4.2 (rate of convergence in the original norm).

(a) *If $\limsup M_n/m_n = K > 1$, then*

$$\|u_n - u^*\| \leq \lambda^{-1}\|F(u_n)\| \leq \text{const.} \cdot \rho^n$$

with $\rho = (K-1)/(K+1)$.

(b) *In the case* $M_n/m_n \le 1 + c_1\|F(u_n)\|^\gamma$ *(with constants* $c_1 > 0,\ \ 0 < \gamma \le 1$*) there holds*

$$\|u_n - u^*\| \le \lambda^{-1}\|F(u_n)\| \le \text{const.} \cdot \rho^{(1+\gamma)^n}$$

with some constant $0 < \rho < 1$.

Proof of Theorem 4.1. Using (16) and (28), we obtain

$$F(u_{n+1}) = (1-\tau_n)F(u_n) - \tau_n\left(F(u_n) - \frac{2}{M_n+m_n}F'(u_n)B_n^{-1}F(u_n)\right) + R(u_n).$$

Hence

$$\|F(u_{n+1})\|_* \le (1-\tau_n)\|F(u_n)\|_* + \tau_n\left\|\left(I - \frac{2}{M_n+m_n}F'(u_n)B_n^{-1}\right)F(u_n)\right\|_* + \|R(u_n)\|_*\,.$$

Here, using Corollary 3.4 and Lemma 2.3,

$$\left\|\left(I - \frac{2}{M_n+m_n}F'(u_n)B_n^{-1}\right)F(u_n)\right\|_* \le\ (1+\mu(u_n))^{1/2}\frac{M_n-m_n}{M_n+m_n}\|F(u_n)\|_n$$

$$\le\ (1+\mu(u_n))\frac{M_n-m_n}{M_n+m_n}\|F(u_n)\|_*\,,$$

where $\mu(u_n) = L\lambda^{-2}\|F(u_n)\|$. Further, from (12) and (17) there follows

$$\|R(u_n)\|_* \le \frac{L}{2\lambda^{1/2}}\|u_{n+1}-u_n\|^2 = \tau_n^2\frac{2L}{\lambda^{1/2}(M+m)^2}\|B_n^{-1}F(u_n)\|^2;$$

hence, using the estimate preceding (19) and then Corollary 3.4, we obtain

$$\|R(u_n)\|_* \le \tau_n^2\frac{2LM^2}{\lambda^{3/2}(M+m)^2}\|F(u_n)\|_n^2$$

$$\le \tau_n^2(1+\mu(u_n))^{1/2}\frac{2LM^2}{\lambda^{3/2}(M+m)^2}\|F(u_n)\|_n\|F(u_n)\|_*.$$

Summing up, we obtain

$$\|F(u_{n+1})\|_* \le \Bigg(1 - \tau_n + \tau_n(1+\mu(u_n))\frac{M_n-m_n}{M_n+m_n}$$

$$+ \tau_n^2(1+\mu(u_n))^{1/2}\frac{2LM^2}{\lambda^{3/2}(M+m)^2}\|F(u_n)\|_n\Bigg)\ \|F(u_n)\|_*\,.$$

That is,

$$\|F(u_{n+1})\|_* \le \left(1 - \tau_n(1-Q_n) + \tau_n^2\rho_n\right)\ \|F(u_n)\|_*\,, \tag{32}$$

where Q_n and ρ_n are as in condition *(v)*.

There exists $\tilde{Q} < 1$ such that

$$Q_n \le \tilde{Q} \qquad (n \in \mathbf{N}). \tag{33}$$

Namely, the assumption $M_n/m_n \le 1+2/(\varepsilon+K\mu(u_n))$ with $K>1$ and $\varepsilon>0$ implies that

$$1+\varepsilon+K\mu(u_n) \le 1+\frac{2}{(M_n/m_n)-1} = \frac{M_n+m_n}{M_n-m_n};$$

hence

$$1+\mu(u_n) \le \tilde{Q}\,\frac{M_n+m_n}{M_n-m_n}$$

with $\tilde{Q} := \max\{1/K, 1/(1+\varepsilon)\} < 1$.

Let us introduce the function $p:[0,1]\to \mathbf{R}$, $p(t) := 1-(1-Q_n)t+\rho_n t^2$. Here $p'(t) = -(1-Q_n)+2\rho_n t$ yields that τ_n defined in (29) satisfies

$$p(\tau_n) = \min_{t\in[0,1]} p(t) < 1,$$

since $p'(0) = -(1-Q_n) < 0$. Hence from (32)

$$\|F(u_{n+1})\|_* \le p(\tau_n)\|F(u_n)\|_* < \|F(u_n)\|_* \,. \tag{34}$$

Moreover, if $\tau_n = 1$ (i.e., when $1 \le (1-Q_n)/2\rho_n$), then

$$p(\tau_n) = Q_n+\rho_n \le Q_n+(1-Q_n)/2 = (1+Q_n)/2 \le (1+\tilde{Q})/2 < 1.$$

In the case $\tau_n = (1-Q_n)/2\rho_n$ we have

$$p(\tau_n) = 1-(1-Q_n)^2/(4\rho_n) \le 1-(1-\tilde{Q})^2/(4\sup_n \rho_n) =: Q' < 1.$$

The latter holds since by (34) $\|F(u_n)\|_*$ is bounded, and hence

$$\rho_n = \text{const.}\cdot\|F(u_n)\|_n\,(1+\text{const.}\cdot\|F(u_n)\|)^{1/2} \tag{35}$$

is bounded, the three norms being equivalent. Altogether, from (34) we obtain

$$\|F(u_n)\|_* \le \text{const.}\cdot r^n \to 0,$$

where $r = \max\{(1+\tilde{Q})/2, Q'\}$. This also implies that $\rho_n \to 0$ and $\mu(u_n) = L\lambda^{-2}\|F(u_n)\| \to 0$. A brief calculation gives

$$p(\tau_n) = Q_n+\rho_n\left(1-(1-\tau_n)^2\right) \tag{36}$$

(for both $\tau_n = 1$ and $\tau_n < 1$); hence (34) yields

$$\limsup \frac{\|F(u_{n+1})\|_*}{\|F(u_n)\|_*} \le \limsup Q_n = \limsup \frac{M_n-m_n}{M_n+m_n}\,.$$

The bound $M_n/m_n \le 1+2/\varepsilon$ in assumption (iv) implies that

$$\limsup \frac{M_n-m_n}{M_n+m_n} \le \frac{1}{1+\varepsilon} < 1\,.$$

Finally, let $M_n/m_n \le 1+c_1\|F(u_n)\|^\gamma$ with constants $c_1>0$, $0<\gamma\le 1$. Then $M_n/m_n \le 1+c_2\|F(u_n)\|_*^\gamma$ with $c_2 = c_1\Lambda^{1/2}$; hence

$$\frac{M_n-m_n}{M_n+m_n} < \frac{M_n-m_n}{m_n} \le c_2\|F(u_n)\|_*^\gamma\,,$$

and therefore

$$Q_n \le c_3 \|F(u_n)\|_*^\gamma$$

with $c_3 = c_2(1 + \sup_n \mu(u_n))$. Also,

$$\rho_n \le c_4 \|F(u_n)\|_*$$

with some $c_4 > 0$ since $\|F(u_n)\|_*$ is bounded (cf. (35)). With the use of notation $e_n = \|F(u_n)\|_*$, we obtain from (34) and (36) that

$$e_{n+1} \le (Q_n + \rho_n)\, e_n \le (Q_n + c_4 e_n)\, e_n \le \left(c_3 e_n^\gamma + c_4 e_0 \frac{e_n}{e_0} \right) e_n$$

$$\le \left(c_3 e_n^\gamma + c_4 e_0 \left(\frac{e_n}{e_0} \right)^\gamma \right) e_n = d_1 e_n^{1+\gamma}$$

with $d_1 = c_3 + c_4 e_0^{1-\gamma}$.

Remark 2. (a) It is worth mentioning that Theorems 3.1 and 4.1 define descent methods, similarly to the simple iteration. Namely, conditions (i)–(iii) of Theorem 3.1 imply the existence of a potential $\Phi : H \to \mathbf{R}$, i.e., $\Phi'(u) = F(u) \quad (u \in H)$. Then the directions $-B_n^{-1} F(u_n)$ are descent directions, since their angle is acute with the steepest descent direction $-F(u_n)$ owing to

$$\langle B_n^{-1} F(u_n), F(u_n) \rangle > 0\,.$$

We also note that the residuals $\Phi(u_n) - \Phi(u^*)$ are equivalent to $\|u_n - u^*\|^2$ owing to the ellipticity condition (ii) of Theorem 3.1.

(b) The conditions of Theorems 3.1 and 4.1 can be relaxed. Namely, since the constructed sequences are bounded, it suffices to assume (ii) and (iii) on the corresponding ball that (u_n) runs within. Further, the proofs can be repeated with obvious modification if (iii) is replaced by Hölder continuity only. (See [1] for a related damped inexact Newton result.)

(c) The value of τ_n need not necessarily be maximized by 1 as in (29), but suitable overrelaxation is also feasible which may accelerate the convergence.

5. Sobolev space preconditioning for nonlinear elliptic problems. In this section we consider the BVP

$$\begin{cases} -\operatorname{div} f(x, \nabla u) = g(x), \\ u_{|\partial\Omega} = 0 \end{cases} \tag{37}$$

with the following conditions: $\Omega \subset \mathbf{R}^N$ is a bounded domain, $g \in L^2(\Omega)$, f is measurable and $f(x, .) \in C^1(\mathbf{R}^N, \mathbf{R}^N)$ for all $x \in \Omega$, and the Jacobians $\frac{\partial f(x,\eta)}{\partial \eta}$ are Lipschitz continuous in η, symmetric, and satisfy

$$\lambda |\xi|^2 \le \frac{\partial f(x,\eta)}{\partial \eta} \xi \cdot \xi \le \Lambda |\xi|^2, \qquad (x,\eta) \in \Omega \times \mathbf{R}^N, \xi \in \mathbf{R}^N, \tag{38}$$

with constants $\Lambda \ge \lambda > 0$ independent of (x, η).

An important special case of f is of the form $f(x,\eta) = a(|\eta|)\eta$, where $0 < \lambda \le a(r) \le (r a(r))' \le \Lambda \quad (r > 0)$. This kind of operator arises, e.g., in plasticity theory or in connection with magnetic potential (see, e.g., [23, 24]).

Owing to uniform ellipticity, problem (37) has a unique weak solution $u^* \in H_0^1(\Omega)$, and the finite element approximations converge to u^*.

Let $V_h \subset H_0^1(\Omega)$ be a finite element subspace, and denote $u_h \in V_h$ the solution of the discretized problem

$$\int_\Omega f(x, \nabla u_h) \cdot \nabla v = \int_\Omega gv \qquad (v \in V_h). \tag{39}$$

Our aim is to find preconditioners for the iterative solution of (39) based on Theorem 4.1.

5.1. Variable preconditioning operators for nonlinear elliptic problems. In order to fit in the operator framework of Theorem 4.1, we first formulate the properties of the operator corresponding to (39). Namely, let $F : V_h \to V_h$ denote the operator defined by

$$\langle F(u), v\rangle = \int_\Omega \big(f(x, \nabla u) \cdot \nabla v \ - gv\big) \qquad (u, v \in V_h), \tag{40}$$

where $\langle u, v\rangle = \int_\Omega \nabla u \cdot \nabla v$. Then F has a Gâteaux derivative satisfying properties (i)–(iii) of Theorem 3.1. Namely,

$$\langle F'(u)v, w\rangle = \int_\Omega \frac{\partial f}{\partial \eta}(x, \nabla u)\, \nabla v \cdot \nabla w \qquad (u, v, w \in V_h); \tag{41}$$

hence (38) implies that $F'(u)$ is self-adjoint and

$$\lambda \|v\|^2 \le \langle F'(u)v, v\rangle \le \Lambda \|v\|^2 \qquad (u, v \in V_h).$$

Further, (41) implies that F' inherits the Lipschitz continuity of $\frac{\partial f(x,\eta)}{\partial \eta}$ in η.

Our approach is *Sobolev space preconditioning*, which means that we will find preconditioners by considering (39) as the projection of the exact equation (with u^* and all $v \in H_0^1(\Omega)$), instead of viewing the actual form of the nonlinear algebraic system that (39) is. This means defining linear operators in weak form, spectrally equivalent to $F'(u_n)$, relying directly on the properties of the matrices $\frac{\partial f}{\partial \eta}(x, \nabla u_n)$. The required variable preconditioners are the matrices that these linear operators define in V_h. (Since these matrices are the projections of operators in $H_0^1(\Omega)$, defined by the same integral, the spectral bounds can be obtained in a mesh uniform way.)

The above idea is the analogue of the one developed in [21] for fixed preconditioners. We note that the Sobolev space framework is also useful in the context of exact Newton iterations; namely, (41) yields the Jacobians without numerical differentiation. The variable preconditioners lead to systems of possibly simpler structure than those containing exact Jacobians.

THEOREM 5.1. *Let $u_0 \in V_h$ be arbitrary, and let $(u_n) \subset V_h$ be the sequence defined as follows.*

If, for $n \in \mathbf{N}$, u_n is obtained, then we choose constants $M_n \ge m_n > 0$ and a symmetric matrix-valued function $G_n \in L^\infty(\Omega, \mathbf{R}^{N\times N})$ for which there holds

$$m_n \langle G_n(x)\xi, \xi\rangle \le \left\langle \frac{\partial f}{\partial \eta}(x, \nabla u_n(x))\xi, \xi \right\rangle \le M_n \langle G_n(x)\xi, \xi\rangle \qquad (x \in \Omega,\ \xi \in \mathbf{R}^N); \tag{42}$$

further, M_n/m_n *and* τ_n *satisfy the conditions* (iv)–(v) *in Theorem* 4.1. *We define*

$$u_{n+1} = u_n - \frac{2\tau_n}{M_n + m_n} z_n \,, \tag{43}$$

where $z_n \in V_h$ *is the solution of*

$$\int_\Omega G_n(x)\, \nabla z_n \cdot \nabla v = \int_\Omega \big(f(x, \nabla u_n) \cdot \nabla v \; - gv \big) \qquad (v \in V_h). \tag{44}$$

Then u_n *converges to* u_h *according to the estimates of Theorem* 4.1.

Proof. For any $n \in \mathbf{N}$ let $B_n : V_h \to V_h$ denote the linear operator

$$\langle B_n v, w \rangle = \int_\Omega G_n(x)\, \nabla v \cdot \nabla w \qquad (v, w \in V_h).$$

Then B_n is self-adjoint and (42) implies that

$$m_n \langle B_n v, v \rangle \le \langle F'(u_n) v, v \rangle \le M_n \langle B_n v, v \rangle \qquad (v \in V_h).$$

Since F fulfills properties (i)–(iii) of Theorem 3.1, therefore all the conditions are satisfied and the convergence results hold.

The matrices of the linear systems corresponding to problems (44) are the variable preconditioners for the problem (39). They can also be considered as the discretizations of the auxiliary linear operators

$$L_n u = -\text{div}\ (G_n(x) \nabla u)$$

(the strong form of B_n). As mentioned above, the choice of the matrices G_n requires only the properties of the matrices $\frac{\partial f}{\partial \eta}(x, \nabla u_n)$, instead of investigating the actual form of the nonlinear algebraic system (39). This means that the preconditioning for the latter is determined by that for the Jacobians of f; hence the preconditioning matrices are not difficult to compile. A further advantage, as mentioned before, is that the resulting condition numbers can be obtained in a straightforward and mesh independent way.

Remark 3. We note that preconditioning by spectral equivalence is also efficient in other contexts (e.g., by using a rougher mesh for the same operator [5]). The raison d'être for the proposed coupling of spectral equivalence with Sobolev space framework is given by the advantages mentioned above.

We also remark that the idea of Theorem 5.1 can be similarly used in the setting of multilevel iterations, i.e., when the discretization parameter h is redefined in each step n.

The systems (44) are of simpler structure than those containing exact Jacobians if the G_n are appropriately chosen. In what follows we give an appropriate construction and study the properties of the obtained method. We note that other possible choices of such preconditioners are summarized in the book [16] together with the theoretical background for preconditioning operators.

5.2. Variable preconditioning operators with piecewise constant coefficient operators. The following example illustrates an appropriate construction of variable preconditioners: the Jacobians are replaced by the discretizations of piecewise constant coefficient preconditioning operators.

These operators are constructed as follows. Let u_n be fixed. To improve the bounds in (38), the domain Ω is decomposed in subdomains Ω_i $(i = i(n) = 1, \ldots, s_n)$ such that for all i

$$\lambda_i \, |\xi|^2 \le \frac{\partial f}{\partial \eta}(x, \nabla u_n) \, \xi \cdot \xi \le \Lambda_i \, |\xi|^2 \qquad (x \in \Omega_i \, , \xi \in \mathbf{R}^N), \tag{45}$$

with $\lambda < \lambda_i \le \Lambda_i < \Lambda$. We introduce a piecewise constant weight function w_n such that

$$w_{n \, | \Omega_i} \equiv c_i \quad (i = i(n) = 1, \ldots, s_n), \tag{46}$$

where c_i is some mean of λ_i and Λ_i. Let

$$G_n(x) := \, w_n(x) \cdot I \, ,$$

where $I \in \mathbf{R}^{N \times N}$ is the identity matrix. Then G_n defines the linear operator B_n corresponding to the weight function w, i.e.,

$$\langle B_n h, v \rangle = \int_\Omega w_n \, \nabla h \cdot \nabla v \, .$$

Defining

$$m_n := \min_i \lambda_i / c_i \quad \text{and} \quad M_n := \max_i \Lambda_i / c_i \, , \tag{47}$$

the left and right sides of (45) can be estimated further by $m_n w(x) |\xi|^2$ and $M_n w(x) |\xi|^2$, respectively. This and (41) yield

$$m_n \langle B_n v, v \rangle = m_n \int_\Omega w |\nabla v|^2 \le \langle F'(u_n) v, v \rangle \le M_n \int_\Omega w |\nabla v|^2 = M_n \langle B_n v, v \rangle \tag{48}$$

for all $v \in V_h$. The obtained condition number estimate of the operator $B_n^{-1} F'(u_n)$ satisfies

$$M_n / m_n = \max_i \Lambda_i / \lambda_i ; \tag{49}$$

hence it can be decreased by the suitable refinement of decomposition.

This operator B_n was studied in [4] in the context of a standard inner-outer iteration.

The corresponding preconditioning matrix $\mathcal{B}_n$ (the discretization of the operator B_n in the subspace V_h) is the modification of the discrete Laplacian via blockwise multiplication by the corresponding constants c_i. Moreover (see [4, 7]), the matrix $\mathcal{B}_n$ can be decomposed in the product form

$$\mathcal{B}_n = \mathcal{C} \mathcal{W}_n \mathcal{C}^T \, , \tag{50}$$

where the matrices $\mathcal{C}$ and $\mathcal{C}^T$ correspond to the discretization of $-div$ and ∇, respectively, and hence are independent of n; further, $\mathcal{W}_n$ is a diagonal matrix consisting of constants c_i at the entries corresponding to the subdomains Ω_i.

Clearly, the condition number of the preconditioned system is also estimated by M_n / m_n. This bound is mesh independent in the setting of multilevel iterations, i.e., independent of the discretization parameter $h = h_n$ defined in step n.

For instance, the required decompositions are straightforward to define for the special cases of (37) of the form

$$\begin{cases} -\mathrm{div}\,(a(|\nabla u|)\nabla u) \;=\; g(x), \\ \;u_{|\partial\Omega} = 0\,, \end{cases} \tag{51}$$

where $0 < \lambda \le a(r) \le a(r) + a'(r)r \le \Lambda$. Here $f(x,\eta) = a(|\eta|)\eta$ satisfies

$$a(|\nabla u_n|)|\xi|^2 \le \frac{\partial f}{\partial \eta}(x, \nabla u_n)\,\xi\cdot\xi \le (a(|\nabla u_n|) + a'(|\nabla u_n|)|\nabla u_n|)\,|\xi|^2$$

(see, e.g., [24]); hence

$$\lambda_i \;= \inf_{\Omega_i} a(|\nabla u_n|)\,, \qquad \Lambda_i \;= \sup_{\Omega_i}\,(a(|\nabla u_n|) + a'(|\nabla u_n|)|\nabla u_n|)\;. \tag{52}$$

That is, the bounds λ_i and Λ_i are determined only by the values of $|\nabla u_n|$ (in addition to the given function $a(r)$).

The above preconditioner compensates possible sharp gradients, i.e., when Λ_i/λ_i is very large. This is the case, e.g., for magnetic potential (in stator sheets, etc.); see [23]. The construction is also suitable to follow discontinuities of the coefficients. In this case the decomposition is to be chosen such that the boundaries of subdomains fit the discontinuities of $f(x, \nabla u_n)$. A straightforward illustration of this is another special case of (37) of the form

$$\begin{cases} -\mathrm{div}\,(b(x,|\nabla u|)\nabla u) \;=\; g(x), \\ \;u_{|\partial\Omega} = 0, \end{cases}$$

where $\Omega_1 \subset \Omega$,

$$b(x,r) := \begin{cases} a(r) & \text{if } x \in \Omega_1, \\ \alpha & \text{if } x \in \Omega \setminus \Omega_1\,, \end{cases}$$

$a(r)$ is as in (51), and $\alpha > 0$ is a constant. (This nonlinearity arises, e.g., in connection with potential in H-shaped magnets [23].) Then an obviously suitable choice of the weight function w is defined as above for (51) on Ω_1 and as constant α on $\Omega \setminus \Omega_1$.

5.3. Numerical experiment. Let us consider the two-dimensional magnetic potential problem (51) with the following nonlinearity, which characterizes the reluctance of stator sheets in the cross-sections of an electrical motor in the case of isotropic media [23]:

$$a(r) = \frac{1}{\mu_0}\left(\alpha + (1-\alpha)\frac{r^8}{r^8+\beta}\right) \qquad (r \ge 0), \tag{53}$$

where $\alpha = 3\cdot 10^{-4}$ and $\beta = 1.6\cdot 10^4$; further, μ_0 is the vacuum permeability. (For simplicity we can consider $\mu_0 = 1$ since μ_0 does not affect the conditioning.) For the right-hand side we set

$$g(x) \equiv \rho = 4\cdot 10^6, \tag{54}$$

which is a realistic value for the electric current density (see also [23]). For simplicity we consider problem (51) on the unit square domain $\Omega = [0,1]\times[0,1]$.

TABLE 1
The number of iterations to achieve error 10^{-4} *and* 10^{-8} *under different numbers of node points using six subdomains.*

Node points:	2^6	2^8	2^{10}
# iterations for $\varepsilon = 10^{-4}$:	10	10	10
# iterations for $\varepsilon = 10^{-8}$:	16	16	16

We note that the function a varies in several magnitudes over the whole domain: the bounds in (38) are $\lambda = \alpha = 0.0003$ and $\Lambda = \max(a(r) + a'(r)r) = 2.5313$; hence the related condition number is $\Lambda/\lambda = 8437.7$.

We apply the variable preconditioning procedure with the above described piecewise constant coefficient preconditioning operators. For simplicity, we choose V_h as the subspace of piecewise linear elements on a uniform triangulation of Ω. Then the iteration (43)–(44) takes the form

$$u_{n+1} = u_n - \frac{2\tau_n}{M_n + m_n} z_n \tag{55}$$

with $z_n \in V_h$ being the solution of problem

$$\int_\Omega w_n(x)\,\nabla z_n \cdot \nabla v = \int_\Omega \big(a(|\nabla u_n|)\,\nabla u_n \cdot \nabla v \ - \rho v\big) \qquad (v \in V_h), \tag{56}$$

where $a(r)$ and ρ are as above, w_n is a piecewise constant weight function, and M_n and m_n are from (47). In order to define w_n, we first observe that for a given number $\kappa > 1$ satisfying $\kappa > \sup_{r>0}\,(a(r)+a'(r)r)/a(r)$, one can recursively define subintervals $J_i = [r_{i-1}, r_i) \subset [0, \infty)$ such that $r_0 = 0$ and

$$\sup_{r\in J_i}\,(a(r) + a'(r)r) \,/\, \inf_{r\in J_i}\, a(r) \ = \kappa \qquad (i \in \mathbf{N}).$$

Then the subdomains Ω_i are defined as the level sets of $|\nabla u_n|$ corresponding to J_i, and by (52) they satisfy $\Lambda_i/\lambda_i \le \kappa$ for all i; hence by (49) we have $M_n/m_n \le \kappa$. In this way one uses the minimal number of subdomains to achieve the condition number κ, and, conversely, for a given number of subdomains one can find the subdomains themselves that produce the lowest bound κ of M_n/m_n. (For details see [4], where this preconditioner was studied in the context of a standard inner-outer iteration.)

In the numerical experiment we have used a decomposition to six subdomains in each step of the iteration. In each step n we have determined the subdomains that yield the lowest value of κ for $|\nabla u_n|$ according as above using a suitable subroutine. We have chosen c_i to be the arithmetic mean of λ_i and Λ_i for all i.

The error during the iteration was measured by the weighted residual errors corresponding to (40) with the inner product with weight w_n. This error is obtained from the iteration without any extra work as the weighted norm of the actual coefficient vector w.r.t. the Gram matrix. (It is a computable approximation of the $*$-norm (27) that appears in the convergence estimates of Theorem 4.1.)

The experiment was made using 2^k node points of the mesh with $k = 6$, 8, and 10. Table 1 summarizes the number of iterations that decrease the residual error $\|F(u_n)\|$ below 10^{-4} and 10^{-8}, respectively. The results exhibit mesh independence; i.e., the number of iterations remains the same when the number of node points is increased.

We have repeated the experiment with 12 subdomains, and the results were the same (except that for 2^6 node points the number of iterations for $\varepsilon = 10^{-8}$ was only

Table 2

The sequence of errors up to eight digits, using 2^{10} node points and six subdomains.

1.0
0.32290943
0.14549087
0.06899055
0.03014214
0.01194232
0.00414995
0.00120266
0.00027565
0.00005601
0.00001047
0.00000182
0.00000033
0.00000006
0.00000001
0.00000000

15). This means that the smaller number of subdomains already suffices to achieve the available convergence speed.

The distribution of the errors behaved much similarly for the different runs. We give one of them below for illustration.

Finally, in order to compare the results in Table 1, we cite results from other papers where the same or a similar problem is studied. The coefficients (53) and (54) are quoted from [18] and [23], and a similar nonlinearity was first considered in the early paper [11]. In the latter, Newton's method is applied with overrelaxation for finite difference method on a square with 90 and 870 points, and requires 20 (resp., 98) iterations to achieve a residual error $\varepsilon = 10^{-6}$. Successive overrelaxation (or Kacanov's frozen coefficient method) in [11] requires 18 (resp., 58) iterations for the same error, and the variants of this method require 162 iterations for $\varepsilon = 10^{-5}$ with 384 node points in [18] (on a complicated domain) and 15 iterations for $\varepsilon = 10^{-6}$ with 1000 node points in [23], respectively. Compared even to this last fastest result, the iteration (55)–(56) is less costly. Namely, since the auxiliary systems in (56) come from a piecewise constant coefficient operator, their structure is simpler than either for Newton's method or for frozen coefficients. That is, (50) implies that the matrices of the auxiliary systems are the modifications of the discrete Laplacian such that their updating consists of updating the diagonal matrix $\mathcal{W}_n$. In fact, the latter requires only distributing the six constants c_i at the entries corresponding to the subdomains Ω_i.

5.4. Conclusions. We summarize our numerical method for the solution of the nonlinear elliptic problem (37).

Using the quasi-Newton setting in an operator framework, the Jacobians were replaced by the discretizations of suitably chosen piecewise constant coefficient elliptic operators. This has a twofold advantage. First, superlinear convergence can be preserved, and stepwise the condition number of the preconditioned system is mesh independent. Second, our method is less costly than either a Newton or a frozen coefficient iteration due to the decomposition (50), which implies that the matrices of the auxiliary systems are the modifications of the discrete Laplacian such that their updating consists of updating the diagonal factor $\mathcal{W}_n$.

Our numerical experiment has demonstrated these advantages. We have achieved convergence which is in accordance with the theoretical result: Table 2 exhibits the superlinear convergence. Table 1 also suggests mesh independence of the iterations. The convergence results are favorable in comparison with some other cited ones. Concerning computational cost, the experiment with a rather ill-conditioned problem has shown that even a coarse decomposition of the domain has been able to provide the above described favorable convergence. Namely, Tables 1 and 2 were achieved using six subdomains, in which case the matrix of the auxiliary system is very close to the discrete Laplacian. Its updating required only distributing the six constants c_i at the entries corresponding to the subdomains Ω_i, and this structure property slightly increases only the complexity of a Laplacian solver.

Appendix.

Proof of Remark 1(a). We use the notations as earlier:

$$e_0 = \|F(u_0)\|_*\,, \quad \alpha = \frac{M+m}{2}, \quad \beta = \frac{L}{\lambda^2}, \quad Q = \frac{M-m}{M+m}\,.$$

The estimate to be proved is

$$\varphi(e_0) = (1+\beta\Lambda^{1/2}e_0)\left(\tilde{Q} + M^2\beta\alpha^{-2}\lambda^{1/2}(e_0/2)\left(1+\beta\Lambda^{1/2}e_0\right)^{1/2}\right) \, < \, 1\,,$$

where $\beta = L\lambda^{-2}$, under the assumption $K\beta e_0 < 1$. We use that the estimate $(1-x)^{1/2} \le 1-(x/2) \quad (0 \le x \le 1)$ implies that

$$Q^{1/2} = \left(1 - \frac{2m}{M+m}\right)^{1/2} \le \frac{M}{M+m}, \quad Q^{-1/2} \ge \frac{M+m}{M} = 1 + \frac{m}{M}\,.$$

Then for the first term in $\varphi(e_0)$ we have

$$1+\beta\Lambda^{1/2}e_0 < 1 + \frac{\Lambda^{1/2}}{K} \le 1 + \frac{\Lambda^{1/2}}{\Lambda^{1/2}(M/m)} = 1 + \frac{m}{M} \le Q^{-1/2}.$$

From this, using $K\beta e_0 < 1$ and using that $2\alpha^2 QK \ge M\lambda^{1/2}(M+m)/m$, the second term is estimated by

$$\tilde{Q} + \frac{M^2\lambda^{1/2}}{2\alpha^2 Q^{1/2}K} \le Q^{1/2}\left(\frac{M}{M+m} + \frac{m}{M+m}\right) = Q^{1/2}.$$

Multiplying the two terms, $\varphi(e_0) < 1$ is verified.

REFERENCES

[1] O. Axelsson, *On global convergence of iterative methods*, in Iterative Solution of Nonlinear Systems of Equations, Lecture Notes in Math. 953, Springer, Berlin, 1982, pp. 1–19.

[2] O. Axelsson, *A mixed variable finite element method for the efficient solution of nonlinear diffusion and potential flow equations*, in Advances in Multi-grid Methods, Notes Numer. Fluid Mech. 11, D. Braess, W. Hackbusch, and U. Trottenberg, eds., Viewig, Braunschweig, 1985, pp. 1–11.

[3] O. Axelsson, *Iterative Solution Methods,* Cambridge University Press, Cambridge, UK, 1994.

[4] O. Axelsson, I. Faragó, and J. Karátson, *Sobolev space preconditioning for Newton's method using domain decomposition*, Numer. Linear Algebra Appl., 9 (2002), pp. 585–598.

[5] O. Axelsson and I. Gustafsson, *An efficient finite element method for nonlinear diffusion problems*, Bull. Greek Math. Soc., 32 (1991), pp. 45–61.

[6] O. Axelsson and J. Karátson, *Double Sobolev gradient preconditioning for elliptic problems*, Numer. Partial Differential Equations, to appear.

[7] O. Axelsson and J. Maubach, *On the updating and assembly of the Hessian matrix in finite element methods*, Comput. Methods Appl. Mech. Engrg., 71 (1988), pp. 41–67.

[8] O. Axelsson and M. Nikolova, *A generalized conjugate gradient minimum residual method (GCG-MR) with variable preconditioners and a relation between residuals of the GCG-MR and GCG-OR methods*, Commun. Appl. Anal., 1 (1997), pp. 371–388.

[9] M. Al-Baali and R. Fletcher, *On the order of convergence of preconditioned nonlinear conjugate gradient methods*, SIAM J. Sci. Comput., 17 (1996), pp. 658–665.

[10] R.E. Bank and D.J. Rose, *Global approximate Newton methods*, Numer. Math., 37 (1981), pp. 279–295.

[11] P. Concus, *Numerical solution of the nonlinear magnetostatic field equation in two dimensions*, J. Comput. Phys., 1 (1967), pp. 330–342.

[12] R.S. Dembo, S.C. Eisenstat, and T. Steihaug, *Inexact Newton methods*, SIAM J. Numer. Anal., 19 (1982), pp. 400–408.

[13] J.E. Dennis, Jr., and J.J. Moré, *Quasi-Newton methods, motivation and theory*, SIAM Rev., 19 (1977), pp. 46–89.

[14] P. Deuflhard and M. Weiser, *Global inexact Newton multilevel FEM for nonlinear elliptic problems*, in Multigrid Methods V, Lect. Notes Comput. Sci. Eng. 3, Springer, Berlin, 1998, pp. 71–89.

[15] I. Faragó and J. Karátson, *The gradient–finite element method for elliptic problems*, Comput. Math. Appl., 42 (2001), pp. 1043–1053.

[16] I. Faragó and J. Karátson, *Numerical solution of nonlinear elliptic problems via preconditioning operators. Theory and applications*, in Advances and Computation, Vol. 11, NOVA Science Publishers, New York, 2002.

[17] H. Gajewski, K. Gröger, and K. Zacharias, *Nichtlineare Operatorgleichungen und Operatordifferentialgleichungen,* Akademie-Verlag, Berlin, 1974.

[18] R. Glowinski and A. Marrocco, *Analyse numérique du champ magnétique d'un alternateur par éléments finis et sur-relaxation ponctuelle non linéaire*, Comput. Methods Appl. Mech. Engrg., 3 (1974), pp. 55–85.

[19] L.V. Kantorovich and G.P. Akilov, *Functional Analysis,* Pergamon Press, Oxford, UK, 1982.

[20] J. Karátson, *Sobolev space preconditioning of strongly nonlinear* 4*th order elliptic problems*, in Numerical Analysis and Its Applications, Lecture Notes in Comput. Sci. 1988, L. Vulkov, J. Wasniewski, and P. Yalamov, eds., Springer, Berlin, 2001, pp. 459–466.

[21] J. Karátson and I. Faragó, *Preconditioning operators and Sobolev gradients for nonlinear elliptic problems*, Comput. Math. Appl., to appear.

[22] C.T. Kelley, *Iterative Methods for Linear and Nonlinear Equations*, Frontiers Appl. Math. 16, SIAM, Philadelphia, 1995.

[23] M. Křižek and P. Neittaanmäki, *Mathematical and Numerical Modeling in Electrical Engineering: Theory and Applications*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1996.

[24] S.G. Mikhlin, *The Numerical Performance of Variational Methods,* Walters-Noordhoff, Groningen, The Netherlands, 1971.

[25] J.W. Neuberger, *Sobolev Gradients and Differential Equations*, Lecture Notes in Math. 1670, Springer, Berlin, 1997.

[26] W.B. Richardson, Jr., *Sobolev gradient preconditioning for PDE applications*, in Iterative Methods in Scientific Computation IV, D.R. Kincaid and A.C. Elster, eds., IMACS, New Brunswick, NJ, 1999, pp. 223–234.

SIAM J. Numer. Anal.
Vol. 41, No. 4, pp. 1263–1285

TWO-LEVEL METHOD BASED ON FINITE ELEMENT AND CRANK–NICOLSON EXTRAPOLATION FOR THE TIME-DEPENDENT NAVIER–STOKES EQUATIONS*

YINNIAN HE[†]

Abstract. A fully discrete two-level finite element method (the two-level method) is presented for solving the two-dimensional time-dependent Navier–Stokes problem. The method requires a Crank–Nicolson extrapolation solution $(u_{H,\tau_0}, p_{H,\tau_0})$ on a spatial-time coarse grid J_{H,τ_0} and a backward Euler solution $(u^{h,\tau}, p^{h,\tau})$ on a space-time fine grid $J_{h,\tau}$. The error estimates of optimal order of the discrete solution for the two-level method are derived. Compared with the standard Crank–Nicolson extrapolation method (the one-level method) based on a space-time fine grid $J_{h,\tau}$, the two-level method is of the error estimates of the same order as the one-level method in the H^1-norm for velocity and the L^2-norm for pressure. However, the two-level method involves much less work than the one-level method.

Key words. Navier–Stokes equations, mixed finite element, two-level method, Crank–Nicolson extrapolation

AMS subject classifications. 35L70, 65N30, 76D06

DOI. 10.1137/S0036142901385659

1. Introduction.

The two-grid strategy is closely related to the finite element nonlinear Galerkin methods (see [2, 16, 26, 27]) and has been widely studied for steady semilinear elliptic equations (see the work of Xu [34, 35]), for steady Navier–Stokes equations (see, for example, the work of Layton [22], Layton and Lenferink [23], Ervin and Layton [10], Ervin, Layton, and Maubach [11], Layton and Tobiska [24], and Girault and Lions [12]), and for time-dependent Navier–Stokes equations (see the work of Girault and Lions [13] and Olshanskii [28]). Moreover, the Crank–Nicolson scheme with second-order accuracy is adapted to the time discretization of the Navier–Stokes equations by Heywood and Rannacher [18]. The well-known Crank–Nicolson extrapolation scheme with second order accuracy is applied to the time discretization of the Navier–Stokes equations by Girault and Raviart [14] and Simo and Armero [30]. Also, the Crank–Nicolson extrapolation scheme is applied to the time discretization of the nonlinear parabolic equations (see Douglas and Dupont [8], Cannon and Lin [6], and Lin [25]) and the nonlinear dynamics (see Simo, Tarnow, and Wong [31]). Other higher-order time discrete schemes are given by Baker, Dougalis, and Karakashian [3] for the time-dependent Navier–Stokes equations and Dupont, Fairweather, and Johnson [9] for the nonlinear parabolic equations.

In the case of the nonlinear evolution problem, the basic idea of the two-level method is to find an approximation u_H by solving a nonlinear problem on a coarse grid with grid size H and find an approximation u^h by solving a linearized problem about the known approximation u_H on a fine grid with grid size h. In [13], Girault and Lions consider the semidiscretization in space of the three-dimensional time-dependent Navier–Stokes equations by the two-level method, where the error

*Received by the editors February 27, 2001; accepted for publication (in revised form) January 2, 2003; published electronically August 22, 2003. This research was subsidized by the Special Funds for Major State Basic Research Projects G1999032801-07, the National High Technology Research and Development Program of China (863 Program, 2001AA111042), and the NSF of China (19971067).
http://www.siam.org/journals/sinum/41-4/38565.html

[†]Faculty of Science, Xi'an Jiaotong University, Xi'an 710049, People's Republic of China (heyn@mail.xjtu.edu.cn; heyinnian@hotmail.com).

estimates are provided. If $H = O(h^{1/2})$ is chosen, then the two-level method is of the convergence rate of the same order as the standard Galerkin method. Furthermore, Olshanskii in [28] deals mainly with the full discretization in space time of the two-dimensional and three-dimensional time-dependent Navier–Stokes equations by the two-level method, where the local error estimates, stability, and convergence are proved, but the global error estimates are not provided. In fact, this scheme is global first-order accurate with respect to the time step size τ.

In this paper we want to combine this two-grid strategy with the Crank–Nicolson extrapolation scheme with second-order accuracy to solve the two-dimensional time-dependent Navier–Stokes equations. We set J_H as a coarse grid of $\bar{\Omega}$ into triangles or quadrilaterals with mesh size H, assumed to be uniformly regular in the usual sense, and τ_0 denotes a large time step size; this spatial-time grid is denoted J_{H,τ_0}. Next, a fine grid J_h can be thought of as generated from the coarse mesh J_H by a mesh refinement process, and therefore nested, and a small time step size $\tau = \frac{\tau_0}{k}$ (k is a fixed integer), and this space-time grid is denoted $J_{h,\tau}$. Here, the basic idea of our two-level method is to find an approximation $(u_{H,\tau_0}, p_{H,\tau_0})$ by solving the linearized Navier–Stokes problem on a space-time coarse grid J_{H,τ_0} and to then find an approximation $(u^{h,\tau}, p^{h,\tau})$ by solving the Stokes problem on a space-time fine grid $J_{h,\tau}$, where $h \ll H, \ \ \tau \ll \tau_0$; here our scheme is second-order accurate with respect to time discretization.

The finite element space pairs (X_H, M_H) and (X_h, M_h) are constructed based on the coarse mesh J_H and the fine mesh J_h, which possess some approximate properties and the so-called inf-sup condition stated in section 3. Now, let us consider the numerical solution $(u^{h,\tau}, p^{h,\tau})$ of the two-dimensional time-dependent Navier–Stokes equations by the two-level method. In the first step, our approximation $(u_{H,\tau_0}, p_{H,\tau_0})$ is computed on a space-time coarse grid J_{H,τ_0} using the Crank–Nicolson extrapolation scheme. In the second step, an approximation $(u^{h,\tau}, p^{h,\tau})$ is computed on a spatial-time fine grid $J_{h,\tau}$ using the backward Euler scheme "around" u_{H,τ_0}. Then under quite general circumstances, we have obtained the error estimate of optimal order for the one-level finite element solution $(u_{h,\tau}, p_{h,\tau})$:

$$\|u(t) - u_{h,\tau}(t)\|_{H_0^1} \le \kappa h + \kappa\sigma^{-1/2}(t)\tau \quad \forall 0 < t \le T, \tag{1.1}$$

$$\|p(t) - p_{h,\tau}(t)\|_{L^2} \le \kappa\sigma^{-1/2}(t)h + \kappa\sigma^{-3/2}(t)\tau \quad \forall 0 < t \le T; \tag{1.2}$$

and we have obtained the error estimate of optimal order for the two-level finite element solution $(u^{h,\tau}, p^{h,\tau})$:

$$\begin{aligned}\|u(t) - u^{h,\tau}(t)\|_{H_0^1} &\le \kappa h + \kappa H(1-t)\sigma^{-1/2}(t)(\tau + H^{3/2} + \tau_0^{3/2}) \\ &\quad + \kappa H(t-1)(\tau + H^2 + \tau_0^2) \quad \forall 0 < t \le T,\end{aligned} \tag{1.3}$$

$$\|p(t) - p^{h,\tau}(t)\|_{L^2} \le \kappa\sigma^{-1/2}(t)h + \kappa\sigma^{-3/2}(t)(\tau + \tau_0^2 + H^2) \quad \forall 0 < t \le T \tag{1.4}$$

on the fine grid $J_{h,\tau}$, where $\sigma(t) = \min\{1, t\}$, $H(t) = 1$ as $t \ge 0$, and $H(t) = 0$ as $t < 0$. Hence, if one chooses (H, τ_0) such that $\tau_0^{3/2} + H^{3/2} = O(\tau)$ for $t \in [0, 1]$ and $\tau_0^2 + H^2 = O(\tau)$ for $t \in (1, T]$, then the two-level solution $(u^{h,\tau}, p^{h,\tau})$ is of the convergence rate of the same order as the one-level solution $(u_{h,\tau}, p_{h,\tau}$; but the velocity $u^{h,\tau}$ of the two-level solution is less accurate than the velocity $u_{h,\tau}$ of the one-level solution in the L^2-norm); see Theorem 5.6 and Lemma 6.3.

Of course, the computation of $(u^{h,\tau}, p^{h,\tau})$ involves much less "work" than the direct computation of $(u_{h,\tau}, p_{h,\tau})$. In fact, to find $(u^{h,\tau}, p^{h,\tau})$ we need to solve the discrete linearized Navier–Stokes problem by using the Crank–Nicolson extrapolation scheme on a space-time coarse grid J_{H,τ_0} and solve the discrete Stokes problem by using the Euler scheme on a fine space-time grid $J_{h,\tau}$; see section 6. However, the computation of $(u_{h,\tau}, p_{h,\tau})$ needs to solve a discrete linearized Navier–Stokes problem by using the extrapolation Crank–Nicolson scheme on a space-time fine grid $J_{h,\tau}$; see section 5.

The contents of this paper are divided into sections as follows. In section 2, functional setting of the Navier–Stokes problem is given with some basic statements. Finite element Galerkin approximation is recalled in section 3. Some key technical lemmas and known results are provided in section 4. The error estimates of the one-level finite element solution $(u_{h,\tau}, p_{h,\tau})$ will be proven in section 5. The error estimates of the two-level finite element solution $(u^{h,\tau}, p^{h,\tau})$ will be derived in section 6.

2. Functional setting of the Navier–Stokes equations. Let Ω be a bounded domain in R^2 assumed to have a Lipschitz continuous boundary $\partial\Omega$ and to satisfy a further condition stated in (A1) below. We consider the time-dependent Navier–Stokes problem

$$\begin{cases} u_t - \nu\Delta u + (u\cdot\nabla)u + \nabla p = f, \ \operatorname{div} u = 0 \ \ \forall (x,t) \in \Omega\times(0,T]; \\ u(x,0) = u_0(x) \ \ \forall x\in\Omega; \quad u(x,t)|_{\partial\Omega} = 0 \ \ \forall t\in[0,T], \end{cases} \tag{2.1}$$

where $u = u(x,t) = (u_1(x,t), u_2(x,t))$ represents the velocity vector, $p = p(x,t)$ the pressure, $f = f(x,t)$ the prescribed body force, $u_0(x)$ the initial velocity, $\nu > 0$ the viscosity, and $T > 0$ a finite time. For the mathematical setting of problem (2.1), we introduce the following Hilbert spaces:

$$X = H_0^1(\Omega)^2, \quad Y = L^2(\Omega)^2, \quad M = L_0^2(\Omega) = \left\{ q\in L^2(\Omega); \int_\Omega q dx = 0 \right\}.$$

The space $L^2(\Omega)^d$, $d = 1,2,4$, is equipped with the usual L^2-scalar product $(\cdot,\cdot)$ and L^2-norm $\|\cdot\|_{L^2}$. The spaces $H_0^1(\Omega)$ and X are equipped with their usual scalar product and equivalent norm

$$((u,v)) = (\nabla u, \nabla v), \quad \|u\|_{H_0^1} = \|\nabla u\|_{L^2}.$$

Next, let the closed subset V of X be given by

$$V = \{v\in X; d(v,q) = 0 \ \ \forall q\in M\} = \{v\in X; \operatorname{div} v = 0\}$$

and denote H the closed subset of Y, i.e.,

$$H = \{v\in Y; \operatorname{div} v = 0, v\cdot n|_{\partial\Omega} = 0\}.$$

We refer the readers to [1, 15, 17, 32] for more details on these spaces. We also denote the Stokes operator by $A = -P\Delta$, where P is the L^2-orthogonal projection of Y onto H.

As mentioned above, we need a further assumption on Ω provided in [17].

(A1). Assume that Ω is smooth so that the unique solution $(v,q)\in(X,M)$ of the steady Stokes problem

$$-\nu\Delta v + \nabla q = g, \quad \operatorname{div} v = 0 \quad \text{in } \Omega, \quad v|_{\partial\Omega} = 0,$$

for any prescribed $g \in Y$ exists and satisfies

$$\|v\|_{H^2} + \|q\|_{H^1} \leq c\|g\|_{L^2},$$

where $c > 0$ is a generic constant depending on Ω and ν which may stand for different values at its different occurrences.

We remark that the validity of assumption (A1) is known (see [15, 17, 20, 32]) if $\partial\Omega$ is of C^2 or if Ω is a two-dimensional convex polygon. From assumption (A1), it is well known [1, 17, 21] that

$$\|v\|_{H^2} \leq c\|Av\|_{L^2} \quad \forall v \in D(A) = H^2(\Omega)^2 \cap V, \tag{2.2}$$

$$\|v\|_{L^2} \leq \gamma_0\|v\|_{H_0^1} \quad \forall v \in X, \quad \|v\|_{H_0^1} \leq \gamma_0\|Av\|_{L^2} \quad \forall v \in D(A), \tag{2.3}$$

where γ_0 is positive constant depending only on Ω.

We usually make the following assumptions about the prescribed data for problem (2.1).

(A2). The initial velocity $u_0(x)$ and force $f(x,t)$ satisfy that $u_0 \in D(A)$, f, f_t, $f_{tt} \in L^\infty(0,T;Y)$ with

$$\|Au_0\|_{L^2} + \sup_{t\in[0,T]}\{\|f(t)\|_{L^2} + \|f_t(t)\|_{L^2} + \|f_{tt}(t)\|_{L^2}\} \leq C$$

for some positive constant C. We also introduce the following bilinear operator:

$$B(u,v) = (u\cdot\nabla)v + \frac{1}{2}(\mathrm{div}u)v \quad \forall u,v \in X.$$

Moreover, we define the continuous bilinear forms $a(\cdot,\cdot)$ and $d(\cdot,\cdot)$ on $X \times X$ and $X \times M$, respectively, by

$$a(u,v) = \nu((u,v)) \quad \forall u,v \in X,$$

$$d(v,q) = -(v,\nabla p) = (q,\mathrm{div}v) \quad \forall v \in X, \quad q \in M,$$

and a trilinear form on $X \times X \times X$ by

$$\begin{aligned} b(u,v,w) = \langle B(u,v),w\rangle_{X',X} &= ((u\cdot\nabla)v,w) + \frac{1}{2}((\mathrm{div}u)v,w) \\ &= \frac{1}{2}((u\cdot\nabla)v,w) - \frac{1}{2}((u\cdot\nabla)w,v) \quad \forall u,v,w \in X. \end{aligned}$$

With the above notations, the variational formulation of problem (2.1) reads as follows: find $(u,p) \in (X,M)$ for all $t \in [0,T]$ such that for all $(v,q) \in (X,M)$,

$$(u_t,v) + a(u,v) - d(v,p) + d(u,q) + b(u,u,v) = (f,v), \tag{2.4}$$

$$u(0) = u_0. \tag{2.5}$$

3. Finite element Galerkin approximation. Let $h > 0$ be a real positive parameter. Finite element subspace (X_h, M_h) of (X,M) is characterized by $J_h = J_h(\Omega)$, a partitioning of $\bar{\Omega}$ into triangles K or quadrilaterals K, assumed to be uniformly regular as $h \to 0$. For further details, the reader can refer to Ciarlet [7] and Girault and Raviart [15].

We define the subspace V_h of X_h given by

$$V_h = \left\{ v_h \in X_h\,;\, d(v_h, q_h) = 0 \ \ \forall q_h \in M_h \right\}. \tag{3.1}$$

Let $P_h : Y \rightarrow V_h$ denote the L^2-orthogonal projection defined by

$$(P_h v,\ v_h) = (v, v_h) \ \ \forall v \in Y,\ v_h \in V_h.$$

We assume that the couple (X_h, M_h) satisfies the following approximation properties: for each $v \in H^2(\Omega)^2 \cap X$ and $q \in H^1(\Omega) \cap M$, there exist approximations $\pi_h v \in X_h$ and $\rho_h q \in M_h$ such that

$$\begin{cases} d(v - \pi_h v, q_h) = 0 \ \forall q_h \in M_h, \\ \|v - \pi_h v\|_{H_0^1} \le ch\|v\|_{H^2},\ \|q - \rho_h q\|_{L^2} \le ch\|q\|_{H^1}, \end{cases} \tag{3.2}$$

together with the inverse inequality

$$\|v_h\|_{H_0^1} \le ch^{-1}\|v_h\|_{L^2} \ \ \forall v_h \in X_h; \tag{3.3}$$

and we have the so-called inf-sup inequality: for each $q_h \in M_h$, there exists $v_h \in X_h, v_h \neq 0$ such that

$$d(v_h, q_h) \ge \beta\|q_h\|_{L^2}\|v_h\|_{H_0^1}\,, \tag{3.4}$$

where $\beta > 0$ is a constant depending on Ω.

The following properties are classical (see [2, 15, 17, 19]):

$$\|P_h v\|_{H_0^1} \le c\|v\|_{H_0^1} \ \ \forall v \in X, \tag{3.5}$$

$$\|v - P_h v\|_{L^2} + h\|v - P_h v\|_{H_0^1} \le ch^2\|Av\|_{L^2} \ \ \forall v \in D(A), \tag{3.6}$$

$$\|v - P_h v\|_{L^2} \le ch\|v - P_h v\|_{H_0^1} \ \ \forall v \in X. \tag{3.7}$$

The standard finite element Galerkin approximation of (2.4)–(2.5) based on (X_h, M_h) reads as follows: find $(u_h, p_h) \in (X_h, M_h)$ such that for all $0 < t \le T$ and $(v_h, q_h) \in (X_h, M_h)$,

$$(u_{ht}, v_h) + a(u_h, v_h) - d(v_h, p_h) + d(u_h, q_h) + b(u_h, u_h, v_h) = (f, v_h), \tag{3.8}$$

$$u_h(0) = u_{0h} = P_h u_0. \tag{3.9}$$

With the above statements, a discrete analogue $A_h = -P_h\Delta_h$ of the Stokes operator A is defined through the condition that $(-\Delta_h u_h, v_h) = ((u_h, v_h))$ for all $u_h, v_h \in X_h$. The restriction of A_h to V_h is invertible, with inverse denoted A_h^{-1}. Since A_h^{-1} is self-adjoint and positive definite, we may define "discrete" Sobolev norms on V_h, of any order $r \in R$, by setting

$$\|v_h\|_r = \|A_h^{r/2} v_h\|_{L^2} \ \ \forall v_h \in V_h.$$

These norms will be assumed to have various properties similar to their continuous counterparts, an assumption that implicitly imposes conditions on the structure of the spaces X_h and M_h. In particular, there holds

$$\|v_h\|_0 = \|v_h\|_{L^2}, \ \ \|v_h\|_1 = \|v_h\|_{H_0^1} \ \ \forall v_h \in V_h.$$

By the way, we derive from (2.3) that

$$\|v_h\|_0 \le c\|v_h\|_1, \quad \|v_h\|_1 \le c\|A_h v_h\|_0 \quad \forall v_h \in V_h. \tag{3.10}$$

Under the conditions above, and with some further assumptions about the structure of the spaces X_h and M_h, it has been shown in Heywood and Rannacher [17] that

$$\|u(t)-u_h(t)\|_{L^2} + h\|u(t)-u_h(t)\|_{H_0^1} + \sigma^{1/2}(t)h\|p(t)-p_h(t)\|_{L^2} \le \kappa h^2 \tag{3.11}$$

for all $t \in (0,T]$.

4. Technical preliminaries. This section considers preliminary estimates which will be very useful in the error estimates of finite element solution (u_h, p_h). Now we will provide the following estimates of the trilinear form b.

LEMMA 4.1. *The trilinear form b satisfies the following estimates:*

$$b(u_h, v_h, w_h) = -b(u_h, w_h, v_h), \tag{4.1}$$

$$|b(u_h, v_h, w_h)| + |b(w_h, v_h, u_h)| \le c\|u_h\|_0^{1/2}\|u_h\|_1^{1/2}\|v_h\|_1\|w_h\|_1 \tag{4.2}$$

for all $u_h, v_h, w_h \in X_h$ and

$$|b(u_h, v_h, w_h)| + |b(w_h, v_h, u_h)| \le c\|A_h v_h\|_0^{1/2}\|v_h\|_1^{1/2}\|u_h\|_1\|w_h\|_0, \tag{4.3}$$

$$|b(u_h, v_h, w_h)| + |b(w_h, v_h, u_h)| \le c\|u_h\|_{-1}\|A_h v_h\|_0\|A_h w_h\|_0 \tag{4.4}$$

for all $u_h, v_h, w_h \in V_h$.

Proof. It is well known [17, 18] that (4.1)–(4.3) are valid. To prove (4.4), we need some discrete Gagliardo–Nireberg estimates (see [19]):

$$\|v_h\|_{L^4} \le c\|v_h\|_0^{1/2}\|v_h\|_1^{1/2} \quad \forall v_h \in X_h, \tag{4.5}$$

$$\|\nabla v_h\|_{L^4} \le c\|v_h\|_1^{1/2}\|A_h v_h\|_0^{1/2}, \quad \|v_h\|_{L^\infty} \le c\|v_h\|_0^{1/2}\|A_h v_h\|_0^{1/2} \quad \forall v_h \in V_h, \tag{4.6}$$

the proof of which is identical to that given by Heywood and Rannacher [17, inequalities (4.37) and (4.39), p. 298]. From the definition of b, there holds the following estimate:

$$\begin{aligned}
&|b(u_h, v_h, w_h)| + |b(w_h, v_h, u_h)| \\
&\qquad \le \frac{1}{2}\|A_h^{-1/2}u_h\|_{L^2}\|A_h^{1/2}((\nabla v_h)w_h - (\nabla w_h)v_h)\|_{L^2} \\
&\qquad + \left\|A_h^{1/2}((w_h\cdot\nabla)v_h + \frac{1}{2}(\mathrm{div} w_h)v_h)\right\|_{L^2}\|A_h^{-1/2}u_h\|_{L^2} \\
&\qquad \le c\|u_h\|_{-1}(\|\nabla v_h\|_{L^4}\|\nabla w_h\|_{L^4} + \|A_h v_h\|_{L^2}\|w_h\|_{L^\infty} + \|v_h\|_{L^\infty}\|A_h w_h\|_{L^2}).
\end{aligned}$$

Combining this inequality with (4.5)–(4.6) and using (3.10) yields (4.4). □

In order to obtain our error analysis for the time discretization, we will recall the following smooth properties of (u_h, p_h) proved in [18].

Theorem 4.2. *Assume that assumptions* (A$_1$)–(A$_2$) *and* (3.2)–(3.4) *are valid. Then the finite element solution* (u_h, p_h) *satisfies the following estimates:*

$$\|u_h(t)\|_2^2 + \|p_h(t)\|_0^2 \le \kappa, \quad \sigma^r(t)\|u_{ht}(t)\|_r^2 \le \kappa, \quad r = 0, 1, 2, \tag{4.7}$$

$$\sigma^{r+2}\|u_{htt}(t)\|_r^2 \le \kappa, \quad r = -1, 0, 1, \tag{4.8}$$

$$\int_0^t \sigma^r(s)\|u_{ht}(s)\|_{r+1}^2 ds \le \kappa, \quad r = 0, 1, \tag{4.9}$$

$$\int_0^t \sigma^{r+1}(s)\|u_{htt}(s)\|_r^2 ds \le \kappa, \quad r = -1, 0, 1, \tag{4.10}$$

$$\int_0^t \sigma^{r+2}(s)\|u_{httt}(s)\|_{r-1}^2 ds \le \kappa, \quad r = -1, 0, 1 \tag{4.11}$$

for all $t \in [0, T]$.

We will frequently use a discrete version of the Gronwall lemmas used in [18] and [29].

Lemma 4.3. *Let* C, τ, *and* a_n, b_n, c_n, d_n, *for integers* $n \ge 0$, *be nonnegative numbers such that*

$$a_m + \tau\sum_{n=0}^m b_n \le \tau\sum_{n=0}^m a_n d_n + \tau\sum_{k=0}^m c_n + C \quad \forall m \ge 1. \tag{4.12}$$

Suppose that $d_n\tau < 1$, *for all* n, *and set* $\gamma_n \equiv (1 - d_n\tau)^{-1}$. *Then*

$$a_m + \Delta t\sum_{n=0}^m b_n \le \exp\left(\tau\sum_{n=0}^m \gamma_n d_n\right)\left(\tau\sum_{n=0}^m c_n + C\right) \quad \forall m \ge 1. \tag{4.13}$$

Lemma 4.4. *Let* C, τ, *and* a_n, b_n, c_n, d_n, *for integers* $n \ge 0$, *be nonnegative numbers such that*

$$a_m + \tau\sum_{n=0}^m b_n \le \tau\sum_{n=0}^{m-1} a_n d_n + \tau\sum_{n=0}^{m-1} c_n + C \quad \forall m \ge 1. \tag{4.14}$$

Then

$$a_m + \tau\sum_{n=0}^m b_n \le \exp\left(\tau\sum_{n=0}^{m-1} d_n\right)\left(\tau\sum_{n=0}^{m-1} c_n + C\right) \quad \forall m \ge 1. \tag{4.15}$$

5. One-level finite element method. In this section we consider the time discretization of the finite element Galerkin approximation (3.8)–(3.9). Let $t_n = n\tau(n = 0, 1, \ldots, N)$ be the discrete point, $\tau = \frac{T}{N}$ the time step size, and N an integer. The Crank–Nicolson extrapolation scheme on the space-time grid $J_{h,\tau}$ is to determine function pair $(u_{h,\tau}(t), p_{h,\tau}(t))$ on $[0, T]$ such that for $t \in (t_{n-1}, t_n], 1 \le n \le N$, we define one-level finite element solution $(u_{h,\tau}(t), p_{h,\tau}(t))$ as follows:

$$u_{h,\tau}(t) = u_h^{n-1} + (t - t_{n-1})d_t u_h^n, \quad p_{h,\tau}(t) = p_h^n,$$

where $u_{h,\tau}(0) = u_{0h} = u_h^0$, $\{u_h^n\}_{n=1}^N \subset V_h$, $\{p_h^n\}_{n=1}^N \subset M_h$ as the solution of the recursive linear equation:

$$(d_t u_h^n, v_h) + a(\bar{u}_h^n, v_h) - d(v_h, p_h^n) + d(\bar{u}_h^n, q_h) + b(\phi(\bar{u}_h^n), \bar{u}_h^n, v_h) = (\bar{f}(t_n), v_h) \tag{5.1}$$

for all $(v_h, q_h) \in (X_h, M_h)$. Here and after, we often use the following notations:

$$\phi(\bar{u}_h^n) = H(1-n)\bar{u}_h^n + H(n-2)\left(\frac{3}{2}u_h^{n-1} - \frac{1}{2}u_h^{n-2}\right),$$

$$\bar{u}_h^n = \frac{1}{2}(u_h^n + u_h^{n-1}), \quad \bar{u}_h(t_n) = \frac{1}{2}(u_h(t_n) + u_h(t_{n-1})), \quad d_t u_h^n = \frac{u_h^n - u_h^{n-1}}{\tau}.$$

From the definition of ϕ, $(u_h^1, p_h^1) = (u_*^1, p_*^1)$ is defined by the Crank–Nicolson scheme, where the scheme is as follows: find $\{u_*^n\}_{n=1}^N \subset V_h$ and $\{p_*^n\}_{n=1}^N \subset M_h$ such that

$$(5.2)\ \ (d_t u_*^n, v_h) + a(\bar{u}_*^n, v_h) - d(v_h, p_*^n) + d(\bar{u}_*^n, q_h) + b(\bar{u}_*^n, \bar{u}_*^n, v_h) = (\bar{f}(t_n), v_h),$$

with the initial value $u_*^0 = u_{0h}$.

An error argument given in [18] shows the following error estimate results.

Theorem 5.1. *Suppose that the assumptions* (A1)–(A2) *are valid and the couple* (X_h, M_h) *satisfies the approximate properties* (3.2)–(3.4) *and* $\tau \le \tau_T$*, where* τ_T *is a finite constant which can take different values at its different occurrences. Then there hold the following error estimates:*

$$(5.3)\qquad \|u_h(t_m) - u_*^m\|_r^2 + \tau\sum_{n=1}^m \|\bar{u}_h(t_n) - \bar{u}_*^n\|_{r+1}^2 \le \kappa\tau^{2-r}, \quad r = -2, -1, 0, 1,$$

$$(5.4)\qquad \|u_h(t_m) - u_*^m\|_r^2 + \tau\sum_{n=1}^m \|d_t(u_h(t_n) - u_*^n)\|_{r-1}^2 \le \kappa\tau^{2-r}, \quad r = -1, 0, 1, 2,$$

$$(5.5)\ \ \sigma(t_m)\|u_h(t_m) - u_*^m\|_r^2 + \tau\sum_{n=1}^m \sigma(t_n)\|u_h(t_n) - u_*^n\|_{r+1}^2 \le \kappa\tau^{3-r}, \quad r = 0, 1,$$

$$(5.6)\ \ \sigma(t_m)\|u_h(t_m) - u_*^m\|_0 \le \kappa\tau^2, \quad \sigma^{3/2}(t_m)\|p_h(t_m) - p_*^m\|_{L^2} \le \kappa\tau$$

for all $1 \le m \le N$.

Again, in determining the velocity approximation, the pressure term can be eliminated by restricting the test functions in (5.1) and (5.2) to V_h. That is, the velocity can be determined from

$$(5.7)\qquad (d_t u_h^n, v_h) + a(\bar{u}_h^n, v_h) + b(\phi(\bar{u}_h^n), \bar{u}_h^n, v_h) = (\bar{f}(t_n), v_h) \quad \forall v_h \in V_h$$

or

$$(5.8)\qquad (d_t u_*^n, v_h) + a(\bar{u}_*^n, v_h) + b(\bar{u}_*^n, \bar{u}_*^n, v_h) = (\bar{f}(t_n), v_h) \quad \forall v_h \in V_h.$$

In order to analyze the discretization error $(u_h(t_n) - u_h^n, \bar{p}_h(t_n) - p_h^n)$, we will first consider the discrete error $(e_h^n, \mu_h^n) = (u_*^n - u_h^n, p_*^n - p_h^n)$ with $(e_h^1, \mu_h^1) = (0, 0)$ and then combine this with Theorem 5.1. However, this needs the following regularity of functions $\{u_*^n\}_{n=1}^N$.

Lemma 5.2. *Under the assumptions of Theorem* 5.1*, there holds*

$$(5.9)\qquad \|u_*^m\|_r^2 + \tau\sum_{n=1}^m (\|\bar{u}_*^n\|_r^2 + \|d_t u_*^n\|_{r-1}^2) \le \kappa, \quad 1 \le m \le N, \quad r = 0, 1, 2,$$

$$(5.10)\ \ \sigma(t_m)\|d_t u_*^m\|_1^2 + \tau\sum_{n=2}^m (\|d_{tt} u_*^n\|_{-1}^2 + \sigma(t_n)\|d_{tt} u_*^n\|_0^2) \le \kappa, \quad 2 \le m \le N.$$

Proof. First, we use (4.7), (4.9) and Theorem 5.1, obtaining

$$\|u_*^n\|_r^2 + \tau\sum_{n=1}^m \|\bar{u}_*^n\|_r^2 \le 2\|u_*^n - u_h(t_n)\|_r^2 + 2\|u_h(t_n)\|_r^2$$

$$+ 2\tau\sum_{n=1}^m (\|\bar{u}_*^n - \bar{u}_h(t_n)\|_r^2 + \|\bar{u}_h(t_n)\|_r^2) \le \kappa,$$

$$\tau\sum_{n=1}^m \|d_t u_*^n\|_{r-1}^2 \le 2\tau\sum_{n=1}^m (\|d_t u_*^n - d_t u_h(t_n)\|_{r-1}^2 + \|d_t u_h(t_n)\|_{r-1}^2)$$

$$\le \kappa + c\int_0^{t_m} \|u_{ht}\|_{r-1}^2 dt + c\int_0^{t_m} \sigma^2(t)\|u_{htt}\|_{r-1}^2 dt \le \kappa,$$

which yields (5.9).

Next, we derive from (5.8) that

$$(d_{tt}u_*^n, v_h) + a(d_t\bar{u}_*^n, v_h) + b(d_t\bar{u}_*^n, \bar{u}_*^n, v_h) + b(\bar{u}_*^{n-1}, d_t\bar{u}_*^n, v_h)$$

$$= \tau^{-1}\int_{t_{n-2}}^{t_n} (f_t(t), v_h)dt \quad \forall v_h \in V_h.$$

Taking $v_h = 2\tau d_{tt}u_*^n$ in the above relation, we obtain

$$2\tau\|d_{tt}u_*^n\|_0^2 + \nu(\|d_t u_*^n\|_1^2 - \|d_t u_*^{n-1}\|_1^2) + 2\tau b(d_t\bar{u}_*^n, \bar{u}_*^n, d_{tt}u_*^n)$$

$$+ 2\tau b(\bar{u}_*^{n-1}, d_t\bar{u}_*^n, d_{tt}u_*^n) \le \frac{\tau}{2}\|d_{tt}u_*^n\|_0^2 + c\int_{t_{n-2}}^{t_n} \|f_t(t)\|_{L^2}^2 dt. \tag{5.11}$$

In view of Lemma 4.1, there holds

$$2\tau|b(d_t\bar{u}_*^n, \bar{u}_*^n, d_{tt}u_*^n)| + 2\tau|b(d_t\bar{u}_*^n, \bar{u}_*^n, d_{tt}u_*^n)|$$

$$\le \frac{\tau}{2}\|d_{tt}u_*^n\|_0^2 + c\tau\|A_h\bar{u}_*^n\|_0^2\|d_t\bar{u}_*^n\|_1^2.$$

Combining this inequality with (5.11), using (5.9) with $r = 2$ and the fact that $\sigma(t_n) \le \sigma(t_{n-1}) + \tau$ yields

$$\tau\sigma(t_n)\|d_{tt}u_*^n\|_0^2 + \nu(\sigma(t_n)\|d_t u_*^n\|_1^2 - \sigma(t_{n-1})\|d_t u_*^{n-1}\|_1^2)$$

$$\le \nu\tau(\|d_t u_*^{n-1}\|_1^2 + \kappa\|d_t u_*^n\|_1^2) + c\int_{t_{n-2}}^{t_n} \|f_t(t)\|_{L^2}^2 dt. \tag{5.12}$$

Summing this inequality from $n = 2$ to $n = m$ and using (5.9), we arrive at

$$\tau\sum_{n=2}^m \sigma(t_n)\|d_{tt}u_*^n\|_0^2 + \nu\sigma(t_m)\|d_t u_*^m\|_1^2 \le \kappa, \quad 2 \le m \le N. \tag{5.13}$$

Finally, we derive from (4.10) and (5.4) with $r = -1$ that

$$\tau\sum_{n=2}^m \|d_{tt}u_*^n\|_{-1}^2 = \tau^{-1}\sum_{n=2}^m \|d_t u_*^n - d_t u_*^{n-1}\|_{-1}^2 \le c\tau^{-1}\sum_{n=2}^m (\|d_t u_*^n - d_t u_h(t_n)\|_{-1}^2$$

$$+ \|d_t u_*^{n-1} - d_t u_h(t_{n-1})\|_{-1}^2 + \|d_t u_h(t_n) - d_t u_h(t_{n-1})\|_{-1}^2)$$

$$\leq \kappa + c\tau^{-3}\sum_{n=2}^{m}\left\|\int_{t_{n-1}}^{t_n}(t_n-t)u_{htt}dt + \int_{t_{n-2}}^{t_{n-1}}(t-t_{n-2})u_{htt}dt\right\|_{-1}^2$$

$$\leq \kappa + c\int_0^{t_m}\|u_{htt}\|_{-1}^2 dt \leq \kappa,$$

which along with (5.13) gives (5.10). □

Lemma 5.3. *Under the assumptions of Theorem* 5.1*, there holds*

$$\|e_h^m\|_r^2 + \tau\sum_{n=1}^{m}\|\bar{e}_h^n\|_{r+1}^2 \leq \kappa\tau^{3-r}, \quad r=-1,0,1, \quad 1\leq m\leq N, \tag{5.14}$$

$$\|e_h^m\|_1^2 + \tau\sum_{n=1}^{m}\|d_t e_h^n\|_0^2 \leq \kappa\tau^2, \quad 1\leq m\leq N. \tag{5.15}$$

Proof. Subtracting (5.7) from (5.8), we obtain

$$\begin{aligned}&(d_t e_h^n, v_h) + a(\bar{e}_h^n, v_h) + b(\phi(\bar{e}_h^n), \bar{u}_*^n, v_h) + b(\phi(\bar{u}_*^n), \bar{e}_h^n, v_h)\\ &\quad - b(\phi(\bar{e}_h^n), \bar{e}_h^n, v_h) + \frac{1}{2}H(n-2)\tau^2 b(d_{tt}u_*^n, \bar{u}_*^n, v_h) = 0 \quad \forall v_h \in V_h.\end{aligned} \tag{5.16}$$

Taking $v_h = 2\tau\bar{e}_h^n$ in (5.16) and using (4.1), we get

$$\begin{aligned}&\|e_h^n\|_0^2 - \|e_h^{n-1}\|_0^2 + 2\nu\tau\|\bar{e}_h^n\|_1^2 + 2\tau b(\phi(\bar{e}_h^n), \bar{u}_*^n, \bar{e}_h^n)\\ &\quad + H(n-2)\tau^3 b(d_{tt}u_*^n, \bar{u}_*^n, \bar{e}_h^n) = 0.\end{aligned} \tag{5.17}$$

Using (4.3), it follows that

$$2\tau|b(\phi(\bar{e}_h^n), \bar{u}_*^n, \bar{e}_h^n)| \leq \frac{\nu\tau}{2}\|\bar{e}_h^n\|_1^2 + c\tau\|A_h\bar{u}_*^n\|_0^2\|\phi(\bar{e}_h^n)\|_0^2,$$

$$H(n-2)\tau^3|b(d_{tt}u_*^n, \bar{u}_*^n, \bar{e}_h^n)| \leq \frac{\nu\tau}{2}\|\bar{e}_h^n\|_1^2 + cH(n-2)\tau^5\|d_{tt}u_*^n\|_0^2\|A_h\bar{u}_*^n\|_0^2.$$

Combining these inequalities with (5.17) and applying Lemma 5.2 results in

$$\|e_h^n\|_0^2 - \|e_h^{n-1}\|_0^2 + \nu\tau\|\bar{e}_h^n\|_1^2 \leq \kappa\|\phi(\bar{e}_h^n)\|_0^2 + \kappa H(n-2)\tau^5\|d_{tt}u_*^n\|_0^2. \tag{5.18}$$

Summing (5.18) from $n=1$ to $n=m$ leads to the following estimate:

$$\|e_h^m\|_0^2 + \nu\tau\sum_{n=1}^{m}\|\bar{e}_h^n\|_1^2 \leq \kappa\tau\sum_{n=1}^{m-1}\|e_h^n\|_0^2 + \kappa\tau^4\sum_{n=2}^{m}\sigma(t_n)\|d_{tt}u_*^n\|_0^2.$$

Applying Lemma 4.4 and Lemma 5.2 to this inequality yields (5.14) with $r=0$.

Next, by setting $v_h = 2\tau d_t e_h^n$ in (5.16) and using (4.1), we get

$$\begin{aligned}&\nu(\|e_h^n\|_1^2 - \|e_h^{n-1}\|_1^2) + 2\tau\|d_t e_h^n\|_0^2 + 2\tau b(\phi(\bar{e}_h^n), \bar{u}_*^n, d_t e_h^n)\\ &\quad + 2\tau b(\phi(\bar{u}_*^n), e_h^{n-1}, d_t e_h^n) + 4b(\phi(\bar{e}_h^n), \bar{e}_h^n, e_h^{n-1})\\ &\quad + H(n-2)\tau^3 b(d_{tt}u_*^n, \bar{u}_*^n, d_t e_h^n) = 0.\end{aligned} \tag{5.19}$$

From Lemma 4.1 and (3.10), we get

$$
\begin{aligned}
2\tau|b(\phi(\bar{e}_h^n),\bar{u}_*^n,d_te_h^n)| &\le \frac{\tau}{4}\|d_te_h^n\|_0+c\tau\|A_h\bar{u}_*^n\|_0^2\|\phi(\bar{e}_h^n)\|_1^2,\\
2\tau|b(\phi(\bar{u}_*^n),e_h^{n-1},d_te_h^n)| &\le \frac{\tau}{4}\|d_te_h^n\|_0^2+c\tau\|A_h\phi(\bar{u}_*^n)\|_0^2\|e_h^{n-1}\|_1^2,\\
4|b(\phi(\bar{e}_h^n),\bar{e}_*^n,e_h^{n-1})| &\le c\|\phi(\bar{e}_h^n)\|_1\|\bar{e}_h^n\|_1\|e_h^{n-1}\|_1,\\
H(n-2)\tau^3|b(d_{tt}u_*^n,\bar{u}_*^n,A_h^{r-1}d_te_h^n)| &\le \frac{\tau}{4}\|d_te_h^n\|_0^2+cH(n-2)\tau^5\|d_{tt}u_*^n\|_1^2\|A_h\bar{u}_*^n\|_0^2.
\end{aligned}
$$

Combining these inequalities with (5.19) and using Lemma 5.2 and (5.14) with $r=0$, we see that

$$
\begin{aligned}
\nu(\|e_h^n\|_1^2-\|e_h^{n-1}\|_1^2)+\tau\|d_te_h^n\|_0^2 &\le \kappa\tau(\|e_h^{n-1}\|_1^2+\|\phi(\bar{e}_h^n)\|_1^2)\\
&\quad +H(n-2)\kappa\tau^3\|d_tu_*^n-d_tu_*^{n-1}\|_1^2.
\end{aligned}
$$

Summing this inequality from $n=1$ to $n=m$ and applying Lemma 5.2 results in

$$
\|e_h^m\|_1^2+\nu\tau\sum_{n=1}^{m}\|d_te_h^n\|_0^2 \le \kappa\tau\sum_{n=1}^{m-1}\|e_h^n\|_1^2+\kappa\tau^2.
$$

Applying Lemma 4.4 to this inequality yields (5.15).

Next, we take $v_h=2\tau A_h^r\bar{e}_h^n$ in (5.16) with $r=-1,1$, obtaining

$$
\begin{aligned}
&\|e_h^n\|_r^2-\|e_h^{n-1}\|_r^2+2\nu\tau\|\bar{e}_h^n\|_{r+1}^2+2\tau b(\phi(\bar{e}_h^n),\bar{u}_*^n,A_h^r\bar{e}_h^n)+2\tau b(\phi(\bar{u}_*^n),\bar{e}_h^n,A_h^r\bar{e}_h^n)\\
&\qquad +2\tau b(\phi(\bar{e}_h^n),\bar{e}_h^n,A_h^r\bar{e}_h^n)+H(n-2)\tau^3 b(d_{tt}u_*^n,\bar{u}_*^n,A_h^r\bar{e}_h^n)=0.
\end{aligned}
\tag{5.20}
$$

From Lemma 4.1, it follows that

$$
\begin{aligned}
2\tau|b(\phi(\bar{e}_h^n),\bar{u}_*^n,A_h^r\bar{e}_h^n)| &\le \frac{\nu\tau}{4}\|\bar{e}_h^n\|_{r+1}^2+c\tau\|A_h\bar{u}_*^n\|_0^2\|\phi(\bar{e}_h^n)\|_r^2,\\
2\tau|b(\phi(\bar{u}_*^n),\bar{e}_h^n,A_h^r\bar{e}_h^n)| &\le \frac{\nu\tau}{4}\|\bar{e}_h^n\|_r^2+c\tau\|A_h\phi(\bar{u}_*^n)\|_0^2\|\bar{e}_h^n\|_r^2,\\
2\tau|b(\phi(\bar{e}_h^n),\bar{e}_h^n,A_h^r\bar{e}_h^n)| &\le \frac{\nu\tau}{4}\|\bar{e}\|_{r+1}^2\\
&\quad +c\tau\|\phi(\bar{e}_h^n)\|_0^{1-r}\|\phi(\bar{e}_h^n)\|_1^{2+2r}\|\bar{e}_h^n\|^2,\\
\tau^3|b(d_{tt}u_*^n,\bar{u}_*^n,A_h^r\bar{e}_h^n)| &\le \frac{\nu\tau}{4}\|\bar{e}_h^n\|_{r+1}^2+c\tau^5\|d_{tt}u_*^n\|_r^2\|A_h\bar{u}_*^n\|_0^2.
\end{aligned}
$$

Combining these inequalities with (5.20) and applying Lemma 5.2 yields

$$
\begin{aligned}
\|e_h^n\|_r^2-\|e^{n-1}\|_r^2+\nu\tau\|\bar{e}_h^n\|_{r+1}^2 &\le c\tau(\|\phi(\bar{e}_h^n)\|_r^2+\|\bar{e}_h^n\|_r^2)\\
&\quad +c\tau\|\phi(\bar{e}_h^n)\|_0^{1-r}\|\phi(\bar{e}_h^n)\|_1^{2+2r}\|\bar{e}_h^n\|_1^2+cH(n-2)\tau^5\|d_{tt}u_*^n\|_r^2.
\end{aligned}
\tag{5.21}
$$

Summing (5.21) from $n=1$ to $n=m$ and applying Lemma 5.2, (5.15), and (5.14) with $r=0$, we obtain

$$
\|e_h^m\|_r^2+\tau\sum_{n=1}^{m}\|\bar{e}_h^n\|_{r+1}^2 \le \kappa\tau^{3-r}+\kappa\tau\sum_{n=1}^{m}\|e_h^n\|_r^2.
\tag{5.22}
$$

Applying Lemma 4.3 to this inequality with $\tau \le \tau_T$ yields (5.14) with $r = -1, 1$. □

LEMMA 5.4. *Under the assumptions of Theorem* 5.1*, there holds*

$$\sigma(t_m)\|e_h^m\|_0^2 + \tau \sum_{n=1}^{m} \sigma(t_n)\|\bar{e}_h^n\|_1^2 \le \kappa\tau^4, \quad 1 \le m \le N, \tag{5.23}$$

$$\sigma(t_m)\|d_t e_h^m\|_0^2 + \tau \sum_{n=1}^{m} \sigma(t_n)\|d_t\bar{e}_h^n\|_1^2 \le \kappa\tau^2, \quad 1 \le m \le N. \tag{5.24}$$

Proof. Multiplying (5.18) by $\sigma(t_n)$ and using Lemma 5.2 and the fact that

$$\sigma(t_n) \le \sigma(t_{n-1}) + \tau, \quad e_h^{n-1} = \bar{e}^n - \frac{\tau}{2} d_t e^n, \tag{5.25}$$

we obtain

$$\begin{aligned}\sigma(t_n)\|e_h^n\|_0^2 - \sigma(t_{n-1})\|e_h^{n-1}\|_0^2 + \nu\tau\sigma(t_n)\|\bar{e}_h^n\|_1^2 \\ \le 2\tau\|\bar{e}_h^n\|_0^2 + \tau^3\|d_t e_h^n\|_0^2 + \kappa\tau\sigma(t_n)\|\phi(\bar{e}_h^n)\|_0^2 + cH(n-2)\tau^5\sigma(t_n)\|d_{tt}u_*^n\|_0^2.\end{aligned}$$

Summing this inequality from $n = 1$ to $n = m$ and using Lemma 5.2, we obtain

$$\begin{aligned}\sigma(t_m)\|e_h^m\|_0^2 + \nu\tau \sum_{n=1}^{m} \sigma(t_n)\|\bar{e}_h^n\|_1^2 \le \kappa\tau^4 + \kappa\tau \sum_{n=0}^{m-1} \sigma(t_n)\|e_h^n\|_0^2 \\ + \tau \sum_{n=1}^{m} (2\|\bar{e}_h^n\|_0^2 + \tau^3\|d_t e_h^n\|_0^2).\end{aligned} \tag{5.26}$$

Applying Lemma 4.4 and Lemma 5.3 with $r = -1$ to this inequality yields

$$\sigma(t_m)\|e_h^m\|_0^2 + \nu\tau \sum_{n=1}^{m} \sigma(t_n)\|\bar{e}_h^n\|_1^2 \le \kappa\tau^4,$$

which gives (5.23).

Furthermore, we derive from (5.16) that

$$\begin{aligned}(d_{tt}e_h^n, v_h) + a(d_t\bar{e}_h^n, v_h) + b(\phi(d_t\bar{e}_h^n), \bar{u}_*^n, v_h) \\ + b(\phi(\bar{e}_h^{n-1}), d_t\bar{u}_*^n, v_h) + b(\phi(d_t\bar{u}_*^n), \bar{e}_h^n, v_h) \\ + b(\phi(\bar{u}_*^{n-1}), d_t\bar{e}_h^n, v_h) - b(\phi(d_t\bar{e}_h^n), \bar{e}_h^n, v_h) \\ - b(\phi(\bar{e}_h^{n-1}), d_t\bar{e}_h^n, v_h) + \frac{1}{2}H(n-2)\tau b(d_{tt}u_*^n, \bar{u}_*^n, v_h) \\ - \frac{1}{2}H(n-3)\tau b(d_{tt}u_*^{n-1}, \bar{u}_*^{n-1}, v_h) = 0\end{aligned} \tag{5.27}$$

for all $v_h \in V_h$. We take $v_h = 2\tau d_t\bar{e}_h^n$ in (5.27) and use (4.1), obtaining

$$\begin{aligned}\|d_t e_h^n\|_0^2 - \|d_t e_h^{n-1}\|_0^2 + 2\nu\tau\|d_t\bar{e}_h^n\|_1^2 + 2\tau b(\phi(d_t\bar{e}_h^n), \bar{u}_*^n, d_t\bar{e}_h^n) \\ + 2\tau b(\phi(\bar{e}_h^{n-1}), d_t\bar{u}_*^n, d_t\bar{e}_h^n) + 2\tau b(\phi(d_t\bar{u}_*^n), \bar{e}_h^n, d_t\bar{e}_h^n) \\ - 2\tau b(\phi(d_t\bar{e}_h^n), \bar{e}_h^n, d_t\bar{e}_h^n) + H(n-2)\tau^2 b(d_{tt}u_*^n, \bar{u}_*^n, d_t\bar{e}_h^n) \\ - H(n-3)\tau^2 b(d_{tt}u_*^{n-1}, \bar{u}_*^{n-1}, d_t\bar{e}_h^n) = 0.\end{aligned} \tag{5.28}$$

From Lemma 4.1, it follows that

$$\begin{aligned}
2\tau|b(\phi(d_t\bar{e}_h^n),\bar{u}_*^n,d_t\bar{e}_h^n)| &\le \frac{\nu\tau}{8}\|d_t\bar{e}_h^n\|_1^2+c\tau\|A_h\bar{u}_*^n\|_0^2\|\phi(d_t\bar{e}_h^n)\|_0^2,\\
2\tau|b(\phi(\bar{e}_h^{n-1}),d_t\bar{u}_*^n,d_t\bar{e}_h^n)| &\le \frac{\nu\tau}{8}\|d_t\bar{e}_h^n\|_1^2+c\tau\|d_t\bar{u}_*^n\|_1^2\|\phi(\bar{e}_h^{n-1})\|_1^2,\\
2\tau|b(\phi(d_t\bar{u}_*^n),\bar{e}_h^n,d_t\bar{e}_h^n)| &\le \frac{\nu\tau}{8}\|d_t\bar{e}_h^n\|_1^2+c\tau^{-1}\|A_h\phi(\bar{u}_*^n)-A_h\phi(\bar{u}_*^{n-1})\|_0^2\|\bar{e}_h^n\|_0^2,\\
2\tau|b(\phi(d_t\bar{e}_h^n),\bar{e}_h^n,d_t\bar{e}_h^n)| &\le \frac{\nu\tau}{4}\|d_t\bar{e}_h^n\|_1^2+c\tau\|\phi(d_t\bar{e}_h^n)\|_1^2\|\bar{e}_h^n\|_1^2,\\
H(n-2)\tau^2|b(d_{tt}u_*^n,\bar{u}_*^n,d_t\bar{e}_h^n)| &+ H(n-3)\tau^2|b(d_{tt}u_*^{n-1},\bar{u}_*^{n-1},d_t\bar{e}_h^n)|\\
&\le \frac{\nu\tau}{4}\|d_t\bar{e}_h^n\|_1^2+c\tau^3H(n-2)\|d_{tt}u_*^n\|_0^2\|A_h\bar{u}_*^n\|_0^2\\
&\quad+c\tau^3H(n-3)\|d_{tt}u_*^{n-1}\|_0^2\|A_h\bar{u}_*^{n-1}\|_0^2.
\end{aligned}$$

Combining these inequalities with (5.28) and using Lemma 5.2 yields

$$\begin{aligned}
\sigma(t_n)\|d_te_h^n\|_0^2 &- \sigma(t_{n-1})\|d_te_h^{n-1}\|_0^2+\nu\tau\sigma(t_n)\|d_t\bar{e}_h^n\|_1^2\\
&\le \tau\|d_te_h^{n-1}\|_0^2+c\sigma(t_n)\tau\|\phi(d_t\bar{e}_h^n)\|_0^2+c\tau\|d_t\bar{u}_*^n\|_1^2\|\bar{e}_h^n\|_1^2\\
&\quad+c\tau^{-1}\|A_h\phi(\bar{u}_*^n)-A_h\phi(\bar{u}_*^{n-1})\|_0^2\|\bar{e}_h^n\|_0^2\\
&\quad+c\tau^3(H(n-2)\sigma(t_n)\|d_{tt}u_*^n\|_0^2+H(n-3)\sigma(t_{n-1})\|d_{tt}u_*^{n-1}\|_0^2).
\end{aligned}$$

Summing this inequality from $n=1$ to $n=m$ and using Lemma 5.2 and Lemma 5.3, we obtain (5.24). □

LEMMA 5.5. *Under the assumptions of Theorem* 5.1, *there holds*

$$\|p_h^m-\bar{p}(t_m)\|_{L^2}\le\kappa\sigma^{-3/2}(t_m)\tau,\quad 1\le m\le N. \tag{5.29}$$

Proof. First, we derive from (5.1)–(5.2) that

$$\begin{aligned}
d(v_h,\mu_h^m) &= (d_te_h^m,v_h)+a(\bar{e}_h^m,v_h)+b(\phi(\bar{e}_h^m),\bar{u}_*^m,v_h)+b(\phi(\bar{u}_*^m),\bar{e}_h^m,v_h)\\
&\quad-b(\phi(\bar{e}_h^m),\bar{e}_h^m,v_h)+\frac{1}{2}H(m-2)\tau^2b(d_{tt}u_*^m,\bar{u}_*^m,v_h)\quad\forall v_h\in X_h,
\end{aligned}\tag{5.30}$$

where $\mu_h^m=p_*^m-p_h^m$. Hence, by using (3.4) and (4.2), we derive from (5.30) that

$$\begin{aligned}
\sigma(t_m)\|\mu_h^m\|_{L^2} &\le c\sigma(t_m)\|d_te_h^m\|_0+c(1+\|\phi(\bar{u}_*^m)\|_1+\|\phi(\bar{e}_h^m)\|_1)\|\bar{e}_h^m\|_1\\
&\quad+c\|\bar{u}_*^m\|_1\|\phi(\bar{e}_h^m)\|_1+\kappa\tau^2\sigma(t_m)\|d_{tt}u_*^m\|_0.
\end{aligned}$$

Using Lemmas 5.2–5.4 in the above estimate yields

$$\sigma(t_m)\|\mu_h^m\|_{L^2}\le\kappa\tau,\quad 1\le m\le N.$$

Combining this estimate with (5.6) in Theorem 5.1 yields (5.29). □

THEOREM 5.6. *Under the assumptions of Theorem* 5.1, *there holds*

$$\|u_h(t)-u_{h,\tau}(t)\|_0\le\kappa\sigma^{-1}(t)\tau^2,\|u_h(t)-u_{h,\tau}(t)\|_1\le\kappa\sigma^{-1/2}(t)\tau, \tag{5.31}$$

$$\|p_h(t)-p_{h,\tau}(t)\|_{L^2}\le\kappa\sigma^{-3/2}(t)\tau \tag{5.32}$$

for all $t \in (0, T]$.

Remark. Combining (5.31)–(5.32) with (3.11) yields (1.1)–(1.2).

Proof. First, for $t \in (0, t_1]$ there holds

$$\begin{aligned}
&\|u_{ht}(s)\|_0 + \|d_t u_h^1\|_0 \le \kappa,\\
&\left\|\int_0^t (u_{ht}(s) - d_t u_h^1) ds\right\|_1 \le \tau^{1/2}\left(\int_0^t \|u_{ht}(s)\|_1^2 ds^{1/2} + \tau^{1/2}\sigma^{1/2}(t_1)\|d_t u_h^1\|_1\right)\\
&\qquad \le \kappa\tau^{1/2}.
\end{aligned}$$

Hence, we have

$$\begin{aligned}
\sigma^{\frac{2-r}{2}}(t)\|u_h(t) - u_{h,\tau}(t)\|_r &= \sigma^{\frac{2-r}{2}}(t)\left\|\int_0^t (u_{ht}(s) - d_t u_h^1) ds\right\|_r\\
&\le \kappa\tau^{2-r}, \quad r = 0, 1. \qquad (5.33)
\end{aligned}$$

Moreover, we note that for $t \in (t_{n-1}, t_n], \quad 2 \le n \le N$, the error $u_h(t) - u_{h,\tau}(t)$ satisfies

$$\begin{aligned}
u_h(t) - u_{h,\tau}(t) &= u_h(t_{n-1}) - u_h^{n-1} + (t - t_{n-1})(d_t u_h(t_n) - d_t u_h^n)\\
&\quad + \int_{t_{n-1}}^t (u_{ht}(s) - d_t u_h(t_n)) ds. \qquad (5.34)
\end{aligned}$$

Thus,

$$\begin{aligned}
\sigma^{1-r/2}(t)\|u_h(t) - u_{h,\tau}(t)\|_r &\le \sigma^{1-r/2}(t)\|u_h(t_{n-1}) - u_h^{n-1}\|_r\\
&\quad + (t - t_{n-1})\sigma^{2-r/2}(t)\|d_t u_h(t_n) - d_t u_h^n\|_r\\
&\quad + \sigma^{1-r/2}(t)\left\|\int_{t_{n-1}}^t (u_{ht}(s) - d_t u_h(t_n)) ds\right\|_r, \quad r = 0, 1. \qquad (5.35)
\end{aligned}$$

From Theorem 4.2 and Theorem 5.1 and Lemma 5.3 and Lemma 5.4 it follows that

$$\begin{aligned}
\sigma^{1-r/2}(t)\left\|\int_{t_{n-1}}^t (u_{ht}(s) - d_t u_h(t_n)) ds\right\|_r &\le c\tau^{2-r}\sigma^{1+r/2}(t_{n-1})\|u_{htt}(\xi)\|_r\\
&\le \kappa\tau^{2-r},\\
\sigma^{1-r/2}(t)\|u_h(t_{n-1}) - u_h^{n-1}\|_r &\le \sigma^{1-r/2}(t_{n-1})(\|u_h(t_{n-1}) - u_*^{n-1}\|_r + \|e_h^{n-1}\|_r)\\
&\le \kappa\tau^{2-r},\\
\sigma(t)\|d_t u_h(t_n) - d_t u_h^n\|_0 &\le \sigma(t_n)(\|d_t u_h(t_n) - d_t u_*^n\|_0 + \|d_t e_h^n\|_0) \le \kappa\tau,\\
\sigma^{1/2}(t)\|d_t u_h(t_n) - d_t u_h^n\|_1 &\le \sigma^{1/2}(t_n)\|d_t u_h(t_n)\|_1 + \sigma^{1/2}(t_n)\|d_t u_h^n\|_1 \le \kappa.
\end{aligned}$$

Combining (5.35) with these inequalities yields

$$\sigma^{\frac{2-r}{2}}(t)\|u_h(t) - u_{h,\tau}(t)\|_r \le \kappa\tau^{2-r}, \quad r = 0, 1, \quad t \in (t_1, T], \qquad (5.36)$$

which along with (5.33) completes the proof of (5.31).

Now, we need the smoothness of p_{ht}. From (3.8) we have

$$d(v_h,p_{ht}) = (u_{htt},v_h) + a(u_{ht},v_h) + b(u_{ht},u_h,v_h) + b(u_h,u_{ht},v_h) \\ - (f_t,v_h) \quad \forall v_h \in X_h,$$

which along with (3.4), (3.10), (4.2), and (4.7)–(4.8) implies that

$$\sigma(t)\|p_{ht}(t)\|_{L^2} \le \kappa, \quad t \in (0,T]. \tag{5.37}$$

Now, we will prove (5.32). From (4.7) we derive that for $t \in (0,t_1]$,

$$\sigma^{3/2}(t)\|p_h(t) - p_{h,\tau}(t)\|_{L^2} \le \sigma^{3/2}(t)\|p_h(t) - \bar{p}_h(t_1)\|_{L^2} \\ + \sigma^{3/2}(t_1)\|\bar{p}_h(t_1) - p_h^1\|_{L^2} \le \kappa\tau. \tag{5.38}$$

Moreover, by using (5.37), we obtain that for $t \in (t_{n-1},t_n],\ 2 \le n \le N$,

$$\sigma(t)\|p_h(t) - \bar{p}_h(t_n)\|_{L^2} = \frac{1}{2}\sigma(t)\left\|\int_{t_{n-1}}^{t} p_{ht}(s)ds - \int_{t}^{t_n} p_{ht}(s)ds\right\|_{L^2} \\ \le \int_{t_{n-1}}^{t} \sigma(s)\|p_{ht}(s)\|_{L^2}ds + \frac{1}{2}\int_{t}^{t_n} \sigma(s)\|p_{ht}(s)\|_{L^2}ds \le \kappa\tau.$$

Hence, by virtue of Lemma 5.5, we obtain that for $t \in (t_{n-1},t_n],\ 2 \le n \le N$,

$$\sigma^{3/2}(t)\|p_h(t) - p_{h,\tau}(t)\|_{L^2} \le \sigma^{3/2}(t)\|p_h(t) - \bar{p}_h(t)\|_{L^2} \\ + \sigma^{3/2}(t_n)\|\bar{p}_h(t_n) - p_h^n\|_{L^2} \le \kappa\tau, \tag{5.39}$$

which along with (5.38) gives (5.32). □

THEOREM 5.7. *Under the assumptions of Theorem* 5.1*, there holds*

$$\int_0^T \|u_h(t) - u_{h,\tau}(t)\|_r^2 dt \le \kappa\tau^{3-r}, \quad r = -1,0,1, \tag{5.40}$$

$$\int_0^T \sigma^{r+1}(t)\|u_h(t) - u_{h,\tau}(t)\|_r^2 dt \le \kappa\tau^4, \quad r = 0,1. \tag{5.41}$$

Proof. First, using the integration by parts and the fact that

$$u_h(t_n) - u_h^n = \bar{u}_h(t_n) - \bar{u}_h^n + \frac{\tau}{2}(d_t u_h(t_n) - d_t u_h^n),$$

we derive

$$\int_{t_{n-1}}^{t_n} \|u_h(t) - u_{h,\tau}(t)\|_r^2 dt \le c\tau\|u_h(t_n) - u_h^n\|_r^2 + c\tau^3\|u_{ht}(t_n) - d_t u_h^n\|_r^2 \\ + c\tau^2\int_{t_{n-1}}^{t_n} (t-t_{n-1})^2\|u_{htt}(t)\|_r^2 dt \\ \le c\tau\|\bar{u}_h(t_n) - \bar{u}_*^n\|_r^2 + c\tau\|\bar{e}_h^n\|_r^2 + c\tau^3\|d_t(u_h(t_n) - u_*^n)\|_r^2 + c\tau^3\|d_t e_h^n\|_r^2 \\ + c\tau^{3-r}\int_{t_{n-1}}^{t_n} \sigma^{1+r}(t)\|u_{htt}(t)\|_r^2 dt. \tag{5.42}$$

Summing this inequality from $n = 1$ to $n = m$, we obtain

$$\int_0^T \|u_h(t) - u_{h,t}(t)\|_r^2 dt \le c\tau \sum_{n=1}^N \|\bar{u}_h(t_n) - \bar{u}_*^n\|_r^2 + c\tau^3 \sum_{n=1}^N \|d_t u_h(t_n) - d_t u_*^n\|_r^2$$

$$+ c\tau \sum_{n=1}^N \|\bar{e}_h^n\|_r^2 + c\tau^3 \sum_{n=1}^N \|d_t e_h^n\|_r^2 + c\tau^{3-r} \int_0^T \sigma^{1+r}(t)\|u_{htt}(t)\|_r^2 dt. \tag{5.43}$$

Applying Theorem 4.2 and Theorem 5.1 and Lemma 5.3 and Lemma 5.4 in (5.43), we obtain (5.40).

Moreover, we also have

$$\begin{aligned}\int_{t_{n-1}}^{t_n} \sigma^{r+1}(t)\|u_h(t) - u_{h,\tau}(t)\|_r^2 dt &\le c\tau\sigma^{r+1}(t_n)\|u_h(t_n) - u_h^n\|_r^2 \\ &+ c\tau^3\sigma^{r+1}(t_n)\|u_{ht}(t_n) - d_t u_h^n\|_r^2 \\ &+ c\tau^2 \int_{t_{n-1}}^{t_n} (t - t_{n-1})\sigma^{r+1}(t)\|u_{htt}(t)\|_r^2 dt \\ &\le c\tau\sigma^{r+1}(t_n)\|\bar{u}_h(t_n) - \bar{u}_*^n\|_r^2 + c\tau\sigma^{r+1}(t_n)\|\bar{e}_h^n\|_r^2 \\ &+ c\tau^3\sigma^{r+1}(t_n)\|d_t(u_h(t_n) - u_*^n)\|_r^2 + c\tau^3\sigma^{r+1}(t_n)\|d_t e_h^n\|_r^2 \\ &+ c\tau^4 \int_{t_{n-1}}^{t_n} \sigma^{r+1}(t)\|u_{htt}(t)\|_r^2 dt.\end{aligned} \tag{5.44}$$

Summing (5.44) from $n = 1$ to $n = m$, we obtain

$$\int_0^T \sigma^{r+1}(t)\|u_h(t) - u_{h,t}(t)\|_r^2 dt \le c\tau \sum_{n=1}^N \sigma^{r+1}(t_n)\|\bar{u}_h(t_n) - \bar{u}_*^n\|_r^2$$

$$+ c\tau^3 \sum_{n=1}^N \sigma^{r+1}(t_n)\|d_t u_h(t_n) - d_t u_*^n\|_r^2 + c\tau \sum_{n=1}^N \sigma^{r+1}(t_n)\|\bar{e}_h^n\|_r^2$$

$$+ c\tau^3 \sum_{n=1}^N \sigma^{r+1}(t_n)\|d_t e_h^n\|_r^2 + c\tau^4 \int_0^T \sigma^{r+1}(t)\|u_{htt}(t)\|_r^2 dt. \tag{5.45}$$

A further extension of some arguments provided by Heywood and Rannacher in [18] can yield the following estimates:

$$\tau \sum_{n=1}^N \sigma^{r+1}(t_n)\|\bar{u}_h(t_n) - \bar{u}_*^n\|_r^2 \le \kappa\tau^4, \tag{5.46}$$

$$\tau \sum_{n=1}^N \sigma^{r+1}(t_n)\|d_t u_h(t_n) - d_t u_*^n\|_r^2 \le \kappa\tau^2 \tag{5.47}$$

for $r = 0, 1$. Using Theorem 4.2, Lemma 5.3, Lemma 5.4, and (5.46)–(5.47) in (5.45), we obtain (5.41). □

6. Two-level finite element method.

In this section we consider a two-level finite element method for the time-dependent Navier–Stokes equations.

Given a coarse grid J_H with $H > h$ and a large time step size $\tau_0 = k\tau = \frac{T}{N_0}$ with $N_0 = \frac{N}{k}$ for some fixed integer k, the two-level finite element method applied to the finite element Galerkin approximation (3.8)–(3.9) is described as follows:

- Calculate the one-level finite element solution $u_{H,\tau_0}(t)$ on a time-spatial coarse grid J_{H,τ_0} by (5.1) with $h = H, \quad \tau = \tau_0, \quad N = N_0$.
- Given $u_{H,\tau_0}(t) \in X_H$, $t \in [0, T]$, calculate the two-level finite element solution

  $$u^{h,\tau}(t) = u^{h,n-1} + (t - t_{n-1}) d_t u^{h,n}, \;\; p^{h,\tau}(t) = p^{h,k} \;\; \forall t \in (t_{n-1}, t_n], \;\; 1 \le n \le N,$$

  on the spatial-time fine grid $J_{h,\tau}$, where $\{(u^{h,n}, p^{h,n})\}_{n=1}^N$ is mainly defined by the backward Euler scheme as follows: for all $(v_h, q_h) \in (X_h, M_h)$ with the starting value $u^{h,0} = u_{0h}$,

$$
\begin{aligned}
&(d_t u^{h,n}, v_h) + a(\bar{u}^{h,n}, v_h) - d(v_h, p^{h,n}) + d(\bar{u}^{h,n}, q_h) \\
&\quad + b(\phi(\bar{u}^{h,n}), \bar{u}_{H,\tau_0}(t_n), v_h) = (\bar{f}(t_n), v_h), \quad 1 \le n \le N,
\end{aligned}
\tag{6.1}
$$

where (6.1) is not the backward Euler scheme in the case of $n = 1$.

By using an exact similar argument as in the proof of Lemma 5.2, we can obtain the following regularity result of the finite element solutions $\{(u_h^n, p_h^n)\}_{n=1}^N$ and $\{(u^{h,n}, p^{h,n})\}_{n=1}^N$.

LEMMA 6.1. *Under the assumptions of Theorem* 5.1, *there holds*

$$\|\varphi_h^m\|_r^2 + \tau \sum_{n=1}^m (\|\bar{\varphi}_h^n\|_r^2 + \|d_t \varphi_h^n\|_{r-1}^2) \le \kappa, 1 \le m \le N, \quad r = 0, 1, 2, \tag{6.2}$$

$$\sigma(t_m)\|d_t \varphi_h^m\|^2 + \tau \sum_{n=2}^m (\|d_{tt}\varphi_h^n\|_{-1}^2 + \sigma(t_n)\|d_{tt}\varphi_h^n\|_0^2) \le \kappa, \quad 2 \le m \le N \tag{6.3}$$

for $\varphi_h^n = u_h^n, u^{h,n}$.

LEMMA 6.2. *Under the assumptions of Theorem* 5.1, *there holds*

$$\|e_h^m\|_r^2 + \tau \sum_{n=1}^m \|\bar{e}_h^n\|_{r+1}^2 \le \kappa(H^{4-r(r+1)} + \tau_0^{3-r}), \quad r = -1, 0, 1, \tag{6.4}$$

$$\|e_h^m\|_r^2 + \tau \sum_{n=1}^m \|d_t e_h^n\|_{r-1}^2 \le \kappa(H^2 + \tau_0^2)^{2-r}, \quad r = 1, 2 \tag{6.5}$$

for $1 \le m \le N$, *where* $e_h^n = u_h^n - u^{h,n}$.

Proof. Subtracting (6.1) from (5.1), we obtain

$$
\begin{aligned}
&(d_t e_h^n, v_h) + a(\bar{e}_h^n, v_h) - d(v_h, \mu_h^n) + b(\phi(\bar{e}_h^n), \bar{u}_h^n, v_h) \\
&\quad + b(\phi(\bar{u}^{h,n}), \bar{u}_{h,\tau}(t_n) - \bar{u}_{H,\tau_0}(t_n), v_h) = 0 \;\; \forall v_h \in X_h,
\end{aligned}
\tag{6.6}
$$

where $u_h^n = u_{h,\tau}(t_n)$ and $\mu_h^n = p_h^n - p^{h,n}$. We take $v_h = 2\tau A_h^r \bar{e}_h^n \in V_h$ in (6.6) with $r = -1, 0, 1$, obtaining

$$
\begin{aligned}
&\|e_h^n\|_r^2 - \|e_h^{n-1}\|_r^2 + 2\nu\tau\|\bar{e}_h^n\|_{r+1}^2 + 2\tau b(\phi(\bar{e}_h^n), \bar{u}_h^n, A_h^r \bar{e}_h^n) \\
&\quad + 2\tau b(\phi(\bar{u}^{h,n}), \bar{u}_{h,\tau}(t_n) - \bar{u}_{H,\tau_0}(t_n), A_h^r \bar{e}_h^n) = 0.
\end{aligned}
\tag{6.7}
$$

In view of Lemma 4.1, there holds

$$2\tau|b(\phi(\bar{e}_h^n),\bar{u}_h^n,A_h^r\bar{e}_h^n)|\le\frac{\nu\tau}{4}\|\bar{e}_h^n\|_{r+1}^2+c\tau\|A_h\bar{u}_h^n\|_0^2\|\phi(\bar{e}_h^n)\|_r^2,$$
$$2\tau|b(\phi(\bar{u}^{h,n}),\bar{u}_{h,\tau}(t_n)-\bar{u}_{H,\tau_0}(t_n),A_h^r\bar{e}_h^n)|$$
$$\le\frac{\nu\tau}{4}\|\bar{e}_h^n\|_{r+1}^2+c\tau\|A_h\phi(\bar{u}^{h,n})\|_0\|\bar{u}_{h,\tau}(t_n)-\bar{u}_{H,\tau_0}(t_n)\|_r^2.$$

Combining these inequalities with (6.7) and using Lemma 5.2 and Lemma 6.1 yields

$$\|e_h^n\|_r^2-\|e_h^{n-1}\|_r^2+\nu\tau\|\bar{e}_h^n\|_{r+1}^2\le\kappa\tau\|\phi(\bar{e}_h^n)\|_r^2+\kappa\tau\|\bar{u}_{h,\tau}(t_n)-\bar{u}_{H,\tau_0}(t_n)\|_r^2. \tag{6.8}$$

Summing (6.8) from $n=1$ to $n=m$ yields

$$\|e_h^m\|_r^2+\nu\tau\sum_{n=1}^{m}\|\bar{e}_h^n\|_{r+1}^2\le\kappa\tau\sum_{n=1}^{N}\|\phi(\bar{e}_h^n)\|_r^2+\kappa\tau\sum_{n=1}^{N}\|\bar{u}_H(t_n)-\bar{u}_{H,\tau_0}(t_n)\|_r^2. \tag{6.9}$$

From integration by parts there holds the following formula:

$$\bar{u}_h(t_n)=\frac{1}{\tau}\int_{t_{n-1}}^{t_n}u_h(t)dt+\frac{1}{2\tau}\int_{t_{n-1}}^{t_n}(t_n-t)(t-t_{n-1})u_{htt}(t)dt. \tag{6.10}$$

Hence, we derive from Theorem 5.7 and (3.10)–(3.11) that

$$\tau\sum_{n=1}^{N}\|u_{h,\tau}(t_n)-u_{H,\tau_0}(t_n)\|_r^2\le\sum_{n=1}^{N}\int_{t_{n-1}}^{t_n}\|u_{h,\tau}(t)-u_{H,\tau_0}(t)\|_r^2dt$$
$$\le c\int_0^T(\|u_h(t)-u_{h,\tau}(t)\|_r^2+\|u_H(t)-u_{H,\tau_0}(t)\|_r^2+\|u_h(t)-u_H(t)\|_r^2)dt \tag{6.11}$$
$$\le\kappa(H^{4-r(r+1)}+\tau_0^{3-r}),$$

which along with (6.9) yields

$$\|e_h^m\|_r^2+\nu\tau\sum_{n=1}^{m}\|\bar{e}_h^n\|_{r+1}^2\le\kappa\tau\sum_{n=1}^{m}\|\phi(\bar{e}_h^n)\|_r^2+\kappa(H^{4-r(r+1)}+\tau_0^{3-r}). \tag{6.12}$$

Applying Lemma 4.3 with $m=1$ and Lemma 4.4 with $2\le m\le N$ to (6.12) yields (6.4).

Next, we take $v_h=2\tau A_h^{r-1}d_te_h^n$ in (6.6) with $r=1,2$, obtaining

$$2\tau\|d_te_h^n\|_{r-1}^2+\nu(\|e_h^n\|_r^2-\|e_h^{n-1}\|_r^2)+2\tau b(\phi(\bar{e}_h^n),\bar{u}_h^n,A_h^{r-1}d_te_h^n)$$
$$+2\tau b(\phi(\bar{u}^{h,n}),\bar{u}_{h,\tau}(t_n)-\bar{u}_{H,\tau_0}(t),A_h^{r-1}d_te_h^n)=0. \tag{6.13}$$

In view of Lemma 4.1, there holds

$$2\tau|b(\phi(\bar{e}_h^n),\bar{u}_h^n,A_h^{r-1}d_te_h^n)|\le\frac{\tau}{4}\|d_te_h^n\|_{r-1}^2+c\tau\|A_h\bar{u}_h^n\|_0^2\|\phi(\bar{e}_h^n)\|_r^2$$
$$2\tau|b(\phi(\bar{u}^{h,n}),\bar{u}_{h,\tau}(t_n)-\bar{u}_{H,\tau_0}(t),A_h^{r-1}d_te_h^n)|$$
$$\le\frac{\tau}{4}\|d_te_h^n\|_{r-1}^2+\kappa\tau\|A_h\phi(\bar{u}^{h,n})\|_0^2\|\bar{u}_{h,\tau}(t_n)-\bar{u}_{H,\tau_0}(t_n)\|^2.$$

Combining these inequalities with (6.13) and using Lemma 5.2 yields

$$\tau\|d_te_h^n\|_{r-1}^2+\nu(\|e_h^n\|_r^2-\|e_h^{n-1}\|_r^2)\le\kappa\tau\|\phi(\bar e_h^n)\|_r^2+\kappa\tau\|\bar u_h^n-\bar u_{H,\tau_0}(t_n)\|_r^2. \tag{6.14}$$

Summing this inequality from $n=1$ to $n=m$ and using (6.11) with $r=1$ and Lemma 6.1 with $r=2$, we obtain

$$\nu\|e_h^m\|_r^2+\tau\sum_{n=1}^m\|d_te_h^n\|_{r-1}^2\le\kappa\tau\sum_{n=1}^m\|\phi(\bar e_h^n)\|_r^2+\kappa(H^2+\tau_0^2)^{2-r}. \tag{6.15}$$

Applying Lemma 4.3 with $m=1$ and Lemma 4.4 with $2\le m\le N$ to (6.15) yields (6.5). □

LEMMA 6.3. *Under the assumptions of Theorem* 5.1*, there holds*

$$\sigma(t_m)\|e_h^m\|_0^2+\tau\sum_{n=1}^m\sigma(t_n)\|\bar e_h^n\|_1^2\le\kappa(H^4+\tau_0^4)\quad\forall 1\le m\le N. \tag{6.16}$$

Proof. For $n=1$ we derive from Lemma 6.2 that

$$\sigma(t_n)\|e_h^n\|_0^2+\tau\sigma(t_n)\|\bar e_h^n\|_1^2\le\kappa(H^4+\tau_0^4). \tag{6.17}$$

Multiplying (6.8) with $r=0$ by $\sigma(t_n)$ and noting

$$\sigma(t_n)\le\sigma(t_{n-1})+\tau,\quad e_h^{n-1}=\bar e_h^n-\frac{\tau}{2}d_te_h^n, \tag{6.18}$$

for $2\le n\le N$, we obtain

$$\begin{aligned}\sigma(t_n)\|e_h^n\|_0^2-\sigma(t_{n-1})\|e_h^{n-1}\|_0^2+2\nu\tau\|\bar e_h^n\|_1^2&\le 2\tau(\|\bar e^n\|_0^2+\tau^2\|d_te_h^n\|_0^2)\\&+\kappa\tau\sigma(t_n)\|\phi(\bar e_h^n)\|_0^2+\kappa\tau\sigma(t_n)\|\bar u_{h,\tau}(t_n)-\bar u_{H,\tau_0}(t_n)\|_0^2.\end{aligned} \tag{6.19}$$

Summing (6.19) from $n=2$ to $n=m$ and using Lemma 6.2 yields

$$\begin{aligned}&\sigma(t_m)\|e_h^m\|_0^2+\nu\tau\sum_{n=2}^m\sigma(t_n)\|\bar e_h^n\|_1^2\le\kappa\sum_{n=2}^{m-1}(\|\bar e^n\|_0^2+\tau^2\|d_te_h^n\|_0^2)\\&+\kappa\tau\sum_{n=2}^N\sigma(t_n)\|\bar u_{h,\tau}(t_n)-\bar u_{H,\tau_0}(t_n)\|_0^2+\kappa\tau\sum_{n=1}^{m-1}\sigma(t_n)\|e_h^n\|_0^2+\kappa(H^4+\tau_0^4).\end{aligned} \tag{6.20}$$

Using Theorem 5.7, we obtain

$$\begin{aligned}\tau\sum_{n=2}^N\sigma(t_n)\|u_{h,\tau}(t_n)-u_{H,\tau_0}(t_n)\|_0^2&\le\sum_{n=1}^N\int_{t_{n-1}}^{t_n}\sigma(t)\|u_{h,\tau}(t)-u_{H,\tau_0}(t)\|_0^2dt\\&\le c\int_0^T(\sigma(t)\|u_h(t)-u_{h,\tau}(t)\|_0^2+\sigma(t)\|u_H(t)-u_{H,\tau_0}(t)\|_0^2)dt\\&+\int_0^T\sigma(t)\|u_h(t)-u_{H,\tau_0}(t)\|_0^2dt\le\kappa(H^4+\tau_0^4).\end{aligned} \tag{6.21}$$

Applying Lemma 4.4 and (6.21) to (6.20) and using (6.17) yields (6.16). □

Lemma 6.4. *Under the assumptions of Theorem 5.1, there holds*

$$\sigma^i(t_m)\|e_h^m\|_1^2 + \tau\sum_{n=1}^{m}\sigma^i(t_n)\|d_te_h^n\|_0^2 \le \kappa(\tau^2 + H^{2+i} + \tau_0^{2+i}), \quad i = 1,2, \tag{6.22}$$

for all $1 \le m \le N$.

Proof. Multiplying (6.13) with $r = 1$ by $\sigma^i(t_n)$, we obtain

$$\begin{aligned}&\tau\sigma^i(t_n)\|d_te_h^n\|_0^2 + \nu(\sigma^i(t_n)\|e_h^n\|_1^2 - \sigma^i(t_{n-1})\|e_h^{n-1}\|_1^2)\\ &\le i\nu\tau\sigma^{i-1}(t_n)\|e_h^{n-1}\|_1^2 + \kappa\tau\sigma^i(t_n)\|\phi(\bar{e}_h^n)\|_1^2 + \kappa\tau\sigma^i(t_n)\|\bar{u}_h^n - \bar{u}_{H,\tau_0}(t_n)\|_1^2.\end{aligned} \tag{6.23}$$

Summing this inequality from $n = 1$ to $n = m$ and using Lemma 6.2 and Lemma 6.3, Theorem 5.7, and (6.10), we obtain

$$\begin{aligned}&\sigma^i(t_m)\|e_h^m\|_1^2 + \tau\sum_{n=1}^{m}\sigma^i(t_n)\|d_te_h^n\|_0^2\\ &\le \kappa\tau\sum_{n=1}^{m}\sigma^i(t_n)\|\phi(\bar{e}_h^n)\|_1^2 + \kappa\tau\sum_{n=1}^{m}\sigma^{i-1}(t_n)(\|\bar{e}_h^n\|_1^2 + \tau^2\|d_te_h^n\|_1^2)\\ &+ \kappa\tau\sum_{n=1}^{m}\sigma^i(t_n)\|\bar{u}_h^n - \bar{u}_{H,\tau_0}(t_n)\|_1^2 \le \kappa(\tau^2 + H^{2+i} + \tau_0^{2+i}),\end{aligned}$$

which is (6.22). □

Lemma 6.5. *Under the assumptions of Theorem* 5.1*, there holds*

$$\|\mu_h^m\|_{L^2} \le \kappa\sigma^{-3/2}(t_m)(\tau + H^2 + \tau_0^2), \quad 1 \le m \le N. \tag{6.24}$$

Proof. From (6.6) we derive

$$\begin{aligned}&(d_{tt}e_h^n, v_h) + a(d_t\bar{e}_h^n, v_h) + b(\phi(d_t\bar{e}_h^n), \bar{u}_h^n, v_h)\\ &\quad + b(\phi(d_t\bar{u}^{h,n}), \bar{u}_{h,\tau}(t_n) - u_{H,\tau_0}(t_n), v_h) + b(\phi(\bar{e}_h^{n-1}), d_t\bar{u}_h^n, v_h)\\ &\quad + b(\phi(\bar{u}_{h,n-1}), d_t\bar{u}_{h,\tau}(t_n) - d_tu_{H,\tau_0}(t_n), v_h) \quad \forall v_h \in V_h.\end{aligned} \tag{6.25}$$

Next, we take $v_h = 2\tau d_t\bar{e}_h^n$ in (6.25), obtaining

$$\begin{aligned}&\|d_te_h^n\|_0^2 - \|d_te_h^{n-1}\|_0^2 + 2\tau\nu\|d_t\bar{e}_h^n\|_1^2 + 2\tau b(\phi(d_t\bar{e}_h^n), \bar{u}_h^n, d_t\bar{e}_h^n)\\ &\quad + 2\tau b(\phi(d_t\bar{u}^{h,n}), \bar{u}_{h,\tau}(t_n) - u_{H,\tau_0}(t_n), d_t\bar{e}_h^n) + 2\tau b(\phi(\bar{e}_h^{n-1}), d_t\bar{u}_h^n, d_t\bar{e}_h^n)\\ &\quad + 2\tau b(\phi(\bar{u}^{n-1,h}), d_t\bar{u}_{h,\tau}(t_n) - d_tu_{H,\tau_0}(t_n), d_t\bar{e}_h^n) = 0.\end{aligned} \tag{6.26}$$

From Lemma 4.1, Lemma 5.2, and Lemma 6.1, we have

$$\begin{aligned}2\tau|b(\phi(d_t\bar{e}_h^n), \bar{u}_h^n, d_t\bar{e}_h^n)| &\le \frac{\nu\tau}{4}\|d_t\bar{e}_h^n\|_1^2 + \kappa\tau\|\phi(d_t\bar{e}_h^n)\|_0^2,\\ 2\tau|b(\phi(d_t\bar{u}^{h,n}), \bar{u}_h^n &- \bar{u}_{H,\tau_0}(t), d_t\bar{e}_h^n)|\\ &\le \frac{\nu\tau}{4}\|d_t\bar{e}_h^n\|_1^2 + \kappa\tau\|d_t\bar{u}^{h,n}\|_1^2\|\bar{u}_h^n - \bar{u}_{H,\tau_0}(t_n)\|_1^2,\\ 2\tau|b(\phi(\bar{e}_h^{n-1}), d_t\bar{u}_h^n, d_t\bar{e}_h^n)| &\le \frac{\nu\tau}{4}\|d_t\bar{e}_h^n\|_1^2 + c\|d_t\bar{u}_h^n\|_1^2\|\phi(\bar{e}_h^{n-1})\|_1^2,\\ 2\tau|b(\phi(\bar{u}^{h,n-1}), d_t\bar{u}_{h,\tau}(t_n) &- d_tu_{H,\tau_0}(t_n), d_t\bar{e}_h^n)| \le \frac{\nu\tau}{4}\|d_t\bar{e}_h^n\|_1^2\\ &+ c\tau\|A_h\phi(\bar{u}^{h,n-1})\|_0^2\|d_t\bar{u}_{h,\tau}(t_n) - d_t\bar{u}_{H,\tau_0}(t_n))\|_0^2.\end{aligned}$$

Combining these inequalities with (6.26) yields

$$\begin{aligned}\|d_te_h^n\|_0^2 - \|d_te_h^{n-1}\|_0^2 + \nu\tau\|d_t\bar{e}_h^n\|_1^2 \leq{}& \kappa\tau\|\phi(d_t\bar{e}_h^n)\|_0^2\\ &+ c\tau\|d_t\bar{u}_{h,\tau}(t_n) - d_t\bar{u}_{H,\tau_0}(t_n)\|_0^2 + c\tau\|\phi(\bar{e}_h^{n-1})\|_1^2\|d_t\bar{u}_h^n\|_1^2\\ &+ c\tau\|\phi(d_t\bar{u}^{h,n})\|_1^2\|\bar{u}_h^n - \bar{u}_{H,\tau_0}(t_n)\|_1^2.\end{aligned}\tag{6.27}$$

Multiplying (6.27) by $\sigma^3(t_n)$ and applying Lemma 5.2 and Lemma 6.1 yields

$$\begin{aligned}\sigma^3(t_n)\|d_te_h^n\|_0^2 - \sigma^3(t_{n-1})\|d_te_h^{n-1}\|_0^2 + \nu\tau\sigma^3(t_n)\|d_t\bar{e}_h^n\|_1^2&\\ \leq 2\tau\sigma^2(t_{n-1})\|d_te_h^{n-1}\|_0^2 + \kappa\tau\sigma^3(t_n)\|\phi(d_t\bar{e}_h^n)\|_0^2&\\ + c\tau\sigma^3(t_n)\|d_t\bar{u}_{h,\tau}(t_n) - d_t\bar{u}_{H,\tau_0}(t_n)\|_0^2 + c\tau\sigma^2(t_n)\|\phi(\bar{e}_h^{n-1})\|_1^2&\\ + c\tau\sigma^2(t_n)\|\bar{u}_{h,\tau}(t_n) - \bar{u}_{H,\tau_0}(t_n)\|_1^2.&\end{aligned}\tag{6.28}$$

Summing (6.28) from $n = 1$ to $n = m$ and applying Lemma 6.3 and Lemma 6.4 with $i = 2$ and Theorem 5.7, we obtain

$$\begin{aligned}\sigma^3(t_m)\|d_te_h^m\|_0^2 + \tau\sum_{n=1}^{m}\sigma^3(t_n)\|d_t\bar{e}_h^n\|_1^2&\\ \leq \kappa(\tau^2 + H^4 + \tau_0^4) + \kappa\tau\sum_{n=1}^{m}\sigma^3(t_n)\|\phi(d_t\bar{e}_h^n)\|_0^2.&\end{aligned}\tag{6.29}$$

Applying Lemma 4.3 with $m = 1$ and Lemma 4.4 with $2 \leq m \leq N$ to (6.29) yields

$$\sigma^3(t_m)\|d_te_h^m\|_0^2 + \tau\sum_{n=1}^{m}\sigma^3(t_n)\|d_t\bar{e}_h^n\|_1^2 \leq \kappa(\tau^2 + H^4 + \tau_0^4).\tag{6.30}$$

Finally, we derive from (3.4), (6.6), (6.30), Theorem 5.7, Lemma 6.4, and (6.30) that

$$\begin{aligned}\|\mu_h^m\|_{L^2} &\leq \kappa(\|d_te_h^m\|_0 + \|\bar{e}_h^m\|_1 + \|\bar{u}_{h,\tau}(t_m) - \bar{u}_{H,\tau_0}(t_m)\|_1 + \|\phi(\bar{e}_h^m)\|_1)\\ &\leq \sigma^{-3/2}(t_m)(\tau + H^2 + \tau_0^2), \quad 1 \leq m \leq N,\end{aligned}$$

which is (6.24). □

Theorem 6.6. *Under the assumptions of Theorem* 5.1, *the following error estimates hold:*

$$\begin{aligned}\|u_h(t) - u^{h,\tau}(t)\|_{H_0^1} \leq{}& \kappa H(1-t)\sigma^{-1/2}(t)(\tau + \tau_0^{3/2} + H^{3/2})\\ &+ \kappa H(t-1)(\tau + \tau_0^2 + H^2), \quad t \in (0, T],\end{aligned}\tag{6.31}$$

$$\|p_h(t) - p^{h,\tau}(t)\|_{L^2} \leq \kappa\sigma^{-3/2}(t)(\tau + \tau_0^2 + H^2), \quad t \in (0, T].\tag{6.32}$$

This proof can be completed by combining Theorem 5.6, Lemma 6.5, and Lemma 6.4 with $i = 1$ for $t \in (0, 1]$ and $i = 2$ for $t \in (1, T]$; it can be omitted.

Remark. Combining Theorem 6.6 with (3.11) yields (1.3)–(1.4).

Acknowledgments. The author would like to thank the editor and referees for their criticism, valuable comments, and suggestions which helped to improve the results of this paper.

REFERENCES

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

[2] A. Ait Ou Ammi and M. Marion, *Nonlinear Galerkin methods and mixed finite elements: Two-grid algorithms for the Navier-Stokes equations*, Numer. Math., 68 (1994), pp. 189–213.

[3] G. A. Baker, V. A. Dougalis, and O. A. Karakashian, *On a high order accurate fully discrete Galerkin approximation to the Navier-Stokes equations*, Math. Comp., 39 (1982), pp. 339–375.

[4] J. Bercovier and O. Pironneau, *Error estimates for finite element solution of the Stokes problem in the primitive variables*, Numer. Math., 33 (1979), pp. 211–224.

[5] C. Bernardi and G. Raugel, *A conforming finite element method for the time-dependent the Navier–Stokes equations*, SIAM J. Numer. Anal., 22 (1985), pp. 455–473.

[6] J. R. Cannon and Yanping Lin, *A priori L^2 error estimates for finite-element methods for nonlinear diffusion equations with memory*, SIAM J. Numer. Anal., 27 (1990), pp. 595–607.

[7] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North-Holland, Amsterdam, 1978.

[8] J. Douglas, Jr., and T. Dupont, *Galerkin methods for parabolic equations*, SIAM J. Numer. Anal., 7 (1970), pp. 575–626.

[9] T. Dupont, G. Fairweather, and J. P. Johnson, *Three-level Galerkin methods for parabolic equations*, SIAM J. Numer. Anal., 11 (1974), pp. 392–410.

[10] V. J. Ervin and W. J. Layton, *A posteriori error estimation for two-level discretizations of flows of electrically conducting, incompressible fluids*, Comput. Math. Appl., 31 (1996), pp. 105–114.

[11] V. Ervin, W. Layton, and J. Maubach, *A posteriori error estimators for a two-level finite element method for the Navier-Stokes equations*, Numer. Methods Partial Differential Equations, 12 (1996), pp. 333–346.

[12] V. Girault and J. L. Lions, *Two-grid finite-element schemes for the steady Navier-Stokes problem in polyhedra*, Port. Math. (N.S.), 58 (2001), pp. 25–57.

[13] V. Girault and J. L. Lions, *Two-grid finite-element schemes for the transient Navier-Stokes problem*, M2AN Math. Model. Numer. Anal., 35 (2001), pp. 945–980.

[14] V. Girault and P. A. Raviart, *Finite Element Approximation of the Navier-Stokes Equations*, Lecture Notes in Math. 749, Springer-Verlag, Berlin, New York, 1974.

[15] V. Girault and P. A. Raviart, *Finite Element Method for Navier-Stokes Equations: Theory and algorithms*, Springer-Verlag, Berlin, Heidelberg, 1987.

[16] Y. N. He and K. T. Li, *Convergence and stability of finite element nonlinear Galerkin method for the Navier-Stokes equations*, Numer. Math., 79 (1998), pp. 77–106.

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

[18] J. G. Heywood and R. Rannacher, *Finite-element approximations of the nonstationary Navier–Stokes problem. Part* IV: *Error estimates for second-order time discretization*, SIAM J. Numer. Anal., 27 (1990), pp. 353–384.

[19] A. T. Hill and E. Süli, *Approximation of the global attractor for the incompressible Navier-Stokes equations*, IMA J. Numer. Anal., 20 (2000), pp. 633–667.

[20] R. B. Kellogg and J. E. Osborn, *A regularity result for the Stokes problem in a convex polygon*, J. Functional Anal., 21 (1976), pp. 397–431.

[21] S. Larsson, *The long-time behavior of finite-element approximations of solutions to semilinear parabolic problems*, SIAM J. Numer. Anal., 26 (1989), pp. 348–365.

[22] W. Layton, *A two level discretization method for the Navier-Stokes equations*, Comput. Math. Appl., 26 (1993), pp. 33–38.

[23] W. Layton and W. Lenferink, *Two-level Picard, defect correction for the Navier-Stokes equations*, Appl. Math. Comput., 80 (1995), pp. 1–12.

[24] W. Layton and L. Tobiska, *A two-level method with backtracking for the Navier–Stokes equations*, SIAM J. Numer. Anal., 35 (1998), pp. 2035–2054.

[25] Yanping Lin, *Galerkin methods for nonlinear parabolic integrodifferential equations with nonlinear boundary conditions*, SIAM J. Numer. Anal., 27 (1990), pp. 608–621.

[26] M. Marion and R. Temam, *Nonlinear Galerkin methods: The finite elements case*, Numer. Math., 57 (1990), pp. 205–226.

[27] M. Marion and J. Xu, *Error estimates on a new nonlinear Galerkin method based on two-grid finite elements*, SIAM J. Numer. Anal., 32 (1995), pp. 1170–1184.

[28] M. A. Olshanskii, *Two-level method and some a priori estimates in unsteady Navier-Stokes calculations*, J. Comput. Appl. Math., 104 (1999), pp. 173–191.

[29] J. Shen, *Long time stability and convergence for fully discrete nonlinear Galerkin methods*, Appl. Anal., 38 (1990), pp. 201–229.

[30] J. C. Simo and F. Armero, *Unconditional stability and long-term behavior of transient algorithms for the incompressible Navier-Stokes and Euler equations*, Comput. Methods Appl. Mech. Engrg., 111 (1994), pp. 111–154.

[31] J. C. Simo, N. Tarnow, and K. K. Wong, *Exact energy-momentum conserving algorithms and symplectic schemes for nonlinear dynamics*, Comput. Methods Appl. Mech. Engrg., 100 (1992), pp. 63–116.

[32] R. Temam, *Navier-Stokes Equations, Theory and Numerical Analysis*, 3rd ed., North-Holland, Amsterdam, 1983.

[33] R. Temam, *Infinite-Dimensional Dynamical Systems in Mechanics and Physics*, Springer, New York, 1988.

[34] J. Xu, *A novel two-grid method for semilinear elliptic equations*, SIAM J. Sci. Comput., 15 (1994), pp. 231–237.

[35] J. Xu, *Two-grid discretization techniques for linear and nonlinear PDEs*, SIAM J. Numer. Anal., 33 (1996), pp. 1759–1777.

SIAM J. NUMER. ANAL.
Vol. 41, No. 4, pp. 1286–1300

STUDY OF A MAXIMAL MONOTONE MODEL WITH A DELAY TERM*

CLAUDE-HENRI LAMARQUE[†], JÉRÔME BASTIEN[‡], AND MATTHIEU HOLLAND[†]

Abstract. We present a model involving a friction term and a delay term. By using results of existence and uniqueness for maximal monotone differential inclusions, we give theoretical results for this model. An implicit Euler numerical scheme and results related to order of convergence are also provided from both a theoretical and a numerical point of view.

Key words. delay, friction, maximal monotone, numerical scheme

AMS subject classifications. 34A60, 65L99, 34G25, 37L05

DOI. 10.1137/S0036142902402547

1. Introduction.

In this paper, we study models including maximal monotone terms and delay terms. Differential inclusions involving only maximal monotone terms have been studied in many references from the point of view of mathematics (e.g., [7, 6, 10, 13]), mathematics and mechanics (e.g., [20, 1, 22, 2, 5]), numerical analysis (e.g., [25, 2, 5]), mechanics (e.g., [27, 17, 8]), and recently with friction and impact [30]. Indeed, many applications are concerned since it could be very convenient to use set-valued force laws in order to write models for impacts, friction, or elastoplastic constitutive laws, for example. Existence and uniqueness for stochastic differential inclusions have been investigated [24, 26, 11, 12] and numerical schemes of Euler type have been considered [26, 11].

Models including smooth nonlinear terms and delay terms have been also studied (see, e.g., [19]). But models including both maximal monotone terms and delay terms have not been investigated from the mathematical or numerical point of view. This is the main topic of this paper.

Motivations for introduction of delay terms can be found in [19, 29] or occur from applications: the control of structures (see, e.g., [31] for a survey of strategies in the frame of civil engineering) may include friction forces or elastoplastic terms together with a control law depending on the state of the structure delayed applied control force. The stability of systems including delay terms is also important for applications [23].

Recently, many studies have been devoted to the analysis of the behavior of cutter-tools (see, e.g., [18, 28, 23]). In [28], hysteresis terms lead to both maximal monotone terms (of sign type) and delay terms. The present work is based on the Master's thesis of Holland [21]. The paper is organized as follows. In section 2, the models considered are described. In section 3, mathematical background and existence and uniqueness results are provided. In section 4, a numerical scheme is built, its convergence is investigated, and numerical results given for a simple example.

*Received by the editors February 13, 2002; accepted for publication (in revised form) January 2, 2003; published electronically August 22, 2003.

http://www.siam.org/journals/sinum/41-4/40254.html

[†]URA 1652 CNRS, Département Génie Civil et Bâtiment, Laboratoire Géomatériaux, École Nationale des Travaux Publics de l'Etat, Rue Maurice Audin, 69518 Vaulx-en-Velin Cedex, France (Claude.Lamarque@entpe.fr).

[‡]Laboratoire Mécatronique 3M, Université de Technologie de Belfort-Montbéliard, 90010 Belfort cedex, France (jerome.bastien@utbm.fr, Matthieu.Holland@equipement.gouv.fr).

2. Description of the model. In the field of mechanics, one frequently considers the following differential inclusion:

$$(2.1)\qquad \begin{cases} m\ddot{x}(t) + f\big(\dot{x}(t), x(t), t\big) + g\big(\dot{x}(t-\tau), x(t-\tau), t\big) \ni 0 \quad \text{on } [0,T], \\ x_0, \dot{x}_0 \text{ given functions on } [-\tau, 0]. \end{cases}$$

In the simplest case $f\big(\dot{x}(t), x(t), t\big)$ can be written as

$$(2.2)\qquad f\big(\dot{x}(t), x(t), t\big) \in c\dot{x}(t) + kx(t) - H(t) + \alpha\sigma\big(\dot{x}(t)\big),$$

with viscous damping c, stiffness k, external force $H(t)$ (exerted on the mass m), and friction force of Coulomb type $\alpha\sigma(\dot{x}(t))$, where σ is the full graph “sign”; it is defined by (see Figure 1)

$$\sigma(x) = \begin{cases} \{-1\} & \text{if } x < 0, \\ \{+1\} & \text{if } x > 0, \\ [-1, 1] & \text{if } x = 0. \end{cases}$$

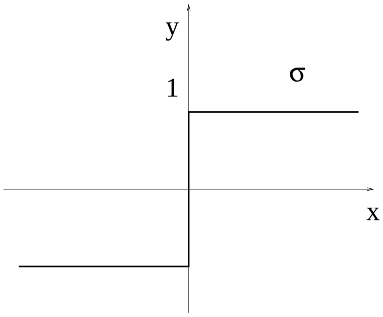


Fig. 1. *The graph σ.*

In (2.2), α depends on the normal force and Coulomb coefficient (see [4, 5]). The function g contains delay terms. In the simplest case g is given by

$$(2.3)\qquad g\big(\dot{x}(t-\tau), x(t-\tau), t\big) = \beta(t)\dot{x}(t-\tau) + \gamma(t)x(t-\tau),$$

with delay τ and smooth functions $\beta(t)$ and $\gamma(t)$. This model is represented in Figure 2.

In the general case with n dof the following models are considered for applications. They correspond to vectorial differential inclusions with several delays τ_i, $1 \le i \le N$, written as

$$\begin{cases} M\ddot{X}(t) + f\left(\dot{X}(t), X(t), t\right) + \sum_{i=1}^{N} g_i\left(\dot{X}(t-\tau_i), X(t-\tau_i), t\right) + \text{friction like forces} \ni 0, \\ X_0,\ \dot{X}_0 \text{ given vector functions on } [-\tau_N, 0], \end{cases}$$

with mass matrix M (size $n \times n$), $X(t) \in \mathbb{R}^n$, f and g_i $(1 \le i \le n)$ similar to the previous one dof case except that $f : \mathbb{R}^n \times \mathbb{R}^n \times [0, +\infty[\longrightarrow \mathbb{R}^n$ and $g : \mathbb{R}^n \times \mathbb{R}^n \times [0, +\infty[\longrightarrow \mathbb{R}^n$. For practical investigation one chooses frequently $N = 1$.

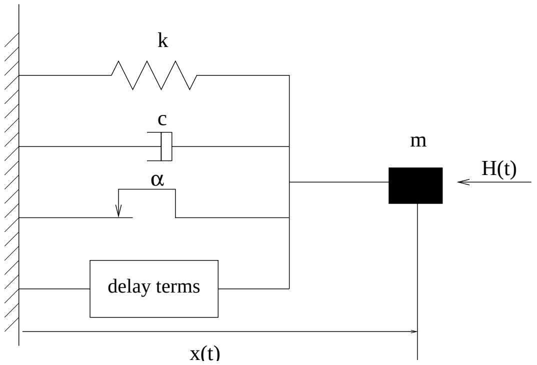


FIG. 2. *The studied mechanical model, with one dof, for f defined by* (2.2).

Let H be a separable Hilbert space equipped with a scalar product denoted $(.,.)$ and a norm denoted $|.|$, with T a strictly positive number. A is a maximal monotone multivalued operator on H, whose domain is denoted $D(A)$. In this paper, a mathematical problem of the form

$$\dot{u}(t)+A\big(u(t)\big)+B\big(t,u(t)\big)+G\big(u(t-\tau)\big)\ni 0 \text{ a.e. on }]0,T[\tag{2.4}$$

$$\forall t\in[-\tau,0],\quad u(t)=z(t), \tag{2.5}$$

is considered, where $\tau>0$, and

$$z\in W^{1,\infty}(-\tau,0;H). \tag{2.6a}$$

G a mapping from H to H whose differential is locally bounded on H, i.e.,

$$\forall R\geq 0,\quad \Psi(R)=\sup\left\{\|G'(x)\| : |x|\leq R\right\}<+\infty. \tag{2.6b}$$

B is a mapping from $[0,T]\times H$ from H, Lipschitz continuous with respect to its second argument and whose derivative maps the bounded sets of $L^2(0,T;H)$ into bounded sets of $L^2(0,T;H)$, i.e.,

$$\exists L\geq 0 :\quad \forall t\in[0,T],\quad \forall x_1,x_2\in H,\quad |B(t,x_1)-B(t,x_2)|\leq\omega\,|x_1-x_2|\,, \tag{2.6c}$$

and let us denote

$$\forall R\geq 0,\quad \Phi(R)=\sup\left\{\left\|\frac{\partial B}{\partial t}(.,v)\right\|_{L^2(0,T;H)} : \|v\|_{L^2(0,T;H)}\leq R\right\}<+\infty. \tag{2.6d}$$

The case with N delay terms would be treated in the same way. For applications, H is equal to $\mathbb{R}^n$.

For numerical simulations, we consider a class of one dof mechanical systems that can occur from a simple control problem with delayed applied control force or from a dynamic cutting process (see, e.g., [18, 28, 29, 23]) governed by (2.1), (2.2), and (2.3) with $m=1$, $\beta\equiv 0$, and $\gamma(t)=\gamma$; so we studied the system:

$$\begin{cases} \ddot{x}(t)+c\dot{x}(t)+kx(t)+\alpha\sigma\big(\dot{x}(t)\big)+\gamma x(t-\tau)\ni H(t) & \text{on } [0,T],\\ x(t)=x_0(t) & \text{on } [-\tau,0],\\ \dot{x}(t)=\dot{x}_0(t) & \text{on } [-\tau,0]. \end{cases} \tag{2.7}$$

3. Existence and uniqueness.

3.1. Summary of an existence and uniqueness result. In this section, we give an existence and uniqueness result that corresponds to the generalization of results obtained by Brezis [6] and [7]. A proof of this result can be found in [3], [2], [5], or in [15]. We proved the existence and uniqueness of the solution to a differential inclusion, as the by-product of convergence results for a numerical scheme. We were in the more general frame of a Gelfand triple $V \hookrightarrow H \hookrightarrow V'$, where we denote $\hookrightarrow$ a dense and continuous inclusion. Here, it is enough to make the identification $V = H = V'$ for the mechanical models studied. This result generalizes the result of [7] for a differential inclusion with a maximal monotone term equal to the subdifferential of the indicatrix of a nonempty closed convex set and the result of [6, Prop. 3.13, p. 107] for a maximal monotone term whose domain is not necessarily with a nonempty interior.

We assume now that f is a function from $[0,T] \times H$ to H, Lipschitz continuous with respect to its second argument and whose derivative maps the bounded sets of $L^2(0,T;H)$ into bounded sets of $L^2(0,T;H)$, i.e.,

$$\exists L \geq 0 : \quad \forall t \in [0,T], \quad \forall x_1, x_2 \in H, \quad |f(t,x_1) - f(t,x_2)| \leq L\,|x_1 - x_2| \tag{3.1}$$

and

$$\forall R \geq 0, \quad \Phi(R) = \sup\left\{ \left\| \frac{\partial f}{\partial t}(.,v) \right\|_{L^2(0,T;H)} : \|v\|_{L^2(0,T;H)} \leq R \right\} < +\infty. \tag{3.2}$$

The existence and uniqueness result is the following proposition.

Proposition 3.1. *If A is a multivalued maximal monotone operator from H and if assumptions* (3.1) *and* (3.2) *hold, then there exists a unique solution $u \in W^{1,\infty}(0,T;H)$ of the differential inclusion*

$$\dot{u}(t) + A\big(u(t)\big) \ni f\big(t, u(t)\big) \text{ a.e. on }]0,T[, \tag{3.3a}$$

$$u(0) = u_0. \tag{3.3b}$$

3.2. Existence and uniqueness results. We now apply Proposition 3.1 to a maximal monotone inclusion with a delay term.

Proposition 3.2. *Under assumptions* (2.6), *there exists a unique function $u \in W^{1,\infty}(-\tau,T;H)$ solution of the differential inclusion with delay*

$$\dot{u}(t) + A\big(u(t)\big) + B\big(t, u(t)\big) + G\big(u(t-\tau)\big) \ni 0 \text{ a.e. on }]0,T[\tag{3.4}$$

$$\forall t \in [-\tau, 0], \quad u(t) = z(t). \tag{3.5}$$

Proof. The delay term $G(u(t-\tau))$ is a smooth term and the maximal monotone inclusion (3.4) is similar to the differential inclusion:

$$\dot{u}(t) + A\big(u(t)\big) \ni \widetilde{f}\big(t, u(t)\big) \text{ a.e. on }]0,T[, \tag{3.6}$$

$$u(0) = u_0. \tag{3.7}$$

Let τ_1 be equal to $\tau_1 = \min(T, \tau)$. Define the mapping f_1 from $[0,\tau_1] \times H$ to H by

$$\forall t \in [0,\tau_1], \quad \forall x \in H, \quad f_1(t,x) = -B(t,x) - G\big(z(t-\tau)\big).$$

Over the interval $[0, \tau_1]$, problem (3.4)–(3.5) is then equivalent to the problem

$$\dot{u}(t) + A\big(u(t)\big) \ni f_1\big(t, u(t)\big) \text{ a.e. on }]0, \tau_1[, \tag{3.8}$$

$$u(0) = z(-\tau). \tag{3.9}$$

We are in the frame of Proposition 3.1. Indeed, according to assumption (2.6c), we have

$$\forall t \in [0, \tau_1], \quad \forall x_1, x_2 \in H, \quad |f_1(t, x_1) - f_1(t, x_2)| \leq \omega\, |x_1 - x_2|\,;$$

thus, assumption (3.1) holds on $[0, \tau_1]$. Otherwise, we have

$$\forall t \in [0, \tau_1], \quad \forall v \in L^2(0, \tau_1; H), \quad \frac{\partial f_1}{\partial t}(t, v) = -\frac{\partial B}{\partial t}(t, v) - \Big(G'\big(z(t-\tau)\big), \dot{z}(t-\tau)\Big),$$

and then, by integration over the interval $[0, \tau_1]$,

$$\int_0^{\tau_1} \left|\frac{\partial f_1}{\partial t}(t, v)\right|^2 dt \leq 2 \left(\int_0^{\tau_1} \left|\frac{\partial B}{\partial t}(t, v)\right|^2 dt + \int_0^{\tau_1} \left|G'\big(z(t-\tau)\big)\right|^2 |\dot{z}(t-\tau)|^2 dt\right),$$

and according to assumptions (2.6a), (2.6b), and (2.6d), if $\|v\|_{L^2(0,\tau_1;H)} \leq R$,

$$\int_0^{\tau_1} \left|\frac{\partial f_1}{\partial t}(t, v)\right|^2 dt \leq 2 \left(\Phi^2(R) + \tau_1 \|\dot{z}\|^2_{L^\infty(-\tau,0,H)} \Psi^2\left(\|z\|_{C^0([-\tau,0],H)}\right)\right).$$

Thus, assumption (3.2) holds on $[0, \tau_1]$. Then, according to Proposition 3.1, there exists a unique solution $u_1 \in W^{1,+\infty}(0, \tau_1; H)$ of (3.8)–(3.9). If $\tau_1 = T$, the proposition is proved. If $\tau_1 < T$, by setting $\tau_2 = \min(2\tau, T)$, we denote

$$\forall t \in [\tau_1, \tau_2], \quad \forall x \in H, \quad f_2(t, x) = -B(t, x) - G\big(u_1(t-\tau)\big).$$

As previously, problem (3.4)–(3.5) on $]\tau_1, \tau_2[$ is equivalent to the problem

$$\dot{u}(t) + A\big(u(t)\big) \ni f_2\big(t, u(t)\big) \text{ a.e. on }]\tau_1, \tau_2[, \tag{3.10}$$

$$u(\tau_1) = u_1(\tau_1 - 0), \tag{3.11}$$

where $u_1(\tau_1 - 0)$ is the value at time τ_1 of the continuous function u_1. So, thanks to Proposition 3.1, there exists a unique solution $u_2 \in W^{1,\infty}(\tau_1, \tau_2; H)$ of (3.10)–(3.11). Finally, we can construct by induction two sequences $(\tau_i)_{0 \leq i \leq q}$: $0 = \tau_0 < \tau_1 < \tau_2 < \cdots < \tau_q = T$ and $(u_i)_{1 \leq i \leq q}$ such that, for all $i \in \{1, \ldots, q\}$, u_i belongs to $W^{1,\infty}(\tau_{i-1}, \tau_i; H)$ and is the unique solution of

$$\dot{u}_i(t) + A\big(u_i(t)\big) + B\big(t, u_i\big) + G\big(u_{i-1}(t-\tau)\big) \ni 0 \text{ a.e. on }]\tau_{i-1}, \tau_i[, \tag{3.12}$$

$$u_i(\tau_{i-1}) = u_{i-1}(\tau_{i-1} - 0), \tag{3.13}$$

where u_0 is equal to z on $[-\tau, 0]$. We consider the unique function u from $[0, T]$ to H whose restriction to each interval $[\tau_{i-1}, \tau_i[$ is equal to u_i. It is clear that u is the unique solution of (3.4)–(3.5). Moreover, each function u_i is continuous on the interval $[\tau_{i-1}, \tau_i[$. According to (3.13), u is continuous on $[0, T]$. By construction, each function $\dot{u}_i$ belongs to $L^\infty(\tau_{i-1}, \tau_i; H)$; thus, the restriction of the function $\dot{u}$ to $[0, T]$ belongs to $L^\infty(0, T; H)$ and the restriction of the function u to $[0, T]$ belongs to $W^{1,\infty}(0, T; H)$, which concludes this proof. □

4. Numerical scheme.

4.1. Summary of results for numerical schemes. In this section, we recall two error estimates, which are generalizations of results by Lippold [25] proved in [3, 2, 5]; we proved that an implicit Euler numerical scheme for differential inclusion (3.3) is of order $1/2$ in the general case and of order one if the multivalued term is equal to the subdifferential of the indicatrix of a nonempty convex set of H.

Results of convergence can also be found in [16], [15], or [9], based on general properties of consistence. Elliot described some cases where this consistency is true (for example if the derivative of the solution possesses only a finite number of discontinuities). Our results [3], [2], or [5] are more general. In [14], a survey on numerical methods for differential inclusions can be read as follows: results for order of convergence are provided. They are obtained if the nonlinear term is compact and Lipschitz (in the sense of a one-sided Lipschitz condition). We did not assume such a condition in [3], [2], or [5]. References about the numerical point of view are given in [3] or [5].

Let N be an integer. Let $h = T/N$, and let U^p be the solution of the numerical scheme

$$\forall p \in \{0, \dots, N-1\}, \quad \frac{U^{p+1} - U^p}{h} + A\left(U^{p+1}\right) \ni f\left(ph, U^p\right), \tag{4.1}$$

$$U^0 = u_0. \tag{4.2}$$

Denote $u_h \in C^0\left([0,T], H\right)$ the linear interpolation at times $t_p = hp$ of the U^p. The solution U^p of the numerical scheme (4.1) exists and is unique since A is maximal monotone; indeed, in this case the operator $(I + hA)^{-1}$, where I is the identity of H, is a single-valued operator defined by all the space H, and U^{p+1} is defined by

$$\forall p \in \{0, \dots, N-1\}, \quad U^{p+1} = (I + hA)^{-1}\big(hf\left(t_p, U^p\right) + U^p\big). \tag{4.3}$$

See [6]. The first result of convergence reads as the following proposition.

Proposition 4.1. *Under the assumptions* (3.1) *and* (3.2)*, the numerical scheme* (4.1)–(4.2) *is of order* $1/2$*; i.e., there exists* C *such that, for all* h*,*

$$\forall t \in [0,T], \quad |u(t) - u_h(t)| \le C\sqrt{h}, \tag{4.4}$$

where u *is the solution of* (3.3).

Let K be a closed convex nonempty subset of H, and let $\partial\psi_K$ be the subdifferential of the indicatrix of K, which is given by

$$\forall (x,y) \in K \times H, \quad y \in \partial\psi_K(x) \Longleftrightarrow \forall z \in K, \quad \langle y, x - z\rangle \ge 0, \tag{4.5}$$

and

$$\forall x \notin K, \quad \partial\psi_K(x) = \emptyset. \tag{4.6}$$

This result of convergence can be improved by the following proposition.

Proposition 4.2. *Under the assumptions* (3.1) *and* (3.2) *and if* A *is equal to* $\partial\psi_K$*, the numerical scheme* (4.1)–(4.2) *is of order one; i.e., there exists* C *such that, for all* h*,*

$$\forall t \in [0,T], \quad |u(t) - u_h(t)| \le Ch, \tag{4.7}$$

where u *is the solution of* (3.3).

For numerical results, we will need a more general result, when the initial condition of the numerical scheme is not equal to u_0.

LEMMA 4.3. *Let v_0 belong to $D(A)$, and let V^p be the solution of the numerical scheme*

$$\forall p \in \{0, \dots, N-1\}, \quad \frac{V^{p+1} - V^p}{h} + A\left(V^{p+1}\right) \ni f\left(ph, V^p\right), \tag{4.8}$$

$$V^0 = v_0. \tag{4.9}$$

Denote $v_h \in C^0([0,T], H)$ the linear interpolation at times $t_p = hp$ of the V^p. In the general case, there exists C such that, for all h,

$$\forall t \in [0,T], \quad |u(t) - v_h(t)| \le |u_0 - v_0| e^{LT} + C\sqrt{h},$$

and if A is equal to $\partial\psi_K$, there exists C such that, for all h,

$$\forall t \in [0,T], \quad |u(t) - v_h(t)| \le |u_0 - v_0| e^{LT} + Ch.$$

Proof. Since we have, by triangle inequality,

$$\forall t \in [0,T], \quad |u(t) - v_h(t)| \le |u(t) - u_h(t)| + |u_h(t) - v_h(t)|,$$

it is enough to prove, according to Propositions 4.1 and 4.2, that we have

$$\forall t \in [0,T], \quad |u_h(t) - v_h(t)| \le |u_0 - v_0| e^{LT},$$

which is true if

$$\forall p \in \{0, \dots, N\}, \quad |U^p - V^p| \le |u_0 - v_0| e^{LT}.$$

Numerical schemes (4.1) and (4.8) can be rewritten as

$$U^{p+1} + hA\left(U^{p+1}\right) \ni hf\left(t_p, U^p\right) + U^p, \tag{4.10}$$

$$V^{p+1} + hA\left(V^{p+1}\right) \ni hf\left(t_p, V^p\right) + V^p. \tag{4.11}$$

If we subtract (4.10) and (4.11), we obtain, by multiplication of $U^{p+1} - V^{p+1}$ by monotonicity of A,

$$\left|U^{p+1} - V^{p+1}\right|^2 \le h\left(f\left(t_p, U^p\right) - f\left(t_p, V^p\right), U^{p+1} - V^{p+1}\right) + \left(U^p - V^p, U^{p+1} - V^{p+1}\right),$$

which implies, thanks to the Cauchy–Schwarz inequality and assumption (3.1),

$$\left|U^{p+1} - V^{p+1}\right|^2 \le \left|U^{p+1} - V^{p+1}\right| |U^p - V^p| \left(hL + 1\right).$$

We can then infer

$$\left|U^{p+1} - V^{p+1}\right| \le |U^p - V^p| \left(hL + 1\right) \le e^{hL} |U^p - V^p|; \tag{4.12}$$

by multiplying (4.12) for $k \in \{0, \dots, p\}$, we obtain

$$\forall p \in \{0, \dots, N\}, \quad |U^p - V^p| \le \left|U^0 - V^0\right| e^{hNL} \le |u_0 - v_0| e^{TL}. \qquad \square$$

4.2. Definition and order of the numerical scheme. Let us now apply the results of Propositions 4.1 and 4.2 to differential inclusion (3.4)–(3.5). In order to simplify the presentation, we assume that

$$\tau < T \text{ and } \exists Q \in \mathbb{N}^* : \quad T = Q\tau. \tag{4.13}$$

Let N be an integer, and let

$$h = \frac{\tau}{N}. \tag{4.14}$$

So we have $T = Q\tau = QhN$. We then have

$$h = \frac{T}{M}, \text{ where } M = QN. \tag{4.15}$$

Let $(U^p)_{-N\leq p\leq M}$ be the solution of the numerical scheme

$$\forall p \in \{0, \dots, M-1\}, \quad \frac{U^{p+1} - U^p}{h} + A\left(U^{p+1}\right) + B\left(ph, U^p\right) + G\left(U^{p-N}\right) \ni 0, \tag{4.16}$$

$$\forall p \in \{-N, \dots, 0\}, \quad U^p = z\left(ph\right). \tag{4.17}$$

Denote $u_h \in C^0\left([-\tau, T], H\right)$ the linear interpolation at times $t_p = hp$ of the U^p for $-N \leq p \leq M$. The solution U^p of the numerical scheme (4.16) is also defined by

(4.18)

$$\forall p \in \{0, \dots, M-1\}, \quad U^{p+1} = (I + hA)^{-1}\left(h\Big(-B\left(ph, U^p\right) - G\left(U^{p-N}\right)\Big) + U^p\right).$$

Proposition 4.4. *Under assumptions* (2.6)*, the numerical scheme* (4.16)*–*(4.17) *is of order* 1/2*; i.e., there exists C such that, for all h,*

$$\forall t \in [0, T], \quad |u(t) - u_h(t)| \leq C\sqrt{h}, \tag{4.19}$$

in the general case and if A is equal to $\partial\psi_K$, then it is of order one; i.e., there exists C such that, for all h,

$$\forall t \in [0, T], \quad |u(t) - u_h(t)| \leq Ch. \tag{4.20}$$

Proof. We use again the notations of the proof of Proposition 3.2: we rewrite problem (3.4)–(3.5) on the interval $[0, \tau_1] = [0, \tau]$ under the form (3.8)–(3.9), discretized by the numerical scheme

$$\forall p \in \{0, \dots, N-1\}, \quad \frac{U^{p+1} - U^p}{h} + A\left(U^{p+1}\right) \ni f_1\left(t_p, U^p\right),$$
$$U^0 = z(0),$$

which is equivalent to

$$\forall p \in \{0, \dots, N-1\}, \quad \frac{U^{p+1} - U^p}{h} + A\left(U^{p+1}\right) + B\left(ph, U^p\right) + G\left(U^{p-N}\right) \ni 0, \tag{4.21}$$

$$U^0 = z(0). \tag{4.22}$$

According to Propositions 4.1 and 4.2, we then have

$$\forall t \in [0, \tau_1], \quad |u(t) - u_h(t)| \leq C h^{\alpha}, \tag{4.23}$$

where $\alpha = 1/2$ in the general case and $\alpha = 1$ if $A = \partial \phi_K$. We then rewrite problem (3.4)–(3.5) on the interval $[\tau_1, \tau_2] = [\tau, \tau_2]$ under the form (3.10)–(3.11), discretized by the numerical scheme with discrete initial condition

$$\begin{aligned}&\forall p \in \{N, \dots, 2N-1\}, \quad \frac{U^{p+1} - U^p}{h} + A\left(U^{p+1}\right) \ni f_2\left(t_p, U^p\right),\\ &U^N = u_h(\tau_1),\end{aligned}$$

which is equivalent to

$$\forall p \in \{N, \dots, 2N-1\}, \quad \frac{U^{p+1} - U^p}{h} + A\left(U^{p+1}\right) + B\left(ph, U^p\right) + G\left(U^{p-N}\right) \ni 0, \tag{4.24}$$

$$U^N = u_h(\tau_1). \tag{4.25}$$

According to Lemma 4.3 applied on the interval $[\tau_1, \tau_2]$ we then have

$$\forall t \in [\tau_1, \tau_2], \quad |u(t) - u_h(t)| \leq |u(\tau_1) - u_h(\tau_1)| \, e^{\tau_1 L} + C h^{\alpha},$$

where $\alpha = 1/2$ in the general case and $\alpha = 1$ if $A = \partial \phi_K$. Thanks to estimate (4.23), we obtain

$$\forall t \in [\tau_1, \tau_2], \quad |u(t) - u_h(t)| \leq C h^{\alpha} \left(1 + e^{\tau_1 L}\right).$$

By induction, we prove easily that the numerical scheme (4.16)–(4.17) is equivalent to, for all $r \in \{0, \dots, Q-1\}$,

$$\begin{aligned}&\forall p \in \{rN, \dots, (r+1)N - 1\}, \quad \frac{U^{p+1} - U^p}{h} + A\left(U^{p+1}\right) + B\left(ph, U^p\right) + G\left(U^{p-N}\right) \ni 0,\\ &U^{rN} = u_h(\tau_r),\end{aligned}$$

and according to Lemma 4.3 applied on the interval $[\tau_r, \tau_{r+1}]$ we then have

$$\forall t \in [\tau_r, \tau_{r+1}], \quad |u(t) - u_h(t)| \leq \left(e^{\tau_1 L} + e^{(\tau_2 - \tau_1) L} + \dots + e^{(T - \tau_{Q-1})}\right) C h^{\alpha},$$

which allows us to conclude this proof since this implies that there exists M not depending on h such that

$$\forall t \in [0, T], \quad |u(t) - u_h(t)| \leq M h^{\alpha}. \qquad \square$$

4.3. Numerical simulations. We study numerically system (2.7). We assume that

$$\alpha \geq 0, \quad x_0 \in W^{2,\infty}(-\tau, 0), \quad g \in H^1(0, T). \tag{4.26}$$

We consider Hilbert space $H = \mathbb{R}^2$ equipped with its canonical scalar product. We define the multivalued operator A on H by

$$\forall u = (u_1, u_2) \in \mathbb{R}^2, \quad A(u_1, u_2) = \{0\} \times \alpha \sigma\left(u_2\right).$$

Since $\alpha \geq 0$ and σ is maximal monotone on $\mathbb{R}$ (see, for example, [6]), A is maximal monotone on H. We define the functions B, G, and z by, for all $u = (u_1, u_2) \in \mathbb{R}^2$, for all $t \in [0, T]$,

$$\begin{aligned} B(t,u) &= \begin{pmatrix} -u_2 \\ ku_1 + cu_2 - H(t) \end{pmatrix}, \\ G(u) &= \begin{pmatrix} 0 \\ \gamma u_1 \end{pmatrix}, \\ z(t) &= \begin{pmatrix} x_0(t) \\ \dot{x}_0(t) \end{pmatrix}. \end{aligned}$$

By setting $u(t) = (x(t), \dot{x}(t))$, we see that system (2.7) is equivalent to systems (3.4) and (3.5). According to assumption (4.26), we prove easily that assumptions (2.6) hold. Then, thanks to Proposition 3.2, system (2.7) admits a unique solution, whose restriction to $[0, T]$ belongs to $W^{2,\infty}(0, T)$. Moreover, thanks to Proposition 4.4, the numerical scheme (4.16)–(4.17) is of order 1/2, i.e.,

$$\forall t \in [0,T], \quad |u(t) - u_h(t)|_{\mathbb{R}^2} \leq C\sqrt{h}. \tag{4.27}$$

We recall that the subdifferential of a convex proper and lower semicontinuous function ϕ from H to $]-\infty, +\infty]$ is defined by

$$\forall (x,y) \in H \times H, \quad y \in \partial\phi(x) \Longleftrightarrow \forall z \in H, \quad \phi(z) - \phi(x) \geq \langle y, z - x \rangle.$$

So we see that

$$\forall x \in \mathbb{R}, \quad \sigma(x) = \partial |x|.$$

Then

$$\forall u = (u_1, u_2) \in \mathbb{R}^2, \quad A(u_1, u_2) = \partial \left(\alpha |u_2| \right).$$

Function $\alpha|.|$ is not equal to the subdifferential of a indicatrix of a closed convex set of $\mathbb{R}^2$: we cannot conclude from Proposition 4.4 that the order of the numerical scheme is equal to 1.

We consider now the numerical scheme defined in section 4.2 and we set

$$\forall p \in \{-N, \dots, M\}, \quad U^p = (u^p, v^p).$$

After computation, we obtain, for all h,

$$\forall \left(u_1, u_2\right) \in \mathbb{R}^2, \quad \left(I + hA\right)^{-1} \left(u_1, u_2\right) = \begin{pmatrix} u_1 \\ \left(I + h\alpha\sigma\right)^{-1} \left(u_2\right) \end{pmatrix},$$

where (see Figure 3)

$$\forall x \in \mathbb{R}, \quad \left(I + h\alpha\sigma\right)^{-1}(x) = \begin{cases} x - \alpha h & \text{if } x \geq \alpha h, \\ x + \alpha h & \text{if } x \leq -\alpha h, \\ 0 & \text{if } x \in [-\alpha h, \alpha h]. \end{cases}$$

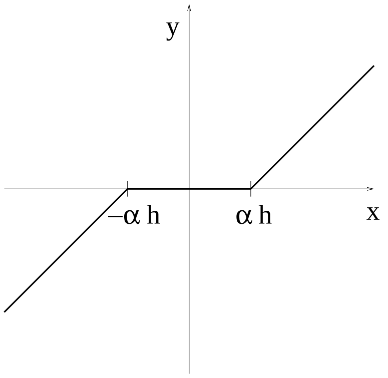


FIG. 3. *The function* $(I + h\alpha\sigma)^{-1}$.

Numerical scheme (4.16)–(4.17) is equivalent to (4.18)–(4.17) and can then be rewritten as

$$\forall p \in \{0,\dots,M-1\},\quad u^{p+1} = hv^p + u^p, \tag{4.28a}$$

$$v^{p+1} = (I+h\alpha\sigma)^{-1}\left(-chv^p - khu^p - \gamma hu^{p-N} + hH(ph) + v^p\right) \tag{4.28b}$$

$$\forall p \in \{-N,\dots,0\},\quad u^p = x_0(ph), \tag{4.28c}$$

$$v^p = \dot{x}_0(ph). \tag{4.28d}$$

We choose three values of parameters c, k, γ, α, τ and functions H and x_0:

$$c=0.1,\quad k=0.5,\quad \gamma=1.1,\quad \alpha=0.99,\quad \tau=1,\quad H(t)=\sin(t),\quad x_0(t)=0, \tag{4.29}$$

$$c=0.1,\quad k=0.5,\quad \gamma=1.1,\quad \alpha=0.10,\quad \tau=1,\quad H(t)=\sin(t),\quad x_0(t)=0, \tag{4.30}$$

$$c=0.1,\quad k=0.5,\quad \gamma=1.1,\quad \alpha=0.99,\quad \tau=1,\quad H(t)=\sin(t),\quad x_0(t)=0.5\sin(t). \tag{4.31}$$

For these three simulations, we choose

$$Q = 100,\quad N = 10^4. \tag{4.32}$$

We plot the discrete abscissa and velocity computed with numerical scheme (4.28) in Figures 4–6. In Figure 4, we observe, as in [21], a stable periodical regime after a transition; this regime is composed of statical phases (with $\dot{x} \equiv 0$) and dynamical phases (with $\operatorname{sign}(\dot{x}) \in \{-1, 1\}$). In Figures 5 and 6, the behavior seems unstable and we see only transient phases.

As in [3], we look second for an empirical order of convergence of the numerical scheme. We expect the error to be of the form

$$\|u - u_h\|_{C^0([0,T],\mathbb{R}^2)} \approx Ch^\delta, \tag{4.33}$$

and we try to identify the numbers C and δ. According to (4.13), (4.14), and (4.15), we rewrite (4.33) under the form

$$\|u - u_N\|_{C^0([0,T],\mathbb{R}^2)} \approx \frac{C}{N^\delta}. \tag{4.34}$$

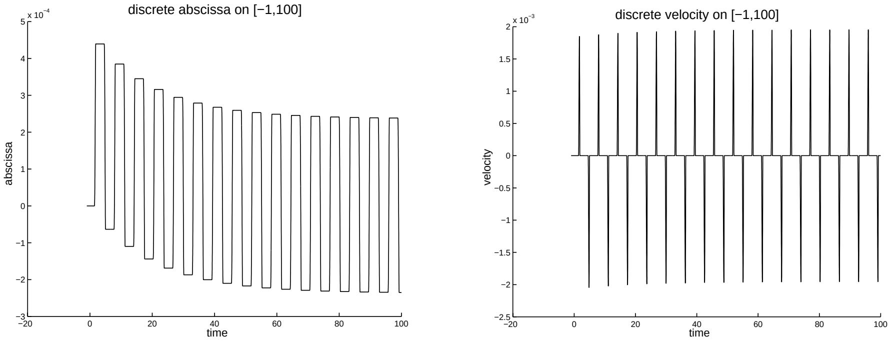


FIG. 4. *Discrete abscissa and velocity for differential inclusion* (2.7) *with* (4.29) *and* (4.32).

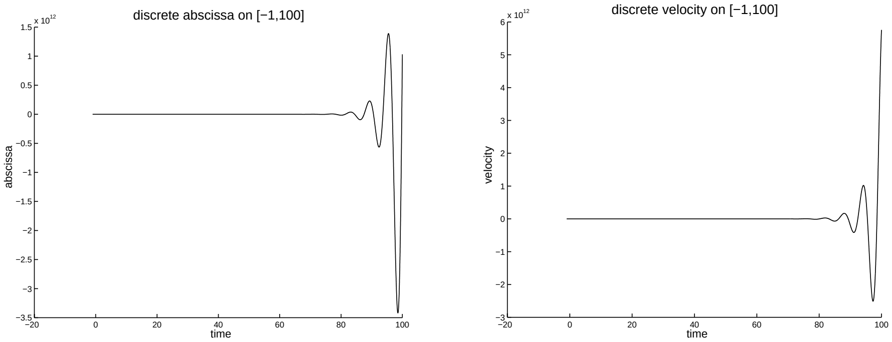


FIG. 5. *Discrete abscissa and velocity for differential inclusion* (2.7) *with* (4.30) *and* (4.32).

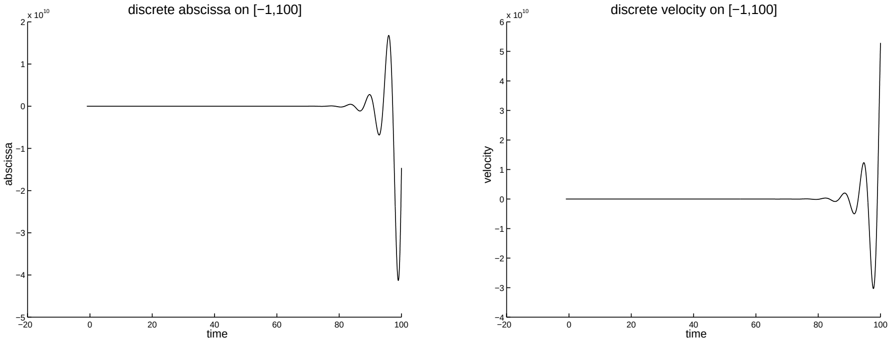


FIG. 6. *Discrete abscissa and velocity for differential inclusion* (2.7) *with* (4.31) *and* (4.32).

Define

$$\varepsilon(N) = \|u_N - u_{2N}\|_{C^0([0,T],\mathbb{R}^2)};$$

then, formally,

$$\log\left(\varepsilon(N)\right) \approx -\delta \log(N) + \log(2C).$$

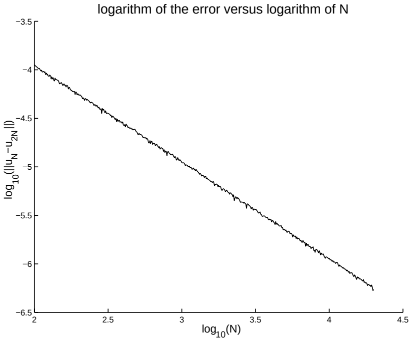


FIG. 7. *Log-log curves* $\varepsilon(N)$ *versus* N *for differential inclusion* (2.7) *with* (4.29)*,* $Q = 100$*, and* (4.35)*.*

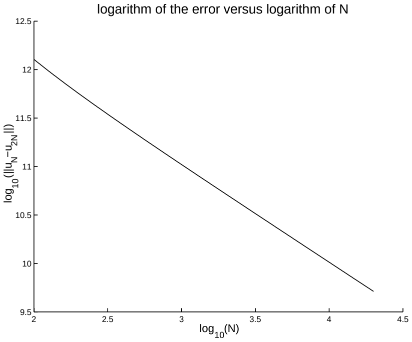


FIG. 8. *Log-log curves* $\varepsilon(N)$ *versus* N *for differential inclusion* (2.7) *with* (4.30)*,* $Q = 100$*, and* (4.35)*.*

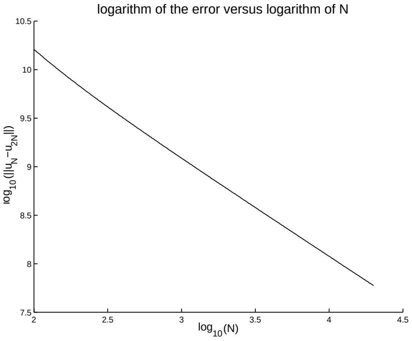


FIG. 9. *Log-log curves* $\varepsilon(N)$ *versus* N *for differential inclusion* (2.7) *with* (4.31)*,* $Q = 100$*, and* (4.35)*.*

A log-log plot of $\varepsilon(N)$ versus N gives an estimate of C and δ. We choose

(4.35)
$$p = 500, \quad N_{\min} = 100, \quad N_{\max} = 20000 \quad \forall i \in \{1, \dots, p\}, \quad N_i = N_{\min}^{\frac{p-i}{p-1}} N_{\max}^{\frac{i-1}{p-1}},$$

and the same physical parameters as above ($Q = 100$ and (4.29), (4.30), and (4.31)). Log-log curves $\varepsilon(N)$ versus N are plotted in Figures 7–9, and the values of δ and the

TABLE 1
Values δ and r for different values of parameters.

Parameter defined by	δ	r
(4.29)	0.995894134	0.999887705
(4.30)	1.0283258	0.999836087
(4.31)	1.04108214	0.999677062

correlation of set of points r are given in Table 1.

We see that the empirical order δ and the correlation r are close to one, which is more accurate than (4.27). The numerical scheme is then of order one, but in this case we proved only that the order is equal to 1/2. To prove that numerical scheme (4.1) and (4.2) is of order one for all maximal monotone graphs A (not necessarily equal to the subdifferential of the indicatrix of a closed set) is an open problem to our knowledge.

5. Conclusion. In this paper existence and uniqueness results have been provided for differential inclusions with both maximal monotone terms and delay terms. Even if in many interesting applications H has finite dimension, the results are very general ones. Based on the previous works, a numerical scheme of Euler implicit type has been proposed. This scheme has been proved to be of order 1/2 for general maximal monotone graphs and possesses order one in the case of the subdifferential of the indicatrix of a convex set. This numerical scheme has been tested in a previous work [21]: a good agreement is obtained with the exact solution known. Here it can be noticed that the numerical behavior is even better than those forecasted by theory. This is an important point for applications: the proposed scheme is the only one available in the literature with rigorous mathematical results (existence, uniqueness, convergence, and order), but 1/2 is a rather weak order. So it is useful to have indeed order one in practice. This good behavior has been observed in all our simulations. As we said, the proof of better theoretical results in the general case is an open problem.

Acknowledgment. We want to thank the reviewers for indicating a few useful references.

REFERENCES

[1] H.D. ALBER, *Materials with Memory*, Lecture Notes in Math. 1682, Springer-Verlag, Berlin, Heidelberg, 1998.

[2] J. BASTIEN AND M. SCHATZMAN, *Schéma numérique pour des inclusions différentielles avec terme maximal monotone*, C. R. Acad. Sci. Paris Sér. I Math., 330 (2000), pp. 611–615.

[3] J. BASTIEN AND M. SCHATZMAN, *Numerical precision for differential inclusions with uniqueness*, M2AN Math. Model. Numer. Anal., 36 (2002), pp. 427–460.

[4] J. BASTIEN, M. SCHATZMAN, AND C.-H. LAMARQUE, *Study of some rheological models with a finite number of degrees of freedom*, Eur. J. Mech. A Solids, 19 (2000), pp. 277–307.

[5] J. BASTIEN, *Étude théorique et numérique d'inclusions différentielles maximales monotones. Applications à des modèles élastoplastiques*, Ph.D. thesis, number 96-2000, Université Lyon I, Villeurbanne Cedex, France, 2000.

[6] H. BREZIS, *Opérateurs Maximaux Monotones et Semi-Groupes de Contractions dans les Espaces de Hilbert*, North-Holland Math. Stud. 5, Notas de Matemática (50), North-Holland, Amsterdam, 1973.

[7] H. BREZIS, *Problèmes unilatéraux*, J. Math. Pures Appl. (9), 51 (1972), pp. 1–168.

[8] B. BROGLIATO, *Nonsmooth Impact Mechanics, Models, Dynamics and Control*, Springer-Verlag, London, 1996.

[9] B. COUNT AND C.M. ELLIOT, *Analysis of a wave power device*, in Industrial Numerical Analysis, C. M. Elliot and S. McKee, eds., Oxford University Press, New York, 1986.

[10] M.G. Crandall and L.C. Evans, *On the relation of the operator $\partial/\partial s + \partial/\partial\tau$ to evolution governed by accretive operators*, Israel J. Math., 21 (1975), pp. 261–278.

[11] E. Cépa, *Equations différentielles stochastiques multivoques*, Ph.D. thesis, Université d'Orléans, Orleans Cedex, France, 1995.

[12] E. Cépa, *Équations différentielles stochastiques multivoques*, in Seminaire de Probabilités XXIX, Lecture Notes in Math. 1613, Springer-Verlag, Berlin, 1995, pp. 86–107.

[13] K. Deimling, *Multivalued Differential Equations*, Walter de Gruyter, Berlin, 1992.

[14] A. Dontchev and Frank Lempio, *Difference methods for differential inclusions: A survey*, SIAM Rev., 34 (1992), pp. 263–294.

[15] C.M. Elliot, *On the convergence of a one-step method for the solution of an ordinary differential inclusion*, IMA J. Numer. Anal., 5 (1985), pp. 3–27.

[16] C.M. Elliot and S. McKee, *On the numerical solution of an integrodifferential equation arising from wave power hydraulics*, BIT, (1981), pp. 318–325.

[17] C. Glocker, *Set-Valued Force Laws*, Lecture Notes Appl. Mech. 1, Springer-Verlag, Berlin, Heidelberg, 2001.

[18] A.M. Gouskov, A.S. Voronov, H. Paris, and S.A. Batzer, *Cylindrical workpiece turning using multiple-cutter tool heads*, in Proceedings of the ASME Design Engineering Technical Conferences, Vol. DETC2001/VIB-21431, Pittsburgh, PA, CD-ROM, ASME, New York, 2001.

[19] J.K. Hale and S.M. Verduyn Lunel, *Introduction to Functional Differential Equations*, Appl. Math. Sci. 99, Springer-Verlag, New York, 1993.

[20] I. Hlavaček, J. Haslinger, J. Nečas, and J. Lovišek, *Solution of Variational Inequalities in Mechanics*, Appl. Math. Sci. 66, Springer-Verlag, New York, 1988.

[21] M. Holland, *Prise en compte d'effets différés dans des systèmes discrets comportant des noms linéarités irrégulières*, Master's thesis, École Nationale des Travaux Publics de l'État—École doctorale Mécanique Énergétique Génie Civil Acoustique MEGA, Vaulx-en-Velin Cedex, France, 2001, Mémoire de Diplôme d'Études Approfondies, soutenu le 19 septembre, 2001.

[22] S. Hu and N.S. Papageorgiou, *Time-dependent subdifferential evolution inclusions and optimal control*, Mem. Amer. Math. Soc., 133 (1998), no. 632.

[23] T. Insperger and G. Stépán, *Semi-discretization of delayed dynamical systems*, in Proceedings of the ASME Design Engineering Technical Conferences, Vol. DETC2001/VIB-21446, Pittsburgh, PA, CD-ROM, ASME, New York, 2001.

[24] P. Kree, *Diffusion equation for multivalued stochastic differential equations*, J. Funct. Anal., 49 (1982), pp. 73–90.

[25] G. Lippold, *Error estimates for the implicit Euler approximation of an evolution inequality*, Nonlinear Anal., 15 (1990), pp. 1077–1089.

[26] D. Lépingle and C. Marois, *Equations différentielles stochastiques multivoques unidimensionnelles*, Sémin. Proba., 21 (1987), pp. 520–533.

[27] J.J. Moreau, *Evolution problem associated with a moving convex set in a hilbert space*, J. Differential Equations, 26 (1977), pp. 252–264.

[28] J. Pan and C.-Y. Su, *Modeling and chatter suppression with ultra-precision in dynamic turning metal cutting process*, in Proceedings of the ASME Design Engineering Technical Conferences, Vol. DETC2001/VIB-21435, Pittsburgh, PA, CD-ROM, ASME, New York, 2001.

[29] G. Stépán, *Retarded Dynamical Systems*, Longman, Harlow, 1989.

[30] D.E. Stewart, *Rigid-body dynamics with friction and impact*, SIAM Rev., 42 (2000), pp. 3–39.

[31] K. Yamada and T. Kobori, *Fundamental dynamics and control strategies for aseismic structural control*, Internat. J. Solids Structures, 38 (2001), pp. 6079–6121.

SIAM J. Numer. Anal.
Vol. 41, No. 4, pp. 1301–1317

A FINITE VOLUME SCHEME FOR TWO-PHASE IMMISCIBLE FLOW IN POROUS MEDIA*

ANTHONY MICHEL†

Abstract. In this paper, we prove the convergence of a numerical method for solving two-phase immiscible, incompressible flow in porous media. The method combines an upwind time implicit finite volume scheme for the saturation equation (hyperbolic-parabolic type) and a centered finite volume scheme for the Chavent global pressure equation (elliptic type). The capillary pressure is not neglected, and we study the case when the diffusion term in the saturation equation is weakly degenerated. Estimates on the approximate solution are proven; then by using compactness theorems we obtain a limit when the size of the discretization goes to zero, and we prove that this limit is the unique weak solution of the problem that we study.

Key words. multiphase flow, Darcy's law, porous media, degenerate elliptic parabolic system, finite volume scheme

AMS subject classifications. 35K65, 76S05, 65M12

DOI. 10.1137/S0036142900382739

1. Introduction.

In this paper, we define and analyze a finite volume method for a mathematical model for the flow of two immiscible incompressible fluids in a porous medium:

$$u_t - \operatorname{div}(k_1(u)\nabla p) = q_1 \text{ in } \Omega \times (0,T), \tag{1}$$

$$(1-u)_t - \operatorname{div}(k_2(u)\nabla(p + p_c(u))\,) = q_2 \text{ in } \Omega \times (0,T), \tag{2}$$

where Ω is a bounded domain of $\mathbb{R}^d$ ($d = 1, 2$ or 3) modeling the reservoir, $\{u, p\}$ are the saturation and the pressure of the wetting fluid (the water in the oil recovery context), k_1 and k_2 are the reduced mobilities of the wetting and the nonwetting fluid, and $\{q_1, q_2\}$ are source terms which model injection or production wells inside the reservoir.

If we introduce the global pressure of Chavent,

$$\theta = p + \int_0^u \frac{k_2(s)}{k_1(s)+k_2(s)}\, {p_c}'(s)ds, \tag{3}$$

and the total velocity flow,

$$\mathbf{F} = -(k_1(u) + k_2(u))\nabla p - k_2(u)\nabla p_c(u), \tag{4}$$

then the system (1)–(2) is equivalent to the following system of two coupled partial differential equations for the unknowns u and θ:

$$u_t + \operatorname{div}(f(u)\mathbf{F}) - \Delta\varphi(u) = q_1, \tag{5}$$

$$\operatorname{div}(M(u)\nabla\theta) = q_1 + q_2, \tag{6}$$

$$(\mathbf{F} = -M(u)\nabla\theta), \tag{7}$$

*Received by the editors December 19, 2000; accepted for publication (in revised form) December 23, 2002; published electronically August 22, 2003.

http://www.siam.org/journals/sinum/41-4/38273.html

†Institut Français du Pétrole, 1 et 4 avenue du Bois Préau, BP 311, 92852 Rueil Malmaison Cedex, France (anthony.michel@ifp.fr).

where $M = k_1 + k_2$ is the total mobility of the two phase fluid, $f = \frac{k_1}{k_1+k_2}$ is the fractional flow of the wetting fluid, and $\varphi(u) = -\int_0^u \frac{k_1(s)k_2(s)}{k_1(s)+k_2(s)} {p_c}'(s)ds$ (introduced by Arbogast in [1]) is a nonlinear function which is closely related to the capillary pressure p_c .

This model has been well known in the reservoir simulation community for more than 30 years [3], [4], [7], [26], [24], [12]. It has been studied from the theoretical point of view [1], [23], [8], and results on existence or uniqueness and some estimates on the weak solution have been proved in different cases. The interest in this problem for oil companies has resulted in early motivated investigations in numerical methods [12], [7]. In particular, numerous works [25], [2], [15], [14], [5], [13], [16] have been done on convergence and error estimates of numerical schemes using a mixed finite element method [27] to approximate the total velocity flow and the pressure. Both miscible and immiscible cases have been extensively treated where the capillary pressure is neglected (the saturation equation is then of hyperbolic type). It is only recently [25], [8] that equivalent studies have been done in the case of immiscible fluids with degenerate capillary pressure.

The aim of this paper is to show that if we discretize the two equations (5)–(6) using first order finite volume schemes (see [18], [28]), then the numerical approximate solution converges to the exact solution of problem (5)–(6). This work takes its originality from the fact that we consider a full finite volume method on a single mesh, and we allow any weak degeneracy of the function φ. The convergence proof is done without assuming any existence or regularity on the weak solution.

The paper is organized as follows. In section 2, we give some assumptions. In section 3, we present the finite volume scheme and prove a priori estimates on the discrete solution. As a corollary we obtain that the implicit scheme is well defined. In section 4.1, we use these estimates to obtain compactness properties on the corresponding piecewise constant approximate solution. In section 5, we state and prove the main result: the convergence theorem, Theorem 5.1. Then we discuss the last step of the proof which consists of passing to the limit in the weak formulation. In section 6, we present some numerical results, and we end in section 7 with some concluding remarks.

2. Assumptions.

In order to close the system (5)–(6), we prescribe the following Neumann boundary conditions:

$$\nabla\theta \cdot \mathbf{n} = 0 \text{ on } \partial\Omega \times (0,T), \tag{8}$$

$$\nabla\varphi(u) \cdot \mathbf{n} = 0 \text{ on } \partial\Omega \times (0,T); \tag{9}$$

the following initial condition:

$$u(\cdot,0) = u_0 \text{ on } \Omega; \tag{10}$$

and, since the pressure is defined only up to a constant, we prescribe the following arbitrary condition:

$$\int_\Omega \theta(x,\cdot)dx = 0 \text{ on } (0,T). \tag{11}$$

The source terms depend on the saturation u and on the injection concentration c. They are defined by

$$q_1 = c\overline{s} - u\underline{s}, \tag{12}$$

$$q_2 = (1-c)\overline{s} - (1-u)\underline{s}. \tag{13}$$

We make the following assumptions on the data:

- φ is an increasing Lipschitz continuous function on $[0,1]$, Φ will denote its Lipschitz constant, and $\varphi^* = \max_{x\in[0,1]} |\varphi(x)|$.
- f is a nondecreasing Lipschitz continuous function on $[0,1]$, $\mathbf{F}$ will denote its Lipschitz constant, $f(0)=0$, and $f(1)=1$.
- M is a continuous function on $[0,1]$ with $0 < M_* \leq M(u) \leq M^* < \infty$.
- The functions $\overline{s}$ and $\underline{s}$ belong to $L^\infty(0,T,L^2(\Omega))$. $\overline{s} \geq 0$ and $\underline{s} \geq 0$ a.e. $x \in \Omega \times (0,T)$, $\int_\Omega \overline{s}(x) - \underline{s}(x)dx = 0$ for a.e. $t \in (0,T)$.
- $u_0 \in L^\infty(\Omega)$, $0 \leq u_0(x) \leq 1$ a.e. $x \in \Omega$.
- c is a constant, $0 \leq c \leq 1$.

Remark 2.1. *The function c can also be taken in $L^\infty(Q)$ with $0 \leq c \leq 1$ without any difficulty. One needs only to replace c by an approximate in the finite volume scheme.*

3. The finite volume scheme.

Assume Ω is a polygonal bounded domain of $\mathbb{R}^d$ and $\mathcal{T}$ is a mesh of Ω consisting of convex polygons. The finite volume method (cf. [18]) consists of integrating the equations over a control volume $K \in \mathcal{T}$ and obtaining a relation between mean values over K and fluxes on the edges of K by using the Stokes formula. Thanks to the Neumann boundary conditions, the normal fluxes on the boundary $\partial\Omega$ are equal to zero, so we need only to consider the interfaces between two control volumes. For convection terms, a simple way to get stability is to compute an upwind scheme, but the results also extend to a general monotone scheme (see [6]). For diffusion terms, we have to approximate the normal derivative of $\varphi(u)$ on the interface. Without additional assumptions on the mesh, we can take into account the values of u on control volumes other than the two neighbors K and L of the interface (see, for example, the VF9 method in [21]). Here we make some assumptions (see Definition 3.1) in order to ensure that the cheap discretization of $\frac{\partial\varphi(u)}{\partial n}$ by $\frac{\varphi(U_L^{n+1})-\varphi(U_K^{n+1})}{d_{K|L}}$ is consistent.

3.1. Definitions and notations.

Definition 3.1 (admissible mesh of Ω). *An admissible mesh $\mathcal{T}$ of Ω is given by a set of open bounded polygonal convex subsets of Ω called control volumes, a family $\mathcal{E}$ of subsets of $\bar{\Omega}$ contained in hyperplanes of $\mathbb{R}^d$ with strictly positive measure, and a family of points (the "centers" of control volumes) satisfying the following properties:*

(i) *The closure of the union of all control volumes is $\bar{\Omega}$.*

(ii) *For any $(K,L) \in \mathcal{T}^2$ with $K \neq L$, either the length of $\bar{K}\cap\bar{L}$ is 0 or $\bar{K}\cap\bar{L} = \bar{\sigma}$ for some $\sigma \in \mathcal{E}$. Then we will denote $\sigma = K|L$.*

(iii) *For any $K \in \mathcal{T}$, there exists a subset $\mathcal{E}(K)$ of $\mathcal{E}$ such that $\partial K = \bar{K}\backslash K = \cup_{\sigma\in\mathcal{E}(K)}\bar{\sigma}$. Furthermore, $\mathcal{E} = \cup_{K\in\mathcal{T}}\mathcal{E}(K)$, and we will denote $\mathcal{N}(K)$ the set of boundary control volumes of K, that is, $\mathcal{N}(K) = \{L \in \mathcal{T}, K|L \in \mathcal{E}(K)\}$.*

(iv) *The family of points $(x_K)_{K\in\mathcal{T}}$ is such that $x_K \in K$ (for all $K \in \mathcal{T}$), and, if $\sigma = K|L$, it is assumed that the straight line (x_K, x_L) is orthogonal to σ.*

For a control volume $K \in \mathcal{T}$, we denote $m(K)$ its measure. If $L \in \mathcal{N}(K)$, then we denote $m(K|L)$ the measure of the interface $K|L$ in $\mathbb{R}^{d-1}$, $d_{K|L}$ the distance between the centers of the control volumes K and L, $T_{K|L} = \frac{m(K|L)}{d_{K|L}}$ the discrete transmissibility, and $\mathbf{n}_{K,L}$ the normal vector of $K|L$ outward to K. We denote $d_{K,K|L}$ the distance between the center x_K of K and the interface $K|L$, and we define the size of the mesh by

$$size(\mathcal{T}) = \max_{K\in\mathcal{T}} diam(K).$$

To prove the convergence theorem, Theorem 5.1, we need uniform regularity properties on meshes in the following sense.

DEFINITION 3.2. *An admissible mesh $\mathcal{T}$ is ξ-regular if for all $K \in \mathcal{T}$,*

$$\sum_{L\in\mathcal{N}(K)} m(K|L)d_{K|L} \le m(K)\,\xi.$$

3.2. The scheme. Let $\mathcal{T}$ be an admissible mesh, and let δt be a time step such that $T = (N+1)\delta t$ with $N \in \mathbb{N}$. We define $\overline{S}_K^{n+1}$ and $\underline{S}_K^{n+1}$ by

$$\overline{S}_K^{n+1} = \frac{1}{\delta t}\int_{n\delta t}^{(n+1)\delta t}\int_K \overline{s},$$
$$\underline{S}_K^{n+1} = \frac{1}{\delta t}\int_{n\delta t}^{(n+1)\delta t}\int_K \underline{s}.$$

With the notations previously introduced, one may define a finite volume scheme as the following set of equations for the discrete unknowns (U,Θ), where $U = (U_K^n)_{K\in\mathcal{T},n\in[\![0,N+1]\!]}$ and $\Theta = (\Theta_K^n)_{K\in\mathcal{T},n\in[\![1,N+1]\!]}$:

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

$$U_K^0 = \frac{1}{m(K)}\int_K u_0(x)dx; \tag{14}$$

$\forall K \in \mathcal{T}, \forall n \in [\![0,N]\!]$,

$$\begin{aligned}\frac{U_K^{n+1}-U_K^n}{\delta t}m(K) - \sum_{L\in\mathcal{N}(K)} T_{K|L}(\varphi(U_L^{n+1}) - \varphi(U_K^{n+1}))\\ + \sum_{L\in\mathcal{N}(K)} \mathbf{F}_{K,L}^{n+1} f(u)_{K|L}^{n+1} = c\overline{S}_K^{n+1} - U_K^{n+1}\,\underline{S}_K^{n+1};\end{aligned} \tag{15}$$

$\forall K \in \mathcal{T}, \forall L \in \mathcal{N}(K), \forall n \in [\![0,N]\!]$,

$$\mathbf{F}_{K,L}^{n+1} = -M(u)_{K|L}^{n+1} T_{K|L}(\Theta_L^{n+1} - \Theta_K^{n+1}); \tag{16}$$

$\forall K \in \mathcal{T}, \forall n \in [\![0,N]\!]$,

$$\sum_{L\in\mathcal{N}(K)} \mathbf{F}_{K,L}^{n+1} = \overline{S}_K^{n+1} - \underline{S}_K^{n+1}; \tag{17}$$

$\forall n \in [\![0,N]\!]$,

$$\sum_{K\in\mathcal{T}} m(K)\Theta_K^{n+1} = 0, \tag{18}$$

where $f(u)_{K|L}^{n+1}$ and $M(u)_{K|L}^{n+1}$ are, respectively, an upwind discretization of $f(u)$ and a consistent approximation of $M(u)$ on the interface $K|L$ given by

$$f(u)_{K|L}^{n+1} = \begin{cases} f(U_K^{n+1}) & \text{if} \quad \mathbf{F}_{K,L}^{n+1} > 0, \\ f(U_L^{n+1}) & \text{if} \quad \mathbf{F}_{K,L}^{n+1} < 0, \end{cases}$$

$$M(u)_{K|L}^{n+1} = \frac{d_{K|L}}{\frac{d_{K,K|L}}{M(U_K^{n+1})} + \frac{d_{L,K|L}}{M(U_L^{n+1})}}. \tag{19}$$

REMARK 3.1. *Definition* (19) *of $M(u)_{K|L}^{n+1}$ by a harmonic mean ensures consistency property in the general case when the function $M(u)$ is discontinuous on the interface $K|L$ (see* [22], [18]*). However, in our proof of convergence we need only $M(u)_{K|L}^{n+1}$ to be in the interval $[M(U_K^{n+1}), M(U_L^{n+1})]$ since the functions used are more regular.*

3.3. A priori estimates. The scheme (14)–(18) is time implicit, so the existence of a solution must be proven. We will first prove a priori estimates assuming existence of a solution and then prove the existence by using the Leray–Schauder theorem (see [11]). We also use these estimates to obtain compactness properties.

PROPOSITION 3.1. *Assume that $((U_K^n)_{K\in\mathcal{T},n\in[\![0,N+1]\!]}, (\Theta_K^{n+1})_{K\in\mathcal{T},n\in[\![0,N]\!]})$ is a solution to* (14)–(18)*; then*

$$0 \le U_K^{n+1} \le 1 \;\forall K \in \mathcal{T}, \forall n \in [\![0, N+1]\!]. \tag{20}$$

Moreover, there exist $C_1(u_0, \overline{s}, \underline{s}, \Phi) \ge 0$ and $C_2(M_*, \overline{s}, \underline{s}) > 0$ such that

$$\sum_{n=0}^{N} \delta t \frac{1}{2} \sum_{K\in\mathcal{T}} \sum_{L\in\mathcal{N}(K)} T_{K|L}(\varphi(U_L^{n+1}) - \varphi(U_K^{n+1}))^2 \le C_1 \tag{21}$$

and

$$\sum_{K\in\mathcal{T}} \sum_{L\in\mathcal{N}(K)} T_{K|L}(\Theta_L^{n+1} - \Theta_K^{n+1})^2 \le C_2 \;\forall n \in [\![0, N]\!]. \tag{22}$$

Proof. Rewriting the discrete convection flux in a nondivergence form using (19) and (17), we obtain

$\forall K \in \mathcal{T}, \forall n \in [\![0, N]\!],$

(23)
$$\sum_{L\in\mathcal{N}(K)} \mathbf{F}_{K,L}^{n+1} f(u)_{K|L}^{n+1} = - \sum_{L\in\mathcal{N}(K)} \mathbf{F}_{K,L}^{n+1\,-}(f(U_L^{n+1}) - f(U_K^{n+1})) + f(U_K^{n+1})(\overline{S}_K^{n+1} - \underline{S}_K^{n+1}),$$

where one denotes $x^- = \max(0, -x)$.

In order to prove the discrete maximum principle (20), we follow the continuous case. If U attains its bounds on $\Omega \times \{0\}$, i.e., at points of type $(K, 0)$, the definition of U_K^0 in (14) gives the conclusion. By contradiction, if, for example, $\max(U) > 1$, then necessarily U attains its maximum at an interior point of the parabolic domain $Q = \Omega \times [0,T)$, i.e., at a point of type $(K, n+1)$. In that case, by (23) and (16) we have

$$\begin{aligned} \frac{U_K^{n+1} - U_K^n}{\delta t} m(K) \quad &+ \sum_{L \in \mathcal{N}(K)} T_{K|L}(\varphi(U_K^{n+1}) - \varphi(U_L^{n+1})) \\ &+ \sum_{L \in \mathcal{N}(K)} \mathbf{F}_{K,L}^{n+1^-}(f(U_K^{n+1}) - f(U_L^{n+1})) \\ &+ (f(U_K^{n+1}) - c)\overline{S}_K^{n+1} + (U_K^{n+1} - f(U_K^{n+1}))\,\underline{S}_K^{n+1} = 0, \end{aligned} \tag{24}$$

and φ and f are nondecreasing functions, so we have $U_K^n \geq U_K^{n+1}$. Consequently, also U attains its maximum at point (K, n) and by induction the maximum is attained on $\Omega \times \{0\}$, which leads to a contradiction.

Proofs of the discrete energy estimates (21) and (22) also mimic continuous ones. Multiplying (24) by $\delta t\, U_K^{n+1}$ and summing the result over $K \in \mathcal{T}$ and $n \in [\![0, N]\!]$ yields $E1 + E2 + E3 + E4 = 0$ with

$$\begin{aligned} E1 &= \sum_{n=0}^{N} \sum_{K \in \mathcal{T}} m(K)(U_K^{n+1} - U_K^n)U_K^{n+1}, \\ E2 &= \sum_{n=0}^{N} \delta t \sum_{K \in \mathcal{T}} \sum_{L \in \mathcal{N}(K)} T_{K|L}(\varphi(U_K^{n+1}) - \varphi(U_L^{n+1}))U_K^{n+1}, \\ E3 &= \sum_{n=0}^{N} \delta t \sum_{K \in \mathcal{T}} \sum_{L \in \mathcal{N}(K)} \mathbf{F}_{K,L}^{n+1^-}(f(U_K^{n+1}) - f(U_L^{n+1}))U_K^{n+1}, \\ E4 &= \sum_{n=0}^{N} \delta t \sum_{K \in \mathcal{T}} (f(U_K^{n+1}) - c)U_K^{n+1}\overline{S}_K^{n+1} + (U_K^{n+1} - f(U_K^{n+1}))U_K^{n+1}\,\underline{S}_K^{n+1}. \end{aligned}$$

By a discrete time integration by parts, we obtain

$$\begin{aligned} E1 &= \frac{1}{2}\sum_{K \in \mathcal{T}} m(K)(U_K^{N+1})^2 - \frac{1}{2}\sum_{K \in \mathcal{T}} m(K)(U_K^0)^2 + \frac{1}{2}\sum_{n=0}^{N} \sum_{K \in \mathcal{T}} m(K)(U_K^{n+1} - U_K^n)^2 \\ &\geq -\frac{1}{2}\|u_0\|^2_{L^2(\Omega \times (0,T))}. \end{aligned}$$

Gathering by edges we get for $E2$

$$\begin{aligned} E2 &= \sum_{n=0}^{N} \delta t \frac{1}{2} \sum_{K \in \mathcal{T}} \sum_{L \in \mathcal{N}(K)} T_{K|L}(\varphi(U_K^{n+1}) - \varphi(U_L^{n+1}))(U_K^{n+1} - U_L^{n+1}) \\ &\geq \frac{1}{\Phi}\sum_{n=0}^{N} \delta t \frac{1}{2} \sum_{K \in \mathcal{T}} \sum_{L \in \mathcal{N}(K)} T_{K|L}(\varphi(U_K^{n+1}) - \varphi(U_L^{n+1}))^2. \end{aligned}$$

To deal with $E3$, we need a technical lemma which is proved, for example, in [18].

LEMMA 3.1. *Let f be a nondecreasing continuous function on $\mathbb{R}$, and define g by $g(u) = uf(u) - \int_0^u f(\tau)d\tau$. Then for every $(a, b) \in \mathbb{R}^2$,*

$$(f(a) - f(b))a \geq g(a) - g(b).$$

By using Lemma 3.1 and the local conservation property $\mathbf{F}_{K,L}^{n+1} + \mathbf{q}_{L,K}^{n+1} = 0$ we get

$$\begin{aligned} E3 &\geq \sum_{n=0}^{N} \delta t \sum_{K\in\mathcal{T}} \mathbf{F}_{K,L}^{n+1^-}(g(U_L^{n+1}) - g(U_K^{n+1})) \\ &= \sum_{K\in\mathcal{T}} g(U_K^{n+1}) \sum_{L\in\mathcal{N}(K)} \mathbf{F}_{K,L}^{n+1} \\ &= -\sum_{K\in\mathcal{T}} g(U_K^{n+1})(\overline{S}_K^{n+1} - \underline{S}_K^{n+1}). \end{aligned}$$

Hence,

$$\begin{aligned} E3 + E4 &\geq \sum_{n=0}^{N} \delta t \sum_{K\in\mathcal{T}} \overline{S}_K^{n+1}(f(U_K^{n+1})U_K^{n+1} - g(U_K^{n+1}) - cU_K^{n+1}) \\ &\quad + \sum_{n=0}^{N} \delta t \sum_{K\in\mathcal{T}} \underline{S}_K^{n+1}(g(U_K^{n+1}) - f(U_K^{n+1})U_K^{n+1} + (U_K^{n+1})^2) \\ &\geq \sum_{n=0}^{N} \delta t \sum_{K\in\mathcal{T}} -2(\overline{S}_K^{n+1} + \underline{S}_K^{n+1}) \\ &\geq -2(\|\overline{s}\|_{L^1(\Omega\times(0,T))} + \|\underline{s}\|_{L^1(\Omega\times(0,T))}). \end{aligned}$$

Collecting the previous inequalities yields exactly (21) with

$$C_1 = \Phi\left(\frac{1}{2}\|u_0\|^2_{L^2(\Omega\times(0,T))} + 2(\|\overline{s}\|_{L^1(\Omega\times(0,T))} + \|\underline{s}\|_{L^1(\Omega\times(0,T))})\right).$$

Now let us multiply (17) by Θ_K^{n+1} and sum over $K \in \mathcal{T}$. Gathering by edges, we obtain

$$\begin{aligned} \sum_{K\in\mathcal{T}} \frac{1}{2} \sum_{L\in\mathcal{N}(K)} M(u)_{K|L}^{n+1}(\Theta_K^{n+1} - \Theta_L^{n+1})^2 &= \sum_{K\in\mathcal{T}} (\overline{S}_K^{n+1} - \underline{S}_K^{n+1})\Theta_K^{n+1} \\ &\leq \|\overline{s} - \underline{s}\|_{L^\infty(0,T,L^2(\Omega))} \left(\sum_{K\in\mathcal{T}} m(K)(\Theta_K^{n+1})^2\right)^{\frac{1}{2}}. \end{aligned}$$

And by the discrete Poincaré inequality (see [10]), there exists $C(\Omega)$ such that

$$\sum_{K\in\mathcal{T}} m(K)(\Theta_K^{n+1})^2 \leq C(\Omega)^2 \sum_{K\in\mathcal{T}} \frac{1}{2} \sum_{L\in\mathcal{N}(K)} T_{K|L}(\Theta_K^{n+1} - \Theta_L^{n+1})^2.$$

This gives (22) with $C_2 = C(\Omega)\frac{1}{M_*}\|\overline{s} - \underline{s}\|^2_{L^2(\Omega\times(0,T))}$. □

3.4. Existence of the approximate solution. Let $E = \mathbb{R}^{[\![0,N+1]\!]\times\mathcal{T}}\mathbb{R}^{[\![1,N+1]\!]\times\mathcal{T}}$ and $\mathcal{G} : E \to E$ such that $\mathcal{G}(U,\theta) = (\tilde{U}, \tilde{\theta})$, where $(\tilde{U}, \tilde{\theta})$ is the solution of the following set of equations:

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

$$\tilde{U}_K^0 = \frac{1}{m(K)} \int_K u_0(x)dx;$$

$\forall K \in \mathcal{T}$ and $\forall n \in [\![0, N]\!]$,

$$\frac{\tilde{U}_K^{n+1} - \tilde{U}_K^n}{\delta t} m(K) - \sum_{\sigma \in \mathcal{E}(K)} T_{K|L}(\varphi(U_L^{n+1}) - \varphi(U_K^{n+1}))$$
$$- \sum_{\sigma \in \mathcal{E}(K)} \tilde{\mathbf{F}}_{K,L}^{n+1+} f(U_K^{n+1}) - \tilde{\mathbf{F}}_{K,L}^{n+1-} f(U_L^{n+1}) = c\overline{S}_K^{n+1} - U_K^{n+1}\,\underline{S}_K^{n+1};$$

$\forall K \in \mathcal{T}$, $\forall L \in \mathcal{N}(K)$, $\forall n \in [\![0, N]\!]$,

$$\tilde{\mathbf{F}}_{K,L}^{n+1} = -M(u)_{K|L}^{n+1} T_{K|L}(\tilde{\Theta}_L^{n+1} - \tilde{\Theta}_K^{n+1});$$

$\forall K \in \mathcal{T}, \forall n \in [\![0, N]\!]$,

$$\sum_{L \in \mathcal{N}(K)} \tilde{\mathbf{F}}_{K,L}^{n+1} = \overline{S}_K^{n+1} - \underline{S}_K^{n+1};$$

$\forall n \in [\![0, N]\!]$,

$$\sum_{K \in \mathcal{T}} m(K)\tilde{\Theta}_K^{n+1} = 0.$$

The function $\mathcal{G}$ is well defined only if there exists a unique solution to this set of equations. In fact, this system of equations can be easily solved iteratively (on the contrary to system (14)–(18)), because the diffusion term on the elliptic equation on $\tilde{\Theta}$, which is $M(u)_{K|L}^{n+1}$, is given and cannot degenerate since $M(u)_{K|L}^{n+1} \in [M(U_K^{n+1}), M(U_L^{n+1})]$ and $M(u) \geq M_* > 0$ by assumption.

By using the continuity of $U \to M(u)_{K|L}^{n+1}$, $x \to x^+$, and $x \to x^-$, we obtain in the same time that $\mathcal{G}$ is a continuous function. Now, by construction, for any $\alpha \in [0, 1]$, the problem $(U, \Theta) = \alpha\mathcal{G}(U, \Theta)$ has exactly the same solutions as the numerical scheme (14)–(18) with $\alpha\varphi$, αu_0, $\alpha\overline{s}$, and $\alpha\underline{s}$ instead of φ, u_0, $\overline{s}$, and $\underline{s}$.

But $\alpha \in [0, 1]$, so Φ is also a Lipschitz constant for $\alpha\varphi$, and we have $\|\alpha\overline{s}\| \leq \|\overline{s}\|$, $\|\alpha\underline{s}\| \leq \|\underline{s}\|$, $\|\alpha u_0\| \leq \|u_0\|$. Thus estimates given in Proposition 3.1 are uniformly satisfied for any $\alpha \in [0, 1]$ and any solution of $(U, \Theta) = \alpha\mathcal{G}(U, \Theta)$.

Now all the assumptions of the Leray–Schauder theorem are satisfied, so there exists a fixed point for $\mathcal{G}$; i.e., there exists at least a solution to the scheme (14)–(18).

4. Compactness results. Each solution $(U, \Theta)_{\mathcal{T},\delta t}$ of (14)–(18) for an admissible mesh $\mathcal{T}$ and a time step δt corresponds to an approximate solution $(u_{\mathcal{T},\delta t}, \theta_{\mathcal{T},\delta t})$ of problem (5)–(6) defined a.e. on $\Omega \times (0, T)$ by

$$u_{\mathcal{T},\delta t}(x, t) = U_K^{n+1}, x \in K, t \in (n\delta t, (n+1)\delta t),$$
$$\theta_{\mathcal{T},\delta t}(x, t) = \Theta_K^{n+1}, x \in K, t \in (n\delta t, (n+1)\delta t).$$

The first step toward the convergence theorem, Theorem 5.1, consists of the proof of compactness properties on $u_{\mathcal{T},\delta t}$ and $\theta_{\mathcal{T},\delta t}$, by using a priori estimates on the discrete solution obtained in Proposition 3.1.

4.1. Compactness of $\boldsymbol{u_{\mathcal{T},\delta t}}$. We shall prove that $\varphi(u_{\mathcal{T},\delta t})$ is relatively compact in $L^2(\Omega \times (0, T))$ for the strong topology by using Kolmogorov's theorem and that when $size(\mathcal{T}) \to 0$ and $\delta t \to 0$, the limit of each convergent sequence of approximate solutions belongs to $L^2(0,T,H^1(\Omega))$. Let us first recall Kolmogorov's compactness theorem, which is a consequence of the Ascoli compactness theorem.

THEOREM 4.1 (Fréchet–Kolmogorov). *Let $\mathcal{F}$ be a bounded subset of $L^2(\mathbb{R}^d)$, and let Ω be a bounded domain of $\mathbb{R}^d$; then $\mathcal{F}$ is relatively compact in $L^2(\Omega)$ if and only if*

$$\lim_{|\xi|\to 0}\sup_{f\in\mathcal{F}}\|f(\cdot+\xi)-f(\cdot)\|_{L^2(\mathbb{R}^d)}=0.$$

In our case, to apply this theorem on $Q=\Omega\times(0,T)$, we need to study the space and time translates of $\varphi(u_{\mathcal{T},\delta t})$. As a direct consequence of (21) (see [17], [18]) we already have the following result.

PROPOSITION 4.1 (space translates). *Let C_1 be defined in Proposition 3.1; then for all $\xi\in\mathbb{R}^d$,*

$$\int_0^T\int_{\Omega_\xi}[\varphi(u_{\mathcal{T},\delta t}(x+\xi,\cdot)-\varphi(u_{\mathcal{T},\delta t}(x,\cdot))]^2dx\le C\Phi|\xi|(2m(\mathcal{T})+|\xi|),$$

where $\Omega_\xi=\{x\in\Omega,[x,x+\xi]\subset\Omega\}$ and $|\xi|$ is the Euclidean norm on $\mathbb{R}^d$.

We can establish an analogue but with slightly different results for time translates estimates. We now adapt the method of [19], but since this method is not quite well known, we shall give the complete proof of it. Let us first state exactly the result we shall prove.

PROPOSITION 4.2 (time translates). *There exist $C'(\varepsilon,\varphi,f,\mathbf{q},u_0,\Omega,T)>0$ such that for every $s\in\mathbb{R}^+$,*

$$\int_0^{T-s}\int_\Omega(\varphi(u_{\mathcal{T},\delta t}(x,t+s))-\varphi(u_{\mathcal{T},\delta t}(x,t))\,)^2dxdt\le C's.$$

Proof. Let us define $A(t)=\int_\Omega(u_{\mathcal{T},\delta t}(x,t+s)-u_{\mathcal{T},\delta t}(x,t))(\varphi(u_{\mathcal{T},\delta t}(x,t+s)-\varphi(u_{\mathcal{T},\delta t}(x,t)))dxdt$. Then

$$\int_\Omega(\varphi(u_{\mathcal{T},\delta t}(x,t+s))-\varphi(u_{\mathcal{T},\delta t}(x,t))\,)^2dxdt\le A(t)\,\Phi.$$

If for any $t\in\mathbb{R}$ we denote $n(t)$ the integer part of $\frac{t}{\delta t}$, then for any $K\in\mathcal{T}$ and $x\in K$,

$$\begin{aligned}u_{\mathcal{T},\delta t}(x,t+s)-u_{\mathcal{T},\delta t}(x,t)&=\sum_{n=n(t)}^{n(t+s)-1}U_K^{n+1}-U_K^n\\&=\frac{1}{m(K)}\sum_{n=n(t)}^{n(t+s)-1}\delta t\sum_{L\in\mathcal{N}(K)}T_{K|L}(\varphi(U_K^{n+1})\\&\quad-\varphi(U_L^{n+1}))-\mathbf{F}_{K,L}^{n+1}f(u)_{K|L}^{n+1}.\end{aligned}$$

So $A(t)=A_1(s,t)-A_1(0,t)-A_2(s,t)+A_2(0,t)$, where for $\rho=0$ or $\rho=s$ we denote

$$A_1(\rho,t)=\sum_{n=n(t)}^{n(t+s)-1}\delta t\sum_{K\in\mathcal{T}}\sum_{L\in\mathcal{N}(K)}T_{K|L}(\varphi(U_K^{n+1})-\varphi(U_L^{n+1}))\varphi(U_K^{n(t+\rho)}),$$

$$A_2(\rho,t)=\sum_{n=n(t)}^{n(t+s)-1}\delta t\sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{E}(K)}\mathbf{F}_{K,L}^{n+1}f(u)_{K|L}^{n+1}\varphi(U_K^{n(t+\rho)}).$$

Gathering by edges and using the local conservation of the discrete fluxes, we get

$$A_1(\rho,t)=\sum_{n=n(t)}^{n(t+s)-1}\delta t\frac{1}{2}\sum_{K\in\mathcal{T}}\sum_{L\in\mathcal{N}(K)}T_{K|L}(\varphi(U_K^{n+1})-\varphi(U_L^{n+1}))(\varphi(U_K^{n(t+\rho)})-\varphi(U_L^{n(t+\rho)})),$$
$$A_2(\rho,t)=\sum_{n=n(t)}^{n(t+s)-1}\delta t\frac{1}{2}\sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{E}(K)}\mathbf{F}_{K,L}^{n+1}f(u)_{K|L}^{n+1}(\varphi(U_K^{n(t+\rho)})-\varphi(U_L^{n(t+\rho)})).$$

Now by using the Young inequality, we have

$$|A_1(\rho,t)|\le\frac{1}{2}\sum_{n=n(t)}^{n(t+s)-1}\delta t\,(S(n)+S(n(t+\rho))),$$
$$|A_2(\rho,t)|\le\frac{1}{2}\sum_{n=n(t)}^{n(t+s)-1}\delta t\,(R(n)+S(n(t+\rho))),$$

where

$$S(n)=\sum_{n=0}^{N}\delta t\frac{1}{2}\sum_{K\in\mathcal{T}}\sum_{L\in\mathcal{N}(K)}T_{K|L}(\varphi(U_K^{n+1})-\varphi(U_L^{n+1}))^2,$$
$$R(n)=\sum_{n=0}^{N}\delta t\frac{1}{2}\sum_{K\in\mathcal{T}}\sum_{L\in\mathcal{N}(K)}\frac{1}{T_{K|L}}(\mathbf{F}_{K,L}^{n+1}f(u)_{K|L}^{n+1})^2.$$

Let us assume the following technical results.

Lemma 4.1. *Let $B:\mathbb{N}\to\mathbb{R}^+$ such that $\sum_{n=0}^{N}\delta tB(n)\le C$; then for all $s\in\mathbb{R}^+$, $\int_0^{T-s}\sum_{n=n(t)}^{n(t+s)-1}\delta tB(n)dt\le C\,s$.*

Lemma 4.2. *Let $B:\mathbb{N}\to\mathbb{R}^+$ such that $\sum_{n=0}^{N}\delta tB(n)\le C$; then for all $s\in\mathbb{R}^+$, $\int_0^{T-s}\sum_{n=n(t)}^{n(t+s)-1}\delta tB(n(t+\rho))dt\le C\,s$.*

According to estimates (21) and (22), S and R satisfy the assumptions of the two lemmas, so we obtain

$$\|A_1(\rho,t)\|\le C',$$
$$\|A_2(\rho,t)\|\le C'',$$

and collecting the previous inequalities completes the proof of Proposition 4.2. Yet, it remains only to show the two lemmas.

For Lemma 4.1, $\int_0^{T-s}\sum_{n=n(t)}^{n(t+s)-1}\delta tB(n)=\sum_{n=0}^{N}\delta tB(n)\int_0^{T-s}\mathbb{1}_{n\in[\![n(t),n(t+s)-1]\!]}dt$ and $n\in[\![n(t),n(t+s)-1]\!]$ if and only if $t\in[(n+1)\delta t-s,(n+1)\delta t)$, so $\int_0^{T-s}\mathbb{1}_{n\in[\![n(t),n(t+s)-1]\!]}dt\le s$.

For Lemma 4.2, $\int_0^{T-s}\sum_{n=n(t)}^{n(t+s)-1}\delta tB(n(t+\rho))dt=\sum_{n=0}^{N}\delta tB(n)\int_0^{T-s}(n(t+s)-n(t))\,\mathbb{1}_{n(t+\rho)=n}dt$, but $n(t+s)-n(t)$ is periodic with period δt, so $\int_0^{T-s}(n(t+s)-n(t))\,\mathbb{1}_{n(t+\rho)=n}dt\le\int_{n\delta t-\rho}^{(n+1)\delta t-\rho}(n(t+s)-n(t))dt\le\int_0^{\delta t}n(t+s)dt\le s$. □

To prove compactness of $\varphi(u_{\mathcal{T},\delta t})$ in $L^2(\Omega)$, let us check the hypothesis of Theorem 4.1. From Propositions 4.1 and 4.2 we easily deduce that if we extend $u_{\mathcal{T},\delta t}$ by zero outside of $\Omega \times (0,T)$, then for every $\xi \in \mathbb{R}^d$ and $s \in \mathbb{R}^+$ one has

$$\|\varphi(u_{\mathcal{T},\delta t}(\cdot+\xi,\cdot+s)) - \varphi(u_{\mathcal{T},\delta t}(\cdot,\cdot))\|_{L^2(\mathbb{R}^{m+1})} \leq 2C|\xi|(|\xi|+2h) + 2C's \\ + (4T|\xi|m(\partial\Omega) + 2m(\Omega)s)(\varphi^*)^2.$$

Let $u_m = u_{\mathcal{T}_m,\delta t_m}$ be a sequence of approximate solutions with $size(\mathcal{T}_m) \to 0$ when m tends to infinity, and suppose that $\varphi(u_m)$ tends to $\bar{\varphi}$ in $L^2(\Omega \times (0,T))$. It remains to show that $\bar{\varphi}$ is in $L^2(0,T,H^1(\Omega))$. In order to prove it, we shall use space translate estimates in the interior of Ω.

Let $\omega \subset\subset \Omega$. From Proposition 4.1, if $(\xi, z) \in \mathbb{R}^d \times \mathbb{R}$ satisfies $|\xi| \leq d(\omega, \mathbb{R}^d - \Omega)$ and $-1 \leq z \leq 1$,

$$\left\| \frac{\varphi(u_m(\cdot+z\xi,\cdot)) - \varphi(u_m(\cdot,\cdot))}{z} \right\|_{L^2(\omega\times(0,T))} \leq |\xi|\sqrt{C\Phi} + \sqrt{2h_m\frac{|\xi|}{|z|}}.$$

So by letting m tend to infinity, it holds that

$$\left\| \frac{\bar{\varphi}(\cdot+z\xi,\cdot) - \bar{\varphi}(\cdot,\cdot)}{z} \right\|_{L^2(\omega\times(0,T))} \leq |\xi|\sqrt{C\Phi},$$

and by letting z tend to zero we obtain finally

$$\|\nabla\bar{\varphi}\cdot\xi\|_{L^2(\omega\times(0,T))} \leq |\xi|\sqrt{C\Phi}. \tag{25}$$

By homogeneity, inequality (25) is true for every $\xi \in \mathbb{R}^d$, so $\bar{\varphi} \in L^2(0,T,H^1(\Omega))$ and $\|\nabla\bar{\varphi}\|_{L^2((0,T)\times\Omega)} \leq \sqrt{C\Phi}$. This regularity property will be useful, for example, for writing a weak formulation for the problem (5)–(6).

4.2. Compactness of $\boldsymbol{\theta_{\mathcal{T},\delta t}}$. In the same way as for $\varphi(u_{\mathcal{T},\delta t})$, we can show space translate estimates on $\theta_{\mathcal{T},\delta t}$, but since the equation satisfied by θ does not include time derivatives relative to θ, we do not have any time translate estimate. Hence we cannot apply the same method to obtain compactness on $\theta_{\mathcal{T},\delta t}$. However, by Poincaré inequality, $(\theta_{\mathcal{T},\delta t})_{\mathcal{T},\delta t}$ is bounded in $L^\infty(0,T,L^2(\Omega))$. Therefore $\theta_{\mathcal{T},\delta t}$ is sequentially weakly relatively compact in $L^2(\Omega\times(0,T))$, and by the same arguments as those used in the previous section, every possible limit when $size(\mathcal{T})$ tends to zero belongs to $L^\infty(0,T,H^1(\Omega))$. This is sufficient for convergence under the hypothesis that φ' is strictly nondecreasing, since we get in that case the strong convergence of $u_{\mathcal{T},\delta t}$ directly from the strong convergence of $\varphi(u_{\mathcal{T},\delta t})$.

5. Convergence of the scheme.

DEFINITION 5.1 (weak solution). *(u,θ) is a weak solution of problem* (5)–(6) *if $u \in L^\infty(\Omega\times(0,T))$, $0 \leq u(x,t) \leq 1$ a.e $(x,t) \in \Omega\times(0,T)$, $\varphi(u) \in L^2(0,T,H^1(\Omega))$, $\theta \in L^\infty(0,T,H^1(\Omega))$, and for any $\psi \in (\mathbb{R}^d\times[0,T))^2$ there hold*

$$\int_0^T\int_\Omega u\psi_t - \int_0^T\int_\Omega \nabla\varphi(u)\cdot\nabla\psi - \int_0^T\int_\Omega f(u)M(u)\nabla\theta\cdot\nabla\psi + \int_\Omega u_0\psi(\cdot,0) \\ = -\int_0^T\int_\Omega c\bar{s}\psi + \int_0^T\int_\Omega u\underline{s}\psi, \tag{26}$$

$$\int_0^T\int_\Omega M(u)\nabla\theta\cdot\nabla\psi = \int_0^T\int_\Omega(\bar{s}-\underline{s})\psi. \tag{27}$$

We shall now give and prove the main result of this paper.

Theorem 5.1 (the convergence theorem). *Let* $(u_m,\theta_m)=(u_{\mathcal{T}_m,\delta t_m},\theta_{\mathcal{T}_m,\delta t_m})$ *be a sequence of approximate solutions given by scheme* (14)–(18) . *Let us assume that there exists* $\xi>0$ *such that for every* $m\in\mathbb{N}$, $\mathcal{T}_m$ *is a* ξ*-regular admissible mesh. Assume also that* $size(\mathcal{T}_m)\to 0$ *and* $\delta t_m\to 0$ *when* m *tends to* ∞. *Then there exists a weak solution* (u,θ) *of problem* (5)–(6) *such that up to a subsequence,*

$$u_m\to u,\ \textit{strongly in } L^2(\Omega\times(0,T))\ \textit{as } m\to\infty,$$
$$\theta_m\to\theta,\ \textit{weakly in } L^2(\Omega\times(0,T))\ \textit{as } m\to\infty.$$

Remark 5.1. *Under the assumption that* $f(\varphi^{-1})$ *is Holder continuous with exponent* $\frac{1}{2}$ *and with the additional hypothesis* $\|\nabla\theta\|\in L^\infty(\Omega\times(0,T))$, *we can prove that the weak solution is unique (see* [9]*). So the whole sequence of approximate solutions is convergent.*

Proof. From compactness properties of u_m and θ_m, we already know that up to a subsequence $\varphi(u_m)\to\bar\varphi$ strongly in $L^2(\Omega\times(0,T))$, $\theta_m\to p$ weakly in $L^2(\Omega\times(0,T))$ as $m\to\infty$. Since $size(\mathcal{T}_m)\to 0$, $\bar\varphi\in L^2(0,T,H^1(\Omega))$ and $\theta\in L^\infty(0,T,H^1(\Omega))$. Now since φ is strictly nondecreasing, by using, for example, the dominated convergence theorem we deduce that u_m tends to $u=\varphi^{-1}(\bar\varphi)$ strongly in $L^2(\Omega\times(0,T))$. It remains to show that u is a weak solution of problem (5)–(6).

Let $(\mathcal{T},\delta t)=(\mathcal{T}_m,\delta t_m)$. We define the discretization and approximate of ψ denoted, respectively, Ψ and $\psi_{\mathcal{T},\delta t}$ by the following formulas:

$$\Psi_K^{n+1}=\psi(x_K,n\delta t),K\in\mathcal{T},n\in[\![0,N]\!], \tag{28}$$
$$\psi_{\mathcal{T}}(x,t)=\Psi_K^{n+1},x\in K,t\in(n\delta t,(n+1)\delta t). \tag{29}$$

In order to prove that (u,θ) is a weak solution, we multiply (16) and (17) by $\delta t\Psi_K^{n+1}$ and sum over $n\in[\![0,N]\!]$ and $K\in\mathcal{T}$. Then we let m tend to infinity and show that we obtain (26) and (27) when passing to the limit.

As in [19], [28], [6], by using the consistency of fluxes, we obtain at the limit the following terms:

$$\int_0^T\int_\Omega u\psi_t+\int_\Omega u_0,\int_0^T\int_\Omega\nabla\varphi(u)\cdot\nabla\psi,\int_0^T\int_\Omega c\overline{s}\psi,\int_0^T\int_\Omega u\underline{s}\psi,\ \text{and}\int_0^T\int_\Omega(\overline{s}-\underline{s})\psi.$$

But we encounter original difficulties to obtain the two remaining terms, namely

$$\int_0^T\int_\Omega f(u)M(u)\nabla\theta\ \text{and}\int_0^T\int_\Omega M(u)\nabla\theta.$$

Indeed, heuristically, we have to prove the weak convergence of $f(u_m)M(u_m)\nabla\theta_m$ and $M(u_m)\nabla\theta_m$ to $f(u)M(u)\nabla\theta$ and $M(u)\nabla\theta$, with u_m strongly converging to u and θ_m bounded in $L^2(0,T,H^1(\Omega))$ and weakly converging to θ. In the continuous case, this problem can be solved since a product of two functions, the first converging strongly and the other weakly, is weakly converging to the product of the limits. However, the gradient of discrete function is not in general a function, so we need to use a regularization argument that we shall detail.

Remark 5.2. *Our heuristic argument justifies why the strong convergence of* u_m *is crucial. Indeed, a product of weak convergent sequences of functions in general does not converge to the product of the limit, even if it has a weak limit. In our case we obtain this strong convergence by using the hypothesis that* φ' *is a strictly increasing* *function, but our method would also work in all cases where we were able to prove that $f(u_m)$ and $M(u_m)$ converge strongly.*

We will restrict ourselves to the case when $A = \int_\Omega \int_0^T f(u)M(u)\nabla\theta \cdot \nabla\psi$ because the other integral is a special case with $f = 1$. Let us first define $A_{\mathcal{T},\delta t}$ by

$$A_{\mathcal{T},\delta t} = \sum_{n=0}^{N} \delta t \frac{1}{2} \sum_{K\in\mathcal{T}} \sum_{L\in\mathcal{N}(K)} f(u)_{K|L}^{n+1} M(u)_{K|L}^{n+1} (\Theta_L^{n+1} - \Theta_K^{n+1})(\Psi_L^{n+1} - \Psi_K^{n+1}).$$

$A_{\mathcal{T},\delta t}$ is the term corresponding to A when we multiply (16) by $\delta t \Psi_K^{n+1}$, sum over $K \in \mathcal{T}$ and $n \in [\![0, N]\!]$, and gather by edges. Assume first that $f(u)$ and $M(u)$ are in $\mathcal{D}(\Omega \times (0,T))$, and $M(U)$, $m(u_{\mathcal{T},\delta t})$ and $f(U)$, $f(u_{\mathcal{T},\delta t})$ are discretizations and approximations of $f(u)$ and $M(u)$ defined in the same way as Ψ and $\psi_{\mathcal{T},\delta t}$ by (28)–(29). By the Stokes formula,

$$A = -\int_0^T \int_\Omega \theta \ \mathrm{div}(f(u)M(u)\nabla\psi).$$

So because of the weak convergence of θ_m to θ, $A = \lim_{m\to\infty} B_{\mathcal{T}_m,\delta t_m}$, where $B_{\mathcal{T},\delta t}$ is given by

$$B_{\mathcal{T},\delta t} = \int_0^T \int_\Omega \theta_{\mathcal{T},\delta t} \mathrm{div}(f(u)M(u)\nabla\psi),$$

and by definition of the piecewise constant function $\theta_{\mathcal{T},\delta t}$ we get

$$\begin{aligned} B_{\mathcal{T},\delta t} &= -\sum_{n=0}^{N} \sum_{K\in\mathcal{T}} \sum_{L\in\mathcal{N}(K)} \Theta_K^{n+1} \int_{n\delta t}^{(n+1)\delta t} \int_{K|L} f(u)M(u)\nabla\psi \cdot \mathbf{n}_{K,L} \\ &= \sum_{n=0}^{N} \frac{1}{2} \sum_{K\in\mathcal{T}} \sum_{L\in\mathcal{N}(K)} (\Theta_L^{n+1} - \Theta_K^{n+1}) \int_{n\delta t}^{(n+1)\delta t} \int_{K|L} f(u)M(u)\nabla\psi \cdot \mathbf{n}_{K,L}. \end{aligned}$$

Now, let us compare $B_{\mathcal{T},\delta t}$ and $A_{\mathcal{T},\delta t}$, using the consistency of the flux on the interfaces $K|L$ for regular functions. By the Cauchy–Schwarz inequality, we have

$$\begin{aligned} &|A_{\mathcal{T},\delta t} - B_{\mathcal{T},\delta t}|^2 \\ &\leq \sum_{n=0}^{N} \delta t \frac{1}{2} \sum_{K\in\mathcal{T}} \sum_{L\in\mathcal{N}(K)} T_{K|L} (\Theta_K^{n+1} - \Theta_L^{n+1})^2 \sum_{n=0}^{N} \delta t \frac{1}{2} \sum_{K\in\mathcal{T}} \sum_{L\in\mathcal{N}(K)} \frac{1}{T_{K|L}} R_{K|L}^2, \end{aligned}$$

where

$$R_{K|L} = |f(u)_{K|L}^{n+1} M(u)_{K|L}^{n+1} T_{K|L} (\Psi_K^{n+1} - \Psi_L^{n+1}) - \frac{1}{\delta t} \int_{n\delta t}^{(n+1)\delta t} \int_{K|L} f(u)M(u)\nabla\psi \cdot \mathbf{n}_{K,L}|.$$

By using the regularity of $f(u)$, $M(u)$, and ψ and the orthogonality property $x_K - x_L = d_{K|L}\mathbf{n}_{K,L}$, we easily get the existence of $C_3 > 0$ depending only on $f(u)$, $M(u)$, and ψ such that $R_{K|L} \leq C_3 d_{K|L} m(\sigma)$. Then using estimate (22) we obtain $\lim_{m\to\infty} A_{\mathcal{T}_m,\delta t_m} - B_{\mathcal{T}_m,\delta t_m} = 0$ and $\lim_{m\to\infty} A_{\mathcal{T}_m,\delta t_m} = A$, in the particular case considered here. To extend the result to the general case by density, it suffices to remark the continuity of A with respect to $M(u), f(u)$ and the uniform continuity of

$A_{\mathcal{T},\delta t}$ for the $L^2(\Omega\times(0,T))$ norm. The continuity of A is clear. To show uniform continuity of $A_{\mathcal{T},\delta t}$, we use first the Cauchy–Schwarz inequality to get the following inequalities:

$$\|A_{\mathcal{T},\delta t}\|^2 \le \|\nabla\psi\|^2_{L^2(0,T,L^\infty(\Omega))} C_2 \sum \delta t \frac{1}{2} \sum_{K\in\mathcal{T}} \sum_{L\in\mathcal{N}(K)} d_{K|L} m(K|L)(M(u)^{n+1}_{K|L})^2,$$

$$\|A_{\mathcal{T},\delta t}\|^2 \le \|\nabla\psi\|^2_{L^2(0,T,L^\infty(\Omega))} (M^*)^2 C_2 \sum \delta t \frac{1}{2} \sum_{K\in\mathcal{T}} \sum_{L\in\mathcal{N}(K)} d_{K|L} m(K|L)(f(u)^{n+1}_{K|L})^2.$$

Then, since $M(u)^{n+1}_{K|L}$ belongs to the interval $[M(U^{n+1}_K), M(U^{n+1}_L)]$ and $f(u)^{n+1}_{K|L} \in [f(U^{n+1}_K), f(U^{n+1}_L)]$, we have

$$\sum_{K\in\mathcal{T}} \sum_{L\in\mathcal{N}(K)} (M(u)^{n+1}_{K|L})^2 d_{K|L} m(K|L) \le 2 \sum_{K\in\mathcal{T}} M(U^{n+1}_K)^2 \left(\sum_{L\in\mathcal{N}(K)} d_{K|L} m(K|L) \right),$$

$$\sum_{K\in\mathcal{T}} \sum_{L\in\mathcal{N}(K)} (f(u)^{n+1}_{K|L})^2 d_{K|L} m(K|L) \le 2 \sum_{K\in\mathcal{T}} f(U^{n+1}_K)^2 \left(\sum_{L\in\mathcal{N}(K)} d_{K|L} m(K|L) \right).$$

By using the uniform ξ-regularity of meshes, we finally get

$$\|A_{\mathcal{T},\delta t}\| \le \sqrt{C_2\xi}\|\psi\|_{L^2(0,T,L^\infty(\Omega))}\|M(u_{\mathcal{T},\delta t})\|,$$
$$\|A_{\mathcal{T},\delta t}\| \le \sqrt{C_2\xi}\|\psi\|_{L^2(0,T,L^\infty(\Omega))} M^* \|f(u_{\mathcal{T},\delta t})\|,$$

and we use the bilinearity of $A_{\mathcal{T},\delta t}$ to conclude. This completes the proof of Theorem 5.1. □

6. Numerical results. As an example of application, we perform numerical experiments with the following data, which are realistic in the study of oil and water flow in a homogeneous porous media:

$$k_1(x) = \frac{x^3}{2}, \quad k_2(x) = \frac{(1-x)^3}{3},$$
$$p_c(x) = -0.5\sqrt{\frac{1-x}{x}}.$$

As an initial condition we take uniformly the value $u_0 = 0.5$ and prescribe $c = 0.8$. We represent in Figure 1 the behavior of k_1, k_2, M, f, and p_c. We can verify that the hypotheses that we made on the data in section 2 are satisfied.

The domain of study is the open subset $\Omega = (0,1)^2$ of $\mathbb{R}^2$. If we denote $D_1 = \{(x,y)\in\mathbb{R}^2, (x-0.5)^2+(y-0.8)^2 \le 0.01\}$, $D_2 = \{(x,y)\in\mathbb{R}^2, (x-0.2)^2+(y-0.2)^2 \le 0.01\}$, $D_3 = \{(x,y)\in\mathbb{R}^2, (x-0.8)^2+(y-0.5)^2 \le 0.01\}$, we can take sources and sinks terms as follows:

$$\overline{s}(x,y) = 10\,\mathbb{1}_{D_1}(x,y) + 20\,\mathbb{1}_{D_2}(x,y),$$
$$\underline{s}(x,y) = 30\,\mathbb{1}_{D_3}(x,y).$$

Figures 2 and 3 represent the numerical results at time $t = 2.1$ and $t = 9.1$. As we could expect, the saturation in the reservoir increases in mean because $u_0 \le c$ and

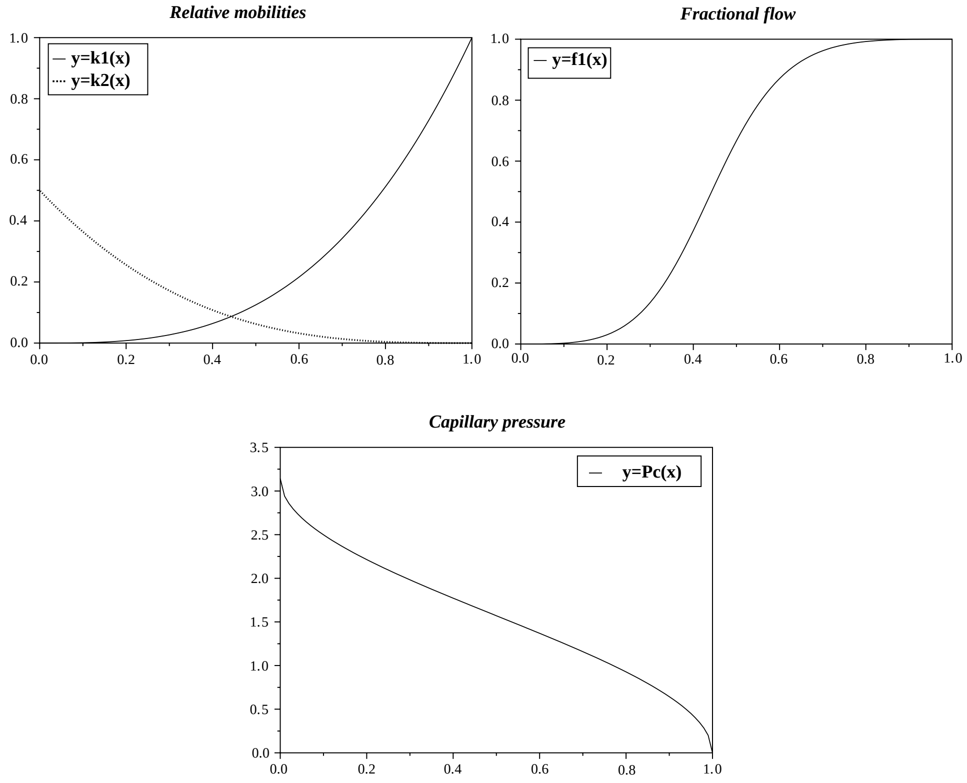


Fig. 1. *Behavior of the functions* k_1, k_2, f_1, p_c.

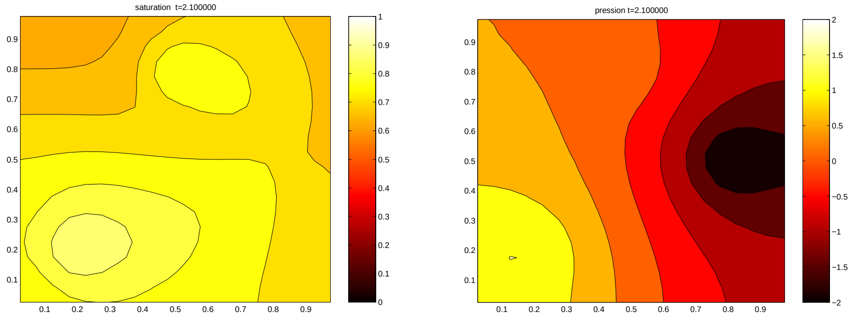


Fig. 2. *Saturation and pressure at time* $t = 2.1$.

the gradient of the pressure is oriented from the sources to the sink. The pressure is nearly stationary, which is not surprising since sources and sinks are stationary and the variations of the diffusion coefficient $M(u)$ is not very large in the interval $[0, 1]$. The flow is more important at the beginning, when the difference between the injection saturation and the mean saturation in the reservoir is the largest.

7. Concluding remarks. We have proved the convergence of an implicit fully finite volume method using the elliptic parabolic structure of the problem after a transformation due to Chavent (see [7]). We have made two-dimensional computations

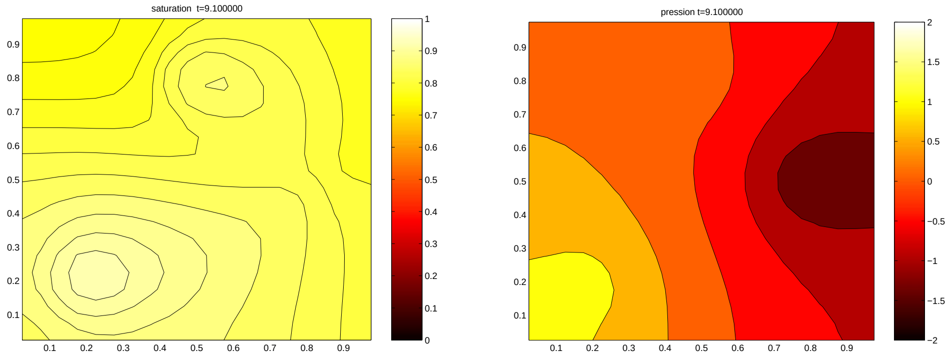


FIG. 3. *Saturation and pressure at time $t = 9.1$.*

that showed that it works well. It should be interesting to compare our method with other methods in terms of accuracy and performance. In particular, we aim to compare it to the phase by phase upwind scheme (see [20]) which consists of directly discretizing the two mass conservation equations.

Acknowledgment. I would like to thank M. Olhberger for helpful discussions.

REFERENCES

[1] T. ARBOGAST, *The existence of weak solutions to single porosity and simple dual-porosity models of two-phase incompressible flow*, Nonlinear Anal., 19 (1992), pp. 1009–1031.

[2] T. ARBOGAST, M. F. WHEELER, AND N.-Y. ZHANG, *A nonlinear mixed finite element method for a degenerate parabolic equation arising in flow in porous media*, SIAM J. Numer. Anal., 33 (1996), pp. 1669–1687.

[3] J. BEAR, *Dynamic of Flow in Porous Media*, Dover, New York, 1967.

[4] J. BEAR, *Modeling transport phenomena in porous media*, in Environmental Studies (Minneapolis, MN, 1992), Springer-Verlag, New York, 1996, pp. 27–63.

[5] F. BREZZI AND M. FORTIN, *Mixed and Hybrid Finite Element Methods*, Springer Ser. Comput. Math. 15, Springer-Verlag, New York, 1991.

[6] C. CHAINAIS-HILLAIRET, *Finite volume schemes for a nonlinear hyperbolic equation. Convergence towards the entropy solution and error estimate*, M2AN Math. Model. Numer. Anal., 33 (1999), pp. 129–156.

[7] G. CHAVENT AND J. JAFFRÉ, *Mathematical Models and Finite Elements for Reservoir Simulation*, Elsevier, Amsterdam, 1986.

[8] Z. CHEN, *Degenerate two-phase flow incompressible flow* 1: *Existence, uniqueness and regularity of a weak solution*, J. Differential Equations, 171 (2001), pp. 203–232.

[9] Z. CHEN AND R. EWING, *Mathematical analysis for reservoir models*, SIAM J. Math. Anal., 30 (1999), pp. 431–453.

[10] Y. COUDIÈRE, J. VILA, AND P. VILLEDIEU, *Convergence rate of a finite volume scheme for a two-dimensional convection-diffusion problem*, M2AN Math. Model. Numer. Anal., 33 (1999), pp. 493–516.

[11] K. DEIMLING, *Nonlinear Functional Analysis*, Springer-Verlag, Berlin, 1985.

[12] J. DOUGLAS, JR., *Finite difference methods for two-phase incompressible flow in porous media*, SIAM J. Numer. Anal., 20 (1983), pp. 681–696.

[13] J. DOUGLAS, JR., R. E. EWING, AND M. F. WHEELER, *The approximation of the pressure by a mixed method in the simulation of miscible displacement*, RAIRO Anal. Numér., 17 (1983), pp. 17–33.

[14] L. J. DURLOFSKY, *Coarse scale models of two phase flow in heterogeneous reservoirs: Volume averaged equations and their relationship to existing upscaling techniques*, Comput. Geosci., 2 (1998), pp. 73–92.

[15] R. E. EWING AND M. F. WHEELER, *Galerkin methods for miscible displacement problems with point sources and sinks—unit mobility ratio case*, in Mathematical Methods in Energy

Research (Laramie, WY, 1982/1983), SIAM, Philadelphia, 1984, pp. 40–58.

[16] R. Eymard and T. Gallouët, *Convergence d'un schéma de type éléments finis–volumes finis pour un système formé d'une équation elliptique et d'une équation hyperbolique*, RAIRO Modél. Math. Anal. Numér., 27 (1993), pp. 843–861.

[17] R. Eymard, T. Gallouët, and R. Herbin, *Convergence of finite volume schemes for semilinear convection diffusion equations*, Numer. Math., 82 (1999), pp. 91–116.

[18] R. Eymard, T. Gallouët, and R. Herbin, *The finite volume method*, in Handbook for Numerical Analysis, Ph. Ciarlet and J. L. Lions, eds., North-Holland, Paris, 2000, pp. 715–1022.

[19] R. Eymard, T. Gallouët, D. Hilhorst, and Y. Naït Slimane, *Finite volumes and nonlinear diffusion equations*, RAIRO Modél. Math. Anal. Numér., 32 (1998), pp. 747–761.

[20] R. Eymard, R. Herbin, and A. Michel, *Mathematical study of a petroleum-engineering scheme*, M2AN Math. Model. Numer. Anal., submitted.

[21] I. Faille, *A control volume method to solve an elliptic equation on a two-dimensional irregular meshing*, Comput. Methods Appl. Mech. Engrg., 100 (1992), pp. 275–290.

[22] R. Herbin, *Finite volume methods for diffusion convection equations on general meshes*, in Proceedings of the 1st International Symposium on Finite Volumes for Complex Applications, Problems and Perspectives (Rouen, 1996), F. Benkhaldoun and R. Vilsmeier, eds., Hermes, Paris, 1996, pp. 153–160.

[23] D. Kroener and S. Luckhaus, *Flow of oil and water in a porous medium*, J. Differential Equations, 55 (1984), pp. 276–288.

[24] S. N. Kruzkov and S. M. Sukorjanskii, *Boundary value problems for systems of equations of two-phase filtration type; formulation of problems, questions of solvability, justification of approximate methods*, Mat. Sb. (N.S.), 104(146) (1977), pp. 69–88, 175–176.

[25] M. Ohlberger, *Convergence of a mixed finite elements–finite volume method for the two phase flow in porous media*, East-West J. Numer. Math., 5 (1997), pp. 183–210.

[26] D. W. Peaceman, *Fundamentals of Numerical Reservoir Simulation*, Elsevier Scientific, Amsterdam, 1977.

[27] P.-A. Raviart, *Mixed finite element methods*, in The Mathematical Basis of Finite Element Methods (London, 1983), Inst. Math. Appl. Conf. Ser. New Ser. 2, Oxford University Press, New York, 1984, pp. 123–156.

[28] M. H. Vignal, *Convergence of a finite volume scheme for an elliptic-hyperbolic system*, RAIRO Modél. Math. Anal. Numér., 30 (1996), pp. 841–872.

SIAM J. NUMER. ANAL.
Vol. 41, No. 4, pp. 1318–1338

AN INTERIOR ESTIMATE OF SUPERCONVERGENCE FOR FINITE ELEMENT SOLUTIONS FOR SECOND-ORDER ELLIPTIC PROBLEMS ON QUASI-UNIFORM MESHES BY LOCAL PROJECTIONS[*]

HONGSEN CHEN[†] AND JUNPING WANG[‡]

Abstract. This paper establishes some superconvergence estimates for finite element solutions of second-order elliptic problems by a projection method depending only on local properties of the domain and the finite element solution. The projection method is a postprocessing procedure that constructs a new approximation by using the method of least squares. In particular, some local superconvergence estimates in the L^2 and L^∞ norms are derived for the local projections of the Galerkin finite element solution. The results have two prominent features. First, they are established for any quasi-uniform meshes, which are of practical importance in scientific computation. Second, they are derived on the basis of local properties of the domain and the solution for the second-order elliptic problem. Therefore, the result of this paper can be employed to provide useful a posteriori error estimators in practical computing.

Key words. finite element method, superconvergence, local estimate, projection method, second-order elliptic equation

AMS subject classifications. Primary, 65N30, 65N15, 65N12, 76S05; Secondary, 41A25, 35B45, 35J20

DOI. 10.1137/S0036142902410039

1. Introduction.

In this paper, we are concerned with local estimates of superconvergence for the Galerkin finite element solution of second-order elliptic equations. The Galerkin finite element method is known to provide numerical solutions for partial differential equations with superconvergence upon the use of appropriately defined postprocessing procedures which are computationally feasible. Because the superconvergence property of the finite element solution can be used to construct a new approximate solution with a higher order of accuracy, it naturally provides a posteriori error estimators in the quality assessment of the finite element approximations in scientific computing.

The research on superconvergence phenomena has been actively conducted by many numerical analysts for over 30 years. Among a large number of literatures, we mention Douglas and Dupont [5], Bramble and Schatz [2], Zlamal [38], Krizek and Neittaanmaki [10], Wahlbin [29], Zienkiewicz and Zhu [37], Ewing, Lazarov, and Wang [7], Zhu and Lin [35], Wheeler and Whiteman [32], Schatz, Sloan, and Wahlbin [24], Zhang [33], Zhu [34], and the references therein. For some technical reasons in the theoretical analysis, all the above-mentioned results on superconvergence require that the underlying finite element meshes be uniform or almost uniform or symmetric about a point. Recently, Wang [30] obtained a superconvergence for general quasi-uniform meshes by using the L^2 projection in the solution postprocessing. Although the results in [30] no longer assume any mesh uniformity or symmetry, they still rely strongly on

[*]Received by the editors June 19, 2002; accepted for publication (in revised form) January 22, 2003; published electronically August 22, 2003. The research was completed while the second author was visiting NASA Center for Computational Sciences (NCCS) at GSFC through the NASA-ASEE Summer Faculty Fellowship Program.

http://www.siam.org/journals/sinum/41-4/41003.html

[†]Department of Mathematics, University of Wyoming, Laramie, WY 82070 (hchen@uwyo.edu).

[‡]Department of Mathematical and Computer Sciences, Colorado School of Mines, Golden, CO 80401 (jwang@mines.edu).

the smoothness of the exact solution and a certain a priori regularity of the underlying problem globally over the whole domain. In general, the required a priori regularity holds true for problems with sufficiently smooth data and domains. Consequently, the superconvergence results developed in [30] have a theoretical limitation in practical applications.

Our objective in this paper is to derive some local superconvergence error estimates for the projection method in which the L^2 projection is defined locally on subdomains. The new results require the exact solution to be only locally smooth, and no global a priori regularity of the problem is assumed. However, the local superconvergence estimate contains a pollution term which is the estimate of the error measured in some negative Sobolev norms. The superconvergence error estimates of this paper are derived in both the L^2 and the maximum norms.

A brief outline of this paper follows. In section 2 we review the Galerkin finite element method for a model second-order elliptic problem. In section 3 we derive a global superconvergence in the L^2 norm for the projection method. Section 4 contains some global estimate in the L^∞ norm. In sections 5 and 6, we establish some local superconvergence error estimates in both the L^2 and the L^∞ norms. A case discussion is made in section 7.

2. Preliminaries. Let Ω be an open bounded domain in R^2. Consider the second-order elliptic boundary value problem that seeks an unknown function $u = u(x)$ satisfying

$$\mathcal{L}u \equiv -\sum_{i,j=1}^{2} \frac{\partial}{\partial x_j}\left(a_{ij}\frac{\partial u}{\partial x_i}\right) + \sum_{i=1}^{2} b_i \frac{\partial u}{\partial x_i} + cu = f \qquad \text{in } \Omega \tag{2.1}$$

and the homogeneous Neumann boundary condition:

$$\mathcal{B} \equiv \sum_{i,j=0}^{2} a_{ij}\frac{\partial u}{\partial x_i} n_j = 0 \qquad \text{on } \partial\Omega,$$

where the coefficients a_{ij}, b_i, and c are given smooth functions and f is a prescribed function; (n_1, n_2) denotes the unit outward normal vector on $\partial\Omega$.

We use the standard notation for Sobolev spaces and norms (see, e.g., Adams [1]). For nonnegative integer k and real number $p \in [1, \infty]$ and subdomain $D \subset \Omega$, denote

$$W^{k,p}(D) = \left\{v : \ \|v\|_{W^{k,p}(D)} < \infty\right\}$$

with

$$\|v\|_{W^{k,p}(D)} = \left(\sum_{|\alpha|\le k} \int_D \left|\frac{\partial^\alpha v(x)}{\partial x^\alpha}\right|^p dx\right)^{1/p} \qquad \text{if } p < \infty,$$

$$\|v\|_{W^{k,\infty}(D)} = \max_{|\alpha|\le k} \sup_{x\in D} \left|\frac{\partial^\alpha v(x)}{\partial x^\alpha}\right| \qquad \text{if } p = \infty.$$

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

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

Denote $(\cdot,\cdot)_D$ the standard inner product in $L^2(D)$ given by

$$(u,v)_D = \int_D uvdx.$$

When $D = \Omega$, we write $(\cdot,\cdot) = (\cdot,\cdot)_\Omega$. The negative Sobolev norm $\|\cdot\|_{H^{-k}(D)}$ is defined as follows:

$$\|v\|_{H^{-k}(D)} = \sup_{\varphi \in C_0^\infty(D)} \frac{(v,\varphi)}{\|\varphi\|_{H^k(D)}},$$

where $\langle v,\varphi\rangle$ is the value of the linear functional v at φ.

To describe a weak formulation for the problem (2.1), we introduce a bilinear form $A_D(\cdot,\cdot)$ for any $D \subset \Omega$ as follows:

$$A_D(u,v) = \int_D \left(\sum_{i,j=1}^2 a_{ij}(x) \frac{\partial u}{\partial x_i}\frac{\partial v}{\partial x_j} + \sum_{i=1}^2 b_i(x)\frac{\partial u}{\partial x_i} v + c(x)uv \right) dx,$$

and we write $A(u,v) = A_\Omega(u,v)$. The weak formulation of the problem (2.1) now seeks $u \in H^1(\Omega)$ such that

$$A(u,v) = (f,v) \quad \forall\, v \in H^1(\Omega). \tag{2.2}$$

Assume that $\mathcal{L}$ is elliptic: there is a constant $\lambda_0 > 0$ such that

$$\sum_{i,j=1}^2 a_{ij}(x)\xi_i\xi_j \geq \lambda_0 \sum_{i=1}^2 \xi_i^2 \quad \forall\, \xi = (\xi_1,\xi_2) \in R^2.$$

Furthermore, assume that the problem (2.2) has a unique solution in $H^1(\Omega)$.

For simplicity of local analysis, we assume that A is $H^1(\Omega)$-coercive: there is a constant $\lambda_1 > 0$ such that

$$A(v,v) \geq \lambda_1 \|v\|^2_{H^1(\Omega)} \quad \forall\, v \in H^1(\Omega).$$

The standard finite element method for numerically solving (2.1) is associated with the weak formulation (2.2) and a finite dimensional subspace $S_r^h \subset H^1(\Omega)$ with two parameters $h \in (0,1)$ and $r \geq 1$. The space S_r^h is associated with a prescribed finite element partition $\mathcal{T}_h$ of the domain Ω and comprises continuous piecewise polynomials of degree no more than r. The finite element approximation, denoted $u_h \in S_r^h$, for u is determined by

$$A(u_h,v) = (f,v) \quad \forall\, v \in S_r^h. \tag{2.3}$$

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

$$A(u-u_h,v) = 0 \quad \forall\, v \in S_r^h, \tag{2.4}$$

which, due to the coercivity of A, implies the optimal order error estimate in H^1 (see Ciarlet [4] or Brenner and Scott [3] for details):

$$\|u-u_h\|_{H^1(\Omega)} \leq C \inf_{v \in V_h} \|u-v\|_{H^1(\Omega)}. \tag{2.5}$$

Here and throughout the paper, C stands for a generic constant independent of the functions and parameters involved.

For any subset $D \subset \Omega$, let $S_r^h(D)$ be the functional space consisting of the restrictions of functions in S_r^h on D. Furthermore, we introduce the notation

$$L^2_<(D) = \{v \in L^2(D) : \ \mathrm{dist}(\mathrm{supp}(v), \partial D \backslash \partial\Omega) > 0\},$$

where $\mathrm{supp}(v)$ denotes the support of v. The finite element space S_r^h is constructed so that the following three assumptions are satisfied:

A.1 (*approximation properties*). Let $D_1 \subset D_2 \subset \Omega$ with $\mathrm{dist}(D_1, \partial D_2 \backslash \partial\Omega) \geq C_0 h$ for some constant $C_0 > 0$ and $0 \leq i \leq 1 \leq j \leq r+1$, $1 \leq p \leq \infty$. Then, for any $v \in W^{j,p}(D_2)$, there exists a $\chi \in S_r^h(D_2)$ such that

$$\|v - \chi\|_{W^{i,p}(D_1)} \leq C h^{j-i} \|v\|_{W^{j,p}(D_2)}.$$

Moreover, if $v \in L^2_<(D_1)$, then $\chi \in L^2_<(D_2)$. The above results also hold true in the case of $D_1 = D_2 = \Omega$.

A.2 (*inverse properties*). Let $D_1 \subset D_2 \subset \Omega$ with $\mathrm{dist}(D_1, \partial D_2 \backslash \partial\Omega) \geq C_0 h$ for some constant $C_0 > 0$. Then, for $0 \leq i \leq j \leq 1$, $1 \leq q \leq p \leq \infty$, and any $v_h \in S_r^h(D_2)$,

$$\|v_h\|_{W^{j,p}(D_1)} \leq C h^{-(j-i)-2(1/q-1/p)} \|v_h\|_{W^{i,q}(D_2)}.$$

Furthermore, if $v_h \in S_r^h(\Omega)$, then

$$\|v_h\|_{W^{j,p}(\Omega)} \leq C h^{-(j-i)-2(1/q-1/p)} \|v_h\|_{W^{i,q}(\Omega)}.$$

A.3 (*superapproximation properties*). Let $D_1 \subset D_2 \subset D_3 \subset D_4 \subset \Omega$ with $\mathrm{dist}(D_1, \partial D_2 \backslash \partial\Omega) \geq C_0 h$, $\mathrm{dist}(D_2, \partial D_3 \backslash \partial\Omega) \geq C_0 h$, $\mathrm{dist}(D_3, \partial D_4 \backslash \partial\Omega) \geq C_0 h$ for some constant $C_0 > 0$. Let $\omega \in L^2_<(D_3) \cap C^\infty(D_3)$ satisfy $\omega \equiv 1$ on D_2. Then, for $j = 0, 1$ and any $v_h \in S_r^h(D_4)$, there is an $\eta \in S_r^h(D_4) \cap L^2_<(D_4)$ such that

$$\|\omega v_h - \eta\|_{L^2(D_3)} \leq C h^{1+j} \|v_h\|_{H^j(D_3 \backslash D_1)}.$$

Let us now briefly review the global and local error estimates for $u - u_h$ in L^2 and L^∞ norms. Under the assumptions A.1, A.2, and A.3, one has the following optimal or suboptimal order error estimates in the L^2 and L^∞ norms for $i = 0, 1$:

$$\|u - u_h\|_{H^i(\Omega)} \leq C h^{r+1-i} \|u\|_{H^{r+1}(\Omega)} \tag{2.6}$$

and

$$\|u - u_h\|_{W^{i,\infty}(\Omega)} \leq C h^{r+1-i} |\ln h|^{\bar{r}} \|u\|_{W^{r+1,\infty}(\Omega)}, \tag{2.7}$$

where $\bar{r}$ is defined by

$$\bar{r} = \begin{cases} 1 & \text{if } r = 1 \text{ and } i = 0, \\ 0 & \text{if } r > 1 \text{ or } i = 1. \end{cases} \tag{2.8}$$

For the estimate (2.6) in L^2, see Ciarlet [4]. For the estimate (2.7) in L^∞, see Natterer [16], Nitsche [17], Scott [28], Rannacher [20], Rannacher and Scott [21], and Schatz and Wahlbin [25, 26]. The error estimates (2.6) and (2.7) require that the solution $u \in$

$H^{r+1}(\Omega)$ and $u \in W^{r+1,\infty}(\Omega)$, in order to achieve the optimal order of convergence in H^i and $W^{i,\infty}$, respectively.

It was shown in Nitsche and Schatz [18] and Schatz and Wahlbin [25] that if $\Omega_0 \subset\subset \Omega_1 \subset \Omega$, then for $i = 0$ or 1 we have the following interior estimates:

$$\|u - u_h\|_{H^i(\Omega_0)} \leq C \left(h^{r+1-i} \|u\|_{H^{r+1}(\Omega_1)} + \|u - u_h\|_{H^{-m}(\Omega)} \right),$$
$$\|u - u_h\|_{W^{i,\infty}(\Omega_0)} \leq C \left(h^{r+1-i} |\ln h|^{\bar{r}} \|u\|_{W^{r+1,\infty}(\Omega_1)} + \|u - u_h\|_{H^{-m}(\Omega)} \right).$$

Here and throughout the paper, $\Omega_0 \subset\subset \Omega_1$ means that $\operatorname{dist}(\Omega_0, \partial\Omega_1) \geq C$ and $\bar{r}$ is defined as above by (2.8). Although these estimates were established for interior regions Ω_0, they can be extended to regions up to the boundary (see Schatz and Wahlbin [26]). We also mention that more localized error estimates have been obtained recently by Schatz [22, 23].

We now turn to the superconvergence estimate. Any estimate that indicates a higher order of convergence than the optimal-order is called a superconvergence. A superconvergence is usually obtained by a postprocessing or recovery technique applied to the original finite element approximation u_h by an appropriate projection or interpolation operator. For example, if u_I is the Lagrange linear interpolation of the exact solution u and if the mesh $\mathcal{T}_h$ is uniform or almost uniform, then we have

$$\|u_I - u_h\|_{H^1(\Omega)} \leq C h^2 \|u\|_{H^3(\Omega)} \tag{2.9}$$

for piecewise linear finite element solutions. The estimate (2.9) leads to a superconvergence estimate of order $\mathcal{O}(h^2)$ for the partial derivatives of u_h by postprocessing ∇u_h via a simple local averaging technique. For details about the definition of uniform or almost uniform meshes and various postprocessing techniques, as well as superconvergence results for other finite elements or other problems, see, for example, Krizek and Neittaanmaki [9, 10], Levine [11], Lin and Wang [12], Lin and Xu [14], Zhu and Lin [35], Wahlbin [29], Wang [31], Douglas and Wang [6], Ewing, Liu, and Wang [8], Li and Zhang [13], Lin and Zhou [15], and the references therein.

To achieve a superconvergence, the exact solution u is often assumed to be more regular than what is needed in the optimal-order error estimate. For example, the superconvergence estimate (2.9) requires that $u \in H^3(\Omega)$, as opposed to $u \in H^2(\Omega)$ in the optimal-order error estimate. Interior or local superconvergence can be found in Wahlbin [29] with additional conditions imposed on the finite element partition $\mathcal{T}_h$. The result shows that one has superconvergence locally in regions where the solution is sufficiently smooth and the finite element partition is either translation invariant or symmetric.

The rest of the paper will investigate the local superconvergence of the Galerkin finite element method by the projection method proposed and studied in [30]. The projection method is essentially an L^2 projection onto a second finite element space based on a high order of polynomials on a coarser grid. This method can be considered as a generalization of the patch recovery technique of Zienkiewicz and Zhu [36, 37] by employing a global patch with smooth functions. More precisely, let $\mathcal{T}_\tau$ be a coarser partition of Ω with $\tau = h^\alpha$ for some $\alpha \in (0, 1)$, and let Q_τ be the L^2 projection operator from $L^2(\Omega)$ onto a finite element space having high order of approximation properties. It was proved in [30] that $u - Q_\tau u_h$ is superconvergent for general quasi-uniform partitions $\mathcal{T}_h$. The superconvergence of [30] requires that both the exact solution be sufficiently smooth and the underlying problem have sufficiently high order of a priori regularity. Although a local L^2 projection was employed to give a superconvergence in [30], the corresponding superconvergence estimates derived in [30] require the same a priori estimates for the global problem just as the global superconvergence does. Our main objective of this paper is to establish some local superconvergence for locally defined projections.

The following regularity on local smooth subdomains shall be assumed in the local superconvergence analysis. Let $D \subset \Omega$ be sufficiently smooth. Then, for any $\varphi \in H^k(D)$, there exists a unique $\Phi \in H^{k+2}(D)$ satisfying

$$\mathcal{L}\Phi = \varphi \quad \text{in } D, \quad \mathcal{B}\Phi = 0 \text{ on } \partial D, \tag{2.10}$$

and

$$\|\Phi\|_{H^{k+2}(D)} \leq C\|\varphi\|_{H^k(D)}, \tag{2.11}$$

where $C > 0$ is a constant independent of φ and Φ. We also assume that the above regularity holds true for the corresponding adjoint operators of $\mathcal{L}$ and $\mathcal{B}$. To be convenient, we say that a subdomain $D \subset \Omega$ is of H^ℓ, $\ell \geq 1$, regularity if (2.11) holds true for any $k \leq \ell - 2$.

We now introduce a second family of finite dimensional subspaces. Let $S^\tau_{m,s} \subset H^m(\Omega)$ be another family of finite dimensional subspaces with $m \geq 0$, $s \geq 0$, $\tau = Ch^\alpha$ for some $\alpha > 0$ to be determined later. The parameter m characterizes the regularity of the fitting space $S^\tau_{m,s}$ and is particularly reserved for this purpose in the rest of the paper. The parameter s indicates the degree of polynomials used in the construction of $S^\tau_{m,s}$. For any $D \subset \Omega$, let $S^\tau_{m,s}(D)$ again be the restriction of functions of $S^\tau_{m,s}$ in D.

Corresponding to A.1, A.2, and A.3, we assume that the following assumptions for spaces $S^\tau_{m,s}$ are satisfied.

B.1 Let $D_1 \subset D_2 \subset \Omega$ with $\text{dist}(D_1, \partial D_2 \backslash \partial\Omega) \geq C_0\tau$ for some constant $C_0 > 0$ and $0 \leq i \leq j \leq s+1$, $1 \leq p \leq \infty$. Then, for any $v \in W^{j,p}(D_2)$, there exists a $\chi \in S^\tau_{m,s}(D_2)$ such that

$$\|v - \chi\|_{W^{i,p}(D_1)} \leq C\tau^{j-i}\|v\|_{W^{j,p}(D_2)}.$$

Moreover, if $v \in L^2_<(D_1)$, then $\chi \in L^2_<(D_2)$.

B.2 (*inverse properties*). Let $D_1 \subset D_2 \subset \Omega$ with $\text{dist}(D_1, \partial D_2 \backslash \partial\Omega) \geq C_0\tau$ for some constant $C_0 > 0$. Then, for $0 \leq i \leq j \leq m$, $1 \leq q \leq p \leq \infty$, and any $v_\tau \in S^\tau_{m,s}(D_2)$,

$$\|v_\tau\|_{W^{j,p}(D_1)} \leq C\tau^{-(j-i)-2(1/q-1/p)}\|v_\tau\|_{W^{i,q}(D_2)}.$$

Furthermore, if $v_\tau \in S^\tau_{m,s}(\Omega)$,

$$\|v_\tau\|_{W^{j,p}(\Omega)} \leq C\tau^{-(j-i)-2(1/q-1/p)}\|v_\tau\|_{W^{i,q}(\Omega)}.$$

B.3 (*superapproximation properties*). Let $D_1 \subset D_2 \subset D_3 \subset D_4 \subset \Omega$ with $\text{dist}(D_1, \partial D_2 \backslash \partial\Omega) \geq C_0\tau$, $\text{dist}(D_2, \partial D_3 \backslash \partial\Omega) \geq C_0\tau$, $\text{dist}(D_3, \partial D_4 \backslash \partial\Omega) \geq C_0\tau$ for some constant $C_0 > 0$. Let $\omega \in L^2_<(D_3) \cap C^\infty(D_3)$ satisfy $\omega \equiv 1$ on D_2. Then, for $0 \leq j \leq \min(m, s+1)$ and any $v_\tau \in S^\tau_{m,s}(D_4)$, there is an $\eta \in S^\tau_{m,s}(D_4) \cap L^2_<(D_4)$ such that

$$\|\omega v_\tau - \eta\|_{L^2(D_4)} \leq C\tau^{1+j}\|v_\tau\|_{H^j(D_4 \backslash D_1)}.$$

Next, we introduce the notation of L^2 projection. For any $D \subset \Omega$, let

$$Q_\tau^D : \; L^2(D) \to S_{m,s}^\tau(D)$$

denote the L^2 projection operator defined by

$$(Q_\tau^D v, \varphi)_D = (v, \varphi)_D \quad \forall\, v \in L^2(D),\ \varphi \in S_{m,s}^\tau(D).$$

When $D = \Omega$, we shall ignore Ω and use the notation $Q_\tau = Q_\tau^\Omega$.

Throughout this paper, we assume that assumptions A.1, A.2, and A.3 and assumptions B.1, B.2, and B.3 are satisfied.

3. A global superconvergence in L^2. In this section we establish a global superconvergence in the L^2 norm for the general second-order elliptic problem (2.1). The following theorem can be considered as a generalization of the results of Wang [30] from Laplacian to general second-order elliptic problems.

THEOREM 3.1. *Assume that u and u_h satisfy* (2.4) *and that Ω is of H^{k+2} regularity, $k \geq 0$. If $u \in H^{1+r}(\Omega)$ when $r > s$ or $u \in H^{1+s}(\Omega)$ when $r \leq s$, then for*

$$\tau = O(h^\alpha), \qquad \text{with } \alpha = \frac{1 + r + \min(r-1, m, k)}{1 + s + \min(r-1, m, k)}, \tag{3.1}$$

we have

$$\|u - Q_\tau u_h\|_{H^i(\Omega)} \leq C h^{\frac{1+r+\min(r-1,m,k)}{1+\theta_i}} \left(\|u\|_{H^{1+r}(\Omega)} + \|u\|_{H^{1+s}(\Omega)} \right),$$

where θ_i is given by

$$\theta_i = \frac{i + \min(m, r-1, k)}{s + 1 - i}. \tag{3.2}$$

Proof. The proof follows the same line as in Wang [30]. For completeness, it is outlined as follows. Since Q_τ is the L^2 projection, we have for any $\varphi \in C_0^\infty(\Omega)$,

$$(Q_\tau(u - u_h), \varphi) = (u - u_h, Q_\tau \varphi). \tag{3.3}$$

Let $\Phi \in H^1(\Omega)$ be the unique solution of

$$A(\psi, \Phi) = (Q_\tau \varphi, \psi) \quad \forall\, \psi \in H^1(\Omega).$$

Using assumption A.1, there is an $\chi \in S_r^h$ such that

$$\|\Phi - \chi\|_{H^1(\Omega)} \leq C h^{i-1} \|\Phi\|_{H^i(\Omega)}, \quad 1 \leq i \leq 1 + r. \tag{3.4}$$

Then it follows from (3.3), Hölder's inequality, and inequality (3.4) that

$$\begin{aligned} (Q_\tau(u - u_h), \varphi) &= A(u - u_h, \Phi) = A(u - u_h, \Phi - \Pi_h \Phi) \\ &\leq C \|u - u_h\|_{H^1(\Omega)} \|\Phi - \Pi_h \Phi\|_{H^1(\Omega)} \\ &\leq C h^{1+\min(m,r-1,k)} \|u - u_h\|_{H^1(\Omega)} \|\Phi\|_{H^{2+\min(r-1,m,k)}(\Omega)}. \end{aligned} \tag{3.5}$$

An application of the a priori H^{k+2} regularity, inverse property B.2, and the stability of L^2 projection yields

$$\begin{aligned} \|\Phi\|_{H^{2+\min(r-1,m,k)}(\Omega)} &\leq C \|Q_\tau \varphi\|_{H^{\min(r-1,m,k)}(\Omega)} \\ &\leq C \tau^{-\min(r-1,m,k)} \|Q_\tau \varphi\|_{L^2(\Omega)} \\ &\leq C \tau^{-\min(r-1,m,k)} \|\varphi\|_{L^2(\Omega)}, \end{aligned}$$

which, combined with (3.5), leads us to

$$\|Q_\tau(u-u_h)\|_{L^2(\Omega)} = \sup_{\varphi\in C_0^\infty(\Omega),\varphi\neq 0} \frac{(Q_\tau(u-u_h),\varphi)}{\|\varphi\|_{L^2(\Omega)}} \tag{3.6}$$
$$\le Ch^{1+\min(r-1,m,k)}\tau^{-\min(r-1,m,k)}\|u-u_h\|_{H^1(\Omega)}.$$

Using the optimal-order error estimate for $\|u-u_h\|_{H^1(\Omega)}$ (Ciarlet [4]),

$$\|u-u_h\|_{H^1(\Omega)} \le Ch^r\|u\|_{H^{1+r}(\Omega)},$$

the best approximation property of the L^2 projection, and the approximation property B.1,

$$\|u-Q_\tau u\|_{L^2(\Omega)} \le C\inf_{\chi\in S_{m,s}^\tau}\|u-\chi\|_{L^2(\Omega)} \le C\tau^{s+1}\|u\|_{H^{1+s}(\Omega)},$$

we conclude that

$$\|u-Q_\tau u_h\|_{L^2(\Omega)} \le \|u-Q_\tau u\|_{L^2(\Omega)} + \|Q_\tau(u-u_h)\|_{L^2(\Omega)} \tag{3.7}$$
$$\le C\tau^{s+1}\|u\|_{H^{1+s}(\Omega)} + Ch^{1+r+\min(r-1,m,k)}\tau^{-\min(r-1,m,k)}\|u\|_{H^{1+r}(\Omega)}.$$

Now, choose τ so that

$$\tau^{1+s} = \mathcal{O}(h^{1+r+\min(r-1,m,k)}\tau^{-\min(r-1,m,k)});$$

then τ satisfies (3.1), and the desired estimate follows from (3.7). This completes the proof. □

4. A global superconvergence in L^∞.

This section is devoted to the derivation of a global superconvergence estimate in the L^∞ norm. A traditional and standard approach for pointwise error estimates of finite element methods will be employed here for superconvergence. This approach is based on special weighted Sobolev norms. The main result of this section is stated as follows.

THEOREM 4.1. *Assume that u and u_h satisfy (2.4) and that the domain Ω is of H^{k+2} regularity with $k \ge 0$. If τ satisfies (3.1) and $u \in W^{1+r,\infty}(\Omega)$ when $r > s$ or $u \in W^{1+s,\infty}(\Omega)$ when $r \le s$, then for $i = 0$ or 1 we have*

$$\|u-Q_\tau u_h\|_{W^{i,\infty}(\Omega)} \le Ch^{\frac{1+r+\varrho}{1+\theta_i}}|\ln h|\left(\|u\|_{W^{1+r,\infty}(\Omega)}+\|u\|_{W^{1+s,\infty}(\Omega)}\right),$$

where θ_i is defined by (3.2) and

$$\varrho = \min(r-1,m,k). \tag{4.1}$$

The proof of this theorem is based on some technical tools in weighted Sobolev norms. We shall develop the tools first and postpone the proof to the end of this section.

Let us start with the weight function definition. For any fixed $z \in \bar\Omega$, we define

$$\sigma = \sigma(x,z) = (|x-z|^2+\gamma h^2)^{1/2},$$

with $\gamma > 0$ being a constant. Associated with the function σ, we introduce the following weighted Sobolev norms:

$$\|v\|_{L^2_{\sigma^\beta}(D)} = \left(\int_D \sigma^\beta|v|^2dx\right)^{1/2},$$

$$\|v\|_{H^k_{\sigma^\beta}(D)} = \left(\sum_{|\alpha|\le k} \left\| \frac{\partial^k v}{\partial x^\alpha} \right\|^2_{L^2_{\sigma^\beta}(D)} \right)^{1/2}.$$

We now collect some well-known estimates associated with the weighted norms. The first one is an analogy of A.1 in the weighted norm.

LEMMA 4.2. *For $0 \le i \le 1 \le j \le r+1$ and any real β, $v \in H^j(\Omega)$, there exists $\chi \in S^h_r$ such that*

$$\|v - \chi\|_{H^i_{\sigma^\beta}(\Omega)} \le C h^{j-i} \|v\|_{H^j_{\sigma^\beta}(\Omega)}.$$

LEMMA 4.3. *For $0 \le i \le j$ and any $\varphi_\tau \in S^\tau_{m,s}(\Omega)$,*

$$\|\varphi_\tau\|_{H^j_{\sigma^\beta}(\Omega)} \le C \tau^{i-j} \|\varphi_\tau\|_{H^i_{\sigma^\beta}(\Omega)}.$$

A proof of Lemmas 4.2 and 4.3 can be found in Brenner and Scott [3]. The following lemma gives the a priori estimates for the differential operator $\mathcal{L}$ in the weighted norms.

LEMMA 4.4. *Suppose $\Phi \in H^{k+2}(\Omega)$ with $k \ge 0$. Then*

$$\|\Phi\|_{H^{k+2}_{\sigma^2}(\Omega)} \le C \|\mathcal{L}\Phi\|_{H^k_{\sigma^2}(\Omega)} + C \|\Phi\|_{H^{k+1}(\Omega)}.$$

Proof. Let $x = (x_1, x_2)$ and $z = (z_1, z_2)$. For any nonnegative integers k_1 and k_2 such that $k_1 + k_2 = k + 2$, we have

$$(x_1 - z_1) \frac{\partial^{k+2}\Phi}{\partial x_1^{k_1} \partial x_2^{k_2}} = \frac{\partial^{k+2}((x_1 - z_1)\Phi)}{\partial x_1^{k_1} \partial x_2^{k_2}} - k_1 \frac{\partial^{k+1}\Phi}{\partial x_1^{k_1-1} \partial x_2^{k_2}}.$$

Hence, it follows that

$$\begin{aligned}
&\int_\Omega (x_1 - z_1)^2 \left| \frac{\partial^{k+2}\Phi}{\partial x_1^{k_1} \partial x_2^{k_2}} \right|^2 dx \\
&\le C \int_\Omega \left| \frac{\partial^{k+2}((x_1 - z_1)\Phi)}{\partial x_1^{k_1} \partial x_2^{k_2}} \right|^2 dx + C \left\| \frac{\partial^{k+1}\Phi}{\partial x_1^{k_1-1} \partial x_2^{k_2}} \right\|^2_{L^2(\Omega)} \\
&\le C \|(x_1 - z_1)\Phi\|^2_{H^{k+2}(\Omega)} + C \|\Phi\|^2_{H^{k+1}(\Omega)} \\
&\le C \|\mathcal{L}((x_1 - z_1)\Phi)\|^2_{H^k(\Omega)} + C \|\Phi\|^2_{H^{k+1}(\Omega)} \\
&\le C \|(x_1 - z_1)\mathcal{L}\Phi\|^2_{H^k(\Omega)} + C \|\Phi\|^2_{H^{k+1}(\Omega)} \\
&\le C \|\mathcal{L}\Phi\|^2_{H^k_{\sigma^2}(\Omega)} + C \|\Phi\|^2_{H^{k+1}(\Omega)}.
\end{aligned}$$

Similarly, we have

$$\int_\Omega (x_2 - z_2)^2 \left| \frac{\partial^{k+2}\Phi}{\partial x_1^{k_1} \partial x_2^{k_2}} \right|^2 dx \le C \|\mathcal{L}\Phi\|^2_{H^k_{\sigma^2}(\Omega)} + C \|\Phi\|^2_{H^{k+1}(\Omega)}.$$

Consequently,

$$\int_\Omega |x - z|^2 \left| \frac{\partial^{k+2}\Phi}{\partial x_1^{k_1} \partial x_2^{k_2}} \right|^2 dx \le C \|\mathcal{L}\Phi\|^2_{H^k_{\sigma^2}(\Omega)} + C \|\Phi\|^2_{H^{k+1}(\Omega)}. \tag{4.2}$$

Furthermore,

$$\int_\Omega \gamma h^2 \left| \frac{\partial^{k+2}\Phi}{\partial x_1^{k_1} \partial x_2^{k_2}} \right|^2 dx \le C\gamma h^2 \|\mathcal{L}\Phi\|^2_{H^k(\Omega)} \le C\|\mathcal{L}\Phi\|^2_{H^k_{\sigma^2}(\Omega)}. \tag{4.3}$$

Adding (4.2) and (4.3) completes the proof. □

For $z \in \bar\Omega$, let $D_{\tau,z} \subset \Omega$ be such that

$$z \in \bar D_{\tau,z}, \quad \operatorname{diam}(D_{\tau,z}) \le C\tau.$$

Construct $\delta_{\tau,z} \in C_0^\infty(D_{\tau,z})$ so that

$$(\delta_{\tau,z}, \varphi_\tau) = \varphi_\tau(z) \quad \forall\ \varphi_\tau \in S^\tau_{m,s}(\Omega)$$

and

$$\|\delta_{\tau,z}\|_{W^{i,\infty}(D_{\tau,z})} \le C\tau^{-2-i}.$$

(See Rannacher and Scott [21] or Brenner and Scott [3].)

Further, let $G_{\tau,z}$, $\partial_\nu G_{\tau,z} \in H^1(\Omega)$ be the solution of

$$A(\varphi, G_{\tau,z}) = (Q_\tau \delta_{\tau,z}, \varphi) \quad \forall\ \varphi \in H^1(\Omega)$$

and

$$A(\varphi, \partial_\nu G_{\tau,z}) = -(Q_\tau(\nu \cdot \nabla \delta_{\tau,z}), \varphi) \quad \forall\ \varphi \in H^1(\Omega),$$

where ν is any fixed vector. Then we have the following $W^{1,1}$ estimates for $G_{\tau,z}$, $\partial_\nu G_{\tau,z}$, and their finite element approximations.

LEMMA 4.5. *Assume that Ω is of H^{k+2} regularity for $k \ge 0$. Then there exist $\Pi_h G_{\tau,z} \in S^\tau_{m,s}$ and $\Pi_h(\partial_\nu G_{\tau,z}) \in S^\tau_{m,s}$ such that*

$$\|G_{\tau,z} - \Pi_h G_{\tau,z}\|_{W^{1,1}(\Omega)} \le Ch^{1+\varrho}\,\tau^{-\varrho}|\ln h|, \tag{4.4}$$

$$\|\partial_\nu G_{\tau,z} - \Pi_h(\partial_\nu G_{\tau,z})\|_{W^{1,1}(\Omega)} \le Ch^{1+\varrho}\,\tau^{-1-\varrho}|\ln h|, \tag{4.5}$$

where ϱ is defined in (4.1) *and ν is any fixed vector.*

Proof. According to Lemma 4.2, there is an $\Pi_h G_{\tau,z} \in S^\tau_{m,s}$ satisfying

$$\|G_{\tau,z} - \Pi_h G_{\tau,z}\|_{H^1_{\sigma^2}(\Omega)} \le Ch^{1+\min(r-1,m,k)}\|G_{\tau,z}\|_{H^{2+\min(r-1,m,k)}_{\sigma^2}(\Omega)},$$

which, along with Lemma 4.4, implies

$$\|G_{\tau,z} - \Pi_h G_{\tau,z}\|_{H^1_{\sigma^2}(\Omega)} \le Ch^{1+\varrho}\left(\|\delta_{\tau,z}\|_{H^\varrho_{\sigma^2}(\Omega)} + \|G_{\tau,z}\|_{H^{1+\varrho}(\Omega)}\right). \tag{4.6}$$

If $\varrho = \min(r-1, m, k) \ge 1$, using the assumption of H^{k+2} a priori regularity and the inverse property B.2 we have

$$\begin{aligned} \|G_{\tau,z}\|_{H^{1+\varrho}(\Omega)} &\le C\|Q_\tau \delta_{\tau,z}\|_{H^{-1+\varrho}(\Omega)} \\ &\le C\tau^{1-\varrho}\|\delta_{\tau,z}\|_{L^2(\Omega)} \le C\tau^{-\varrho}. \end{aligned} \tag{4.7}$$

If $\varrho = \min(r-1, m, k) = 0$, using the coercivity of A and the definition of $G_{\tau,z}$, we get

$$\begin{aligned} \lambda_1 \|G_{\tau,z}\|^2_{H^{1+\varrho}(\Omega)} &= \lambda_1 \|G_{\tau,z}\|^2_{H^1(\Omega)} \le A(G_{\tau,z}, G_{\tau,z}) \\ &= (Q_\tau \delta_{\tau,z}, G_{\tau,z}) = Q_\tau G_{\tau,z}(z) \\ &\le \|Q_\tau G_{\tau,z}\|_{L^\infty(\Omega)} \\ &\le C|\ln \tau|^{1/2} \|Q_\tau G_{\tau,z}\|_{H^1(\Omega)}. \end{aligned} \tag{4.8}$$

Here, we have also used the following estimate:

$$\|Q_\tau G_{\tau,z}\|_{L^\infty(\Omega)} \le C|\ln \tau|^{1/2} \|Q_\tau G_{\tau,z}\|_{H^1(\Omega)},$$

which can be found, e.g., in Ciarlet [4]. With the estimates (4.7) and (4.8) and noting that

$$\|\delta_{\tau,z}\|_{H^\varrho_{\sigma^2}(\Omega)} \le C\tau^{-\varrho},$$

we conclude from (4.6) that

$$\|G_{\tau,z} - \Pi_h G_{\tau,z}\|_{H^1_{\sigma^2}(\Omega)} \le Ch^{1+\varrho}\tau^{-\varrho}|\ln \tau|^{1/2}. \tag{4.9}$$

From Hölder's inequality, it follows that

$$\begin{aligned} \|G_{\tau,z} - \Pi_h G_{\tau,z}\|_{W^{1,1}(\Omega)} &\le \left(\int_\Omega \sigma^{-2} dx\right)^{1/2} \|G_{\tau,z} - \Pi_h G_{\tau,z}\|_{H^1_{\sigma^2}(\Omega)} \\ &\le C|\ln h|^{1/2} \|G_{\tau,z} - \Pi_h G_{\tau,z}\|_{H^1_{\sigma^2}(\Omega)}. \end{aligned} \tag{4.10}$$

Since $|\ln \tau|$ is proportional to $|\ln h|$, substituting (4.9) into (4.10) yields (4.4). The estimate (4.5) can be derived in a similar way. This completes the proof. □

Now we are in a position to prove the main result of this section.

Proof of Theorem 4.1. In view of the definition of $G_{\tau,z}$ and (2.4), we have

$$\begin{aligned} Q_\tau(u-u_h)(z) &= (Q_\tau(u-u_h), \delta_{\tau,z}) = (u-u_h, Q_\tau \delta_{\tau,z}) \\ &= A(u-u_h, G_{\tau,z}) = A(u-u_h, G_{\tau,z} - \Pi_h G_{\tau,z}) \\ &\le C\|u-u_h\|_{W^{1,\infty}(\Omega)} \|G_{\tau,z} - \Pi_h G_{\tau,z}\|_{W^{1,1}(\Omega)}, \end{aligned}$$

where $\Pi_h G_{\tau,z}$ is chosen according to Lemma 4.5. Applying Lemma 4.5 and the following well-known estimate (e.g., [3], [4]),

$$\|u-u_h\|_{W^{1,\infty}(\Omega)} \le Ch^r \|u\|_{W^{1+r,\infty}(\Omega)}, \tag{4.11}$$

we see that

$$|Q_\tau(u-u_h)(z)| \le Ch^{1+r+\varrho}\tau^{-\varrho}|\ln \tau| \|u\|_{W^{1+r,\infty}(\Omega)}. \tag{4.12}$$

Hence, from the approximation property B.1 and (4.12), it follows that

$$\begin{aligned} |(u - Q_\tau u_h)(z)| &\le |(u - Q_\tau u)(z)| + |Q_\tau(u-u_h)(z)| \\ &\le C\tau^{1+s}\|u\|_{W^{1+s,\infty}(\Omega)} \\ &\quad + Ch^{1+r+\varrho}\tau^{-\varrho}|\ln \tau| \|u\|_{W^{1+r,\infty}(\Omega)}, \end{aligned} \tag{4.13}$$

which proves Theorem 4.1 for $i = 0$ according to the relationship (3.1) between h and τ. To show Theorem 4.1 for $i = 1$, we let ν be a fixed vector and note that (with possibly a minor modification of $\delta_{\tau,z}$)

$$\begin{aligned}\nu \cdot \nabla Q_\tau(u-u_h)(z) &= (\nu \cdot \nabla Q_\tau(u-u_h), \delta_{\tau,z}) = -(Q_\tau(u-u_h), \nu \cdot \nabla \delta_{\tau,z}) \\ &= -(u-u_h, Q_\tau(\nu \cdot \nabla \delta_{\tau,z})) \\ &= a(u-u_h, \partial_\nu G_{\tau,z}) = a(u-u_h, \partial_\nu G_{\tau,z} - \Pi_h(\partial_\nu G_{\tau,z})) \\ &\leq C\|u-u_h\|_{W^{1,\infty}(\Omega)} \|\partial G_{\tau,z} - \Pi_h(\partial G_{\tau,z})\|_{W^{1,1}(\Omega)},\end{aligned}$$

which, along with Lemma 4.5, yields

$$|\nu \cdot \nabla Q_\tau(u-u_h)(z)| \leq Ch^{1+\rho}\tau^{-1-\rho}\|u-u_h\|_{W^{1,\infty}(\Omega)}. \tag{4.14}$$

Using (4.14) and the well-known estimate for L^2 projection,

$$|\nu \cdot \nabla(u - Q_\tau u)(z)| \leq C\|u - Q_\tau u\|_{W^{1,\infty}(\Omega)} \leq C\tau^s \|u\|_{W^{1+s,\infty}(\Omega)},$$

and (4.11) we get

$$\begin{aligned}&|\nu \cdot \nabla(u - Q_\tau u_h)(z)| \\ &\leq |\nu \cdot \nabla(u - Q_\tau u)(z)| + |\nu \cdot \nabla Q_\tau(u-u_h)(z)| \\ &\leq C\tau^s\|u\|_{W^{1+s,\infty}(\Omega)} + Ch^{1+r+\varrho}\tau^{-1-\varrho}|\ln \tau| \|u\|_{W^{1+r,\infty}(\Omega)}.\end{aligned} \tag{4.15}$$

Thus, choosing a direction ν and a point z such that

$$\|\nabla(u - Q_\tau u_h)\|_{L^\infty(\Omega)} = |\nu \cdot \nabla(u - Q_\tau u_h)(z)|,$$

and using (4.15), we complete the proof of Theorem 4.1. □

5. A local superconvergence in L^2.

In this section, we establish a local superconvergence in the L^2 and H^1 norms. The main result is stated as follows.

THEOREM 5.1. *Suppose u and u_h satisfy* (2.4) *and the subdomain $\Omega_0 \subset \Omega_1 \subset \Omega$ satisfies dist$(\Omega_0, \partial\Omega_1 \backslash \partial\Omega) \geq C_1 h$ for sufficiently large $C_1 > 1$, and $\partial\Omega_1 \cap \partial\Omega$ is sufficiently smooth. Then for τ given by* (3.1) *and any $0 \leq q \leq m+1$,*

$$\begin{aligned}\|u - Q_\tau^{\Omega_1} u_h\|_{H^i(\Omega_0)} \leq Ch^{\frac{r+1+\min(m,k,r-1)}{1+\theta_i}} &\left(\|u\|_{H^{1+r}(\Omega_1)} + \|u\|_{H^{1+s}(\Omega_1)}\right) \\ &+ C\|u-u_h\|_{H^{-q}(\Omega_1)},\end{aligned} \tag{5.1}$$

where θ_i is given by (3.2).

Proof. Observe that

$$\|u - Q_\tau^{\Omega_1} u_h\|_{H^i(\Omega_0)} \leq \|u - Q_\tau^{\Omega_1} u\|_{H^i(\Omega_0)} + \|Q_\tau^{\Omega_1}(u-u_h)\|_{H^i(\Omega_0)}. \tag{5.2}$$

Let $\Omega_0 \subset\subset \Omega_2 \subset\subset \Omega_1$. Using Lemma 5.3 and assumption B.1, we have

$$\begin{aligned}\|u - Q_\tau^{\Omega_1} u\|_{L^2(\Omega_0)} &\leq C \inf_{\chi \in S_{m,s}^\tau} \|u - \chi\|_{L^2(\Omega_2)} + \tau^{1+s}\|u\|_{L^2(\Omega_1)} \\ &\leq C\tau^{1+s}\|u\|_{H^{1+s}(\Omega_1)}.\end{aligned} \tag{5.3}$$

Applying a well-known interior estimate in H^1 norm (Nitsche and Schatz [18])

$$\|u-u_h\|_{H^1(\Omega_2)} \leq Ch^r\|u\|_{H^{1+r}(\Omega_1)} + C\|u-u_h\|_{H^{-q}(\Omega_1)}$$

and Lemma 5.5 we obtain

$$\begin{aligned}(5.4)\qquad &\|Q_\tau^{\Omega_1}(u-u_h)\|_{L^2(\Omega_0)}\\ &\le Ch^{1+\min(r-1,m)}\tau^{-\min(r-1,m)}\|u-u_h\|_{H^1(\Omega_2)}+C\|u-u_h\|_{H^{-q}(\Omega_2)}\\ &\le Ch^{1+r}(h/\tau)^{\min(r-1,m)}\|u\|_{H^{1+r}(\Omega_1)}+C\|u-u_h\|_{H^{-q}(\Omega_1)}.\end{aligned}$$

The desired estimate (5.1) with $i=0$ follows from (5.3), (5.4), and (5.2). With the same procedure combined with the inverse inequality in B.2, we obtain (5.1) for $i=1$. This completes the proof. □

The proof of Theorem 5.1 is based on the result of Lemma 5.5, which will be established in the rest of this section.

LEMMA 5.2. *Let $D_1 \subset D_2 \subset\subset D_3 \subset D_4 \subset D \subset \Omega$ with $\operatorname{dist}(D_1, \partial D_2\backslash\partial\Omega) \ge C_0 h$, $\operatorname{dist}(D_2, \partial D_3\backslash\partial\Omega) \ge C_0 h$, $\operatorname{dist}(D_3, \partial D_4\backslash\partial\Omega) \ge C_0 h$, $\operatorname{dist}(D_4, \partial D\backslash\partial\Omega) \ge C_0 h$ for some constant $C_0 > 0$. Let $\omega \in C^\infty(D_3) \cap L^2_<(D_3)$ satisfy $\omega \equiv 1$ on D_2. Then, for $0 \le j \le \min(s+1, m)$ and any $v_\tau \in S^\tau_{m,s}(D_3)$, we have*

$$\|\omega v_\tau - Q_\tau^D(\omega v_\tau)\|_{L^2(D)} \le C\tau^{1+j}\|v_\tau\|_{H^j(D_4\backslash D_1)}.$$

Proof. The lemma follows immediately from assumption B.3 and the best approximation property of the L^2 projection. □

The following lemma provides a local error estimate for the L^2 projection. The result shows that the local error of a "global" L^2 projection Q_τ is bounded by the best local approximation plus a global pollution of order $\mathcal{O}(\tau^M)$ with arbitrary $M > 0$.

LEMMA 5.3. *For $D_0 \subset D_1 \subset D \subset \Omega$ satisfying dist$(D_0, \partial D_1\backslash\partial\Omega) \ge C_1 h$ with sufficiently large $C_1 > 0$, and $M > 0$, there holds*

$$\|v - Q_\tau^D v\|_{L^2(D_0)} \le \inf_{\chi\in S^\tau_{m,s}(D)} \|v-\chi\|_{L^2(D_1)} + C\tau^M\|v\|_{L^2(D)}.$$

Proof. The proof can be found in Nitsche and Schatz [19] and Schatz and Wahlbin [27]. □

We now show a local a priori estimate for the differential operator $\mathcal{L}$.

LEMMA 5.4. *Suppose $D_0 \subset D_1 \subset D \subset \Omega$ with $d = $ dist$(D_0, \partial D_1\backslash\partial\Omega) > 0$ and $\partial D_1 \cap \partial\Omega$ is sufficiently smooth. Then for $k \ge 0$ and any $w \in H^{k+2}(D_1)$ with $\mathcal{L}w \in L^1(D)$ and $\mathcal{B}w = 0$ on ∂D,*

$$\|w\|_{H^{k+2}(D_0)} \le C\left(\|\mathcal{L}w\|_{H^k(D_1)} + \|\mathcal{L}w\|_{L^1(D)}\right).$$

Proof. Let $D_0 \subset D_2 \subset D_1 \subset \Omega$ with $\operatorname{dist}(D_0, \partial D_2\backslash\partial\Omega) = \operatorname{dist}(D_2, \partial D_1\backslash\partial\Omega) = d/2$ and $\omega \in C^\infty(D_2) \cap L^2_<(D_2)$ satisfying $\omega \equiv 1$ on D_0. We assume that ∂D_2 is sufficiently smooth. Then

$$\begin{aligned}(5.5)\qquad \|w\|_{H^{k+2}(D_0)} &\le \|\omega w\|_{H^{k+2}(D_2)} \le C\|\mathcal{L}(\omega w)\|_{H^k(D_2)}\\ &\le C\|\mathcal{L}w\|_{H^k(D_2)} + C\|w\|_{H^{k+1}(D_2)}.\end{aligned}$$

Applying (5.5) for $\|w\|_{H^{k+1}(D_2)}$ and repeating the procedure, we conclude that

$$\begin{aligned}(5.6)\qquad \|w\|_{H^{k+2}(D_0)} &\le C\|\mathcal{L}w\|_{H^k(D_1)} + C\|w\|_{L^2(D_1)}\\ &\le C\|\mathcal{L}w\|_{H^k(D_1)} + C\|\mathcal{L}w\|_{L^1(D)}.\end{aligned}$$

Here, in the last step of (5.6), we have used the following estimate:

$$\|w\|_{L^2(D_1)} \le C\|w\|_{L^2(D)} \le C\|\mathcal{L}w\|_{L^1(D)},$$

which can be verified, e.g., by using the representation of w in terms of the Green's function of the differential operator $\mathcal{L}$ on domain D. This completes the proof. ∎

LEMMA 5.5. *Suppose $\Omega_0 \subset \Omega_1 \subset \Omega$ with $d = dist(\Omega_0, \partial\Omega_1 \backslash \partial\Omega) \geq C_1 h$ for some sufficiently large $C_1 > 1$; then, for any $0 \leq q \leq 1+m$, there holds*

$$(5.7) \qquad \|Q_\tau^{\Omega_1}(u-u_h)\|_{L^2(\Omega_0)} \leq Ch^{1+\min(r-1,m)}\tau^{-\min(r-1,m)}\|u-u_h\|_{H^1(\Omega_1)} + C\|u-u_h\|_{H^{-q}(\Omega_1)},$$

where the constant $C > 0$ depends on d but is independent of h and u.

Proof. Let $\Omega_0 \subset \Omega_2 \subset \Omega_3 \subset \Omega_4 \subset \Omega_5 \subset \Omega_6 \subset \Omega_7 \subset \Omega_1$ satisfy

$$\begin{aligned} &\mathrm{dist}(\Omega_0, \partial\Omega_2\backslash\partial\Omega) = \mathrm{dist}(\Omega_2, \partial\Omega_3\backslash\partial\Omega) = \mathrm{dist}(\Omega_3, \partial\Omega_4\backslash\partial\Omega) = \mathrm{dist}(\Omega_4, \partial\Omega_5\backslash\partial\Omega) \\ &= \mathrm{dist}(\Omega_5, \partial\Omega_6\backslash\partial\Omega) = \mathrm{dist}(\Omega_6, \partial\Omega_7\backslash\partial\Omega) = \mathrm{dist}(\Omega_7, \partial\Omega_1\backslash\partial\Omega) = d/7 \end{aligned}$$

with sufficiently smooth $\partial\Omega_7$, and let $\omega \in C^\infty(\Omega_5) \cap L^2_<(\Omega_5)$ be a function satisfying $\omega \equiv 1$ on Ω_4. For any $\varphi \in C_0^\infty(\Omega_0)$, by some elementary manipulations we have

$$\begin{aligned} (5.8) \quad &(Q_\tau^{\Omega_1}(u-u_h), \varphi) = (\omega Q_\tau^{\Omega_1}(u-u_h), \omega\varphi) \\ &= (Q_\tau^{\Omega_1}(\omega^2 Q_\tau^{\Omega_1}(u-u_h)), \varphi) + (\omega^2 Q_\tau^{\Omega_1}(u-u_h) - Q_\tau^{\Omega_1}(\omega^2 Q_\tau^{\Omega_1}(u-u_h)), \varphi) \\ &= (Q_\tau^{\Omega_1}(u-u_h), \omega^2 Q_\tau^{\Omega_1}\varphi) + (\omega^2 Q_\tau^{\Omega_1}(u-u_h) - Q_\tau^{\Omega_1}(\omega^2 Q_\tau^{\Omega_1}(u-u_h)), \varphi) \\ &= (\omega^2(u-u_h), Q_\tau^{\Omega_1}\varphi) + (u-u_h, Q_\tau^{\Omega_1}(\omega^2 Q_\tau^{\Omega_1}\varphi) - \omega^2 Q_\tau^{\Omega_1}\varphi)_{\Omega_1} \\ &\quad + (\omega^2 Q_\tau^{\Omega_1}(u-u_h) - Q_\tau^{\Omega_1}(\omega^2 Q_\tau^{\Omega_1}(u-u_h)), \varphi). \end{aligned}$$

For the second term on the right-hand side of (5.8), using Lemma 5.2 we have

$$\begin{aligned} (5.9) \qquad &|(u-u_h, Q_\tau^{\Omega_1}(\omega^2 Q_\tau^{\Omega_1}\varphi) - \omega^2 Q_\tau^{\Omega_1}\varphi)_{\Omega_1}| \\ &\leq \|u-u_h\|_{L^2(\Omega_1)}\|\omega^2 Q_\tau^{\Omega_1}\varphi - Q_\tau^{\Omega_1}(\omega^2 Q_\tau^{\Omega_1}\varphi)\|_{L^2(\Omega_1)} \\ &\leq c\tau\|u-u_h\|_{L^2(\Omega_1)}\|Q_\tau^{\Omega_1}\varphi\|_{L^2(\Omega_6\backslash\Omega_3)}. \end{aligned}$$

Using Lemma 5.3, we obtain

$$\begin{aligned} \|Q_\tau^{\Omega_1}\varphi\|_{L^2(\Omega_6\backslash\Omega_3)} &\leq \|\varphi\|_{L^2(\Omega_6\backslash\Omega_3)} + \|\varphi - Q_\tau^{\Omega_1}\varphi\|_{L^2(\Omega_6\backslash\Omega_3)} \\ &\leq 0 + \|\varphi\|_{L^2(\Omega_7\backslash\Omega_2)} + C\tau^{M-1}\|\varphi\|_{L^2(\Omega_0)} \\ &\leq C\tau^{M-1}\|\varphi\|_{L^2(\Omega_0)}, \end{aligned}$$

which, along with (5.9), gives

$$|(u-u_h, Q_\tau^{\Omega_1}(\omega^2 Q_\tau^{\Omega_1}\varphi) - \omega^2 Q_\tau^{\Omega_1}\varphi)_{\Omega_1}| \leq c\tau^M\|u-u_h\|_{L^2(\Omega_1)}\|\varphi\|_{L^2(\Omega_0)}.$$

For the third term on the right-hand side of (5.8), using Lemma 5.2 we have

$$\begin{aligned} (5.10) \qquad &|(\omega^2 Q_\tau^{\Omega_1}(u-u_h) - Q_\tau^{\Omega_1}(\omega^2 Q_\tau^{\Omega_1}(u-u_h)), \varphi)| \\ &\leq \|\omega^2 Q_\tau^{\Omega_1}(u-u_h) - Q_\tau^{\Omega_1}(\omega^2 Q_\tau^{\Omega_1}(u-u_h))\|_{L^2(\Omega_0)}\|\varphi\|_{L^2(\Omega_0)} \\ &\leq C\tau\|Q_\tau^{\Omega_1}(u-u_h)\|_{L^2(\Omega_6)}\|\varphi\|_{L^2(\Omega_0)}. \end{aligned}$$

Substituting (5.9) and (5.10) into (5.8), we obtain

$$\begin{aligned} (5.11) \qquad &\|Q_\tau^{\Omega_1}(u-u_h)\|_{L^2(\Omega_0)} \\ &\leq \sup_{\varphi \in C_0^\infty(\Omega_0)} \frac{(\omega(u-u_h), \omega Q_\tau^{\Omega_1}\varphi)}{\|\varphi\|_{L^2(\Omega_0)}} \\ &\quad + C\tau^M\|u-u_h\|_{L^2(\Omega_1)} + C\tau\|Q_\tau^{\Omega_1}(u-u_h)\|_{L^2(\Omega_6)}. \end{aligned}$$

It remains to estimate the first term on the right-hand side of (5.11). To this end, let $\Phi \in H^1(\Omega_7)$ be the solution of the following auxiliary problem:

$$A_{\Omega_7}(\psi, \Phi) = (\omega Q_\tau^{\Omega_1}\varphi, \psi)_{\Omega_7} \quad \forall\ v \in H^1(\Omega_7).$$

Then, by (2.10), for any $0 \le k \le m$,

$$\|\Phi\|_{H^{k+2}(\Omega_7)} \le C\|\omega Q_\tau^{\Omega_1}\varphi\|_{H^k(\Omega_7)} \le C\|Q_\tau^{\Omega_1}\varphi\|_{H^k(\Omega_5)} \tag{5.12}$$

and

$$\|\Phi\|_{H^{k+2}(\Omega_5\setminus\Omega_4)} \le C\left(\|Q_\tau^{\Omega_1}\varphi\|_{H^k(\Omega_6\setminus\Omega_3)} + \|Q_\tau^{\Omega_1}\varphi\|_{L^2(\Omega_1)}\right). \tag{5.13}$$

Thus,

$$\begin{aligned}(\omega(u-u_h), \omega Q_\tau^{\Omega_1}\varphi_\tau) &= A_{\Omega_1}(\omega(u-u_h), \Phi) = A(\omega(u-u_h), \Phi)\\ &= A(u-u_h, \omega\Phi) + I_\omega(u-u_h, \Phi),\end{aligned} \tag{5.14}$$

where, for the Laplacian operator,

$$I_\omega(u, v) = (u\nabla\omega, \nabla v) - (\nabla u, v\nabla\omega).$$

By assumption A.1, there exists a $\chi \in S_r^h \cap L^2_<(\Omega_6)$ such that

$$\|\omega\Phi - \chi\|_{H^1(\Omega_6)} \le Ch^{i-1}\|\Phi\|_{H^i(\Omega_7)}, \quad 1 \le i \le 1+r. \tag{5.15}$$

Thus, according to (2.4), (5.15), and (5.12), we have

$$\begin{aligned}A(u-u_h, \omega\Phi) &= A(u-u_h, \omega\Phi-\chi)\\ &\le Ch^{1+\min(m,r-1)}\|u-u_h\|_{H^1(\Omega_6)}\|\Phi\|_{H^{2+\min(m,r-1)}(\Omega_7)}\\ &\le Ch^{1+\min(m,r-1)}\|u-u_h\|_{H^1(\Omega_1)}\|Q_\tau^{\Omega_1}\varphi\|_{H^{\min(m,r-1)}(\Omega_5)}\\ &\le Ch^{1+\min(m,r-1)}\tau^{-\min(m,r-1)}\|u-u_h\|_{H^1(\Omega_1)}\|Q_\tau^{\Omega_1}\varphi\|_{L^2(\Omega_6)}\\ &\le Ch^{1+\min(m,r-1)}\tau^{-\min(m,r-1)}\|u-u_h\|_{H^1(\Omega_1)}\|\varphi\|_{L^2(\Omega_0)}.\end{aligned} \tag{5.16}$$

For the second term on the right-hand side of (5.14), we have for $0 \le q \le 1+m$

$$\begin{aligned}&|I_\omega(u-u_h, \Phi)|\\ &\le C\|u-u_h\|_{H^{-q}(\Omega_5)}\|\Phi\|_{H^{q+1}(\Omega_5\setminus\Omega_4)}\\ &\le C\|u-u_h\|_{H^{-q}(\Omega_1)}\left(\|Q_\tau^{\Omega_1}\varphi\|_{H^{q-1}(\Omega_6\setminus\Omega_3)} + \|Q_\tau^{\Omega_1}\varphi\|_{L^2(\Omega_1)}\right)\\ &\le C\|u-u_h\|_{H^{-q}(\Omega_1)}\left(\|\varphi\|_{H^{q-1}(\Omega_7\setminus\Omega_2)} + \tau^M\|Q_\tau^{\Omega_1}\varphi\|_{L^2(\Omega_1)}\right.\\ &\qquad\left. + \|Q_\tau^{\Omega_1}\varphi\|_{L^2(\Omega_1)}\right)\\ &\le C\|u-u_h\|_{H^{-q}(\Omega_1)}\|\varphi\|_{L^2(\Omega_0)}.\end{aligned} \tag{5.17}$$

Substituting (5.17) and (5.16) into (5.14), we have

$$\begin{aligned}\sup_{\varphi\in C_0^\infty(\Omega_0)} \frac{(\omega(u-u_h), \omega Q_\tau^{\Omega_1}\varphi)}{\|\varphi\|_{L^2(\Omega_0)}} &\le Ch^{1+\min(m,r-1)}\tau^{-\min(m,r-1)}\|u-u_h\|_{H^1(\Omega_1)}\\ &\quad + C\|u-u_h\|_{H^{-q}(\Omega_1)},\end{aligned}$$

which, together with (5.11), implies

$$\begin{aligned}(5.18)\qquad &\|Q_\tau^{\Omega_1}(u-u_h)\|_{L^2(\Omega_0)}\\ &\le Ch^{1+\min(m,r-1)}\tau^{-\min(m,r-1)}\|u-u_h\|_{H^1(\Omega_1)}\\ &\quad + C\tau^M\|u-u_h\|_{L^2(\Omega_1)} + \|u-u_h\|_{H^{-q}(\Omega_1)}\\ &\quad + C\tau\|Q_\tau^{\Omega_1}(u-u_h)\|_{L^2(\Omega_6)}.\end{aligned}$$

A repeated use of (5.18) leads to

$$\begin{aligned}(5.19)\qquad &\|Q_\tau^{\Omega_1}(u-u_h)\|_{L^2(\Omega_0)}\\ &\le Ch^{1+\min(m,r-1)}\tau^{-\min(m,r-1)}\|u-u_h\|_{H^1(\Omega_1)}\\ &\quad + C\tau^M\|u-u_h\|_{L^2(\Omega_1)} + \|u-u_h\|_{H^{-q}(\Omega_1)}\\ &\quad + C\tau^M\|Q_\tau^{\Omega_1}(u-u_h)\|_{L^2(\Omega_7)}\\ &\le Ch^{1+\min(m,r-1)}\tau^{-\min(m,r-1)}\|u-u_h\|_{H^1(\Omega_1)}\\ &\quad + C\tau^M\|u-u_h\|_{L^2(\Omega_1)} + \|u-u_h\|_{H^{-q}(\Omega_1)}.\end{aligned}$$

Choosing $M > 0$ sufficiently large so that

$$\tau^M\|u-u_h\|_{L^2(\Omega_1)} \le Ch^{1+\min(m,r-1)}\tau^{-\min(m,r-1)}\|u-u_h\|_{H^1(\Omega_1)}$$

completes the proof. □

6. A local superconvergence in L^∞.

Our objective in this section is to derive a local superconvergence estimate in the maximum norm. The result can be stated as follows.

THEOREM 6.1. *Suppose u and u_h satisfy* (2.4) *and the subdomains $\Omega_0 \subset \Omega_1 \subset \Omega$ satisfy dist$(\Omega_0, \partial\Omega_1\backslash\partial\Omega) \ge C_1 h$ for sufficiently large $C_1 > 1$. If τ satisfies* (3.1)*, then, for any $0 \le q \le m+1$ and $i = 0, 1$, we have*

$$\begin{aligned}(6.1)\qquad &\|u - Q_\tau^{\Omega_1}u_h\|_{W^{i,\infty}(\Omega_0)}\\ &\le Ch^{\frac{r+1+\min(m,r-1)}{1+\theta_i}}\left(\|u\|_{W^{1+r,\infty}(\Omega_1)} + \|u\|_{W^{1+s,\infty}(\Omega_1)}\right)\\ &\quad + C\|u-u_h\|_{H^{-q}(\Omega_1)},\end{aligned}$$

where θ_i is defined by (3.2).

Proof. Observe that from the triangle inequality we have

$$(6.2)\quad \|u - Q_\tau^{\Omega_1}u_h\|_{W^{i,\infty}(\Omega_0)} \le \|u - Q_\tau^{\Omega_1}u\|_{W^{i,\infty}(\Omega_0)} + \|Q_\tau^{\Omega_1}(u-u_h)\|_{W^{i,\infty}(\Omega_0)}.$$

The first term on the right-hand side of (6.2) can be estimated using an analogy of Lemma 5.3 in the L^∞ norm. The second term on the right-hand side of (6.2) can be handled using Lemma 6.2, which will be established in the rest of this section. The desired superconvergence is merely a combination of those results. □

The remaining portion of this section is devoted to the establishment of a result that has been used in the proof of Theorem 6.1.

LEMMA 6.2. *Suppose $\Omega_0 \subset \Omega_1 \subset \Omega$ with dist$(\Omega_0, \partial\Omega_1\backslash\partial\Omega) \ge C_1 h$ for sufficiently large $C_1 > 1$; then, for $0 \le q \le m+1$, there holds*

$$\begin{aligned}(6.3)\qquad &\|Q_\tau^{\Omega_1}(u-u_h)\|_{L^\infty(\Omega_0)}\\ &\le Ch^{1+\min(m,r-1)}\tau^{-\min(m,r-1)}\|u-u_h\|_{W^{1,\infty}(\Omega_1)}\\ &\quad + C\|Q_\tau^{\Omega_1}(u-u_h)\|_{L^2(\Omega_1)} + C\|u-u_h\|_{H^{-q}(\Omega_1)}.\end{aligned}$$

Proof. Let $z \in \bar{\Omega}_0$ be such that

$$|Q_\tau(u-u_h)(z)| = \max_{x\in\Omega_0} |Q_\tau(u-u_h)(x)|.$$

Let $\Omega_0 \subset \Omega_2 \subset \Omega_3 \subset \Omega_4 \subset \Omega_5 \subset \Omega_6 \subset \Omega_7 \subset \Omega_1$ satisfy

$$\begin{aligned}&\mathrm{dist}(\Omega_0, \partial\Omega_2\backslash\partial\Omega) = \mathrm{dist}(\Omega_2, \partial\Omega_3\backslash\partial\Omega) = \mathrm{dist}(\Omega_3, \partial\Omega_4\backslash\partial\Omega) = \mathrm{dist}(\Omega_4, \partial\Omega_5\backslash\partial\Omega)\\ &= \mathrm{dist}(\Omega_5, \partial\Omega_6\backslash\partial\Omega) = \mathrm{dist}(\Omega_6, \partial\Omega_7\backslash\partial\Omega) = \mathrm{dist}(\Omega_7, \partial\Omega_1\backslash\partial\Omega) = d/7\end{aligned}$$

with sufficiently smooth $\partial\Omega_7$. Then, for $\omega \in C^\infty(\Omega_5) \cap L^2_<(\Omega_5)$ with $\omega \equiv 1$ in Ω_4,

$$\begin{aligned}(6.4)\qquad &Q_\tau^{\Omega_1}(u-u_h)(z)\\ &= (Q_\tau^{\Omega_1}(u-u_h), \delta_{\tau,z}) = (\omega Q_\tau^{\Omega_1}(u-u_h), \omega\delta_{\tau,z})\\ &= (\omega^2(u-u_h), Q_\tau^{\Omega_1}\delta_{\tau,z}) + (u-u_h, Q_\tau^{\Omega_1}(\omega^2 Q_\tau^{\Omega_1}\delta_{\tau,z}) - \omega^2 Q_\tau^{\Omega_1}\delta_{\tau,z})\\ &\quad + (\omega^2 Q_\tau^{\Omega_1}(u-u_h) - Q_\tau^{\Omega_1}(\omega^2 Q_\tau^{\Omega_1}(u-u_h)), \delta_{\tau,z}).\end{aligned}$$

By Lemma 5.2, we have

$$\begin{aligned}(6.5)\qquad &(u-u_h, Q_\tau^{\Omega_1}(\omega^2 Q_\tau^{\Omega_1}\delta_{\tau,z}) - \omega^2 Q_\tau^{\Omega_1}\delta_{\tau,z})_{\Omega_1}\\ &\le \|u-u_h\|_{L^2(\Omega_1)} \|\omega^2 Q_\tau^{\Omega_1}\delta_{\tau,z} - Q_\tau^{\Omega_1}(\omega^2 Q_\tau^{\Omega_1}\delta_{\tau,z})\|_{L^2(\Omega_1)}\\ &\le C\tau\|u-u_h\|_{L^2(\Omega_1)} \|Q_\tau^{\Omega_1}\delta_{\tau,z}\|_{L^2(\Omega_6\backslash\Omega_3)}\\ &\le C\tau^{M+1}\|u-u_h\|_{L^2(\Omega_1)} \|\delta_{\tau,z}\|_{L^2(\Omega_1)}\\ &\le C\tau^M\|u-u_h\|_{L^2(\Omega_1)}\end{aligned}$$

and

$$\begin{aligned}(6.6)\qquad &(\omega^2 Q_\tau^{\Omega_1}(u-u_h) - Q_\tau^{\Omega_1}(\omega^2 Q_\tau^{\Omega_1}(u-u_h)), \delta_{\tau,z})\\ &\le \|\omega^2 Q_\tau^{\Omega_1}(u-u_h) - Q_\tau^{\Omega_1}(\omega^2 Q_\tau^{\Omega_1}(u-u_h)\|_{L^2(\Omega_1)} \|\delta_{\tau,z}\|_{L^2(\Omega_1)}\\ &\le C\tau\|Q_\tau^{\Omega_1}(u-u_h)\|_{L^2(\Omega_6)} \|\delta_{\tau,z}\|_{L^2(\Omega_1)}\\ &\le C\|Q_\tau^{\Omega_1}(u-u_h)\|_{L^2(\Omega_6)}.\end{aligned}$$

It remains to estimate the first term on the right-hand side of (6.4). Let $G \in H^1(\Omega_7)$ be the solution of

$$A_{\Omega_7}(\psi, G) = (\omega Q_\tau^{\Omega_1}\delta_{\tau,z}, \psi) \quad \forall\ \psi \in H^1(\Omega_7).$$

Then, similar to (5.14), we have

$$(6.7)\qquad (\omega(u-u_h), \omega Q_\tau^{\Omega_1}\delta_{\tau,z}) = A(u-u_h, \omega G) + I_\omega(u-u_h, G).$$

Moreover, choosing $\chi \in S_r^h \cap L^2_<(\Omega_6)$ such that

$$\|\omega G - \chi\|_{W^{1,1}(\Omega_6)} \le Ch^{i-1}\|G\|_{W^{i,1}(\Omega_7)}, \quad 1 \le i \le 1+r,$$

we have

$$\begin{aligned}(6.8)\qquad &A(u-u_h, \omega G) = A(u-u_h, \omega G - \Pi_h(\omega G))\\ &\le Ch^{1+\min(m,r-1)}\|u-u_h\|_{W^{1,\infty}(\Omega_6)} \|G\|_{W^{2+\min(m,r-1),1}(\Omega_7)}.\end{aligned}$$

To estimate $\|G\|_{W^{2+\min(m,r-1),1}(\Omega_7)}$, we use Lemma 4.4:

$$\begin{aligned}(6.9)\qquad &\|G\|_{W^{2+\min(m,r-1),1}(\Omega_7)} \le C|\ln\tau|^{1/2}\|G\|_{H^{2+\min(m,r-1)}_{\sigma^2}(\Omega_7)}\\ &\le C\Big(\|Q^{\Omega_1}_\tau\delta_{\tau,z}\|_{H^{\min(m,r-1)}_{\sigma^2}(\Omega_7)} + \|G\|_{H^{1+\min(m,r-1)}(\Omega_7)}\Big)\\ &\le C\Big(\|Q^{\Omega_1}_\tau\delta_{\tau,z}\|_{H^{\min(m,r-1)}_{\sigma^2}(\Omega_7)} + \|Q^{\Omega_1}_\tau\delta_{\tau,z}\|_{H^{-1+\min(m,r-1)}(\Omega_7)}\Big).\end{aligned}$$

For $\|Q^{\Omega_1}_\tau\delta_{\tau,z}\|_{H^{\min(m,r-1)}_{\sigma^2}(\Omega_7)}$, we have

$$\begin{aligned}(6.10)\qquad &\|Q^{\Omega_1}_\tau\delta_{\tau,z}\|_{H^{\min(m,r-1)}_{\sigma^2}(\Omega_7)}\\ &\le C\|\delta_{\tau,z}\|_{H^{\min(m,r-1)}_{\sigma^2}(\Omega_7)} + C\|\delta_{\tau,z} - Q^{\Omega_1}_\tau\delta_{\tau,z}\|_{H^{\min(m,r-1)}_{\sigma^2}(\Omega_7)}\\ &\le C\tau^{-\min(m,r-1)} + C\tau^2\|\delta_{\tau,z}\|_{H^{2+\min(m,r-1)}_{\sigma^2}(\Omega_1)}\\ &\le C\tau^{-\min(m,r-1)}.\end{aligned}$$

For $\|Q^{\Omega_1}_\tau\delta_{\tau,z}\|_{H^{-1+\min(m,r-1)}(\Omega_7)}$, we have if $\min(m,r-1)\ge 1$,

$$\begin{aligned}(6.11)\qquad \|Q^{\Omega_1}_\tau\delta_{\tau,z}\|_{H^{-1+\min(m,r-1)}(\Omega_1)} &\le C\tau^{1-\min(m,r-1)}\|Q^{\Omega_1}_\tau\delta_{\tau,z}\|_{L^2(\Omega_7)}\\ &\le C\tau^{-\min(m,r-1)},\end{aligned}$$

and if $\min(m,r-1)=0$,

$$(6.12)\qquad \|Q^{\Omega_1}_\tau\delta_{\tau,z}\|_{H^{-1}(\Omega_7)} \le C|\ln\tau|^{1/2}.$$

Therefore, we have

$$\|G\|_{W^{2+\min(m,r-1),1}(\Omega_7)} \le C\tau^{-\min(m,r-1)}|\ln\tau|^{\bar r/2},$$

which, together with (6.8), implies

$$(6.13)\qquad A(u-u_h,\omega G) \le Ch^{1+\min(m,r-1)}\tau^{-\min(m,r-1)}|\ln\tau|^{\bar r}.$$

Next, we estimate the second term $I_\omega(u-u_h,G)$ in (6.7). According to the formula for I_ω, we have for $0\le q\le 1+m$

$$\begin{aligned}(6.14)\qquad &I_\omega(u-u_h,G)\\ &\le C\|u-u_h\|_{H^{-q}(\Omega_5)}\|G\|_{H^{q+1}(\Omega_5\backslash\Omega_3)}\\ &\le C\|u-u_h\|_{H^{-q}(\Omega_1)}\left(\|Q^{\Omega_1}_\tau\delta_{\tau,z}\|_{H^{q-1}(\Omega_4\backslash\Omega_2)} + \|Q^{\Omega_1}_\tau\delta_{\tau,z}\|_{L^1(\Omega_1)}\right)\\ &\le C\|u-u_h\|_{H^{-q}(\Omega_1)},\end{aligned}$$

where we have employed Lemma 5.4 in the second inequality. Combining (6.13), (6.14), and (6.7), we obtain

$$\begin{aligned}(6.15)\qquad &(\omega(u-u_h),\omega Q^{\Omega_1}_\tau\delta_{\tau,z})\\ &\le Ch^{1+\min(m,r-1)}\tau^{-\min(m,r-1)}|\ln h|^{\bar r}\|u-u_h\|_{W^{1,\infty}(\Omega_1)}\\ &\quad + C\|u-u_h\|_{H^{-q}(\Omega_1)}.\end{aligned}$$

Therefore, (6.5), (6.6), and (6.15) conclude that

$$\begin{aligned}(6.16)\qquad &|Q_\tau^{\Omega_1}(u-u_h)(z)|\\ &\le Ch^{1+\min(m,r-1)}\tau^{-\min(m,r-1)}|\ln h|^{\bar r}\|u-u_h\|_{W^{1,\infty}(\Omega_1)}\\ &\quad + C\|u-u_h\|_{H^{-q}(\Omega_1)} + C\|Q_\tau^{\Omega_1}(u-u_h)\|_{L^2(\Omega_6)}\\ &\quad + C\tau^M\|u-u_h\|_{L^2(\Omega_1)}.\end{aligned}$$

By choosing sufficiently large values of M, we can bound the last term by the first term on the right-hand side of (6.16). This completes the proof. □

7. Case discussions. According to the results derived in the previous sections, the leading term of the global error $u-Q_\tau u_h$ or the local error $u-Q_\tau^{\Omega_1}u_h$ is of order

$$\mathcal{O}\left(h^{(1+r+\min(r-1,m,k))/(1+\theta_0)}\right).$$

Recall that the parameter m is the smoothness of the finite element projection space $S_{m,s}^\tau$, and $k+2$ is the regularity of the second-order elliptic problem defined locally on smooth subdomains. In practical applications, we have $m>\frac{1}{2}$ and $k\ge 0$. In view of the optimal-order error estimate

$$u-u_h=\mathcal{O}\left(h^{1+r}\right),$$

we see that the projected finite element approximation has superconvergence if $\min(r-1,m,k)>\theta_0(1+r)$ or, equivalently, if $r<s$ and $\min(r-1,m,k)>0$. In other words, if $\min(r-1,m,k)>0$ holds true, then any L^2 projection of the finite element solution u_h in a higher (i.e., $s>r$) order finite element space $S_{m,s}^\tau$ will produce a superconvergent new approximation.

For the gradient of the error, the leading term of the global or local errors are bounded by

$$\mathcal{O}\left(h^{(1+r+\min(r-1,m,k))/(1+\theta_1)}\right).$$

Since the optimal-order of error estimate for the gradient is $\mathcal{O}(h^r)$, we then obtain a superconvergence if $\min(r-1,m,k)>\theta_1 r-1$. Equivalently speaking, the global or local L^2 projections produce superconvergent approximations if $\min(r-1,m,k)>-1$ and $r<s$.

Notice that the error $u-u_h$ in negative norms can be shown to be of higher order than $\mathcal{O}(h^{1+r})$ for $k>0$ and of higher order than $\mathcal{O}(h^r)$ for $k>-1$. Thus, the following conclusions can be made without any proof:

1. $u-Q_\tau u_h$ and its local analogy $u-Q_\tau^{\Omega_1}u_h$ are superconvergent if $s>r\ge 2$, $m>0$, and $k>0$.
2. $\nabla(u-Q_\tau u_h)$ and its local analogy $\nabla(u-Q_\tau^{\Omega_1}u_h)$ are superconvergent if $s>r\ge 1$, $m>-1$, and $k>-1$.

A more detailed illustration on the exact order of superconvergence corresponding to different indices of r, s, m can be found in Wang [30].

REFERENCES

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

[2] J. H. Bramble and A. H. Schatz, *Higher order local accuracy by averaging in the finite element method*, Math. Comp., 31 (1977), pp. 94–111.

[3] S. C. Brenner and L. R. Scott, *The Mathematical Theory of Finite Element Methods*, Springer-Verlag, Berlin, Heidelberg, New York, 1994.

[4] P. G. Ciarlet, *The Finite Element Methods for Elliptic Problems*, North-Holland, Amsterdam, New York, Oxford, 1978.

[5] J. Douglas and T. Dupont, *Superconvergence for Galerkin methods for the two-point boundary problem via local projections*, Numer. Math., 21 (1973), pp. 270–278.

[6] J. Douglas and J. Wang, *A new family of spaces in mixed finite element methods for rectangular elements*, Comput. Appl. Math., 12 (1993), pp. 183–197.

[7] R. E. Ewing, R. D. Lazarov, and J. Wang, *Superconvergence of the velocity along the Gauss lines in mixed finite element methods*, SIAM J. Numer. Anal., 28 (1991), pp. 1015–1029.

[8] R. E. Ewing, M. M. Liu, and J. Wang, *Superconvergence of mixed finite element approximations over quadrilaterals*, SIAM J. Numer. Anal., 36 (1999), pp. 772–787.

[9] M. Krizek and P. Neittaanmaki, *Superconvergence phenomenon in the finite element method arising from averaging gradients*, Numer. Math., 45 (1984), pp. 105–116.

[10] M. Krizek and P. Neittaanmaki, *On superconvergence techniques*, Acta Appl. Math., 9 (1987), pp. 175–198.

[11] N. D. Levine, *Superconvergence recovery of the gradient from piecewise linear finite element approximations*, IMA J. Numer. Anal., 5 (1985), pp. 407–427.

[12] Q. Lin and J. Wang, *Some Expansions for Finite Element Approximation*, Research Report IMS-15, Academia Sinica, Chengdu, China, 1984.

[13] B. Li and Z. Zhang, *Analysis of a class of superconvergence patch recovery techniques for linear and bilinear finite elements*, Numer. Methods Partial Differential Equations, 15 (1999), pp. 151–167.

[14] Q. Lin and J. Xu, *Linear finite elements with high accuracy*, J. Comput. Math., 3 (1985), pp. 115–133.

[15] Q. Lin and A. Zhou, *Notes on superconvergence and its related topics*, J. Comput. Math., 11 (1993), pp. 211–214.

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

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

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

[19] J. A. Nitsche and A. H. Schatz, *On local approximation properties of L_2 projection on spline subspaces*, Appl. Anal., 2 (1972), pp. 161–168.

[20] R. Rannacher, *Zur L_∞konvergenz linear finiter elemente*, Math. Z., 149 (1976), pp. 69–77.

[21] R. Rannacher and L. R. Scott, *Some optimal error estimates for piecewise linear finite element approximations*, Math. Comp., 38 (1982), pp. 437–445.

[22] A. H. Schatz, *Pointwise error estimates and asymptotic error expansion inequalities for the finite element method on irregular grids: Part* I. *Global estimates*, Math. Comp., 67 (1998), pp. 877–899.

[23] A. H. Schatz, *Pointwise error estimates and asymptotic error expansion inequalities for the finite element method on irregular grids: Part* II. *Interior estimates*, SIAM J. Numer. Anal., 38 (2000), pp. 1269–1293.

[24] A. H. Schatz, I. H. Sloan, and L. B. Wahlbin, *Superconvergence in finite element methods and meshes that are symmetric with respect to a point*, SIAM J. Numer. Anal., 33 (1996), pp. 505–521.

[25] A. H. Schatz and L. B. Wahlbin, *Interior maximum norm estimates for finite element methods*, Math. Comp., 31 (1977), pp. 414–442.

[26] A. H. Schatz and L. B. Wahlbin, *Maximum norm estimates in the finite element method on plane polygonal domains. Part* I, Math. Comp., 32 (1978), pp. 73–109.

[27] A. H. Schatz and L. B. Wahlbin, *On the finite element for singularly perturbed reaction diffusion problems in two and one dimensions*, Math. Comp., 40 (1983), pp. 47–89.

[28] R. Scott, *Optimal L_∞ estimates for the finite element methods on irregular grids*, Math. Comp., 30 (1976), pp. 681–697.

[29] L. B. Wahlbin, *Superconvergence in Galerkin Finite Element Methods*, Lecture Notes in Math. 1605, Springer-Verlag, New York, 1995.

[30] J. Wang, *A superconvergence analysis for finite element solutions by the least-squares surface fitting on irregular meshes for smooth problems*, J. Math. Study, 33 (2000), pp. 229–243.

[31] J. Wang, *Superconvergence and extrapolation for mixed finite element methods on rectangular domains*, Math. Comp., 56 (1991), pp. 477–503.

[32] M. F. Wheeler and J. R. Whiteman, *Superconvergence recovery of gradient on subdomains from piecewise linear finite element approximations*, Numer. Methods Partial Differential Equations, 3 (1987), pp. 65–82.

[33] Z. Zhang, *Ultrconvergence of the patch recovery technique*, Math. Comp., 65 (1996), pp. 1431–1437.

[34] Q. Zhu, *Natural inner superconvergence*, in Proceedings of the China-France Symposium on Finite Element Methods, Beijing, China, 1982, Science Press, Beijing, China, 1983, pp. 935–960.

[35] Q. Zhu and Q. Lin, *Superconvergence Theory of the Finite Element Methods*, Hunan Science Press, Changsha, China, 1989 (in Chinese).

[36] O. C. Zienkiewicz and J. Z. Zhu, *The superconvergence patch recovery and a posteriori error estimates, Part* 1*, The recovery technique*, Internat. J. Numer. Methods Engrg., 33 (1992), pp. 1331–1364.

[37] O. C. Zienkiewicz and J. Z. Zhu, *The superconvergence patch recovery and a posteriori error estimates, Part* 2*, Error estimates and adaptivity*, Internat. J. Numer. Methods Engrg., 33 (1992), pp. 1365–1382.

[38] M. Zlamal, *Superconvergence and reduced integration in the finite element method*, Math Comp., 32 (1978), pp. 663–685.

SIAM J. Numer. Anal.
Vol. 41, No. 4, pp. 1339–1369

HIGH-ORDER CENTRAL WENO SCHEMES FOR MULTIDIMENSIONAL HAMILTON–JACOBI EQUATIONS*

STEVE BRYSON† AND DORON LEVY‡

Abstract. We present new third- and fifth-order Godunov-type central schemes for approximating solutions of the Hamilton–Jacobi (HJ) equation in an arbitrary number of space dimensions. These are the first central schemes for approximating solutions of the HJ equations with an order of accuracy that is greater than two. In two space dimensions we present two versions for the third-order scheme: one scheme that is based on a genuinely two-dimensional central weighted ENO reconstruction, and another scheme that is based on a simpler dimension-by-dimension reconstruction. The simpler dimension-by-dimension variant is then extended to a multidimensional fifth-order scheme. Our numerical examples in one, two, and three space dimensions verify the expected order of accuracy of the schemes.

Key words. Hamilton–Jacobi equations, central schemes, high order, WENO, CWENO

AMS subject classifications. Primary, 65M06; Secondary, 35L99

DOI. 10.1137/S0036142902408404

1. Introduction.

We are interested in high-order numerical approximations for the solution of multidimensional Hamilton–Jacobi (HJ) equations of the form

$$\phi_t + H(\nabla\phi) = 0, \qquad \vec{x} = (x_1, \dots, x_d) \in \mathbb{R}^d,$$

where H is the Hamiltonian, which we assume depends on $\nabla\phi$ and possibly on x and t. In recent years, the HJ equations have attracted a lot of attention from analysts and numerical analysts due to the important role that they play in applications such as optimal control theory, image processing, geometric optics, differential games, calculus of variations, etc. The main difficulty in treating these equations arises from the discontinuous derivatives that develop in finite time even when the initial data is smooth. Vanishing viscosity solutions provide a good tool for defining weak solutions when the Hamiltonian is convex [15]. The celebrated *viscosity solution* provides a suitable extension of weak solutions for more general Hamiltonians [3, 7, 8, 9, 10, 28, 29].

Given the importance of the HJ equations, there has been relatively little activity in developing numerical tools for approximating their solutions. This is surprising, given that most of the numerical ideas are based on the similarity between hyperbolic conservation laws and the HJ equations, and that the field of numerical methods for conservation laws has been flourishing in recent years.

Converging first-order approximations were introduced by Souganidis in [38]. High-order upwind methods were introduced by Osher and Sethian [34] and Osher and Shu [35]. These methods are based on Harten's essentially nonoscillatory (ENO) reconstruction [13, 37], which is evolved in time with a monotone flux. The weighted

*Received by the editors May 28, 2002; accepted for publication (in revised form) February 16, 2003; published electronically August 22, 2003.
http://www.siam.org/journals/sinum/41-4/40840.html

†Program in Scientific Computing/Computational Mathematics, Stanford University and the NASA Advanced Supercomputing Division, NASA Ames Research Center, Moffett Field, CA 94035-1000 (bryson@nas.nasa.gov).

‡Department of Mathematics, Stanford University, Stanford, CA 94305-2125 (dlevy@math.stanford.edu). This author was supported in part by the National Science Foundation under Career Grant DMS-0133511.

ENO (WENO) interpolant of [18, 32, 36] was used for constructing high-order upwind methods for the HJ equations in [17], and extensions of these methods for triangular meshes were introduced in [1, 40]. We note in passing that there are other approaches for approximating solutions of HJ equations such as discontinuous Galerkin methods [14, 24] and relaxation schemes [20].

A different class of Godunov-type schemes for hyperbolic conservation laws, the so-called central schemes, has recently been applied to the HJ equations. The prototype for these schemes is the Lax–Friedrichs scheme [11]. A second-order staggered central scheme was developed for conservation laws by Nessyahu and Tadmor in [33]. The main advantage of central schemes is their simplicity. Since they do not require any (approximate) Riemann solvers, they are particularly suitable for approximating multidimensional systems of conservation laws. Lin and Tadmor applied these ideas to the HJ equations in [31]. There, first- and second-order staggered schemes versions of [2, 19, 33] were written in one and two space dimensions. An L^1 convergence of order one for this scheme was proved in [30]. After the introduction of a semidiscrete central scheme for hyperbolic conservation laws in [23], a second-order semidiscrete scheme for HJ equations was introduced by the same authors in [22]. While less dissipative, this scheme requires the estimation of the local speed of propagation at every grid-point, a task that is computationally intensive, particularly with problems of high dimensionality. By considering more precise information about the local speed of propagation, an even less dissipative scheme was generated in [21].

Recently we introduced in [5] new and efficient central schemes for multidimensional HJ equations. These nonoscillatory, nonstaggered schemes were first- and second-order accurate and were designed to scale well with an increasing dimension. Efficiency was obtained by carefully choosing the location of the evolution points and by using a one-dimensional projection step. Avoiding staggering by adding an additional projection step is an idea which we already utilized in the framework of conservation laws [16].

In this work we introduce third- and fifth-order accurate schemes for approximating solutions of multidimensional HJ equations. These are the *first* central schemes for such equations of order greater than two. This work is the HJ analogue to the corresponding works in conservation laws: an ENO-based central scheme [4] and the central WENO (CWENO) central schemes [25, 26, 27]. We announced a preliminary version of the one-dimensional results in a recent proceedings publication [6].

The structure of this paper is as follows. We start in section 2 with the derivation of our one-dimensional schemes. A third-order WENO reconstruction scheme is presented in section 2.2. This scheme requires a fourth-order reconstruction of the point-values and a third-order reconstruction of the derivatives at the evolution points. Even though the optimal location of the evolution points in one dimension is in the center of the interval, in order to prepare the grounds for the multidimensional schemes we write a reconstruction for an arbitrary location of the evolution points. A fifth-order method is then presented in section 2.3.

We turn to the multidimensional framework in section 3. Here there is flexibility in the reconstruction step. For simplicity we carry out most of the discussion in two space dimensions. Extensions to more than two space dimensions are presented in section 3.4. First, we provide a brief outline of the general structure of two-dimensional central schemes in section 3.1. The main remaining ingredient, the reconstruction step, is then described in the following two sections. For a two-dimensional third-order scheme we present in section 3.2 two ways to obtain a high-order reconstruction of the approximate solution at the evolution points. The first option in section 3.2.1 is based on a genuinely two-dimensional reconstruction. An alternative dimension-by-dimension approach is based on a sequence of one-dimensional reconstructions and is presented in section 3.2.2. Our numerical results show that both approaches are essentially equivalent. Hence, the rest of the paper deals with the dimension-by-dimension reconstruction. A fifth-order dimension-by-dimension extension of the one-dimensional scheme in section 2.3 to two dimensions is then presented in section 3.3. Since the solution at the next time step is computed at grid-points that are different from those on which the data is given, we reproject the evolved solution back onto the original grid-points. Different ways to approach this reprojection step are discussed in section 3.2.3.

We conclude in section 4 with several numerical examples in one, two, and three space dimensions that confirm the expected order of accuracy and the high-resolution nature of our scheme. We compare our results with the scheme of Jiang and Peng [17]. We also study the convergence rate after the emergence of the discontinuities in the solution.

2. One-dimensional schemes.

2.1. One-dimensional central schemes.

Consider the one-dimensional HJ equation of the form

$$\phi_t(x,t) + H\left(\phi_x\right) = 0, \qquad x \in \mathbb{R}. \tag{2.1}$$

We are interested in approximating solutions of (2.1) subject to the initial data $\phi(x, t = 0) = \phi_0(x)$. For simplicity we assume a uniform grid in space and time with mesh spacings Δx and Δt, respectively. Denote the grid-points by $x_i = i\Delta x$, $t^n = n\Delta t$, and the fixed mesh ratio by $\lambda = \Delta t/\Delta x$. Let φ_i^n denote the approximate value of $\phi\left(x_i, t^n\right)$, and $\left(\varphi_x\right)_i^n$ denote the approximate value of the derivative $\phi_x\left(x_i, t^n\right)$. We define the forward and backward differencing as $\Delta^+\varphi_i^n := \varphi_{i+1}^n - \varphi_i^n$ and $\Delta^-\varphi_i^n := \varphi_i^n - \varphi_{i-1}^n$.

Assume that the approximate solution at time t^n, φ_i^n is given. A Godunov-type scheme for approximating the solution of (2.1) starts with a continuous piecewise-polynomial $\tilde{\varphi}(x, t^n)$ that is reconstructed from the data φ_i^n:

$$\tilde{\varphi}(x,t^n) = \sum_i P_{i+\frac{1}{2}}(x,t^n)\chi_{i+\frac{1}{2}}(x). \tag{2.2}$$

Here, $\chi_{i+1/2}(x)$ is the characteristic function of the interval $[x_i, x_{i+1}]$, and $P_{i+1/2}(x, t^n)$ is a polynomial of a suitable degree that satisfies the interpolation requirements

$$P_{i+\frac{1}{2}}(x_{i+\beta}, t^n) = \varphi_{i+\beta}^n, \quad \beta = 0, 1.$$

The reconstruction (2.2) is then evolved from time t^n to time t^{n+1} according to (2.1) and is sampled at the half-integer grid-points $\{x_{i+1/2}\}$, where the reconstruction is smooth (as long as the CFL condition $\lambda\left|H'\left(\varphi_x\right)\right| \leq 1/2$ is satisfied):

$$\varphi_{i+\frac{1}{2}}^{n+1} = \varphi_{i+\frac{1}{2}}^n - \int_{t^n}^{t^{n+1}} H\left(\tilde{\varphi}_x\left(x_{i+\frac{1}{2}}, \tau\right)\right) d\tau. \tag{2.3}$$

The point-value $\varphi_{i+1/2}^n$ is obtained by sampling (2.2) at $x_{i+1/2}$; i.e., $\varphi_{i+1/2}^n = \tilde{\varphi}(x_{i+1/2}, t^n)$. Since the evolution step (2.3) is done at points where the solution is smooth, we can approximate the time integral on the right-hand side (RHS) of (2.3) using a sufficiently accurate quadrature rule. For example, for a third- and fourth-order method, this integral can be replaced by a Simpson's quadrature,

$$(2.4)\quad \int_{t^n}^{t^{n+1}} H\left(\tilde{\varphi}_x\left(x_{i+\frac{1}{2}},\tau\right)\right)d\tau \approx \frac{\Delta t}{6}\left[H\left(\varphi'^{\,n}_{i+\frac{1}{2}}\right)+4H\left(\varphi'^{\,n+\frac{1}{2}}_{i+\frac{1}{2}}\right)+H\left(\varphi'^{\,n+1}_{i+\frac{1}{2}}\right)\right].$$

The derivative at time t^n, $\varphi'^{\,n}_{i+1/2}$, is obtained by sampling the derivative of the reconstruction (2.2), i.e., $\varphi'^{\,n}_{i+1/2}=\tilde{\varphi}'(x_{i+1/2},t^n)$. The intermediate values of the derivative in time, $\varphi'^{\,n+1/2}_{i+1/2}$ and $\varphi'^{\,n+1}_{i+1/2}$, which are required in the quadrature (2.4), can be predicted using a Taylor expansion or with a Runge–Kutta (RK) method. Alternatively, (2.1) can be treated as a semidiscrete equation by replacing the spatial derivatives with their numerical approximations and integrating in time via an RK method.

The only remaining ingredient to specify is the reconstruction (2.2). Below we present two reconstructions. The first is a fourth-order reconstruction of the point-values and the derivatives, which leads to a third-order scheme, and the second is a sixth-order reconstruction that results in a fifth-order scheme.

Remarks.

1. In order to return to the original grid, we project $\varphi^{n+1}_{i+1/2}$ back onto the integer grid-points $\{x_i\}$ to end up with φ^{n+1}_i. This projection is accomplished with the same reconstruction used to approximate $\varphi^n_{i+1/2}$ from φ^n_i.

2. In order to maximize the size of the time step, the evolution points should be taken as far as possible from the singularities in the reconstructed piecewise polynomial. In one dimension the appropriate evolution point is located at $x_{i+1/2}$. In d dimensions with a uniform grid with spacing Δx, the optimal evolution points are located at $x_{i+\alpha}=x_i+\alpha\Delta x$ in each direction, where $\alpha=1/(d+\sqrt{d})$ (see [5]). One of the multidimensional schemes we present in section 3 is based on one-dimensional reconstructions. Hence, in order to prepare the grounds for the multidimensional setup, we write the one-dimensional reconstruction in this section, assuming that the evolution points are $x_{i\pm\alpha}$. The reader should keep in mind that in one dimension, $\alpha=1/2$.

3. We would like to point out that one does not need to fully reconstruct the polynomials $P_{i+1/2}(x,t^n)$. The only values that the scheme requires are the approximated point-values $\varphi^n_{i+1/2}=\tilde{\varphi}(x_{i+1/2},t^n)$ and the approximated derivatives $\varphi'_{i+1/2}=\tilde{\varphi}'(x_{i+1/2})$. Hence, in the rest of the paper whenever we refer to reconstruction steps we directly treat the recovery of these two quantities.

2.2. A third-order scheme.

A third-order scheme is generated by combining a third-order accurate ODE solver in time, for predicting the intermediate values of the derivatives in (2.4), with a sufficiently high-order reconstruction in space.

Given φ^n_i, in order to invoke (2.3), we should compute two quantities in every time step: the point-values at the evolution points, $\varphi_{i\pm\alpha}$, and the derivatives $\varphi'_{i\pm\alpha}$. In order to obtain a third-order scheme, the approximations of the point-values should be fourth-order accurate, and the approximation of the derivatives should be third-order accurate. In this scheme, the reconstruction of the point-values is done in locations that are staggered with respect to the location of the data. The reconstruction of the derivatives, which is required in every step of the ODE solver, is done at the same points where the data is given. Since we need two types of reconstructions and due to symmetry considerations, we derive a fourth-order approximation of the derivatives.

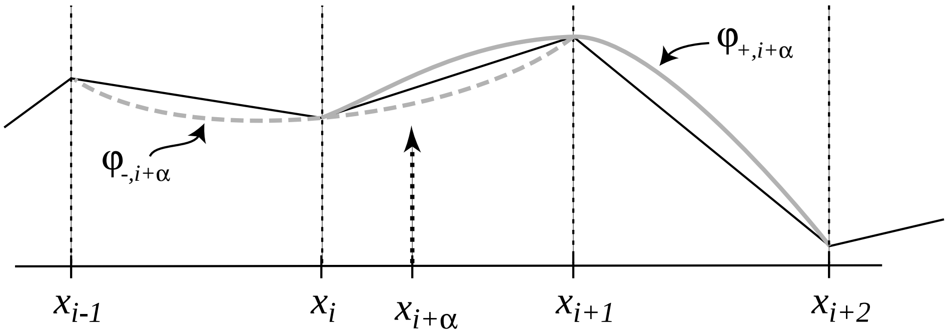


FIG. 2.1. *The two interpolants used for the third-order reconstruction at the evolution point at* $x_{i+\alpha}$.

Obviously, this more accurate reconstruction of the derivatives does not increase the order of accuracy of the scheme, but it does reduce the error.

2.2.1. The reconstruction of $\varphi_{i\pm\alpha}$ from φ_i. A fourth-order reconstruction of $\varphi_{i+\alpha}$ can be obtained by considering a convex combination of two quadratic polynomials, each of which requires the evaluation of φ on a three-point stencil. One quadratic polynomial $\varphi_-(x)$ is constructed on a stencil that is left-biased with respect to $x_{i+\alpha}$, $\{x_{i-1}, x_i, x_{i+1}\}$, while the other polynomial $\varphi_+(x)$ is constructed on a right-biased stencil, $\{x_i, x_{i+1}, x_{i+2}\}$; see Figure 2.1. We set

$$\begin{aligned} \varphi_{-,i+\alpha} &= \left(\frac{-\alpha+\alpha^2}{2}\right)\varphi_{i-1} + \left(1-\alpha^2\right)\varphi_i + \left(\frac{\alpha+\alpha^2}{2}\right)\varphi_{i+1}, \\ \varphi_{+,i+\alpha} &= \left(\frac{2-3\alpha+\alpha^2}{2}\right)\varphi_i + \left(2\alpha-\alpha^2\right)\varphi_{i+1} + \left(\frac{-\alpha+\alpha^2}{2}\right)\varphi_{i+2}. \end{aligned} \tag{2.5}$$

For smooth φ, a straightforward computation shows that $\varphi_{\pm,i+\alpha} = \varphi\left(x_{i+\alpha}\right) + O(\Delta x^3)$ and

$$\frac{1}{3}\left(2-\alpha\right)\varphi_{-,i+\alpha} + \frac{1}{3}\left(1+\alpha\right)\varphi_{+,i+\alpha} = \varphi\left(x_{i+\alpha}\right) + O\left(\Delta x^4\right).$$

Similarly, the reconstruction of $\varphi_{i-\alpha}$ is obtained using the quadratic polynomials $\varphi_-(x)$ based on the left-biased stencil enclosing $x_{i-\alpha}$, $\{x_{i-2}, x_{i-1}, x_i\}$, and $\varphi_+(x)$ based on the right-biased stencil $\{x_{i-1}, x_i, x_{i+1}\}$:

$$\begin{aligned} \varphi_{-,i-\alpha} &= \left(\frac{-\alpha+\alpha^2}{2}\right)\varphi_{i-2} + \left(2\alpha-\alpha^2\right)\varphi_{i-1} + \left(\frac{2-3\alpha+\alpha^2}{2}\right)\varphi_i, \\ \varphi_{+,i-\alpha} &= \left(\frac{\alpha+\alpha^2}{2}\right)\varphi_{i-1} + \left(1-\alpha^2\right)\varphi_i + \left(\frac{-\alpha+\alpha^2}{2}\right)\varphi_{i+1}. \end{aligned} \tag{2.6}$$

This time, $\varphi_{\pm,i-\alpha} = \varphi\left(x_{i-\alpha}\right) + O(\Delta x^3)$ and

$$\frac{1}{3}\left(1+\alpha\right)\varphi_{-,i-\alpha} + \frac{1}{3}\left(2-\alpha\right)\varphi_{+,i-\alpha} = \varphi\left(x_{i-\alpha}\right) + O\left(\Delta x^4\right).$$

A fourth-order WENO estimate of $\varphi_{i\pm\alpha}$ is therefore given by the convex combination

$$\varphi_{i\pm\alpha} = w^-_{i\pm\alpha}\varphi_{-,i\pm\alpha} + w^+_{i\pm\alpha}\varphi_{+,i\pm\alpha}, \tag{2.7}$$

where the weights satisfy $w^-_{i\pm\alpha} + w^+_{i\pm\alpha} = 1$, $w^\pm_{i\pm\alpha} \geq 0$, $\forall i$. In smooth regions we would like to satisfy $w^-_{i+\alpha} = w^+_{i-\alpha} \approx (2-\alpha)/3$ and $w^+_{i+\alpha} = w^-_{i-\alpha} \approx (1+\alpha)/3$ to attain an $O(\Delta x^4)$ error. When the stencil supporting $\varphi_{i\pm\alpha}$ contains a discontinuity, the weight of the more oscillatory polynomial should vanish. Following [18, 32], these requirements are met by setting

$$w^k_{i\pm\alpha} = \frac{\alpha^k_{i\pm\alpha}}{\sum_l \alpha^l_{i\pm\alpha}}, \qquad \alpha^k_{i\pm\alpha} = \frac{c^k_{i\pm\alpha}}{\left(\epsilon + S^k_{i\pm\alpha}\right)^p}, \tag{2.8}$$

where $k, l \in \{+,-\}$. The constants are independent of the grid index i and are given by $c^-_{i+\alpha} = c^+_{i-\alpha} = (2-\alpha)/3$, $c^+_{i+\alpha} = c^-_{i-\alpha} = (1+\alpha)/3$. We choose ϵ as 10^{-6} to prevent the denominator in (2.8) from vanishing, and set $p = 2$ (see [18]). The smoothness measures $S^\pm_i$ should be large when φ is nearly singular. Following [18], we take $S_{i\pm\alpha}$ to be the sum of the squares of the L^2-norms of the derivatives on the stencil supporting $\varphi_\pm$. If we approximate the first derivative at x_i by $\Delta^+\varphi_i/\Delta x$, the second derivative by $\Delta^+\Delta^-\varphi_i/(\Delta x)^2$, and define the smoothness measure

$$S_i\,[r,s] = \Delta x \sum_{j=r}^{s} \left(\frac{1}{\Delta x}\Delta^+\varphi_{i+j}\right)^2 + \Delta x \sum_{j=r+1}^{s} \left(\frac{1}{\Delta x^2}\Delta^+\Delta^-\varphi_{i+j}\right)^2, \tag{2.9}$$

then we have $S^-_{i+\alpha} = S_i\,[-1,0]$, $S^+_{i+\alpha} = S_i\,[0,1]$, $S^-_{i-\alpha} = S_i\,[-2,-1]$, and $S^+_{i-\alpha} = S_i\,[-1,0]$.

For future reference we label the reconstruction in this section with the procedural form

$$\varphi_{i\pm\alpha} = \text{reconstruct\_}\varphi\text{\_1D\_3}\,(i, \pm\alpha, \varphi)\,, \tag{2.10}$$

where φ is the one-dimensional array $(\varphi_1, \ldots, \varphi_N)$. This notation will be used in the dimension-by-dimension reconstructions in section 3.

2.2.2. The reconstruction of $\varphi'_{i\pm\alpha}$ from $\varphi_{i\pm\alpha}$. The values of φ that we recovered in the previous step at the regularly spaced locations $\{x_{i\pm\alpha}\}$ can be used to recover the derivative $\varphi'_{i\pm\alpha}$ via a (noncentral) WENO reconstruction. To obtain a fourth-order WENO approximation of $\varphi'_{i\pm\alpha}$, we write a convex combination of three quadratic interpolants: $\varphi'_{-,i\pm\alpha}$ on the stencil $\{x_{i-2\pm\alpha}, x_{i-1\pm\alpha}, x_{i\pm\alpha}\}$, $\varphi'_{0,i\pm\alpha}$ on $\{x_{i-1\pm\alpha}, x_{i\pm\alpha}, x_{i+1\pm\alpha}\}$, and $\varphi'_{+,i\pm\alpha}$ on $\{x_{i\pm\alpha}, x_{i+1\pm\alpha}, x_{i+2\pm\alpha}\}$. For smooth φ,

$$\begin{aligned}
\varphi'_{-,i\pm\alpha} &= \frac{1}{2\Delta x}(\varphi_{i-2\pm\alpha} - 4\varphi_{i-1\pm\alpha} + 3\varphi_{i\pm\alpha}) = \varphi'\,(x_{i\pm\alpha}) + O\left(\Delta x^2\right), \\
\varphi'_{0,i\pm\alpha} &= \frac{1}{2\Delta x}(\varphi_{i+1\pm\alpha} - \varphi_{i-1\pm\alpha}) = \varphi'\,(x_{i\pm\alpha}) + O\left(\Delta x^2\right), \\
\varphi'_{+,i\pm\alpha} &= \frac{1}{2\Delta x}(-3\varphi_{i\pm\alpha} + 4\varphi_{i+1\pm\alpha} - \varphi_{i+2\pm\alpha}) = \varphi'\,(x_{i\pm\alpha}) + O\left(\Delta x^2\right).
\end{aligned} \tag{2.11}$$

A straightforward computation yields

$$\frac{1}{6}\varphi'_{-,i\pm\alpha} + \frac{2}{3}\varphi'_{0,i\pm\alpha} + \frac{1}{6}\varphi'_{+,i\pm\alpha} = \varphi'\,(x_{i\pm\alpha}) + O\left(\Delta x^4\right).$$

The fourth-order WENO estimate of $\varphi'_{i\pm\alpha}$ from $\varphi_{i\pm\alpha}$ is therefore

$$\varphi'_{i\pm\alpha} = w^-_{i\pm\alpha}\varphi'_{-,i\pm\alpha} + w^0_{i\pm\alpha}\varphi'_{0,i\pm\alpha} + w^+_{i\pm\alpha}\varphi'_{+,i\pm\alpha}, \tag{2.12}$$

where the weights w are of the form (2.8), with $k, l \in \{+, 0, -\}$, $c^- = c^+ = 1/6, c^0 = 2/3$, and the oscillatory indicators are $S^-_{i\pm\alpha} = S_{i\pm\alpha}[-2,-1]$, $S^0_{i\pm\alpha} = S_{i\pm\alpha}[-1,0]$, and $S^+_{i\pm\alpha} = S_{i\pm\alpha}[0,1]$.

For future reference we label the above reconstruction of $\varphi'_{i\pm\alpha}$ with the procedural form

$$\varphi'_{i\pm\alpha} = \text{reconstruct\_}\varphi'\text{\_1D\_3}\left(i, \pm\alpha, \varphi_{\pm\alpha}\right), \tag{2.13}$$

where $\varphi_{\pm\alpha}$ is the one-dimensional array $(\varphi_{1\pm\alpha}, \dots, \varphi_{N\pm\alpha})$.

We would like to summarize the one-dimensional third-order algorithm in the following, where $\text{RK}(\varphi^n_{i\pm\alpha}, \varphi'^{\,n}_{i\pm\alpha}, \Delta t)$ is the third-order Runge–Kutta method that integrates (2.1) and is used to predict the intermediate values of the derivatives. Each internal step of the RK method will require additional reconstructions of $\varphi'_{i\pm\alpha}$ from that step's $\varphi_{i\pm\alpha}$.

ALGORITHM 2.1. *Assume that* $\{\varphi^n_i\}$ *are given.*

(a) *Reconstruct:*

$$\begin{aligned}\varphi^n_{i\pm\alpha} &= \text{reconstruct\_}\varphi\text{\_1D\_3}\left(i, \pm\alpha, \varphi^n\right),\\ \varphi'^{\,n}_{i\pm\alpha} &= \text{reconstruct\_}\varphi'\text{\_1D\_3}(i, \pm\alpha, \varphi^n_{i\pm\alpha}).\end{aligned}$$

(b) *Integrate:*

$$\begin{aligned}\varphi^{n+\frac{1}{2}}_{i\pm\alpha} &= RK\left(\varphi^n_{i\pm\alpha}, \varphi'^{\,n}_{i\pm\alpha}, \Delta t/2\right),\\ \varphi'^{\,n+\frac{1}{2}}_{i\pm\alpha} &= \text{reconstruct\_}\varphi'\text{\_1D\_3}(i, \pm\alpha, \varphi^{n+\frac{1}{2}}_{i\pm\alpha}),\\ \varphi^{n+1}_{i\pm\alpha} &= RK\left(\varphi^n_{i\pm\alpha}, \varphi'^{\,n}_{i\pm\alpha}, \Delta t\right),\\ \varphi'^{\,n+1}_{i\pm\alpha} &= \text{reconstruct\_}\varphi'\text{\_1D\_3}\left(i, \pm\alpha, \varphi^{n+1}_{i\pm\alpha}\right),\\ \varphi^{n+1}_{i\pm\alpha} &= \varphi^n_{i\pm\alpha} + \frac{\Delta t}{6}\left[H\left(\varphi'^{\,n}_{i\pm\alpha}\right) + 4H(\varphi'^{\,n+\frac{1}{2}}_{i\pm\alpha}) + H\left(\varphi'^{\,n+1}_{i\pm\alpha}\right)\right].\end{aligned}$$

(c) *Reproject:*

$$\varphi^{n+1}_i = \text{reconstruct\_}\varphi\text{\_1D\_3}\left(i, \mp\alpha, \varphi^{n+1}_{i\pm\alpha}\right).$$

Remark. It is possible to replace the Simpson's quadrature in the integration step with a single RK time step, $\varphi^{n+1}_{i\pm\alpha} = \text{RK}(\varphi^n_{i\pm\alpha}, \varphi'^{\,n}_{i\pm\alpha}, \Delta t)$. Our simulations show that this choice reduces the complexity of the computation but also reduces its accuracy.

2.3. A fifth-order scheme. In order to obtain a fifth-order scheme, we need a sixth-order approximation of the point-values of φ, a fifth-order approximation of the derivative φ', and a higher-order prediction of the intermediate derivatives which appear in the quadrature formula. Due to arguments similar to those given in section 2.2, we again derive a more accurate reconstruction of the derivatives, which in this case is sixth-order.

We start with the reconstruction of $\varphi_{i+\alpha}$ from φ_i. We write sixth-order interpolants as a convex combination of three cubic interpolants, each of which requires the evaluation of φ on a four-point stencil. We use the polynomials $\varphi_-(x)$ defined on the left-biased stencil $\{x_{i-2}, x_{i-1}, x_i, x_{i+1}\}$, $\varphi_0(x)$ defined on the centered stencil $\{x_{i-1}, x_i, x_{i+1}, x_{i+2}\}$, and $\varphi_+(x)$ defined on the right-biased stencil $\{x_i, x_{i+1}, x_{i+2}, x_{i+3}\}$; see Figure 2.2. For smooth φ,

$$\begin{aligned}\varphi_{-,i+\alpha} &= a_1\varphi_{i-2} + a_2\varphi_{i-1} + a_3\varphi_i + a_4\varphi_{i+1} = \varphi\left(x_{i+\alpha}\right) + O\left(\Delta x^4\right),\\ \varphi_{0,i+\alpha} &= a_5\varphi_{i-1} + a_6\varphi_i + a_7\varphi_{i+1} + a_8\varphi_{i+2} = \varphi\left(x_{i+\alpha}\right) + O\left(\Delta x^4\right),\\ \varphi_{+,i+\alpha} &= a_9\varphi_i + a_{10}\varphi_{i+1} + a_{11}\varphi_{i+2} + a_{12}\varphi_{i+3} = \varphi\left(x_{i+\alpha}\right) + O\left(\Delta x^4\right),\end{aligned} \tag{2.14}$$

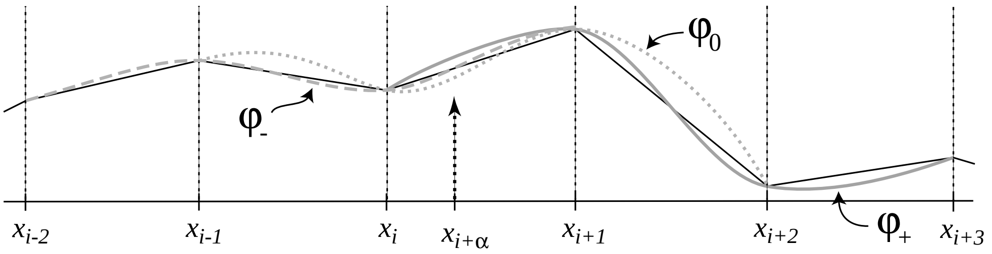


Fig. 2.2. *The three interpolants used for the fifth-order reconstruction $\varphi_{i+\alpha}$ at the evolution point at $x_{i+\alpha}$. In this example, because of the large gradient between x_{i+1} and x_{i+2}, the interpolant φ_- will have the strongest contribution to the CWENO reconstruction at $x_{i+\alpha}$.*

where the constants are given by

$$\begin{aligned}
a_1 &= \frac{1}{6}\alpha - \frac{1}{6}\alpha^3, & a_2 &= -\alpha + \frac{1}{2}\alpha^2 + \frac{1}{2}\alpha^3,\\
a_3 &= 1 + \frac{1}{2}\alpha - \alpha^2 - \frac{1}{2}\alpha^3, & a_4 &= \frac{1}{3}\alpha + \frac{1}{2}\alpha^2 + \frac{1}{6}\alpha^3,\\
a_5 &= -\frac{1}{3}\alpha + \frac{1}{2}\alpha^2 - \frac{1}{6}\alpha^3, & a_6 &= 1 - \frac{1}{2}\alpha - \alpha^2 + \frac{1}{2}\alpha^3,\\
a_7 &= \alpha + \frac{1}{2}\alpha^2 - \frac{1}{2}\alpha^3, & a_8 &= -\frac{1}{6}\alpha + \frac{1}{6}\alpha^3 = -a_1,\\
a_9 &= 1 - \frac{11}{6}\alpha + \alpha^2 - \frac{1}{6}\alpha^3, & a_{10} &= 3\alpha - \frac{5}{2}\alpha^2 + \frac{1}{2}\alpha^3,\\
a_{11} &= -\frac{3}{2}\alpha + 2\alpha^2 - \frac{1}{2}\alpha^3, & a_{12} &= \frac{1}{3}\alpha - \frac{1}{2}\alpha^2 + \frac{1}{6}\alpha^3.
\end{aligned}$$

At $x_{i-\alpha}$ we have

$$\begin{aligned}
\varphi_{-,i-\alpha} &= a_{12}\varphi_{i-3} + a_{11}\varphi_{i-2} + a_{10}\varphi_{i-1} + a_9\varphi_i = \varphi\left(x_{i-\alpha}\right) + O\left(\Delta x^4\right),\\
\varphi_{0,i-\alpha} &= a_8\varphi_{i-2} + a_7\varphi_{i-1} + a_6\varphi_i + a_5\varphi_{i+1} = \varphi\left(x_{i-\alpha}\right) + O\left(\Delta x^4\right),\\
\varphi_{+,i-\alpha} &= a_4\varphi_{i-1} + a_3\varphi_i + a_2\varphi_{i+1} + a_1\varphi_{i+2} = \varphi\left(x_{i-\alpha}\right) + O\left(\Delta x^4\right).
\end{aligned} \tag{2.15}$$

A straightforward computation yields

$$c^-_{i\pm\alpha}\varphi_{-,i\pm\alpha} + c^0_{i\pm\alpha}\varphi_{0,i\pm\alpha} + c^+_{i\pm\alpha}\varphi_{+,i\pm\alpha} = \varphi\left(x_{i\pm\alpha}\right) + O\left(\Delta x^6\right),$$

where

$$\begin{aligned}
c^-_{i+\alpha} = c^+_{i-\alpha} &= \frac{1}{20}\alpha^2 - \frac{1}{4}\alpha + \frac{3}{10},\\
c^0_{i\pm\alpha} &= -\frac{1}{10}\alpha^2 + \frac{1}{10}\alpha + \frac{3}{5},\\
c^+_{i+\alpha} = c^-_{i-\alpha} &= \frac{1}{20}\alpha^2 + \frac{3}{20}\alpha + \frac{1}{10}.
\end{aligned} \tag{2.16}$$

A sixth-order reconstruction of $\varphi_{i\pm\alpha}$ is therefore given by

$$\varphi_{i\pm\alpha} = w^-_{i\pm\alpha}\varphi_{-,i\pm\alpha} + w^0_{i\pm\alpha}\varphi_{0,i\pm\alpha} + w^+_{i\pm\alpha}\varphi_{+,i\pm\alpha}, \tag{2.17}$$

where the weights w^k are given by (2.8) with $k, l \in \{+, 0, -\}$, and the constants c^k are given by (2.16). The oscillatory indicators are given via (2.9) by $S^-_{i\pm\alpha} = S_i\,[-2, 0]$, $S^0_{i\pm\alpha} = S_i\,[-1, 1]$, and $S^+_{i\pm\alpha} = S_i\,[0, 2]$.

A sixth-order approximation of $\varphi'_{i\pm\alpha}$ from $\varphi_{i\pm\alpha}$ is written as a convex combination of four cubic interpolants. This reconstruction is similar to the third-order case and is based on a noncentral WENO reconstruction. We skip the details and summarize the result:

$$\varphi'_{i\pm\alpha} = w^1_{i\pm\alpha}\varphi'_{1,i\pm\alpha} + w^2_{i\pm\alpha}\varphi'_{2,i\pm\alpha} + w^3_{i\pm\alpha}\varphi'_{3,i\pm\alpha} + w^4_{i\pm\alpha}\varphi'_{4,i\pm\alpha}, \tag{2.18}$$

where

$$\begin{aligned}
\varphi'_{1,i\pm\alpha} &= \frac{1}{6\Delta x}(-2\varphi_{i-3\pm\alpha} + 9\varphi_{i-2\pm\alpha} - 18\varphi_{i-1\pm\alpha} + 11\varphi_{i\pm\alpha}),\\
\varphi'_{2,i\pm\alpha} &= \frac{1}{6\Delta x}(\varphi_{i-2\pm\alpha} - 6\varphi_{i-1\pm\alpha} + 3\varphi_{i\pm\alpha} + 2\varphi_{i+1\pm\alpha}),\\
\varphi'_{3,i\pm\alpha} &= \frac{1}{6\Delta x}(-2\varphi_{i-1\pm\alpha} - 3\varphi_{i\pm\alpha} + 6\varphi_{i+1\pm\alpha} - \varphi_{i+2\pm\alpha}),\\
\varphi'_{4,i\pm\alpha} &= \frac{1}{6\Delta x}(-11\varphi_{i\pm\alpha} + 18\varphi_{i+1\pm\alpha} - 9\varphi_{i+2\pm\alpha} + 2\varphi_{i+3\pm\alpha}).
\end{aligned}$$

Here the weights w^k are given by (2.8) with $c_1 = c_4 = 1/20, c_2 = c_3 = 9/20$, $S^1_{i\pm\alpha} = S_{i\pm\alpha}\,[-3, -1]$, $S^2_{i\pm\alpha} = S_{i\pm\alpha}\,[-2, 0]$, $S^3_{i\pm\alpha} = S_{i\pm\alpha}\,[-1, 1]$, and $S^4_{i\pm\alpha} = S_{i\pm\alpha}\,[0, 2]$.

Notation.

1. We label the reconstruction of the point-values, (2.17), as

$$\varphi_{i\pm\alpha} = \text{reconstruct\_}\varphi\text{\_1D\_5}\,(i, \pm\alpha, \varphi)\,, \tag{2.19}$$

where φ is the one-dimensional array $(\varphi_1, \dots, \varphi_N)$.

2. We label the reconstruction of $\varphi'_{i\pm\alpha}$, (2.18), as

$$\varphi'_{i\pm\alpha} = \text{reconstruct\_}\varphi'\text{\_1D\_5}\,(i, \pm\alpha, \varphi_{\pm\alpha})\,, \tag{2.20}$$

where $\varphi_{\pm\alpha}$ is the one-dimensional array $(\varphi_{1\pm\alpha}, \dots, \varphi_{N\pm\alpha})$.

Remarks.

1. To conclude, the fifth-order method is given by Algorithm 2.1, where the fourth-order reconstructions are replaced by the sixth-order reconstructions (2.19)–(2.20). As is, this scheme is only fourth-order in time. A higher-order method in time can be easily obtained by replacing Simpson's quadrature with a more accurate quadrature and computing the sixth-order approximations for the point-values and the derivatives at the new quadrature points.

2. We choose to predict the intermediate values of the derivatives in time using the fourth-order strong stability preserving (SSP) RK scheme of [12]. For $s \in \left\{\frac{1}{2}, 1\right\}$, the SSP-RK scheme is given by

$$\begin{aligned}
\varphi^{(1)} &= \varphi^n - \frac{1}{2}s\Delta t H\left(\varphi^n_x\right),\\
\varphi^{(2)} &= \frac{649}{1600}\varphi^n + \frac{10890423}{25193600}s\Delta t H\left(\varphi^n_x\right) + \frac{951}{1600}\varphi^{(1)} - \frac{5000}{7873}s\Delta t H(\varphi^{(1)}_x),
\end{aligned}$$

$$\begin{aligned}\varphi^{(3)} &= \frac{53989}{2500000}\varphi^{n} + \frac{102261}{5000000}s\Delta tH\left(\varphi_x^n\right) + \frac{4806213}{20000000}\varphi^{(1)}\\ &\quad + \frac{5121}{20000}s\Delta tH(\varphi_x^{(1)}) + \frac{23619}{32000}\varphi^{(2)} + \frac{7873}{10000}s\Delta tH(\varphi_x^{(2)}),\\ \varphi^{n+s} &= \frac{1}{5}\varphi^{n} - \frac{1}{10}s\Delta tH\left(\varphi_x^n\right) + \frac{6127}{30000}\varphi^{(1)} + \frac{1}{6}s\Delta tH(\varphi_x^{(1)}) + \frac{7873}{30000}\varphi^{(2)}\\ &\quad + \frac{1}{3}\varphi^{(3)} - \frac{1}{6}s\Delta tH(\varphi_x^{(3)}).\end{aligned}$$

Alternatively, the natural continuous extension of the RK method [39] can be used to produce the intermediate values $\varphi'^{\,n+\frac{1}{2}}$ and $\varphi'^{\,n+1}$ with a single RK step, though we observe that errors are somewhat larger in this case.

3. Multidimensional schemes.

3.1. Two-dimensional central schemes. Consider the two-dimensional HJ equation of the form

$$\phi_t + H(\nabla\phi) = 0, \qquad \vec{x} = (x_1, x_2) \in \mathbb{R}^2, \tag{3.1}$$

subject to the initial data $\phi(\vec{x}, t = 0) = \phi_0(\vec{x})$. Define $x_{i,j} := (x_1 + i\Delta x_1, x_2 + j\Delta x_2)$. Similarly to the one-dimensional setup, $\varphi_{i,j}$ will denote the approximation of ϕ at $x_{i,j}$. We define the two sets of grid-points, $I_+ = \{x_{i,j}, x_{i+1,j}, x_{i,j+1}\}$ and $I_- = \{x_{i,j}, x_{i-1,j}, x_{i,j-1}\}$, and denote by T_+, T_- the triangles with vertices I_+ and I_-, respectively. For simplicity we assume a uniform grid $\Delta x_1 = \Delta x_2 = \Delta x$.

Assume that the approximate solution at time t^n, $\varphi_{i,j}^n$, is given. Similarly to the one-dimensional setup in section 2.1, a Godunov-type scheme for approximating the solution of (3.1) starts with a continuous piecewise polynomial $\tilde{\varphi}(\vec{x}, t^n)$ that is reconstructed from the data $\varphi_{i,j}^n$,

$$\tilde{\varphi}(\vec{x}, t^n) = \sum_{i,j} P_{i,j}^{T_\pm}(\vec{x}, t^n)\chi_{T_\pm}(\vec{x}). \tag{3.2}$$

As usual, $\chi_{T_\pm}(\vec{x})$ is the characteristic function of the triangle $T_\pm$, and $P_{i,j}^{T_\pm}(\vec{x}, t^n)$ is a polynomial of a suitable degree that satisfies the interpolation requirements

$$P_{i,j}^{T_\pm}(\vec{x}_l, t^n) = \varphi(\vec{x}_l, t^n), \quad \vec{x}_l \in I_\pm$$

(see Figure 3.1). The reconstruction (3.2) is then evolved from time t^n to time t^{n+1} by (3.1) and sampled at the evolution points $\{x_{i\pm\alpha,j\pm\alpha}\}$. In two dimensions the choice $\alpha = 1/(2+\sqrt{2})$ guarantees that the solution remains smooth at the evolution point as long as the CFL condition $\frac{\Delta t}{\Delta x}\left|H'\left(\nabla\varphi\right)\right| < \alpha$ is satisfied. The evolved solution now reads

$$\varphi_{i\pm\alpha,j\pm\alpha}^{n+1} = \varphi_{i\pm\alpha,j\pm\alpha}^{n} - \int_{t^n}^{t^{n+1}} H\left(\nabla\tilde{\varphi}\left(x_{i\pm\alpha,j\pm\alpha}, \tau\right)\right)d\tau. \tag{3.3}$$

The point-values $\varphi_{i\pm\alpha,j\pm\alpha}^n$ are obtained by sampling (3.2) at $x_{i\pm\alpha,j\pm\alpha}$, i.e., $\varphi_{i\pm\alpha,j\pm\alpha}^n = \tilde{\varphi}(x_{i\pm\alpha,j\pm\alpha}, t^n)$. As in the one-dimensional case, the evolution points are in smooth regions, and therefore the integral on the RHS of (3.3) can be replaced with a sufficiently accurate quadrature such as the Simpson rule (2.4), which leads to a scheme that is

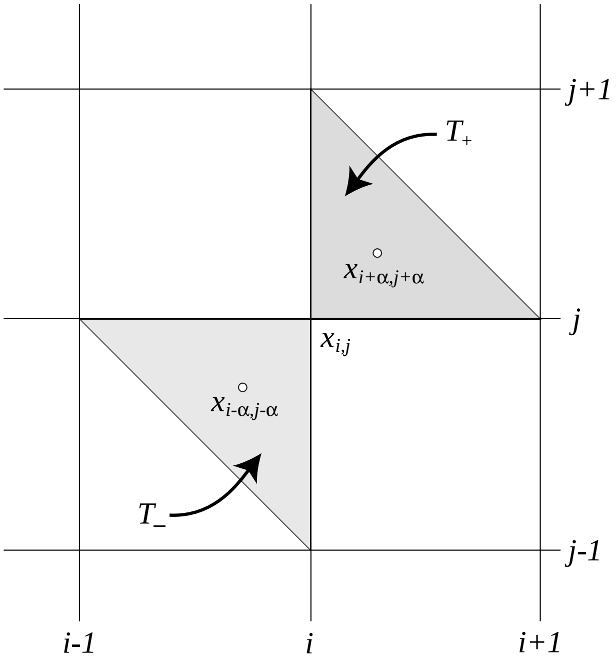


FIG. 3.1. *The location of the evolution points $x_{i\pm\alpha,j\pm\alpha}$ and the domain of definition of the interpolants $\varphi_{i\pm\alpha,j\pm\alpha}$ in two dimensions.*

fourth-order accurate in time. The derivatives at time t^n, $\varphi'^{\,n}_{i\pm\alpha,j\pm\alpha}$, are obtained by sampling the derivative of the reconstruction (3.2), i.e., $\varphi'^{\,n}_{i\pm\alpha,j\pm\alpha} = \tilde{\varphi}'(x_{i\pm\alpha,j\pm\alpha}, t^n)$. The other intermediate values of the derivative in time that are required in the quadrature can be predicted using a Taylor expansion or with a RK method in a way analogous to that for the one-dimensional case.

Remarks.

1. We present two different algorithms for constructing $\varphi_{i\pm\alpha,j\pm\alpha}$: two-dimensional interpolants defined on two-dimensional stencils and a dimension-by-dimension approach. We present both algorithms for the third-order scheme and extend the simpler dimension-by-dimension approach to fifth-order. Our numerical simulations in section 4 indicate that both reconstructions of $\varphi_{i\pm\alpha,j\pm\alpha}$ are of a comparable quality. In both approaches, the reconstruction of the derivatives $\nabla\varphi_{i\pm\alpha,j\pm\alpha}$ is done dimension-by-dimension.

2. We reproject $\varphi^{n+1}_{i+\alpha,j+\alpha}$ and $\varphi^{n+1}_{i-\alpha,j-\alpha}$ back onto the integer grid-points, obtaining $\varphi^{n+1}_{i,j}$. We present several ways to carry out this reprojection: a genuinely two-dimensional approach, a dimension-by-dimension strategy, and a reprojection along the diagonal line through $x_{i-\alpha,j-\alpha}$ and $x_{i+\alpha,j+\alpha}$.

3.2. Two-dimensional third-order schemes.

In order to obtain a third-order scheme, we need a fourth-order reconstruction of the point-values at the evolution points $x_{i\pm\alpha,j\pm\alpha}$.

3.2.1. A two-dimensional reconstruction of $\varphi_{i\pm\alpha,j\pm\alpha}$.

In this section we present a two-dimensional fourth-order reconstruction of the point-values $\varphi_{i\pm\alpha,j\pm\alpha}$. In principle, a two-dimensional cubic interpolant would provide a reconstruction with the desired accuracy. Such an interpolant is based on a ten-point stencil. As usual,

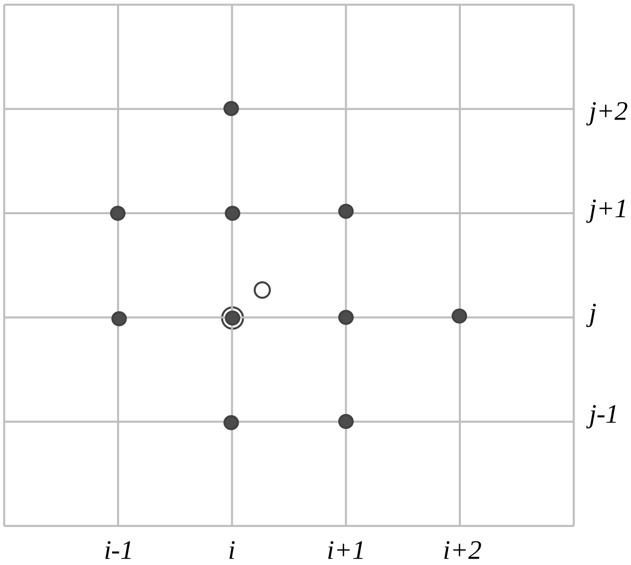


Fig. 3.2. *The ten-point stencil for the two-dimensional reconstruction of $\varphi_{i+\alpha,j+\alpha}$. The open circle shows the location of the evolution point at $x_{i+\alpha,j+\alpha}$.*

solving such a direct interpolation problem is unsatisfactory because spurious oscillations might develop as a result of the lack of smoothness in the solution. Instead, we generate a two-dimensional fourth-order reconstruction as a convex combination of four quadratic interpolants, each of which is based on a six-point stencil. We choose compact quadratic interpolants such that the union of all the six-point stencils is a compact ten-point stencil. Similarly to any WENO-type reconstruction, when singularities are present the six-point stencils containing the singularities are suppressed. In any case, we implicitly assume that the solution is sufficiently resolved such that the singularities in the solution are isolated in the sense that they do not occur along neighboring parallel cell edges. Singularities will in general occur along adjacent cell edges. There is a lot of flexibility in choosing the ten-point stencil as well as the different six-point stencils. Here, for the evolution point $x_{i+\alpha,j+\alpha}$ we choose the ten-point stencil shown in Figure 3.2. We also choose to use the four six-point stencils that are shown in Figure 3.3; obviously, the union of these stencils is the ten-point stencil in Figure 3.2. Furthermore, they all enclose the cell containing the evolution point, and they all cross different edges of the enclosing cell. A singularity along an edge will suppress two of these stencils, while a singularity in a corner will suppress three of these stencils.

Remarks.

1. The stencils for the evolution point at $x_{i-\alpha,j-\alpha}$ are obtained by a rotation of 180 degrees of the stencils in Figures 3.2–3.3.

2. We could use fewer than four stencils and still generate a scheme that will have the desired order of accuracy.

Given the four six-point stencils in Figure 3.3, a straightforward computation shows that third-order approximations for smooth φ at the evolution points $x_{i\pm\alpha,j\pm\alpha}$, $\varphi^k_{i\pm\alpha,j\pm\alpha} = \varphi\left(x_{i\pm\alpha}, y_{j\pm\alpha}\right) + O(\Delta x^3, \Delta y^3)\ \forall k \in \{1,2,3,4\}$ are obtained with

$$\varphi^1_{i\pm\alpha,j\pm\alpha} = a_1\varphi_{i,j} + a_2\varphi_{i\pm1,j} + a_2\varphi_{i,j\pm1} + a_3\varphi_{i\pm1,j\pm1} + a_4\varphi_{i\pm2,j} + a_4\varphi_{i,j\pm2}, \tag{3.4}$$

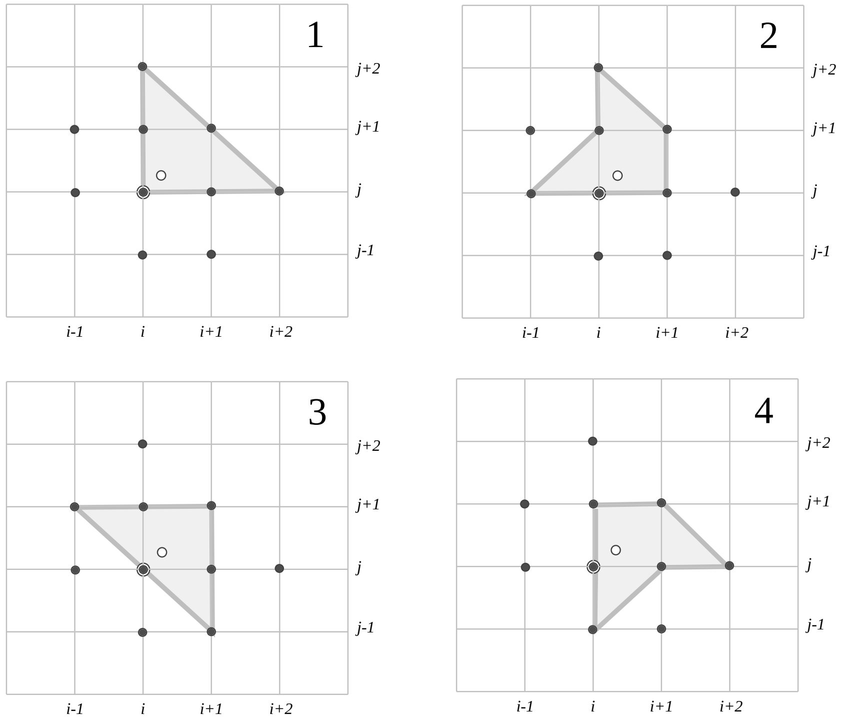


FIG. 3.3. *The four six-point stencils that cover the ten-point stencil for the two-dimensional reconstruction.*

$$\varphi^2_{i\pm\alpha,j\pm\alpha} = a_5\varphi_{i,j} + a_6\varphi_{i\pm1,j} + a_2\varphi_{i,j\pm1} + a_3\varphi_{i\pm1,j\pm1} + a_4\varphi_{i,j\pm2} + a_4\varphi_{i\mp1,j},$$
$$\varphi^3_{i\pm\alpha,j\pm\alpha} = a_7\varphi_{i,j} + a_2\varphi_{i\pm1,j} + a_2\varphi_{i,j\pm1} + a_8\varphi_{i\pm1,j\pm1} + a_4\varphi_{i\pm1,j\mp1} + a_4\varphi_{i\mp1,j\pm1},$$
$$\varphi^4_{i\pm\alpha,j\pm\alpha} = a_5\varphi_{i,j} + a_2\varphi_{i\pm1,j} + a_6\varphi_{i,j\pm1} + a_3\varphi_{i\pm1,j\pm1} + a_4\varphi_{i\pm2,j} + a_4\varphi_{i,j\mp1},$$

where

$$\begin{aligned} &a_1 = 1 - 3\alpha + 2\alpha^2, && a_2 = 2\alpha - 2\alpha^2, && a_3 = \alpha^2, \\ &a_4 = -\frac{1}{2}\alpha + \frac{1}{2}\alpha^2, && a_5 = 1 - \frac{3}{2}\alpha + \frac{1}{2}\alpha^2, && a_6 = \frac{1}{2}\alpha - \frac{1}{2}\alpha^2, \\ &a_7 = 1 - 2\alpha + \alpha^2, && a_8 = -\alpha + 2\alpha^2. \end{aligned} \tag{3.5}$$

The linear combination

$$\sum_{k=1}^{4} c_k \varphi^k_{i\pm\alpha,j\pm\alpha} = \varphi\left(x_{i\pm\alpha}, y_{j\pm\alpha}\right) + O\left(\Delta x^4, \Delta y^4\right)$$

is fourth-order accurate, provided that the constants c_i are taken as

$$c_1 = \frac{1}{3}\left(5\alpha - 1\right), \quad c_2 = c_4 = \frac{2}{3}\left(-2\alpha + 1\right), \quad c_3 = \alpha. \tag{3.6}$$

A two-dimensional CWENO reconstruction is a straightforward generalization of the one-dimensional case (compare with (2.7), (2.8)):

$$\varphi_{i\pm\alpha,j\pm\alpha} = \sum_{k=1}^{4} w^k_{i\pm\alpha,j\pm\alpha}\varphi^k_{i\pm\alpha,j\pm\alpha}.$$

Here

$$w^k_{i\pm\alpha,j\pm\alpha} = \frac{\alpha^k_{i\pm\alpha,j\pm\alpha}}{\sum_{l=1}^{4}\alpha^l_{i\pm\alpha,j\pm\alpha}}, \qquad \alpha^k_{i\pm\alpha,j\pm\alpha} = \frac{c_k}{\left(\epsilon + S^k_{i\pm\alpha,j\pm\alpha}\right)^p},$$

with the constants c_k given by (3.6). As usual, the smoothness measure for every stencil is taken as a normalized sum of the discrete L^2-norms of the derivatives. If we define the forward and backward differences $\Delta_x^+\varphi_{i,j} = \varphi_{i+1,j} - \varphi_{i,j}$, $\Delta_x^-\varphi_{i,j} = \varphi_{i,j} - \varphi_{i-1,j}$, $\Delta_y^+\varphi_{i,j} = \varphi_{i,j+1} - \varphi_{i,j}$, $\Delta_y^-\varphi_{i,j} = \varphi_{i,j} - \varphi_{i,j-1}$, then the smoothness measures for the evolution point $x_{i+\alpha,j+\alpha}$ are given by

$$\begin{aligned}
S^1_{i+\alpha,j+\alpha} &= \left(\Delta_x^+\varphi_{i,j}\right)^2 + \left(\Delta_x^+\varphi_{i+1,j}\right)^2 + \left(\Delta_x^+\varphi_{i,j+1}\right)^2 + \left(\Delta_y^+\varphi_{i,j}\right)^2 + \left(\Delta_y^+\varphi_{i,j+1}\right)^2 \\
&\quad + \left(\Delta_y^+\varphi_{i+1,j}\right)^2 + \frac{1}{\Delta x^2}\left[\left(\Delta_x^+\Delta_x^-\varphi_{i+1,j}\right)^2 + \left(\Delta_y^+\Delta_y^-\varphi_{i,j+1}\right)^2\right], \\
S^2_{i+\alpha,j+\alpha} &= \left(\Delta_x^+\varphi_{i,j}\right)^2 + \left(\Delta_x^+\varphi_{i-1,j}\right)^2 + \left(\Delta_x^+\varphi_{i,j+1}\right)^2 + \left(\Delta_y^+\varphi_{i,j}\right)^2 + \left(\Delta_y^+\varphi_{i,j+1}\right)^2 \\
&\quad + \left(\Delta_y^+\varphi_{i+1,j}\right)^2 + \frac{1}{\Delta x^2}\left[\left(\Delta_x^+\Delta_x^-\varphi_{i,j}\right)^2 + \left(\Delta_y^+\Delta_y^-\varphi_{i,j+1}\right)^2\right], \\
S^3_{i+\alpha,j+\alpha} &= \left(\Delta_x^+\varphi_{i,j}\right)^2 + \left(\Delta_x^+\varphi_{i,j+1}\right)^2 + \left(\Delta_x^+\varphi_{i-1,j+1}\right)^2 + \left(\Delta_y^+\varphi_{i,j}\right)^2 + \left(\Delta_y^+\varphi_{i+1,j}\right)^2 \\
&\quad + \left(\Delta_y^+\varphi_{i+1,j-1}\right)^2 + \frac{1}{\Delta x^2}\left[\left(\Delta_x^+\Delta_x^-\varphi_{i,j+1}\right)^2 + \left(\Delta_y^+\Delta_y^-\varphi_{i+1,j}\right)^2\right], \\
S^4_{i+\alpha,j+\alpha} &= \left(\Delta_x^+\varphi_{i,j}\right)^2 + \left(\Delta_x^+\varphi_{i+1,j}\right)^2 + \left(\Delta_x^+\varphi_{i,j+1}\right)^2 + \left(\Delta_y^+\varphi_{i,j}\right)^2 + \left(\Delta_y^+\varphi_{i,j-1}\right)^2 \\
&\quad + \left(\Delta_y^+\varphi_{i+1,j}\right)^2 + \frac{1}{\Delta x^2}\left[\left(\Delta_x^+\Delta_x^-\varphi_{i+1,j}\right)^2 + \left(\Delta_y^+\Delta_y^-\varphi_{i,j}\right)^2\right].
\end{aligned}$$

The smoothness measures for the evolution point $x_{i-\alpha,j-\alpha}$ are

$$\begin{aligned}
S^1_{i-\alpha,j-\alpha} &= \left(\Delta_x^+\varphi_{i-2,j}\right)^2 + \left(\Delta_x^+\varphi_{i-1,j}\right)^2 + \left(\Delta_x^+\varphi_{i-1,j-1}\right)^2 + \left(\Delta_y^+\varphi_{i,j-2}\right)^2 + \left(\Delta_y^+\varphi_{i,j-1}\right)^2 \\
&\quad + \left(\Delta_y^+\varphi_{i-1,j-1}\right)^2 + \frac{1}{\Delta x^2}\left[\left(\Delta_x^+\Delta_x^-\varphi_{i-1,j}\right)^2 + \left(\Delta_y^+\Delta_y^-\varphi_{i,j-1}\right)^2\right], \\
S^2_{i-\alpha,j-\alpha} &= \left(\Delta_x^+\varphi_{i,j}\right)^2 + \left(\Delta_x^+\varphi_{i-1,j}\right)^2 + \left(\Delta_x^+\varphi_{i-1,j-1}\right)^2 + \left(\Delta_y^+\varphi_{i-1,j}\right)^2 + \left(\Delta_y^+\varphi_{i-1,j-1}\right)^2 \\
&\quad + \left(\Delta_y^+\varphi_{i,j-2}\right)^2 + \frac{1}{\Delta x^2}\left[\left(\Delta_x^+\Delta_x^-\varphi_{i,j}\right)^2 + \left(\Delta_y^+\Delta_y^-\varphi_{i,j-1}\right)^2\right], \\
S^3_{i-\alpha,j-\alpha} &= \left(\Delta_x^+\varphi_{i-1,j}\right)^2 + \left(\Delta_x^+\varphi_{i,j-1}\right)^2 + \left(\Delta_x^+\varphi_{i-1,j-1}\right)^2 + \left(\Delta_y^+\varphi_{i,j-1}\right)^2 + \left(\Delta_y^+\varphi_{i-1,j}\right)^2 \\
&\quad + \left(\Delta_y^+\varphi_{i-1,j-1}\right)^2 + \frac{1}{\Delta x^2}\left[\left(\Delta_x^+\Delta_x^-\varphi_{i,j-1}\right)^2 + \left(\Delta_y^+\Delta_y^-\varphi_{i-1,j}\right)^2\right], \\
S^4_{i-\alpha,j-\alpha} &= \left(\Delta_x^+\varphi_{i-2,j}\right)^2 + \left(\Delta_x^+\varphi_{i-1,j}\right)^2 + \left(\Delta_x^+\varphi_{i-1,j-1}\right)^2 + \left(\Delta_y^+\varphi_{i,j}\right)^2 + \left(\Delta_y^+\varphi_{i,j-1}\right)^2 \\
&\quad + \left(\Delta_y^+\varphi_{i-1,j-1}\right)^2 + \frac{1}{\Delta x^2}\left[\left(\Delta_x^+\Delta_x^-\varphi_{i-1,j}\right)^2 + \left(\Delta_y^+\Delta_y^-\varphi_{i,j}\right)^2\right].
\end{aligned}$$

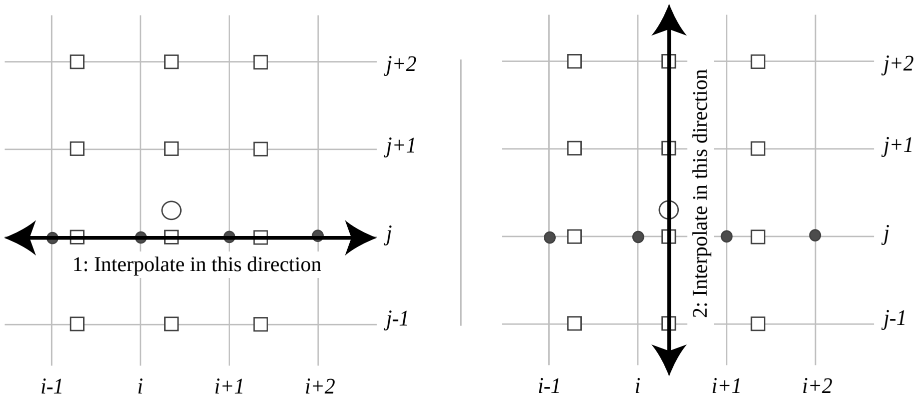

1: Interpolate in this direction
2: Interpolate in this direction
j+2
j+1
j
j-1
i-1
i
i+1
i+2

FIG. 3.4. *The dimension-by-dimension reconstruction process in two dimensions. Left: the first step, where the intermediate interpolants $\varphi_{i+\alpha,j}$ at $x_{i+\alpha,j}$ (open squares) are computed using the data $\varphi_{i,j}$ (black dots). Right: the second step, where $\varphi_{i+\alpha,j}$ is interpolated in the j direction, giving $\varphi_{i+\alpha,j+\alpha}$ at $x_{i+\alpha,j+\alpha}$ (open circle).*

3.2.2. A dimension-by-dimension reconstruction of $\varphi_{i\pm\alpha,j\pm\alpha}$. A different way to obtain high-order approximations for the values of $\varphi_{i\pm\alpha,j\pm\alpha}$ is by carrying out a sequence of one-dimensional reconstructions from section 2.2. This dimension-by-dimension approach for the reconstruction step is similar in spirit to that of [17], but here, in order to generate a Godunov-type scheme (unlike [17]), we are forced to use evolution points that are not positioned in the same locations as the data $x_{i,j}$. An appropriately chosen sequence of one-dimensional reconstructions addresses this problem.

We use the subscript "$*$" to denote the full range of an array, such that $\varphi_{*,j}$ and $\varphi_{i,*}$ denote the one-dimensional arrays $\varphi_{*,j} = (\varphi_{1,j}, \dots, \varphi_{N,j})$ and $\varphi_{i,*} = (\varphi_{i,1}, \dots, \varphi_{i,N})$. With the notation for the one-dimensional third-order reconstruction, (2.10), we can express the dimension-by-dimension reconstruction at $x_{i+\alpha,j+\alpha}$ as

1. for each i, j: $\varphi_{i+\alpha,j} = \text{reconstruct\_}\varphi\text{\_1D\_3}\,(i, \alpha, \varphi_{*,j})$;
2. for each i, j: $\varphi_{i+\alpha,j+\alpha} = \text{reconstruct\_}\varphi\text{\_1D\_3}\,(j, \alpha, \varphi_{i+\alpha,*})$.

Here, we first interpolate along the first coordinate axis and reconstruct φ at $x_{i+\alpha,j}$. The data at $x_{i+\alpha,j}$ is then interpolated along the second coordinate axis to the location $x_{i+\alpha,j+\alpha}$ to give $\varphi_{i+\alpha,j+\alpha}$ (see Figure 3.4). Obviously, the order in which the steps are performed is not important. In a similar way, a dimension-by-dimension reconstruction at $x_{i-\alpha,j-\alpha}$ is given by

1. for each i, j: $\varphi_{i-\alpha,j} = \text{reconstruct\_}\varphi\text{\_1D\_3}\,(i, -\alpha, \varphi_{*,j})$;
2. for each i, j: $\varphi_{i-\alpha,j-\alpha} = \text{reconstruct\_}\varphi\text{\_1D\_3}\,(j, -\alpha, \varphi_{i-\alpha,*})$.

3.2.3. The reprojection step. After evolving the solution to the next time step at the evolution points $x_{i\pm\alpha,j\pm\alpha}$, we would like to reproject $\varphi^{n+1}_{i+\alpha,j+\alpha}$ back onto the integer grid-points $x_{i,j}$ to end up with $\varphi^{n+1}_{i,j}$. There are several different ways to perform this task, out of which we choose to present the following: a two-dimensional reprojection using the two-dimensional reconstruction of section 3.2.1 or the dimension-by-dimension reconstruction of section 3.2.2, and a one-dimensional projection along the diagonal.

I. *A 2D reprojection.* The evolution points at $x_{i\pm\alpha,j\pm\alpha}$ have the same geometrical relationship to $x_{i,j}$ as $x_{i,j}$ has to $x_{i-\alpha,j-\alpha}$. Hence, in order to reconstruct $\varphi^{n+1}_{i,j}$ from $\varphi_{i\pm\alpha,j\pm\alpha}$, we can directly utilize the projections from section 3.2.1 or section 3.2.2,

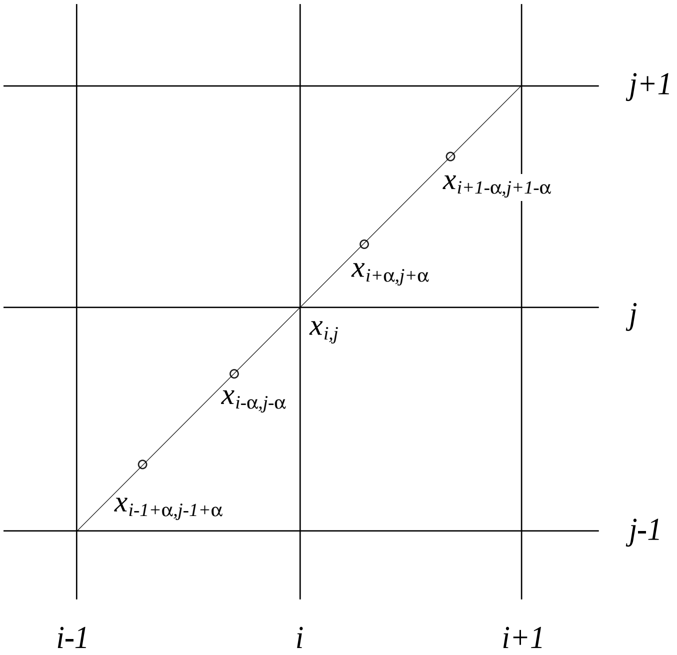


FIG. 3.5. *The evolution points used for the diagonal reconstruction of* $\varphi_{i,j}$.

taking $\varphi_{i\pm\alpha,j\pm\alpha}$ as the input data and reversing the sign of the parameter from $\pm\alpha$ to $\mp\alpha$. The final value $\varphi^{n+1}_{i,j}$ is then taken as the average of the projections of $\varphi_{i+\alpha,j+\alpha}$ and $\varphi_{i-\alpha,j-\alpha}$. Hence, if we denote either the two-dimensional or the dimension-by-dimension reconstruction described in section 3.2.1 or section 3.2.2 as

$$\varphi_{i\pm\alpha,j\pm\alpha} = \text{reconstruct\_}\varphi\text{\_2D\_3}\left(i, j, \pm\alpha, \varphi\right), \tag{3.7}$$

where φ is now the two-dimensional array $\{\varphi_{i,j}\}$, then the reprojection step is

(i) for each i, j: $\varphi^{+}_{i,j} = \text{reconstruct\_}\varphi\text{\_2D\_3}\left(i, -\alpha, \varphi_{i+\alpha,j+\alpha}\right)$;

(ii) for each i, j: $\varphi^{-}_{i,j} = \text{reconstruct\_}\varphi\text{\_2D\_3}\left(i, \alpha, \varphi_{i-\alpha,j-\alpha}\right)$;

(iii) for each i, j: $\varphi^{n+1}_{i,j} = \frac{1}{2}(\varphi^{+}_{i,j} + \varphi^{-}_{i,j})$.

II. *A diagonal reprojection.* In this case we use one-dimensional data along the diagonal, $\{\varphi_{i-1+\alpha,j-1+\alpha}, \varphi_{i-\alpha,j-\alpha}, \varphi_{i+\alpha,j+\alpha}, \varphi_{i+1-\alpha,j+1-\alpha}\}$, to construct a third-order WENO approximation of $\varphi^{n+1}_{i,j}$ (see Figure 3.5).

Define

$$\begin{aligned} \varphi^{-}_{i,j} &:= \frac{\alpha^2}{2\alpha-1}\varphi_{i-1+\alpha,j-1+\alpha} + \frac{\alpha-1}{2(2\alpha-1)}\varphi_{i-\alpha,j-\alpha} \\ &\quad + \frac{1-\alpha}{2}\varphi_{i+\alpha,j+\alpha} = \varphi\left(x_{i,j}\right) + O\left(\Delta x^3, \Delta y^3\right), \\ \varphi^{+}_{i,j} &:= \frac{1-\alpha}{2}\varphi_{i-\alpha,j-\alpha} + \frac{\alpha-1}{2(2\alpha-1)}\varphi_{i+\alpha,j+\alpha} \\ &\quad + \frac{\alpha^2}{2\alpha-1}\varphi_{i+1-\alpha,j+1-\alpha} = \varphi\left(x_{i,j}\right) + O\left(\Delta x^3, \Delta y^3\right). \end{aligned} \tag{3.8}$$

Since $(\varphi^-_{i,j} + \varphi^+_{i,j})/2 = \varphi(x_{i,j}) + O(\Delta x^4, \Delta y^4)$, we can obtain $\varphi^{n+1}_{i,j}$ as

$$\varphi^{n+1}_{i,j} = w^-_{i,j}\varphi^-_{i,j} + w^+_{i,j}\varphi^+_{i,j}, \tag{3.9}$$

where as usual $w^\pm_{i,j} = \alpha^\pm_{i,j}/(\alpha^+_{i,j} + \alpha^-_{i,j})$ and $\alpha^\pm_{i,j} = (2(\epsilon + S^\pm_{i,j})^p)^{-1}$. The smoothness measures are again taken as the sum of the discrete L^2-norm of the derivatives, which in this case is more complicated due to the uneven spacing of the data:

$$S^-_{i,j} = \frac{1}{\Delta x}\left[\left(\frac{\varphi_{i-\alpha,j-\alpha} - \varphi_{i-1+\alpha,j-1+\alpha}}{1-2\alpha}\right)^2 + \left(\frac{\varphi_{i+\alpha,j+\alpha} - \varphi_{i-\alpha,j-\alpha}}{2\alpha}\right)^2\right]$$
$$+ \frac{4}{\Delta x^3}\left(\frac{\varphi_{i-\alpha,j-\alpha} - \varphi_{i-1+\alpha,j-1+\alpha}}{1-2\alpha} - \frac{\varphi_{i+\alpha,j+\alpha} - \varphi_{i-\alpha,j-\alpha}}{2\alpha}\right)^2,$$

$$S^+_{i,j} = \frac{1}{\Delta x}\left[\left(\frac{\varphi_{i+\alpha,j+\alpha} - \varphi_{i-\alpha,j-\alpha}}{2\alpha}\right)^2 + \left(\frac{\varphi_{i+1-\alpha,j+1-\alpha} - \varphi_{i+\alpha,j+\alpha}}{1-2\alpha}\right)^2\right]$$
$$+ \frac{4}{\Delta x^3}\left(\frac{\varphi_{i+\alpha,j+\alpha} - \varphi_{i-\alpha,j-\alpha}}{2\alpha} - \frac{\varphi_{i+1-\alpha,j+1-\alpha} - \varphi_{i+\alpha,j+\alpha}}{1-2\alpha}\right)^2.$$

Remark. Our numerical simulations in section 4.3 indicate that there is little difference between the quality of the two-dimensional reconstruction and the dimension-by-dimension reconstruction of sections 3.2.1 and 3.2.2. We will use this fact when extending our methods to fifth order and higher dimensions. We note that the diagonal reprojection significantly reduces the CFL number (see section 4.4).

3.3. A two-dimensional fifth-order scheme.

Using the dimension-by-dimension approach, it is easy to extend the above scheme to fifth order: simply replace the one-dimensional third-order interpolations by the fifth-order interpolation in section 3.2.2. Using the one-dimensional notation, (2.19), we obtain a fifth-order reconstruction at $x_{i+\alpha,j+\alpha}$ as

1. for each i,j: $\varphi_{i+\alpha,j} = \text{reconstruct\_}\varphi\text{\_1D\_5}\,(i,\alpha,\varphi_{*,j})\,;$
2. for each i,j: $\varphi_{i+\alpha,j+\alpha} = \text{reconstruct\_}\varphi\text{\_1D\_5}\,(j,\alpha,\varphi_{i+\alpha,*})$.

Similarly, at $x_{i-\alpha,j-\alpha}$ we have

1. for each i,j: $\varphi_{i-\alpha,j} = \text{reconstruct\_}\varphi\text{\_1D\_5}\,(i,-\alpha,\varphi_{*,j})\,;$
2. for each i,j: $\varphi_{i-\alpha,j-\alpha} = \text{reconstruct\_}\varphi\text{\_1D\_5}\,(j,-\alpha,\varphi_{i-\alpha,*})$.

We denote this reconstruction as

$$\varphi_{i\pm\alpha,j\pm\alpha} = \text{reconstruct\_}\varphi\text{\_2D\_5}\,(i,j,\pm\alpha,\varphi)\,. \tag{3.10}$$

For the derivatives we have

1. for each i,j: $\varphi'_{i\pm\alpha,j} = \text{reconstruct\_}\varphi'\text{\_1D\_5}\,(i,\pm\alpha,\varphi_{*,j})$,
2. for each i,j: $\varphi'_{i\pm\alpha,j\pm\alpha} = \text{reconstruct\_}\varphi'\text{\_1D\_5}\,(j,\pm\alpha,\varphi_{i\pm\alpha,*})$,

which we denote as

$$\varphi'_{i\pm\alpha,j\pm\alpha} = \text{reconstruct\_}\varphi'\text{\_2D\_5}\,(i,j,\pm\alpha,\varphi)\,. \tag{3.11}$$

Reprojection onto the original grid-points $x_{i,j}$ is performed using the two-dimensional dimension-by-dimension reprojection option described in section 3.2.3.

Remarks.

1. Due to the reduced stability resulting from the use of diagonal reprojection, which is demonstrated in section 4.4, we do not develop a fifth-order analogue to the third-order diagonal reprojection.

2. It is straightforward to develop a fifth-order two-dimensional method involving two-dimensional stencils, extending section 3.2.1. Such a method would involve four interpolants defined on ten-point stencils that cover a 21-point stencil.

We summarize the two-dimensional fifth-order algorithm in the following, where $\mathrm{RK}(\varphi^n_{i\pm\alpha}, \varphi'^{\,n}_{i\pm\alpha}, \Delta t)$ is now the fourth-order RK method which integrates (2.1). As in Algorithm 2.1, each internal step of the RK method will require additional reconstructions of $\varphi'_{i\pm\alpha}$ from that step's $\varphi_{i\pm\alpha}$.

ALGORITHM 3.1. *Let $\alpha = 1/(2+\sqrt{2})$. Assume that $\{\varphi^n_{i,j}\}$ are given.*

(a) *Reconstruct:*

$$\begin{aligned}\varphi_{i\pm\alpha,j\pm\alpha} &= \text{reconstruct\_}\varphi\text{\_2D\_5}\,(i,j,\pm\alpha,\varphi)\,,\\ \varphi'^{\,n}_{i\pm\alpha,j\pm\alpha} &= \text{reconstruct\_}\varphi'\text{\_2D\_5}\,(i,j,\pm\alpha,\varphi)\,.\end{aligned}$$

(b) *Integrate:*

$$\begin{aligned}\varphi^{n+\frac{1}{2}}_{i\pm\alpha,j\pm\alpha} &= RK\left(\varphi^n_{i\pm\alpha,j\pm\alpha}, \varphi'^{\,n}_{i\pm\alpha,j\pm\alpha}, \Delta t/2\right),\\ \varphi'^{\,n+\frac{1}{2}}_{i\pm\alpha,j\pm\alpha} &= \text{reconstruct\_}\varphi'\text{\_2D\_5}(i,\pm\alpha,\varphi^{n+\frac{1}{2}}_{\pm\alpha,\pm\alpha}),\\ \varphi^{n+1}_{i\pm\alpha,j\pm\alpha} &= RK\left(\varphi^n_{i\pm\alpha,j\pm\alpha}, \varphi'^{\,n}_{i\pm\alpha,j\pm\alpha}, \Delta t\right),\\ \varphi'^{\,n+1}_{i\pm\alpha,j\pm\alpha} &= \text{reconstruct\_}\varphi'\text{\_2D\_5}\left(i,\pm\alpha,\varphi^{n+1}_{\pm\alpha,\pm\alpha}\right),\\ \varphi^{n+1}_{i\pm\alpha,j\pm\alpha} &= \varphi^n_{i\pm\alpha,j\pm\alpha} + \frac{\Delta t}{6}\left[H\left(\varphi'^{\,n}_{i\pm\alpha,j\pm\alpha}\right) + 4H(\varphi'^{\,n+\frac{1}{2}}_{i\pm\alpha,j\pm\alpha}) + H\left(\varphi'^{\,n+1}_{i\pm\alpha,j\pm\alpha}\right)\right].\end{aligned}$$

(c) *Reproject:*

$$\varphi^{n+1}_{i,j} = \text{reconstruct\_}\varphi\text{\_2D\_5}\left(i,j,\mp\alpha,\varphi^{n+1}_{\pm\alpha,\pm\alpha}\right).$$

3.4. Multidimensional extensions.

The extension of the dimension-by-dimension approach to more than two space dimensions is straightforward. For example, using the notation of section 3.3, a three-dimensional fifth-order reconstruction is

1. for each i,j,k: $\varphi_{i+\alpha,j,k} = \text{reconstruct\_}\varphi\text{\_1D\_5}\,(i,\alpha,\varphi_{*,j,k})$;
2. for each i,j,k: $\varphi_{i+\alpha,j+\alpha,k} = \text{reconstruct\_}\varphi\text{\_1D\_5}\,(j,\alpha,\varphi_{i+\alpha,*,k})$;
3. for each i,j,k: $\varphi_{i+\alpha,j+\alpha,k+\alpha} = \text{reconstruct\_}\varphi\text{\_1D\_5}\,(k,\alpha,\varphi_{i+\alpha,j+\alpha,*})$.

The reconstruction at $x_{i-\alpha,j-\alpha,k-\alpha}$ is handled similarly, and the same for the reconstruction of $\varphi'_{i+\alpha,j+\alpha,k+\alpha}$. In three dimensions, $\alpha = 1/(3+\sqrt{3})$.

A d-dimensional reconstruction based on d-dimensional stencils quickly becomes very large. It is readily apparent that the dimension-by-dimension approach will scale to high dimensions better than d-dimensional interpolants.

4. Numerical simulations.

In this section we present simulations that test the schemes we developed in this paper. In section 4.1 we demonstrate the third- and fifth-order methods in one dimension. Section 4.2 focuses on the fifth-order method in two and three space dimensions. In section 4.3 we compare the two-dimensional third-order method based on two-dimensional stencils with the dimension-by-dimension approach. In section 4.4 we examine, in detail, stability issues in two dimensions, including comparisons with [17]. Some of these examples are standard test cases that can be found, e.g., in [22, 31, 35].

We do not follow the practice in [17] of masking singular regions from our error measurements.

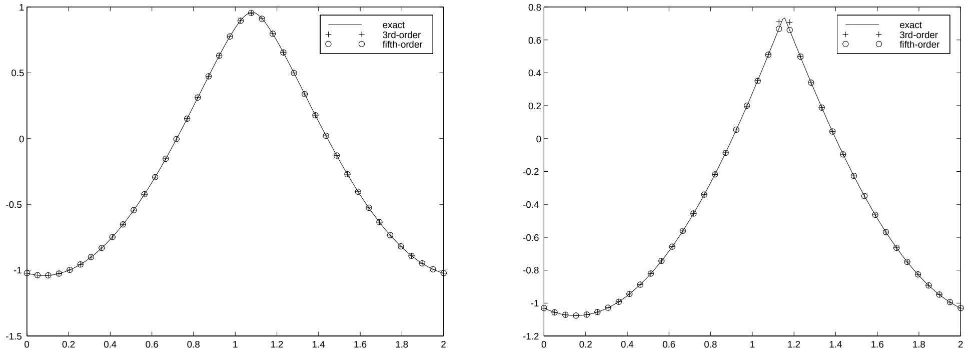


FIG. 4.1. *One-dimensional convex Hamiltonian* (4.1). *Left: the solution before the singularity formation, $T = 0.8/\pi^2$. Right: the solution after the singularity formation, $T = 1.5/\pi^2$. In both panels $N = 40$. Shown are the third- and fifth-order approximations and the exact solution.*

4.1. One-dimensional examples.

A convex Hamiltonian. We start by testing the performance of our schemes on a convex Hamiltonian. We approximate solutions of the one-dimensional equation

$$\phi_t + \frac{1}{2}\left(\phi_x + 1\right)^2 = 0, \tag{4.1}$$

subject to the initial data $\phi(x,0) = -\cos(\pi x)$ with periodic boundary conditions on $[0,2]$. The change of variables $u\,(x,t) = \phi_x\,(x,t) + 1$ transforms the equation into the Burgers equation $u_t + \frac{1}{2}\left(u^2\right)_x = 0$, which can be easily solved via the method of characteristics [35]. As is well known, the Burgers equation generally develops discontinuous solutions even with smooth initial data, and hence we expect the solutions of (4.1) to have discontinuous derivatives. In our case, the solution develops a singularity at time $t = \pi^{-2}$.

The results of our simulations are shown in Figure 4.1. The order of accuracy of these methods is determined from the relative L^1 error (see [30]), defined as the L^1-norm of the error divided by the L^1-norm of the exact solution. These results along with the relative L^∞-norm before the singularity, at $T = 0.8/\pi^2$, are given in Table 4.1, and after the singularity, at $T = 1.5/\pi^2$, in Table 4.2.

A nonconvex Hamiltonian. In this example we deal with nonconvex HJ equations. In one dimension we solve

$$\phi_t - \cos\left(\phi_x + 1\right) = 0, \tag{4.2}$$

subject to the initial data $\phi\,(x,0) = -\cos\,(\pi x)$ with periodic boundary conditions on $[0,2]$. In this case (4.2) has a smooth solution for $t \lesssim 1.049/\pi^2$, after which a singularity forms. A second singularity forms at $t \approx 1.29/\pi^2$. The results are shown in Figure 4.2. The convergence results before and after the singularity formation are given in Tables 4.3–4.4.

A linear advection equation. In this example (from [17], with a misprint corrected in [40]) we solve the one-dimensional linear advection equation, i.e., $H\,(\phi_x) = \phi_x$. We assume periodic boundary conditions on $[-1,1]$ and take the initial data as

Table 4.1

Relative L^1 errors for the one-dimensional convex HJ problem (4.1) *before the singularity formation.* $T = 0.8/\pi^2$.

	Third-order method			
N	Relative L^1 error	L^1-order	Relative L^∞ error	L^∞-order
100	9.41×10^{-5}	–	1.77×10^{-5}	–
200	1.13×10^{-5}	3.06	1.33×10^{-6}	3.73
400	1.39×10^{-6}	3.02	9.35×10^{-8}	3.83
800	1.74×10^{-7}	3.00	5.94×10^{-9}	3.00
	Fifth-order method			
N	Relative L^1 error	L^1-order	Relative L^∞ error	L^∞-order
100	1.41×10^{-5}	–	2.61×10^{-6}	–
200	4.21×10^{-7}	5.07	4.03×10^{-8}	6.02
400	3.31×10^{-8}	5.00	6.53×10^{-10}	5.95
800	4.03×10^{-10}	5.03	1.00×10^{-11}	6.03

Table 4.2

Relative L^1 errors for the one-dimensional convex HJ problem (4.1) *after the singularity formation.* $T = 1.5/\pi^2$.

	Third-order method			
N	Relative L^1 error	L^1-order	Relative L^∞ error	L^∞-order
100	9.10×10^{-4}	–	2.77×10^{-4}	–
200	2.16×10^{-4}	2.07	7.63×10^{-5}	1.86
400	6.84×10^{-5}	1.66	2.68×10^{-5}	1.51
800	2.75×10^{-5}	1.31	2.08×10^{-5}	0.37
	Fifth-order method			
N	Relative L^1 error	L^1-order	Relative L^∞ error	L^∞-order
100	7.85×10^{-4}	–	5.78×10^{-4}	–
200	1.61×10^{-4}	2.29	8.29×10^{-5}	2.29
400	6.71×10^{-5}	1.26	5.09×10^{-5}	1.26
800	3.44×10^{-5}	0.96	3.44×10^{-5}	0.96

$\phi(x,0) = g(x-0.5)$ on $[-1,1]$, where

$$g(x) = -\left(\frac{\sqrt{3}}{2} + \frac{9}{2} + \frac{2\pi}{3}\right)(x+1) + h(x),$$

$$h(x) = \begin{cases} 2\cos\left(\frac{3\pi}{2}x^2\right) - \sqrt{3}, & -1 < x < -\frac{1}{3}, \\ 3/2 + 3\cos(2\pi x), & -\frac{1}{3} < x < 0, \\ 15/2 - 3\cos(2\pi x), & 0 < x < \frac{1}{3}, \\ (28 + 4\pi + \cos(3\pi x))/3 + 6\pi x(x-1), & \frac{1}{3} < x < 1. \end{cases} \tag{4.3}$$

The results of the fifth-order method are shown in Figure 4.3, where it is compared with the fifth-order method of [17]. The reduced dissipation effects of our method are visible in the reduced round-off of the corners.

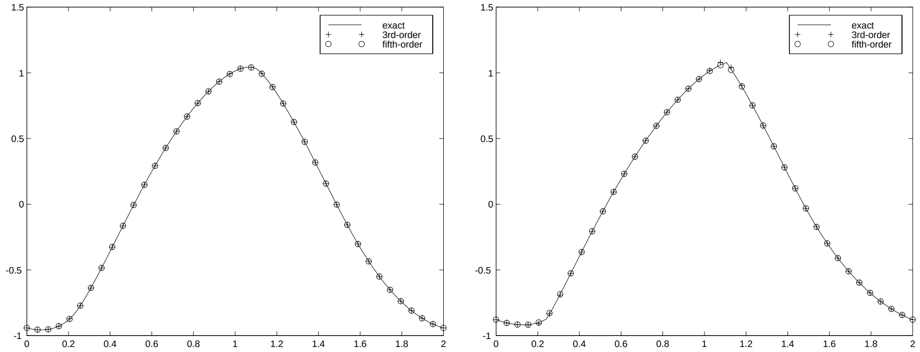


FIG. 4.2. *One-dimensional nonconvex Hamiltonian* (4.2). *Left: The solution before the singularity formation, $T = 0.8/\pi^2$. Right: The solution after the singularity formation, $T = 1.5/\pi^2$. In both panels $N = 40$. Shown are the third- and fifth-order approximations and the exact solution.*

TABLE 4.3

Relative L^1 errors for the one-dimensional nonconvex HJ problem (4.2) *before the singularity formation. $T = 0.8/\pi^2$.*

	Third-order method			
N	Relative L^1 error	L^1-order	Relative L^∞ error	L^∞-order
100	6.47×10^{-5}	–	9.05×10^{-6}	–
200	7.78×10^{-6}	3.06	1.11×10^{-6}	3.03
400	8.77×10^{-7}	3.15	9.27×10^{-8}	3.58
800	9.87×10^{-8}	3.15	6.12×10^{-9}	3.92
	Fifth-order method			
N	Relative L^1 error	L^1-order	Relative L^∞ error	L^∞-order
100	1.29×10^{-5}	–	4.97×10^{-6}	–
200	6.52×10^{-7}	4.31	2.38×10^{-7}	4.38
400	2.10×10^{-8}	4.95	6.13×10^{-9}	5.28
800	5.96×10^{-10}	5.14	1.03×10^{-10}	5.90

TABLE 4.4

Relative L^1 errors for the one-dimensional nonconvex HJ problem (4.2) *after the singularity formation. $T = 1.5/\pi^2$.*

	Third-order method			
N	Relative L^1 error	L^1-order	Relative L^∞ error	L^∞-order
100	2.81×10^{-4}	–	9.64×10^{-5}	–
200	1.32×10^{-4}	1.08	5.05×10^{-5}	0.93
400	2.31×10^{-5}	2.52	6.00×10^{-6}	3.07
800	8.43×10^{-6}	1.46	3.30×10^{-6}	0.86
	Fifth-order method			
N	Relative L^1 error	L^1-order	Relative L^∞ error	L^∞-order
100	1.57×10^{-4}	–	1.12×10^{-4}	–
200	8.34×10^{-5}	0.91	6.60×10^{-5}	0.77
400	1.22×10^{-5}	2.78	8.64×10^{-6}	2.93
800	6.67×10^{-5}	0.87	5.23×10^{-6}	.072

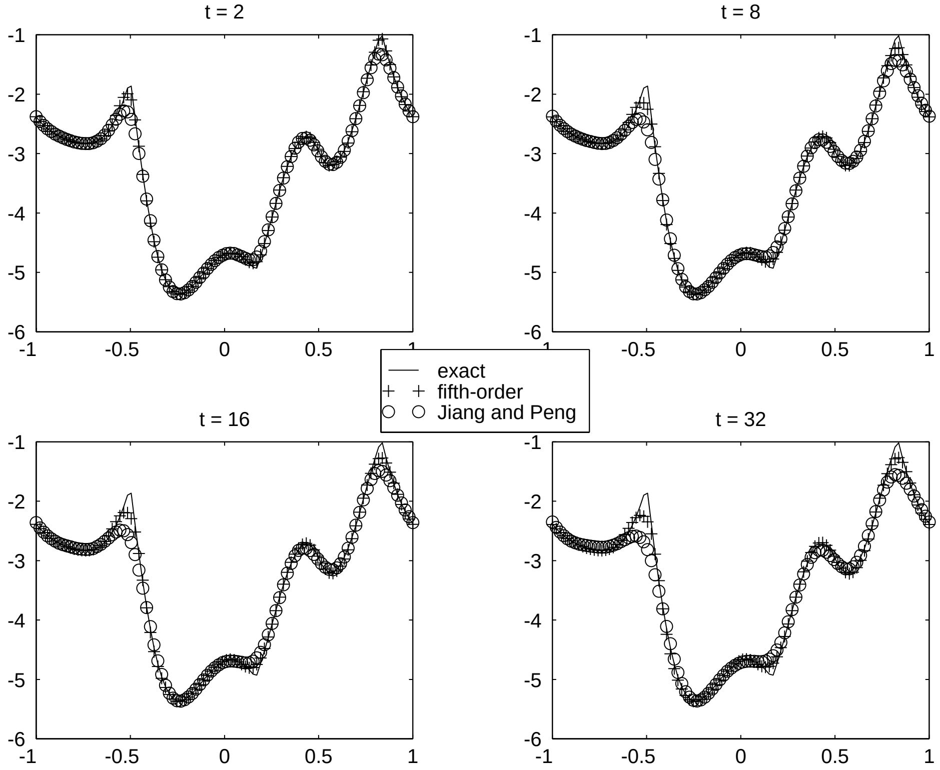


FIG. 4.3. *One-dimensional linear advection,* (4.3)*.* $T = 2, 8, 16, 32$*;* $N = 100$*. Crosses: our fifth-order method. Circles: the fifth-order method of* [17] *with a local Lax–Friedrichs flux. Solid line: the exact solution.*

4.2. Two-dimensional examples.

A convex Hamiltonian. In two dimensions we solve a problem similar to (4.1),

$$\phi_t + \frac{1}{2}\left(\phi_x + \phi_y + 1\right)^2 = 0, \tag{4.4}$$

which can be reduced to a one-dimensional problem via the coordinate transformation $\binom{\xi}{\eta} = \frac{1}{2}\left(\begin{smallmatrix}1 & 1\\ 1 & -1\end{smallmatrix}\right)\binom{x}{y}$. The results of the fifth-order calculations for the initial data $\phi(x, y, 0) = -\cos(\pi(x + y)/2) = -\cos(\pi\xi)$ are shown in Figure 4.4. The convergence rates for the two-dimensional fifth-order scheme before and after the singularity are shown in Table 4.5.

A nonconvex Hamiltonian. The two-dimensional nonconvex problem, which is analogous to the one-dimensional problem (4.2), is

$$\phi_t - \cos(\phi_x + \phi_y + 1) = 0. \tag{4.5}$$

Here we assume initial data, given by $\phi(x, y, 0) = -\cos(\pi(x + y)/2)$, and periodic boundary conditions. The results are shown in Figure 4.5. The convergence results for the two-dimensional fifth-order scheme before and after the singularity formation are given in Table 4.6.

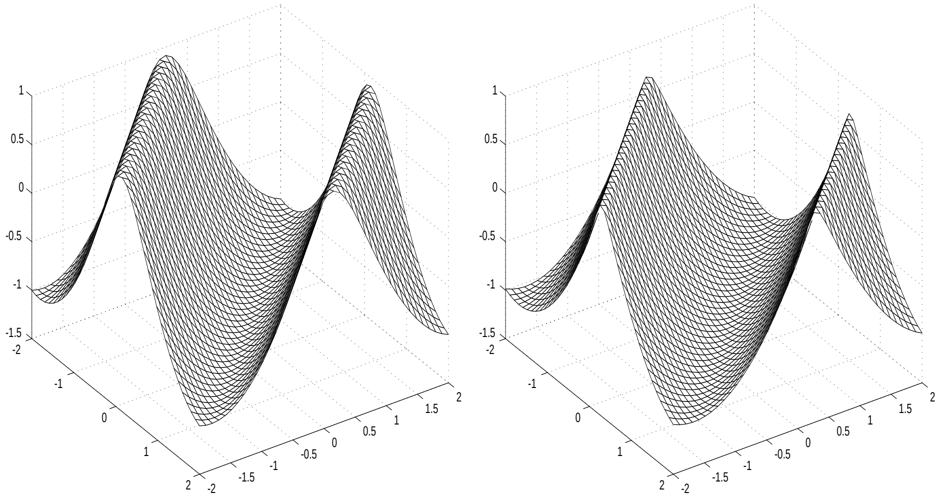


Fig. 4.4. *Two-dimensional convex Hamiltonian,* (4.4). *Left: the solution before the singularity formation, $T = 0.8/\pi^2$. Right: the solution after the singularity formation, $T = 1.5/\pi^2$. In both panels $N = 40 \times 40$. The solution is computed with the fifth-order method.*

Table 4.5

Relative L^1 and L^∞ errors for the two-dimensional convex HJ problem (4.4) *before and after singularity formation, computed via the fifth-order method.*

Before singularity $T = 0.8/\pi^2$				
N	Relative L^1 error	L^1-order	Relative L^∞ error	L^∞-order
50	1.19×10^{-4}	–	7.78×10^{-7}	–
100	6.80×10^{-6}	4.13	1.64×10^{-8}	5.56
200	1.73×10^{-7}	5.30	1.12×10^{-10}	7.20
After singularity $T = 1.5/\pi^2$				
N	Relative L^1 error	L^1-order	Relative L^∞ error	L^∞-order
50	1.32×10^{-3}	–	2.07×10^{-5}	–
100	3.89×10^{-4}	1.76	3.60×10^{-6}	2.52
200	4.86×10^{-5}	3.00	1.69×10^{-7}	4.41

A fully two-dimensional example. The above two-dimensional examples are actually one-dimensional along the diagonal. To check the performance of our methods on fully two-dimensional problems, we solve

$$\phi_t + \phi_x \phi_y = 0 \tag{4.6}$$

on $[-\pi,\pi]\times[-\pi,\pi]$, subject to the initial data $\phi(x,y,0) = \sin(x) + \cos(y)$ with periodic boundary conditions. The exact solution for this problem is given implicitly by $\phi(x,y,t) = -\cos(q)\sin(r) + \sin(q) + \cos(r)$, where $x = q - t\sin(r)$ and $y = r + t\cos(q)$. This solution is smooth for $t < 1$, continuous $\forall t$, and has discontinuous derivatives for $t \geq 1$. The results of our simulations at times $T = 0.8, 1.5$ are shown in Figure 4.6. The convergence results for the fifth-order two-dimensional schemes before the singularity formation are given in Table 4.7 and confirm the expected order of accuracy of our methods.

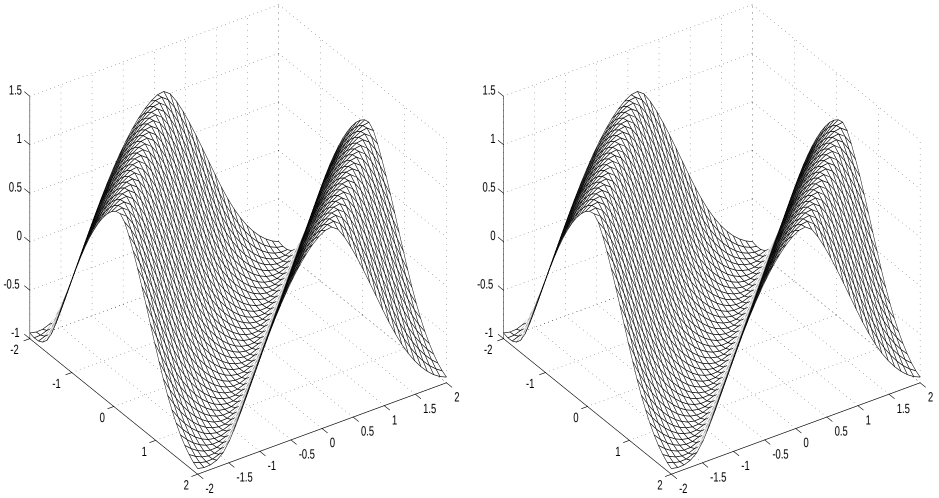


FIG. 4.5. *Two-dimensional nonconvex Hamiltonian,* (4.5). *Left: the solution before the singularity formation, $T = 0.8/\pi^2$. Right: the solution after the singularity formation, $T = 1.5/\pi^2$. $N = 40 \times 40$. The solution is computed with the fifth-order method.*

TABLE 4.6

Relative L^1 and L^∞ errors for the two-dimensional nonconvex HJ problem (4.5) *before and after the singularity formation, computed with the fifth-order method.*

Before singularity $T = 0.8/\pi^2$				
N	Relative L^1 error	L^1-order	Relative L^∞ error	L^∞-order
50	1.11×10^{-4}	–	1.26×10^{-6}	–
100	6.91×10^{-6}	4.00	2.42×10^{-8}	5.70
200	3.85×10^{-7}	4.17	6.27×10^{-10}	5.27
After singularity $T = 1.5/\pi^2$				
N	Relative L^1 error	L^1-order	Relative L^∞ error	L^∞-order
50	1.47×10^{-3}	–	8.58×10^{-6}	–
100	1.93×10^{-4}	2.93	9.27×10^{-7}	3.21
200	8.87×10^{-5}	1.12	3.09×10^{-7}	1.58

An eikonal equation in geometric optics. We consider a two-dimensional nonconvex problem that arises in geometric optics [20]:

$$\begin{cases} \phi_t + \sqrt{\phi_x^2 + \phi_y^2 + 1} = 0, \\ \phi(x, y, 0) = \frac{1}{4}(\cos(2\pi x) - 1)(\cos(2\pi y) - 1) - 1. \end{cases} \tag{4.7}$$

The results of our fifth-order method at time $T = 0.6$ are shown in Figure 4.7, where we see the sharp corners that develop in this problem.

An optimal control problem. We solve an optimal control problem related to cost determination [35]. Here the Hamiltonian is of the form $H(x, y, \nabla\phi)$:

$$\begin{cases} \phi_t - \sin(y)\,\phi_x + \sin(x)\,\phi_y + |\phi_y| - \frac{1}{2}\sin^2(y) - 1 + \cos(x) = 0, \\ \phi(x, y, 0) = 0. \end{cases} \tag{4.8}$$

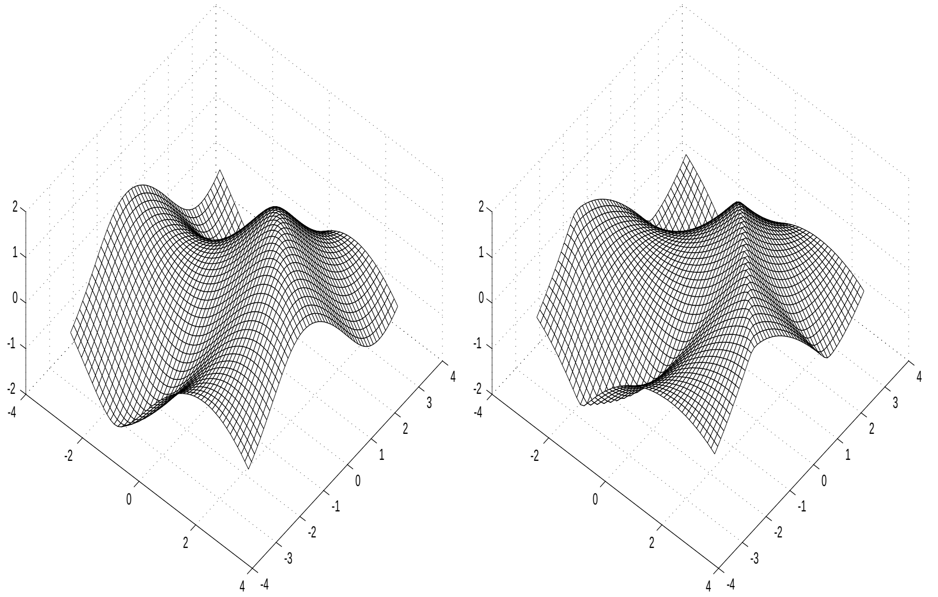


Fig. 4.6. *Fully two-dimensional Hamiltonian,* (4.6). *Left: the solution before the singularity formation, $T = 0.8$. Right: the solution after the singularity formation, $T = 1.5$. In both panels $N = 50 \times 50$. The solution is computed with the fifth-order method.*

Table 4.7
Relative L^1 errors for the two-dimensional HJ problem (4.6) *before singularity formation. $T = 0.8$. The solution is computed with the fifth-order method.*

Before singularity $T = 0.8$				
N	Relative L^1 error	L^1-order	Relative L^∞ error	L^∞-order
50	6.10×10^{-6}	–	8.15×10^{-8}	–
100	2.10×10^{-7}	4.86	7.35×10^{-10}	6.79
200	7.53×10^{-9}	4.80	5.59×10^{-12}	7.04

The result of our fifth-order scheme is presented in Figure 4.8 and is in qualitative agreement with [31].

4.3. A comparison of two-dimensional third-order interpolants. In this section we use the examples (4.4), (4.5), and (4.6) to compare the third-order method of section 3.2.1, based on interpolation via two-dimensional stencils, with that of section 3.2.2, where we used a dimension-by-dimension approach. The results are shown in Table 4.8. The dimension-by-dimension method produces errors that are approximately twice as large as those for the genuinely two-dimensional reconstruction. However, the convergence rate is qualitatively the same in both methods. These results motivated us to base our fifth-order scheme on the much simpler dimension-by-dimension reconstruction.

4.4. A stability study. In this section we present a couple of stability studies that we obtained in our simulations. We start by checking the stability properties of the third-order scheme with different reprojection steps. The reconstruction step

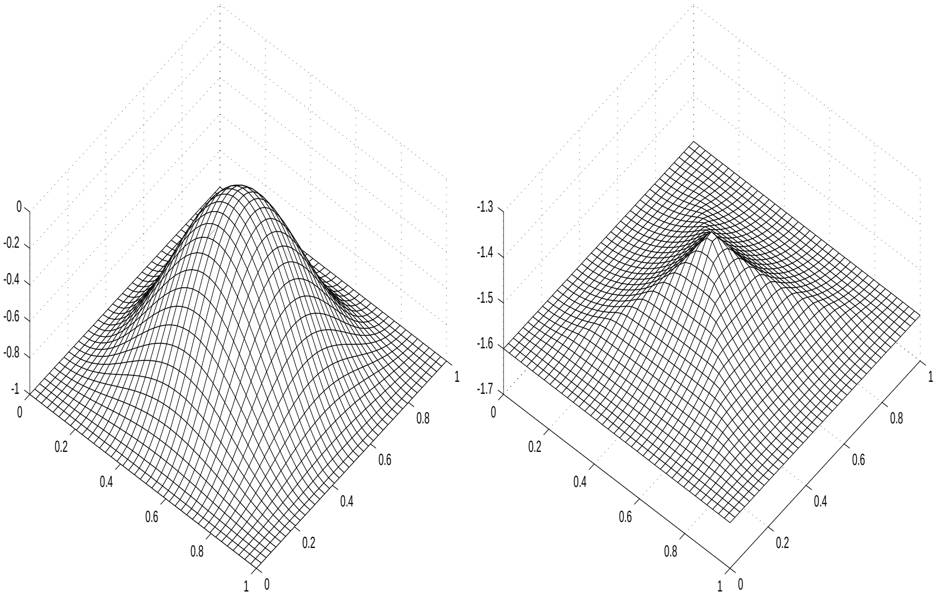


FIG. 4.7. *Two-dimensional eikonal equation,* (4.7)*.* $N = 40 \times 40$*. Left: the initial data. Right: the fifth-order approximation at* $T = 0.6$*.*

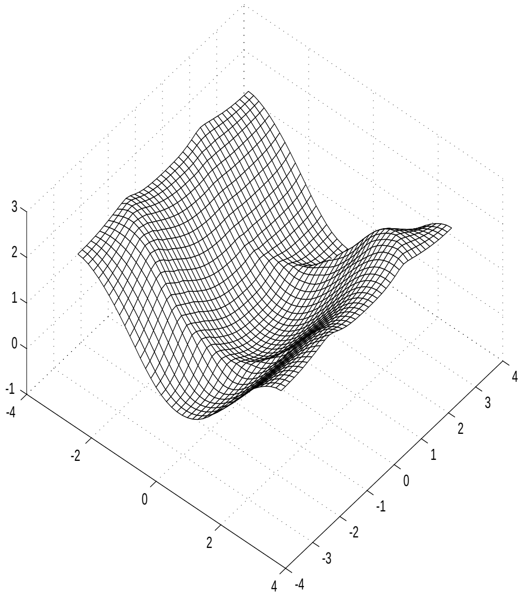


FIG. 4.8. *Two-dimensional optimal control problem,* (4.8)*. An approximation with the fifth-order method is shown at* $T = 1$ *and* $N = 40 \times 40$*.*

TABLE 4.8

Comparison of the third-order method of section 3.2.1, *using an interpolation via two-dimensional stencils, and that of section* 3.2.2, *using the dimension-by-dimension approach.*

	2D stencils		Dimension-by-dimension	
N	Relative L^1 error	L^1-order	Relative L^1 error	L^1-order
Convex Hamiltonian at $T = 0.8/\pi^2$				
50	4.70×10^{-4}	–	6.13×10^{-4}	–
100	7.54×10^{-5}	2.64	9.43×10^{-5}	2.70
200	8.07×10^{-6}	3.23	1.02×10^{-5}	3.21
Convex Hamiltonian at $T = 1.5/\pi^2$				
50	1.23×10^{-3}	–	2.61×10^{-3}	–
100	4.56×10^{-4}	1.44	8.19×10^{-4}	1.67
200	3.70×10^{-5}	3.62	1.22×10^{-4}	2.74
Nonconvex Hamiltonian at $T = 0.8/\pi^2$				
50	2.27×10^{-4}	–	3.92×10^{-4}	–
100	3.75×10^{-5}	2.60	6.97×10^{-5}	2.49
200	3.99×10^{-6}	3.23	7.22×10^{-6}	3.27
Nonconvex Hamiltonian at $T = 1.5/\pi^2$				
50	1.23×10^{-3}	–	1.94×10^{-3}	–
100	2.50×10^{-4}	2.30	4.16×10^{-4}	2.22
200	7.63×10^{-5}	1.71	1.20×10^{-4}	1.79
Fully 2D example at $T = 0.8$				
50	2.01×10^{-4}	–	1.48×10^{-4}	–
100	2.42×10^{-5}	3.05	1.65×10^{-5}	3.16
200	2.95×10^{-6}	3.04	1.95×10^{-6}	3.08

is done in all cases using the dimension-by-dimension interpolant. We compare the dimension-by-dimension reprojection and the diagonal reprojection (of section 3.2.3). In Figure 4.9 we plot the L^1 error as a function of the CFL number. The test problem is (4.6) with the fully two-dimensional Hamiltonian. The solution is computed at $T = 0.8$. We see that the use of a diagonal reprojection significantly reduces the maximum allowed CFL number.

We now turn to checking the stability properties of the two-dimensional fifth-order method of section 3.3 by computing the L^1 errors for various examples while varying the CFL number. In Figure 4.10 we compare the results obtained with our fifth-order scheme with the fifth-order method of [17], for which we used a local Lax–Friedrichs flux. The numerical tests indicate that larger CFL numbers can be used with our method.

4.5. Three-dimensional examples. We proceed with a three-dimensional generalization of the convex Hamiltonian (4.4),

$$\phi_t + \frac{1}{2}\left(\phi_x + \phi_y + \phi_z + 1\right)^2 = 0, \tag{4.9}$$

subject to the initial data $\phi(x, y, z, 0) = -\cos(\pi(x + y + z)/3)$. The convergence results for the three-dimensional fifth-order scheme before and after the singularity formation are given in Table 4.9. We also approximate the solution of the nonconvex

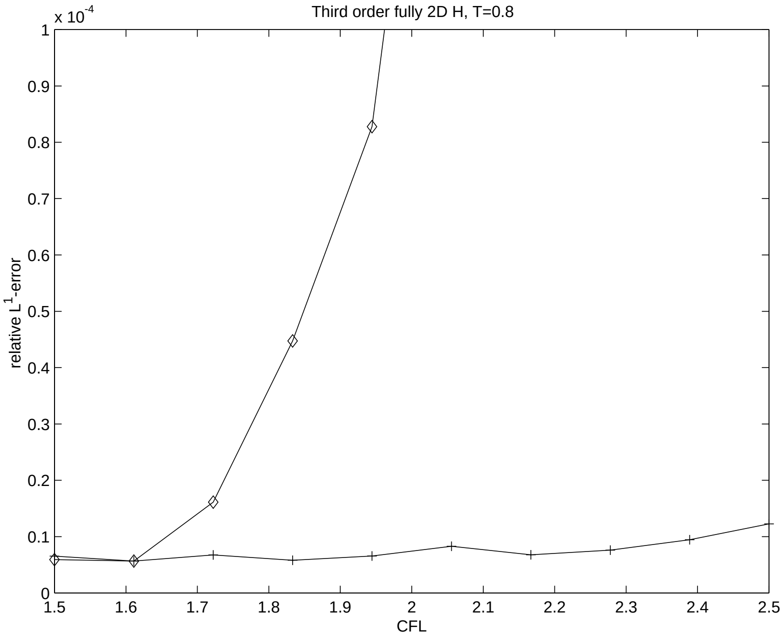


Fig. 4.9. *Stability of the two-dimensional third-order method with a dimension-by-dimension reprojection (crosses) vs. a diagonal reprojection (diamonds). Fully two-dimensional Hamiltonian* (4.6)*. $T = 0.8$ (before singularity), $N = 100 \times 100$.*

Table 4.9

Relative L^1 and L^∞ errors for the three-dimensional convex HJ problem (4.9) *before and after the singularity formation, computed with the fifth-order method.*

Before singularity $T = 0.5/\pi^2$				
N	Relative L^1 error	L^1-order	Relative L^∞ error	L^∞-order
25	2.61×10^{-4}	–	1.07×10^{-7}	–
50	6.40×10^{-6}	5.35	3.16×10^{-10}	8.41
100	1.50×10^{-7}	5.42	9.18×10^{-13}	8.43
After singularity $T = 1.5/\pi^2$				
N	Relative L^1 error	L^1-order	Relative L^∞ error	L^∞-order
25	6.95×10^{-3}	–	1.80×10^{-5}	–
50	1.40×10^{-3}	2.31	4.15×10^{-6}	2.12
100	5.33×10^{-4}	1.39	6.94×10^{-7}	2.58

problem

$$\phi_t - \cos(\phi_x + \phi_y + \phi_z + 1) = 0, \tag{4.10}$$

with the same initial data. The convergence rates for the three-dimensional fifth-order schemes are given in Table 4.10.

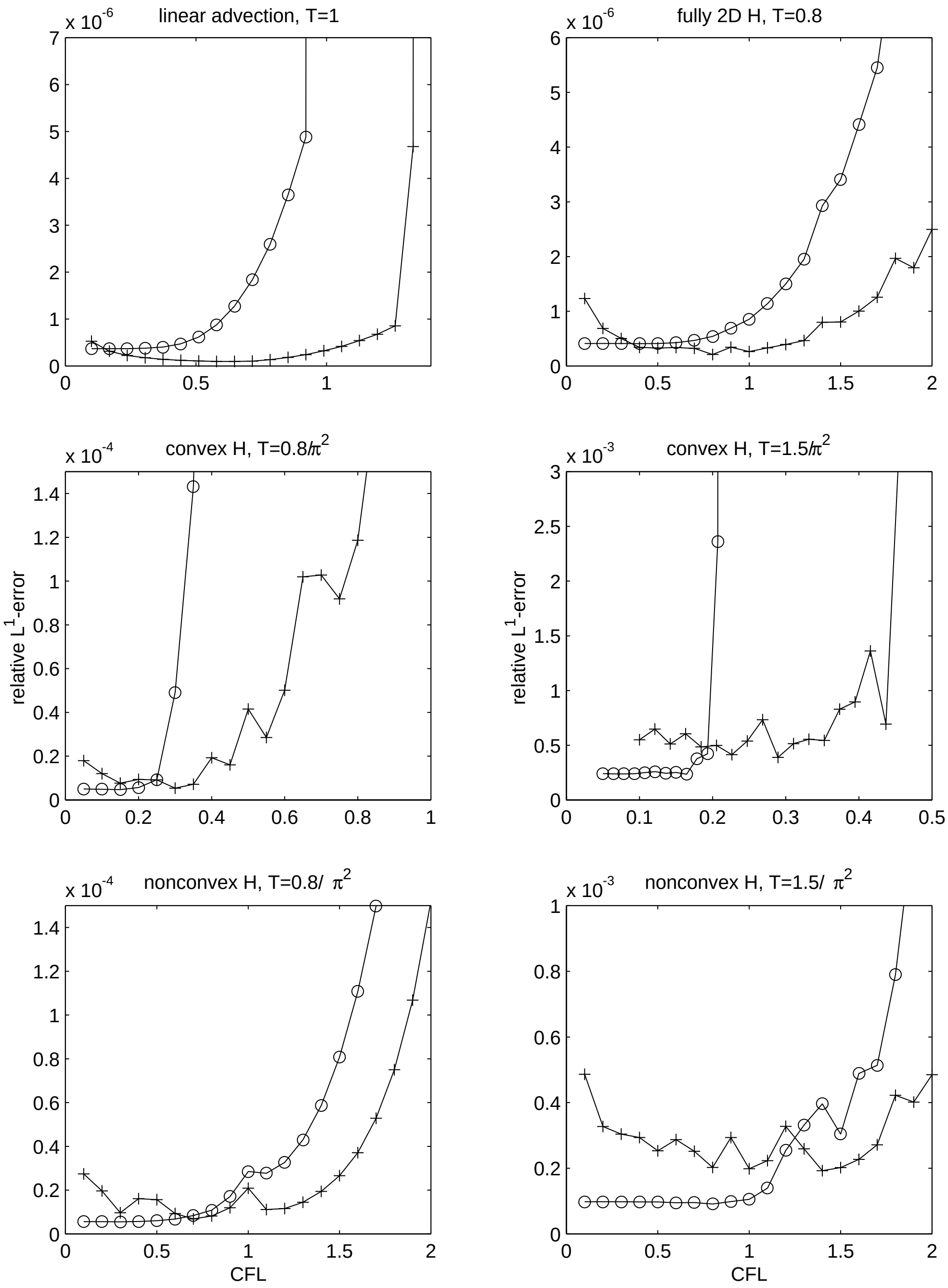


FIG. 4.10. *Stability of the two-dimensional fifth-order method.* $N = 100 \times 100$. *Crosses: our fifth-order method. Circles: the fifth-order method of* [17] *with a local Lax–Friedrichs flux. Upper left: linear advection* ($H(\nabla\varphi) = \nabla\varphi$) *with initial condition* $\phi(x, y, 0) = -\cos(\pi(x + y)/2)$. *Upper right: fully* $2D$ *Hamiltonian* (4.6). *Middle row: convex Hamiltonian* (4.4), *before the singularity (left) and after the singularity (right). Bottom row: nonconvex Hamiltonian* (4.5), *before the singularity (left) and after the singularity (right).*

Table 4.10

Relative L^1 and L^∞ errors for the three-dimensional nonconvex HJ problem (4.10) *before and after the singularity formation, computed with the fifth-order method.*

Before singularity $T = 0.5/\pi^2$				
N	Relative L^1 error	L^1-order	Relative L^∞ error	L^∞-order
25	7.28×10^{-4}	–	3.70×10^{-7}	–
50	3.71×10^{-5}	4.29	4.06×10^{-9}	6.51
100	1.05×10^{-6}	5.14	2.18×10^{-11}	7.54
After singularity $T = 1.5/\pi^2$				
N	Relative L^1 error	L^1-order	Relative L^∞ error	L^∞-order
25	6.74×10^{-3}	–	3.27×10^{-6}	–
50	1.26×10^{-3}	2.42	6.90×10^{-7}	2.25
100	4.21×10^{-4}	1.59	6.84×10^{-8}	3.33

Acknowledgment. We would like to thank Volker Elling for helpful discussions throughout the early stages of this work.

REFERENCES

[1] R. Abgrall, *Numerical discretization of the first-order Hamilton-Jacobi equation on triangular meshes*, Comm. Pure Appl. Math., 49 (1996), pp. 1339–1373.

[2] P. Arminjon and M.-C. Viallon, *Généralisation du schéma de Nessyahu-Tadmor pour une équation hyperbolique à deux dimensions d'espace*, C. R. Acad. Sci. Paris Sér. I, 320 (1995), pp. 85–88.

[3] G. Barles, *Solution de viscosité des équations de Hamilton-Jacobi*, Springer-Verlag, Berlin, 1994.

[4] F. Bianco, G. Puppo, and G. Russo, *High-order central schemes for hyperbolic systems of conservation laws*, SIAM J. Sci. Comput., 21 (1999), pp. 294–322.

[5] S. Bryson and D. Levy, *Central schemes for multidimensional Hamilton–Jacobi equations*, SIAM J. Sci. Comput., to appear.

[6] S. Bryson and D. Levy, *High-order central WENO schemes for* 1*D Hamilton-Jacobi equations*, in Numerical Mathematics and Advanced Applications (Proceedings of ENUMATH 2001, Ischia, Italy), F. Brezzi, A. Buffa, S. Cosaro, and A. Murli, eds., Springer-Verlag, 2003.

[7] M.G. Crandall, L.C. Evans, and P.-L. Lions, *Some properties of viscosity solutions of Hamilton-Jacobi equations*, Trans. Amer. Math. Soc., 282 (1984), pp. 487–502.

[8] M.G. Crandall, H. Ishii, and P.-L. Lions, *User's guide to viscosity solutions of second order partial differential equations*, Bull. Amer. Math. Soc., 27 (1992), pp. 1–67.

[9] M.G. Crandall and P.-L. Lions, *Viscosity solutions of Hamilton-Jacobi equations*, Trans. Amer. Math. Soc., 277 (1983), pp. 1–42.

[10] M.G. Crandall and P.-L. Lions, *Two approximations of solutions of Hamilton-Jacobi equations*, Math. Comp., 43 (1984), pp. 1–19.

[11] K.O. Friedrichs and P.D. Lax, *Systems of conservation equations with a convex extension*, Proc. Natl. Acad. Sci. USA, 68 (1971), pp.1686–1688.

[12] S. Gottlieb, C.-W. Shu, and E. Tadmor, *Strong stability-preserving high-order time discretization methods*, SIAM Rev., 43 (2001), pp. 89–112.

[13] A. Harten, B. Engquist, S. Osher, and S. Chakravarthy, *Uniformly high order accurate essentially non-oscillatory schemes* III, J. Comput. Phys., 71 (1987), pp. 231–303.

[14] C. Hu and C.-W. Shu, *A discontinuous Galerkin finite element method for Hamilton–Jacobi equations*, SIAM J. Sci. Comput., 21 (1999), pp. 666–690.

[15] S.N. Kruzkov, *The Cauchy problem in the large for nonlinear equations and for certain quasilinear systems of the first order with several variables*, Soviet Math. Dokl., 5 (1964), pp. 493–496.

[16] G.-S. Jiang, D. Levy, C.-T. Lin, S. Osher, and E. Tadmor, *High-resolution nonoscillatory central schemes with nonstaggered grids for hyperbolic conservation laws*, SIAM J. Numer. Anal., 35 (1998), pp. 2147–2168.

[17] G.-S. Jiang and D. Peng, *Weighted ENO schemes for Hamilton–Jacobi equations*, SIAM J. Sci. Comput., 21 (2000), pp. 2126–2143.

[18] G.-S. Jiang and D.-W. Shu, *Efficient implementation of weighted ENO schemes*, J. Comput. Phys., 126 (1996), pp. 202–228.

[19] G.-S. Jiang and E. Tadmor, *Nonoscillatory central schemes for multidimensional hyperbolic conservation laws*, SIAM J. Sci. Comput., 19 (1998), pp. 1892–1917.

[20] S. Jin and Z. Xin, *Numerical passage from systems of conservation laws to Hamilton–Jacobi equations, and relaxation schemes*, SIAM J. Numer. Anal., 35 (1998), pp. 2385–2404.

[21] A. Kurganov, S. Noelle, and G. Petrova, *Semidiscrete central-upwind schemes for hyperbolic conservation laws and Hamilton–Jacobi equations*, SIAM J. Sci. Comput., 23 (2001), pp. 707–740.

[22] A. Kurganov and E. Tadmor, *New high-resolution semi-discrete central schemes for Hamilton-Jacobi equations*, J. Comput. Phys., 160 (2000), pp. 720–724.

[23] A. Kurganov and E. Tadmor, *New high-resolution central schemes for nonlinear conservation laws and convection-diffusion equations*, J. Comput. Phys., 160 (2000), pp. 241–282.

[24] O. Lepsky, C. Hu, and C.-W. Shu, *Analysis of the discontinuous Galerkin method for Hamilton-Jacobi equations*, Appl. Numer. Math., 33 (2000), pp. 423–434.

[25] D. Levy, G. Puppo, and G. Russo, *A fourth order central WENO scheme for multidimensional hyperbolic systems of conservation laws*, SIAM J. Sci. Comput., 24 (2002), pp. 480–506.

[26] D. Levy, G. Puppo, and G. Russo, *Central WENO schemes for hyperbolic systems of conservation laws*, Math. Model. Numer. Anal., 33 (1999), pp. 547–571.

[27] D. Levy, G. Puppo, and G. Russo, *Compact central WENO schemes for multidimensional conservation laws*, SIAM J. Sci. Comput., 22 (2000), pp. 656–672.

[28] P.L. Lions, *Generalized Solutions of Hamilton–Jacobi Equations*, Pitman, London, 1982.

[29] P.L. Lions and P.E. Souganidis, *Convergence of MUSCL and filtered schemes for scalar conservation laws and Hamilton-Jacobi equations*, Numer. Math., 69 (1995), pp. 441–470.

[30] C.-T. Lin and E. Tadmor, *L^1-stability and error estimates for approximate Hamilton-Jacobi solutions*, Numer. Math., 87 (2001), pp. 701–735.

[31] C.-T. Lin and E. Tadmor, *High-resolution nonoscillatory central schemes for Hamilton–Jacobi equations*, SIAM J. Sci. Comput., 21 (2000), pp. 2163–2186.

[32] X.-D. Liu, S. Osher, and T. Chan, *Weighted essentially non-oscillatory schemes*, J. Comput. Phys., 115 (1994), pp. 200–212.

[33] H. Nessyahu and E. Tadmor, *Non-oscillatory central differencing for hyperbolic conservation laws*, J. Comput. Phys., 87 (1990), pp. 408–463.

[34] S. Osher and J. Sethian, *Fronts propagating with curvature dependent speed: Algorithms based on Hamilton-Jacobi formulations*, J. Comput. Phys., 79 (1988), pp. 12–49.

[35] S. Osher and C.-W. Shu, *High-order essentially nonoscillatory schemes for Hamilton–Jacobi equations*, SIAM J. Numer. Anal., 28 (1991), pp. 907–922.

[36] J. Shi, C. Hu, and C.-W. Shu, *A technique of treating negative weights in WENO schemes*, J. Comput. Phys., 175 (2002), pp. 108–127.

[37] C.-W. Shu and S. Osher, *Efficient implementation of essentially non-oscillatory shock-capturing schemes*, II, J. Comput. Phys., 83 (1989), pp. 32–78.

[38] P.E. Souganidis, *Approximation schemes for viscosity solutions of Hamilton-Jacobi equations*, J. Differential Equations, 59 (1985), pp. 1–43.

[39] M. Zennaro, *Natural continuous extensions of Runge-Kutta methods*, Math. Comp., 46 (1986), pp. 119–133.

[40] Y.-T. Zhang and C.-W. Shu, *High-order WENO schemes for Hamilton–Jacobi equations on triangular meshes*, SIAM J. Sci. Comput., 24 (2003), pp. 1005–1030.

SIAM J. NUMER. ANAL.
Vol. 41, No. 4, pp. 1370–1387

PSEUDOSPECTRAL LEAST-SQUARES METHOD FOR THE SECOND-ORDER ELLIPTIC BOUNDARY VALUE PROBLEM*

SANG DONG KIM†, HYUNG-CHUN LEE‡, AND BYEONG CHUN SHIN§

Abstract. The least-squares Legendre and Chebyshev pseudospectral methods are presented for a first-order system equivalent to a second-order elliptic partial differential equation. Continuous and discrete homogeneous least-squares functionals using Legendre and Chebyshev weights are shown to be equivalent to the $H^1(\Omega)$ norm and Chebyshev-weighted Div-Curl norm over appropriate polynomial spaces, respectively. The spectral error estimates are derived. The block diagonal finite element preconditioner is developed for the both cases. Several numerical tests are demonstrated on the spectral discretization errors and on performances of the finite element preconditioner.

Key words. pseudospectral method, first-order system least-squares method

AMS subject classifications. 65F10, 65M30

DOI. 10.1137/S0036142901398234

1. Introduction.

Let Ω be the square $(-1,1)^2$. We consider the second-order elliptic boundary value problem

$$\left\{\begin{array}{rcll} -\nabla\cdot\nabla p+\mathbf{b}\cdot\nabla p+c_0\,p &=& f & \text{in } \Omega,\\ p &=& 0 & \text{on } \Gamma_D,\\ \mathbf{n}\cdot\nabla p &=& 0 & \text{on } \Gamma_N, \end{array}\right. \tag{1.1}$$

where $\partial\Omega=\Gamma_D\cup\Gamma_N$ denotes the boundary of Ω, f is a given continuous function, $\mathbf{b}$ and c_0 are given constant vector and scalar, respectively, and $\mathbf{n}$ is the outward unit vector normal to the boundary.

Introducing the flux variable $\mathbf{u}=\nabla p$, (1.1) can be written as an equivalent first-order system of linear equations (see [4], [5], and [21], for example). For the use of finite element methods, the least-squares approach was studied in [12], [13], and [14], for example, and it has been widely used by combining with functionals consisting of appropriate norms of residual equations (see [2], [3], [4], [5], [15], etc.). Then the homogeneous continuous and discrete least-squares functionals were shown to be equivalent to appropriate product norms. In this paper, the success of finite element least-squares methods for the last decade stimulated the usage of pseudospectral methods or Legendre (Chebyshev) spectral elements with a staggered grid, which is known to be a very accurate method (see [1], [6], [9], [10], [11], and [21]). Therefore we believe that it is worthwhile to develop the Legendre and Chebyshev pseudospectral least-squares methods for solving the first-order system corresponding to (1.1). Using the Legendre–Gauss–Lobatto (LGL) and Chebyshev–Gauss–Lobatto (CGL) points with corresponding quadrature weights, we define two discrete Legendre and Chebyshev least-squares functionals. For the continuous Legendre least-squares functional, we

*Received by the editors November 15, 2001; accepted for publication (in revised form) December 4, 2002; published electronically September 9, 2003. This work was supported by Korea Research Foundation grant KRF-2001-015-DP0043.

http://www.siam.org/journals/sinum/41-4/39823.html

†Department of Mathematics, Kyungpook National University, Taegu 702-701, Korea (skim@knu.ac.kr).

‡Department of Mathematics, Ajou University, Suwon 442-749, Korea (hclee@madang.ajou.ac.kr).

§Department of Mathematics, Chonnam National University, Kwangju 500-757, Korea (bcshin@chonnam.ac.kr).

adopt the continuous least-squares functional developed in [5]. The equivalence between the usual L^2-norm and the discrete Legendre spectral norm over an appropriate polynomial space yields that the discrete Legendre least-squares functional is equivalent to a product H^1-norm over a product of polynomial spaces. The continuous Chebyshev-weighted least-squares functional is also defined as the sum of L^2_w-norms of residual equations. Then the continuous and corresponding discrete Chebyshev-weighted least-squares functionals are shown to be equivalent to a product norm $\|\mathbf{u}\|^2_{L^2_w(\Omega)^2} + \|\nabla \cdot \mathbf{u}\|^2_{L^2_w(\Omega)^2} + \|\nabla \times \mathbf{u}\|^2_{L^2_w(\Omega)^2} + \|p\|^2_{H^1_w(\Omega)}$, in which we do not provide its equivalence to the Chebyshev-weighted H^1_w product norm. It is shown that the proposed methods have spectral convergence. Based on a norm equivalence, a block diagonal finite element preconditioner is developed for the use of an iterative method like the conjugate gradient method. Such finite element preconditioning techniques are discussed in [7], [8], [16], [20], and [24], for example. The finite element preconditioner for the Legendre case is optimal in the sense that the condition number of the preconditioned system behaves like $O(1)$. We also consider the Chebyshev weighted finite element preconditioner for the Chebyshev case. Some numerical experiments demonstrate that such a preconditioner seems to be optimal.

This paper consists of the following. In section 2, we provide definitions, notations, and basic known facts. In sections 3 and 4, we present Legendre and Chebyshev pseudospectral least-squares methods, respectively, including the norm equivalences and spectral convergences. In section 5, we explain how the linear system can be set up, and propose a block diagonal finite element preconditioner. Finally, we provide several numerical experiments including the condition numbers of the resulting linear system and preconditioned linear system, and spectral convergence of discretization errors in L^2, L^2_w- and H^1, H^1_w-norms.

2. Preliminaries.

In this section, we provide some preliminaries, definitions, and notations for future use. The standard notations and definitions are used for the weighted Sobolev spaces $H^s_w(\Omega)^2$ equipped with weighted inner products $(\cdot, \cdot)_{s,w}$ and corresponding weighted norms $\|\cdot\|_{s,w}$, $s \geq 0$, where $w(x,y) = \hat{w}(x)\hat{w}(y)$ is the Legendre weight function when $\hat{w}(t) = 1$ or Chebyshev weight function when $\hat{w}(t) = \frac{1}{\sqrt{1-t^2}}$. The space $H^0_w(\Omega)$ coincides with $L^2_w(\Omega)$, in which case the norm and inner product will be denoted by $\|\cdot\|_w$ and $(\cdot, \cdot)_w$, respectively. For the Legendre case, we simply write the notations without the subscript w, for example, $H^s(\Omega) := H^s_w(\Omega)$, $(\cdot,\cdot) := (\cdot,\cdot)_w$, $\|\cdot\| := \|\cdot\|_w$ if $w(x,y) = 1$.

Let $\mathcal{P}_N$ be the space of all polynomials of degree less than or equal to N. Let $\{\xi_i\}_{i=0}^N$ be the LGL or CGL points on $[-1,1]$ such that

$$-1 =: \xi_0 < \xi_1 < \cdots < \xi_{N-1} < \xi_N := 1.$$

For the Legendre case, $\{\xi_i\}_{i=0}^N$ are the zeros of $(1-t^2)L'_N(t)$, where L_N is the Nth Legendre polynomial and the corresponding quadrature weights $\{w_i\}_{i=0}^N$ are given by

$$\begin{aligned} w_j &= \frac{2}{N(N+1)}\frac{1}{[L_N(\xi_j)]^2}, \quad 1 \leq j \leq N-1, \\ w_0 = w_N &= \frac{2}{N(N+1)}. \end{aligned} \tag{2.1}$$

For the Chebyshev case, $\{\xi_i\}_{i=0}^N$ are the zeros of $(1-t^2)T'_N(t)$, where T_N is the Nth Chebyshev polynomial and the corresponding quadrature weights $\{w_i\}_{i=0}^N$ are given by

$$w_j = \frac{\pi}{N}, \quad 1 \le j \le N-1, \qquad w_0 = w_N = \frac{\pi}{2N}. \tag{2.2}$$

Then, we have the following LGL or CGL quadrature formula such that

$$\int_{-1}^{1} p(t)\hat{w}(t)\,dt = \sum_{i=0}^{N} w_i\, p(\xi_i) \quad \forall\, p \in \mathcal{P}_{2N-1}. \tag{2.3}$$

Let $\{\phi_i\}_{i=0}^N$ be the set of Lagrange polynomials of degree N with respect to LGL or CGL points $\{\xi_i\}_{i=0}^N$ which satisfy

$$\phi_i(\xi_j) = \delta_{ij} \quad \forall\, i, j = 0, 1, \dots, N,$$

where δ_{ij} denotes the Kronecker delta. The two-dimensional LGL or CGL nodes $\{\mathrm{x}_{ij}\}$ and weights $\{\mathrm{w}_{ij}\}$ are given by

$$\mathrm{x}_{ij} = (\xi_i, \xi_j), \quad \mathrm{w}_{ij} = w_i w_j, \quad i, j = 0, 1, \dots, N.$$

Let $\mathcal{Q}_N$ be the space of all polynomials of degree less than or equal to N with respect to each single variable x and y. Define the basis for $\mathcal{Q}_N$ as

$$\psi_{ij}(x, y) = \phi_i(x)\phi_j(y), \qquad i, j = 0, 1, \dots, N.$$

For any continuous functions u and v on $\bar{\Omega}$, the associated discrete scalar product and norm are

$$\langle u, v\rangle_{w,N} = \sum_{i,j=0}^{N} \mathrm{w}_{ij}\, u(\mathrm{x}_{ij})\, v(\mathrm{x}_{ij}) \quad \text{and} \quad \|v\|_{w,N} = \langle v, v\rangle_{w,N}^{\frac{1}{2}}. \tag{2.4}$$

Then, by (2.3) we have

$$\langle u, v\rangle_{w,N} = (u, v)_w \quad \text{for} \;\; uv \in \mathcal{Q}_{2N-1}. \tag{2.5}$$

It is well known that

$$\|v\|_w \le \|v\|_{w,N} \le \gamma^* \|v\|_w \quad \forall\, v \in \mathcal{Q}_N, \tag{2.6}$$

where $\gamma^* = 2 + \frac{1}{N}$ for the Legendre case and $\gamma^* = 2$ for the Chebyshev case. For any continuous function v on $\bar{\Omega}$, we denote by $I_N v \in \mathcal{Q}_N$ the interpolant of v at the LGL- or CGL-points $\{\mathrm{x}_{ij}\}$ such that

$$I_N v(\mathrm{x}) = \sum_{i,j=0}^{N} v(\mathrm{x}_{ij})\psi_{ij}(\mathrm{x}) \quad \forall\, \mathrm{x} \in \bar{\Omega}. \tag{2.7}$$

The following results are found in [1], [6], and [23]. The interpolation error estimate is known as

$$\|v - I_N v\|_{k,w} \le C\, N^{k-s} \|v\|_{s,w}, \quad k = 0, 1, \tag{2.8}$$

provided $v \in H_w^s(\Omega)$ for $s \ge 2$. Using (2.5)–(2.8), we can show that for any $u \in H_w^s(\Omega)$, $s \ge 2$, and any $v_N \in \mathcal{Q}_N$

$$|(u, v_N)_w - \langle u, v_N\rangle_{w,N}| \le C\, N^{-s}\, \|u\|_{s,w}\, \|v_N\|_w. \tag{2.9}$$

3. Legendre pseudospectral least-squares method. In this section, we investigate the Legendre pseudospectral least-squares method for the first-order system of linear equations equivalent to problem (1.1). Throughout this section, we set $w(x,y)=1$.

Let

$$V := \{v \in H^1(\Omega) : \, v = 0 \text{ on } \Gamma_D\}$$

and

$$\mathbf{W} := \{\mathbf{v} \in H^1(\Omega)^2 : \, \mathbf{n}\cdot\mathbf{v} = 0 \text{ on } \Gamma_N, \quad \boldsymbol{\tau}\cdot\mathbf{v} = 0 \text{ on } \Gamma_D\},$$

where $\mathbf{n}$ and $\boldsymbol{\tau}$ are unit normal and tangent vector, respectively. Let $\mathbf{W}_N = \mathcal{Q}_N^2 \cap \mathbf{W}$ and $V_N = \mathcal{Q}_N \cap V$. Let $\nabla\times$ denote the curl operator given by $\nabla\times\mathbf{v} = \partial_x v_2 - \partial_y v_1$ for a vector function $\mathbf{v} = (v_1, v_2)^T$.

Setting the flux variable $\mathbf{u} = \nabla p$ and using the identities

$$\nabla\times\mathbf{u} = 0 \quad \text{in } \Omega \quad \text{and} \quad \boldsymbol{\tau}\cdot\mathbf{u} = 0 \quad \text{on } \Gamma_D,$$

we employ the first-order system of linear equations equivalent to (1.1) such that

$$\left\{\begin{array}{rcll} \mathbf{u} - \nabla p &=& \mathbf{0} & \text{in } \Omega, \\ -\nabla\cdot\mathbf{u} + \mathbf{b}\cdot\mathbf{u} + c_0\, p &=& f & \text{in } \Omega, \\ \nabla\times\mathbf{u} &=& 0 & \text{in } \Omega, \\ p &=& 0 & \text{on } \Gamma_D, \\ \mathbf{n}\cdot\mathbf{u} &=& 0 & \text{on } \Gamma_N, \\ \boldsymbol{\tau}\cdot\mathbf{u} &=& 0 & \text{on } \Gamma_D. \end{array}\right. \tag{3.1}$$

Define the least-squares functional for the system (3.1) as

$$G(\mathbf{v},q;f) = \|f + \nabla\cdot\mathbf{v} - \mathbf{b}\cdot\mathbf{v} - c_0\, q\|^2 + \|\mathbf{v} - \nabla q\|^2 + \|\nabla\times\mathbf{v}\|^2 \tag{3.2}$$

for $(\mathbf{v},q) \in \mathbf{W}\times V$. The first-order system least-squares variational problem for (3.1) is to minimize the quadratic functional $G(\mathbf{v},q;f)$ over $\mathbf{W}\times V$: find $(\mathbf{u},p) \in \mathbf{W}\times V$ such that

$$G(\mathbf{u},p;f) = \inf_{(\mathbf{v},q)\in\mathbf{W}\times V} G(\mathbf{v},q;f). \tag{3.3}$$

The corresponding variational problem is to find $(\mathbf{u},p) \in \mathbf{W}\times V$ such that

$$a(\mathbf{u},p;\mathbf{v},q) = f(\mathbf{v},q) \quad \forall\, (\mathbf{v},q) \in \mathbf{W}\times V, \tag{3.4}$$

where the bilinear form $a(\cdot;\cdot)$ is given by

$$\begin{aligned} a(\mathbf{u},p;\mathbf{v},q) &= (\nabla\cdot\mathbf{u} - \mathbf{b}\cdot\mathbf{u} - c_0\, p,\ \nabla\cdot\mathbf{v} - \mathbf{b}\cdot\mathbf{v} - c_0\, q) \\ &\quad + (\mathbf{u} - \nabla p,\ \mathbf{v} - \nabla q) + (\nabla\times\mathbf{u},\ \nabla\times\mathbf{v}) \end{aligned} \tag{3.5}$$

and the linear form $f(\cdot)$ is given by

$$f(\mathbf{v},q) = -(f, \nabla\cdot\mathbf{v} - \mathbf{b}\cdot\mathbf{v} - c_0\, q). \tag{3.6}$$

THEOREM 3.1. *For any $(\mathbf{v},q) \in \mathbf{W}\times V$, there exists a positive constant C such that*

$$\frac{1}{C}\big(\|\mathbf{v}\|_1^2 + \|q\|_1^2\big) \le a(\mathbf{v},q;\mathbf{v},q) \le C\big(\|\mathbf{v}\|_1^2 + \|q\|_1^2\big). \tag{3.7}$$

Proof. The functional (3.2) is a special case of the general form given in [5]. Hence, the continuity and ellipticity of the bilinear form $a(\cdot;\cdot)$ are immediate consequences of [5]. □

Define the discrete least-squares functional using the discrete Legendre spectral norm as

$$G_N(\mathbf{v},q;f)=\|f+\nabla\cdot\mathbf{v}-\mathbf{b}\cdot\mathbf{v}-c_0\,q\|_N^2+\|\mathbf{v}-\nabla q\|_N^2+\|\nabla\times\mathbf{v}\|_N^2 \tag{3.8}$$

for $(\mathbf{v},q)\in\mathbf{W}_N\times V_N$. The discrete least-squares problem associated with (3.8) is then to minimize the quadratic functional $G_N(\mathbf{v},q;f)$ over $\mathbf{W}_N\times V_N$, and the corresponding variational problem (Legendre pseudospectral collocation problem) is to find $(\mathbf{u}_N,\,p_N)\in\mathbf{W}_N\times V_N$ such that

$$a_N(\mathbf{u}_N,\,p_N;\mathbf{v},q)=f_N(\mathbf{v},q)\quad\forall\,(\mathbf{v},q)\in\mathbf{W}_N\times V_N, \tag{3.9}$$

where the discrete bilinear form $a_N(\cdot;\cdot)$ and linear form $f_N(\cdot)$ are given by

$$\begin{aligned}a_N(\mathbf{u}_N,\,p_N;\mathbf{v},q)&=\langle\nabla\cdot\mathbf{u}_N-\mathbf{b}\cdot\mathbf{u}_N-c_0\,p_N,\nabla\cdot\mathbf{v}-\mathbf{b}\cdot\mathbf{v}-c_0\,q\rangle_N\\&\quad+\langle\mathbf{u}_N-\nabla p_N,\mathbf{v}-\nabla q\rangle_N+\langle\nabla\times\mathbf{u}_N,\nabla\times\mathbf{v}\rangle_N\end{aligned} \tag{3.10}$$

and

$$f_N(\mathbf{v},q)=-\langle f,\nabla\cdot\mathbf{v}-\mathbf{b}\cdot\mathbf{v}-c_0\,q\rangle_N. \tag{3.11}$$

The continuity and ellipticity of the discrete functional $G_N(\cdot;0)$ are shown in the following theorem.

THEOREM 3.2. *For any $(\mathbf{v},q)\in\mathbf{W}_N\times V_N$, there exists a constant C such that*

$$\frac{1}{C}\left(\|\mathbf{v}\|_1^2+\|q\|_1^2\right)\le G_N(\mathbf{v},q;0)\le C\left(\|\mathbf{v}\|_1^2+\|q\|_1^2\right). \tag{3.12}$$

Proof. Since $\mathbf{v}-\nabla q\in\mathcal{Q}_N^2$ and $\nabla\cdot\mathbf{v}-\mathbf{b}\cdot\mathbf{v}-c_0\,q$, $\nabla\times\mathbf{v}\in\mathcal{P}_N$, we can easily show from (2.6) that

$$G(\mathbf{v},q;0)\le G_N(\mathbf{v},q;0)\le\left(2+\frac{1}{N}\right)^2G(\mathbf{v},q;0).$$

Hence, the bounds (3.12) are an immediate consequence of the Theorem 3.1. □

We show the spectral convergence in the H^1 product norm for the Legendre pseudospectral least-squares method following the same techniques used in [1] and [23]. Similarly one may show the L^2 convergence (see [1] or [23]).

THEOREM 3.3. *Assume that the solution $(\mathbf{u},p)$ of (3.4) is in $H^s(\Omega)^3$ for some $s\ge1$ and $f\in H^\ell(\Omega)$ for some integer $\ell\ge2$. Let $(\mathbf{u}_N,p_N)\in\mathbf{W}_N\times V_N$ be the discrete solution of the problem (3.9). Then there exists a constant C such that*

$$\|\mathbf{u}-\mathbf{u}_N\|_1+\|p-p_N\|_1\le C\left[N^{1-s}(\|\mathbf{u}\|_s+\|p\|_s)+N^{-\ell}\|f\|_\ell\right]. \tag{3.13}$$

Proof. Using the first Strang lemma, which uses the coercivity of $a_N(\cdot;\cdot)$ and the continuity of $a(\cdot;\cdot)$, one may easily verify from [23] (or see in [1, p. 88]) that

$$\begin{aligned}C\big(\|\mathbf{u}-\mathbf{u}_N\|_1+\|p-p_N\|_1\big)\le\inf_{(\mathbf{v},q)\in\mathbf{W}_N\times V_N}&\Bigg[\|\mathbf{u}-\mathbf{v}\|_1+\|p-q\|_1\\&+\sup_{(\mathbf{w},r)\in\mathbf{W}_N\times V_N}\frac{|a(\mathbf{v},q;\mathbf{w},r)-a_N(\mathbf{v},q;\mathbf{w},r)|}{\|\mathbf{w}\|_1+\|r\|_1}\Bigg]\\&+\sup_{(\mathbf{w},r)\in\mathbf{W}_N\times V_N}\frac{|f(\mathbf{w},r)-f_N(\mathbf{w},r)|}{\|\mathbf{w}\|_1+\|r\|_1}.\end{aligned} \tag{3.14}$$

If we take $\mathbf{v} \in \mathbf{W}_{N-1}$ and $q \in V_{N-1}$, we see from (2.5) that

$$a(\mathbf{v}, q; \mathbf{w}, r) = a_N(\mathbf{v}, q; \mathbf{w}, r) \quad \forall\, (\mathbf{w}, r) \in \mathbf{W}_N \times V_N.$$

Using the inequality (2.9) yields

$$|f(\mathbf{w}, r) - f_N(\mathbf{w}, r)| \le C\, N^{-\ell}\, \|f\|_\ell\, \|\nabla \cdot \mathbf{w} - \mathbf{b} \cdot \mathbf{w} - c_0\, r\|, \quad \ell \ge 2.$$

Since

$$\|\nabla \cdot \mathbf{w} - \mathbf{b} \cdot \nabla r - c_0\, r\| \le C \left(\|\mathbf{w}\|_1 + \|r\|_1\right) \quad \forall\, (\mathbf{w}, r) \in \mathbf{W}_N \times V_N,$$

we have

$$\frac{|f(\mathbf{w}, r) - f_N(\mathbf{w}, r)|}{\|\mathbf{w}\|_1 + \|r\|_1} \le C\, N^{-\ell}\, \|f\|_\ell, \quad \ell \ge 2.$$

Now, using (2.8), we deduce the conclusion (3.13). □

4. Chebyshev pseudospectral least-squares method. In this section, we investigate the Chebyshev pseudospectral least-squares method for the first-order system of linear equations equivalent to problem (1.1). We will use the same notation used in the previous section, but the definitions may be different from those of the previous section. Throughout this section, we set $w(x, y) = \hat{w}(x)\hat{w}(y)$, with $\hat{w}(t) = \frac{1}{\sqrt{1-t^2}}$ and $\Gamma_N = \emptyset$.

We redefine the spaces V and $\mathbf{W}$ such that

$$V := \{v \in H^1_w(\Omega) : \ v = 0 \text{ on } \ \partial\Omega\}$$

and

$$\mathbf{W} := \{\mathbf{v} \in L^2_w(\Omega)^2 : \ \|\mathbf{v}\|_W < \infty \quad \text{and} \quad \boldsymbol{\tau} \cdot \mathbf{v} = 0 \text{ on } \ \partial\Omega\},$$

which is a Hilbert space equipped with the norm

$$\|\mathbf{v}\|_W := \left(\|\mathbf{v}\|_w^2 + \|\nabla \cdot \mathbf{v}\|_w^2 + \|\nabla \times \mathbf{v}\|_w^2\right)^{\frac{1}{2}}.$$

Let $H^{-1}_w(\Omega)$ be the dual space of V equipped with the norm (see [1, p. 18])

$$\|u\|_{-1,w} := \sup_{0 \ne \phi \in V} \frac{(u, \phi)_w}{\|\phi\|_{1,w}}. \tag{4.1}$$

Let $\mathbf{W}_N = Q_N^2 \cap \mathbf{W}$ and $V_N = Q_N \cap V$. The first-order system (3.1) is rewritten as

$$\left\{\begin{array}{rcll} \mathbf{u} - \nabla p & = & \mathbf{0} & \text{in } \Omega, \\ -\nabla \cdot \mathbf{u} + \mathbf{b} \cdot \mathbf{u} + c_0\, p & = & f & \text{in } \Omega, \\ \nabla \times \mathbf{u} & = & 0 & \text{in } \Omega, \\ p & = & 0 & \text{on } \partial\Omega, \\ \boldsymbol{\tau} \cdot \mathbf{u} & = & 0 & \text{on } \partial\Omega. \end{array}\right. \tag{4.2}$$

Define the least-squares functional corresponding to (4.2) using the Chebyshev weighted L^2-norms as

$$G_w(\mathbf{v}, q; f) = \|f + \nabla \cdot \mathbf{v} - \mathbf{b} \cdot \mathbf{v} - c_0\, q\|_w^2 + \|\mathbf{v} - \nabla q\|_w^2 + \|\nabla \times \mathbf{v}\|_w^2 \tag{4.3}$$

for $(\mathbf{v}, q) \in \mathbf{W} \times V$. Then the first-order system least-squares variational problem for (4.2) is to minimize the quadratic functional $G_w(\mathbf{v}, q; f)$ over $\mathbf{W} \times V$: find $(\mathbf{u}, p) \in \mathbf{W} \times V$ such that

$$G_w(\mathbf{u}, p; f) = \inf_{(\mathbf{v},q)\in\mathbf{W}\times V} G_w(\mathbf{v}, q; f), \tag{4.4}$$

and the variational problem for (4.4) is to find $(\mathbf{u}, p) \in \mathbf{W} \times V$ such that

$$a_w(\mathbf{u}, p; \mathbf{v}, q) = f_w(\mathbf{v}, q) \quad \forall\, (\mathbf{v}, q) \in \mathbf{W} \times V, \tag{4.5}$$

where the bilinear form $a_w(\cdot;\cdot)$ is given by

$$\begin{aligned} a_w(\mathbf{u}, p; \mathbf{v}, q) = {} & (\nabla\cdot\mathbf{u} - \mathbf{b}\cdot\mathbf{u} - c_0\, p, \nabla\cdot\mathbf{v} - \mathbf{b}\cdot\mathbf{v} - c_0\, q)_w \\ & + (\mathbf{u} - \nabla p, \mathbf{v} - \nabla q)_w + (\nabla\times\mathbf{u}, \nabla\times\mathbf{v})_w \end{aligned} \tag{4.6}$$

and the linear form $f_w(\cdot)$ is given by

$$f_w(\mathbf{v}, q) = -(f, \nabla\cdot\mathbf{v} - \mathbf{b}\cdot\mathbf{v} - c_0\, q)_w. \tag{4.7}$$

From now on, we will establish the coercivity and continuity for the homogeneous Chebyshev functional $G_w(\cdot; 0)$ over $\mathbf{W} \times V$. We recall the Poincaré–Friedrichs inequality such that (see [6])

$$\|v\|_w \le C\|\nabla v\|_w \quad \forall\, v \in V, \tag{4.8}$$

where C is a positive constant.

LEMMA 4.1. *For $\phi(x, y) \in V$, the following hold:*

(a) *There are two positive constants c and C such that*

$$c\|\phi\|^2_{1,w} \le \int_\Omega \nabla\phi\cdot\nabla(\phi w)\, dxdy \le C\|\phi\|^2_{1,w}.$$

(b) *There is a constant C such that, with $t = x$ or y,*

$$\left| \int_\Omega \phi^2(x, y) \frac{t^2}{(1 - t^2)^2} \hat{w}(x)\hat{w}(y)\, dxdy \right| \le C\|\phi\|^2_{1,w}.$$

Proof. First note that (a) is found in, for example, [6], [10], or [19]. For the proof of (b), recall from Lemma 2 in [19] that

$$\left| \int_{-1}^{1} u^2(t) \frac{t^2}{(1 - t^2)^2} \hat{w}(t)\, dt \right| \le \frac{4}{9}|u|^2_{H^1_w(-1,1)} \quad \text{for} \quad u \in H^1_{0,w}(-1, 1).$$

From this estimate one may prove the conclusion (b). □

LEMMA 4.2. *For any $\mathbf{v} \in \mathbf{W}$, we have*

$$\|\nabla\cdot\mathbf{v}\|_{-1,w} \le C\,\|\mathbf{v}\|_w.$$

Proof. Using the divergence theorem yields that, for $\phi \in V$ and $\mathbf{v} = (v_1, v_2)^T \in \mathbf{W}$,

$$(\nabla\cdot\mathbf{v}, \phi)_w = -\,(\mathbf{v}, \nabla(\phi w)) = -(v_1, \phi_x w + \phi w_x) - (v_2, \phi_y w + \phi w_y), \tag{4.9}$$

where we define $\phi_t = \frac{\partial}{\partial t}\phi(x,y)$ for $t = x$ or y. By the Schwarz inequality, we have

$$|(v_i, \phi_t w)| = |(v_i, \phi_t)_w| \le \|v_i\|_w \, \|\phi\|_{1,w}, \quad i = 1, 2, \quad t = x, y. \tag{4.10}$$

On the other hand, with $t = x$ or y, the Schwarz inequality and Lemma 4.1 yield

$$\begin{aligned} |(v_i, \phi w_t)| &= \left| \int_\Omega v_i(x,y)\phi(x,y)\, \frac{t}{1-t^2}\, w(x,y)\, dxdy \right| \\ &\le \left(\int_\Omega v_i(x,y)^2 w(x,y)\, dxdy \right)^{\frac{1}{2}} \left(\int_\Omega \phi^2(x,y)\, \frac{t^2}{(1-t^2)^2}\, w(x,y)\, dxdy \right)^{\frac{1}{2}} \\ &\le C \, \|v_i\|_w \, \|\phi\|_{1,w}. \end{aligned} \tag{4.11}$$

Combining (4.9) with (4.10) and (4.11), we have

$$(\nabla \cdot \mathbf{v}, \phi)_w \le C \|\mathbf{v}\|_w \, \|\phi\|_{1,w} \quad \forall\, \phi \in V,$$

which completes the conclusion via the definition of $\| \cdot \|_{-1,w}$. □

LEMMA 4.3. *There exists a constant C such that*

$$\|p\|_{1,w} \le C \, \| - \Delta p\|_{-1,w} \quad \forall\, p \in V. \tag{4.12}$$

Proof. It is well known from Theorem 11.1 in [6] that there exists a constant C such that

$$\|p\|_{1,w}^2 \le C \, (\nabla p, \nabla(pw)) = C \, (-\Delta p, p)_w \quad \forall\, p \in V.$$

Hence, by the definition of $\| \cdot \|_{-1,w}$, we have the conclusion (4.12). □

Due to the above lemma, from now on we may assume that there are constant coefficients $\mathbf{b}$ and c_0 satisfying the following a priori estimate:

$$\|p\|_{1,w} \le C \, \| - \Delta p + \mathbf{b} \cdot \nabla p + c_0 \, p\|_{-1,w} \quad \forall\, p \in V. \tag{A_0}$$

Indeed, for the case $\mathbf{b} = \mathbf{0}$ and $c_0 > 0$ it is clear, and for the other case we can find the restrictions to $\mathbf{b}$ and c_0 following the arguments in [6].

Now, under the assumption (A_0), we establish the coercivity and continuity for the homogeneous Chebyshev least-squares functional $G_w(\cdot; 0)$ over $\mathbf{W} \times V$ as follows.

THEOREM 4.4. *Assume that (A_0) holds. Then there exists a positive constant C such that*

$$\frac{1}{C}\big(\|\mathbf{v}\|_W^2 + \|q\|_{1,w}^2\big) \le G_w(\mathbf{v}, q, ; 0) \le C\big(\|\mathbf{v}\|_W^2 + \|q\|_{1,w}^2\big) \quad \forall\, (\mathbf{v}, q) \in \mathbf{W} \times V. \tag{4.13}$$

Proof. The triangle inequality yields the upper bound. For the lower bound, let $(\mathbf{v}, q) \in \mathbf{W} \times V$. By (A_0), the triangle inequality, and Lemma 4.2, we have

$$\begin{aligned} \|q\|_{1,w}^2 &\le C \, \|\Delta q - \mathbf{b} \cdot \nabla q - c_0 \, q\|_{-1,w}^2 \\ &\le C \, \big(\|\nabla \cdot \mathbf{v} - \mathbf{b} \cdot \mathbf{v} - c_0 \, q\|_{-1,w}^2 + \|\nabla \cdot (\mathbf{v} - \nabla q)\|_{-1,w}^2 + \|\mathbf{b} \cdot (\mathbf{v} - \nabla q)\|_{-1,w}^2\big) \\ &\le C \, \big(\|\nabla \cdot \mathbf{v} - \mathbf{b} \cdot \mathbf{v} - c_0 \, q\|_w^2 + \|\mathbf{v} - \nabla q\|_w^2\big) \\ &\le C \, G_w(\mathbf{v}, q; 0). \end{aligned}$$

Using the triangle inequality together with the last inequality, we have

$$\|\mathbf{v}\|_w^2 \le C \, \big(\|\mathbf{v} - \nabla q\|_w^2 + \|\nabla q\|_w^2\big) \le C \, G_w(\mathbf{v}, q; 0)$$

and

$$\begin{aligned}\|\nabla\cdot\mathbf{v}\|_w^2 &\le C\,\left(\|\nabla\cdot\mathbf{v}-\mathbf{b}\cdot\mathbf{v}-c_0\,q\|_w^2+\|\mathbf{b}\cdot\mathbf{v}+c_0\,q\|_w\right)\\ &\le C\,(G_w(\mathbf{v},q;0)+\|\mathbf{v}\|_w^2+\|q\|_w^2)\le C\,G_w(\mathbf{v},q;0).\end{aligned}$$

Thus we have

$$\|\mathbf{v}\|_W^2=\|\mathbf{v}\|_w^2+\|\nabla\cdot\mathbf{v}\|_w^2+\|\nabla\times\mathbf{v}\|_w^2\le C\,G_w(\mathbf{v},q;0),$$

which completes the theorem. □

Define the discrete least-squares functionals using the discrete Chebyshev norm as

$$G_{w,N}(\mathbf{v},q;f)=\|f+\nabla\cdot\mathbf{v}-\mathbf{b}\cdot\mathbf{v}-c_0\,q\|_{w,N}^2+\|\mathbf{v}-\nabla q\|_{w,N}^2+\|\nabla\times\mathbf{v}\|_{w,N}^2 \tag{4.14}$$

for $(\mathbf{v},q)\in\mathbf{W}_N\times V_N$. The discrete least-squares problem associated with (4.14) is then to minimize the quadratic functional $G_{w,N}(\mathbf{v},q;f)$ over $\mathbf{W}_N\times V_N$, and the corresponding variational problem (Chebyshev pseudospectral collocation problem) is to find $(\mathbf{u}_N,\,p_N)\in\mathbf{W}_N\times V_N$ such that

$$a_{w,N}(\mathbf{u}_N,\,p_N;\mathbf{v},q)=f_{w,N}(\mathbf{v},q)\quad\forall\,(\mathbf{v},q)\in\mathbf{W}_N\times V_N, \tag{4.15}$$

where the discrete bilinear form $a_{w,N}(\cdot;\cdot)$ and linear form $f_{w,N}(\cdot)$ are given by

$$\begin{aligned}a_{w,N}(\mathbf{u}_N,\,p_N;\mathbf{v},q)&=\langle\nabla\cdot\mathbf{u}_N-\mathbf{b}\cdot\mathbf{u}_N-c_0\,p_N,\nabla\cdot\mathbf{v}-\mathbf{b}\cdot\mathbf{v}-c_0\,q\rangle_{w,N}\\ &\quad+\langle\mathbf{u}_N-\nabla p_N,\mathbf{v}-\nabla q\rangle_{w,N}+\langle\nabla\times\mathbf{u}_N,\nabla\times\mathbf{v}\rangle_{w,N}\end{aligned} \tag{4.16}$$

and

$$f_{w,N}(\mathbf{v},q)=-\langle f,\nabla\cdot\mathbf{v}-\mathbf{b}\cdot\mathbf{v}-c_0\,q\rangle_{w,N}. \tag{4.17}$$

The continuity and coercivity of the discrete functional $G_{w,N}(\cdot;0)$ are given in the following theorem.

THEOREM 4.5. *Assume that* (A_0) *holds. There exists a constant* C *such that*

$$\frac{1}{C}\left(\|\mathbf{v}\|_W^2+\|q\|_{1,w}^2\right)\le G_{w,N}(\mathbf{v},q;0)\le C\left(\|\mathbf{v}\|_W^2+\|q\|_{1,w}^2\right) \tag{4.18}$$

for all $(\mathbf{v},q)\in\mathbf{W}_N\times V_N$.

Proof. Since $\mathbf{v}-\nabla q\in\mathcal{Q}_N^2$ and $\nabla\cdot\mathbf{v}-\mathbf{b}\cdot\mathbf{v}-c_0\,q$, $\nabla\times\mathbf{v}\in\mathcal{P}_N$, we can easily show from (2.6) that

$$G_w(\mathbf{v},q;0)\le G_{w,N}(\mathbf{v},q;0)\le 4G_w(\mathbf{v},q;0).$$

Now, the bounds (4.18) are an immediate consequence of Theorem 4.4. □

Applying again the same techniques in [1] and [23] to our case, we show the spectral convergence for the Chebyshev pseudospectral least-squares method.

THEOREM 4.6. *Assume that* (A_0) *holds and that the solution* $(\mathbf{u},p)$ *of* (4.5) *is in* $H_w^s(\Omega)^3$ *for some* $s\ge 1$ *and* $f\in H_w^\ell(\Omega)$ *for some integer* $\ell\ge 2$. *Let* $(\mathbf{u}_N,p_N)\in\mathbf{W}_N\times V_N$ *be the discrete solution of problem* (4.15). *Then there exists a constant* C *such that*

$$\|\mathbf{u}-\mathbf{u}_N\|_W+\|p-p_N\|_{1,w}\le C\left[N^{1-s}(\|\mathbf{u}\|_{s,w}+\|p\|_{s,w})+N^{-\ell}\|f\|_{\ell,w}\right]. \tag{4.19}$$

Proof. Using the first Strang lemma, which uses the coercivity of $a_{w,N}(\cdot;\cdot)$ and the continuity of $a_w(\cdot;\cdot)$, one may easily verify from [23] (or see [1, p. 88]) that

$$\begin{aligned} C\big(\|\mathbf{u}-\mathbf{u}_N\|_W + \|p-p_N\|_{1,w}\big) \le \inf_{(\mathbf{v},q)\in \mathbf{W}_N\times V_N} \Bigg[&\|\mathbf{u}-\mathbf{v}\|_W + \|p-q\|_{1,w} \\ &+ \sup_{(\mathbf{w},r)\in \mathbf{W}_N\times V_N} \frac{|a_w(\mathbf{v},q;\mathbf{w},r) - a_{w,N}(\mathbf{v},q;\mathbf{w},r)|}{\|\mathbf{w}\|_W + \|r\|_{1,w}} \Bigg] \\ &+ \sup_{(\mathbf{w},r)\in \mathbf{W}_N\times V_N} \frac{|f_w(\mathbf{w},r) - f_{w,N}(\mathbf{w},r)|}{\|\mathbf{w}\|_W + \|r\|_{1,w}}. \end{aligned} \tag{4.20}$$

If we take $\mathbf{v} \in \mathbf{W}_{N-1}$ and $q \in V_{N-1}$, we see from (2.5) that

$$a_w(\mathbf{v},q;\mathbf{w},r) = a_{w,N}(\mathbf{v},q;\mathbf{w},r) \quad \forall\,(\mathbf{w},r) \in \mathbf{W}_N \times V_N.$$

Using the inequality (2.9) yields

$$|f_w(\mathbf{w},r) - f_{w,N}(\mathbf{w},r)| \le C\,N^{-\ell}\,\|f\|_{\ell,w}\,\|\nabla\cdot\mathbf{w} - \mathbf{b}\cdot\mathbf{w} - c_0\,r\|_w, \quad \ell \ge 2.$$

Since

$$\|\nabla\cdot\mathbf{w} - \mathbf{b}\cdot\mathbf{w} - c_0\,r\|_w \le C\left(\|\mathbf{w}\|_W + \|r\|_{1,w}\right) \quad \forall\,(\mathbf{w},r) \in \mathbf{W}_N \times V_N,$$

we have

$$\frac{|f_w(\mathbf{w},r) - f_{w,N}(\mathbf{w},r)|}{\|\mathbf{w}\|_W + \|r\|_{1,w}} \le C\,N^{-\ell}\,\|f\|_{\ell,w}, \quad \ell \ge 2.$$

Now, using (2.8), we deduce the conclusion (4.19). □

5. Implementation and preconditioning.

5.1. Implementation. The computation for problems (3.9) and (4.15) can be easily implemented by using the one-dimensional pseudospectral matrix D_N associated with the $N+1$ values $\{p(\xi_j)\}_{j=0}^N$ and the $N+1$ values $\{(\partial_N p)(\xi_j)\}_{j=0}^N$ of the pseudospectral derivative of p at LGL or CGL points (see [6], [23]). In this section we give the implementation of Legendre pseudospectral least-squares approximation. One may similarly obtain the implementation for the Chebyshev approximation. First, we reorder the LGL points from bottom to top and then left to right such that $\mathrm{x}_{k(N+1)+l} := \mathrm{x}_{kl} = (\xi_k, \xi_l)$ for $k,l = 0,1,\ldots,N$. The basis functions $\psi_{k(N+1)+l}(x,y) := \psi_{kl}(x,y) = \phi_k(x)\phi_l(y)$ and quadrature weights $\mathrm{w}_{k(N+1)+l} := \mathrm{w}_{kl} = w_k w_l$ are reordered accordingly. Then two-dimensional Legendre pseudospectral matrices S_x and S_y related to $\{(\partial_x p)(\mathrm{x}_j)\}_{j=0}^{(N+1)^2-1}$ and $\{(\partial_y p)(\mathrm{x}_j)\}_{j=0}^{(N+1)^2-1}$ of the pseudospectral partial derivatives of p, respectively, are given by the tensor products of the identity matrix I_N and one-dimensional Legendre pseudospectral matrix D_N such that

$$S_x = D_N \otimes I_N \quad \text{and} \quad S_y = I_N \otimes D_N.$$

Indeed, the (i,j)-entries of S_x and S_y are given by $\partial_x\psi_j(\mathrm{x}_i)$ and $\partial_y\psi_j(\mathrm{x}_i)$, respectively. Let $W = \mathrm{diag}\{\mathrm{w}_i\}$ be the diagonal weight matrix.

Let $\mathcal{A}_N$ be the matrix corresponding to the bilinear form $a_N(\cdot;\cdot)$. Then $\mathcal{A}_N$ is a symmetric 3×3 block matrix. The same basis functions ψ_j, except for the basis functions corresponding to nodes on the boundaries where the solution is required to be 0, are used to approximate all components of the function $(\mathbf{u}_N, p_N) \in \mathbf{W}_N \times V_N$. Thus, it is more convenient to assemble the matrix $\mathcal{A}_N^*$ using all basis functions of $\mathcal{Q}_N$ and ignoring the boundary conditions. Then, the matrix $\mathcal{A}_N$ can be obtained from $\mathcal{A}_N^*$ by eliminating the rows and columns relative to the nodes on the appropriate boundaries. Hence, we will hereafter regard $\mathbf{W}_N \times V_N$ as $\mathcal{Q}_N^3$ in order to give a convenient description of assembly for $\mathcal{A}_N$. We denote by $\hat{\mathbf{s}}$ the vector containing the nodal values of a continuous function $\mathbf{s}$, that is,

$$\hat{\mathbf{s}} = (s(x_0), \ldots, s(x_{(N+1)^2-1}))^T.$$

Using the expressions of

$$\partial_t p(\mathrm{x}_i) = \sum_{j=0}^{(N+1)^2-1} \partial_t \psi_j(\mathrm{x}_i) p(\mathrm{x}_j) = (S_t \hat{\mathbf{p}})_i \quad \text{for } t = x \text{ or } y \text{ and } p \in \mathcal{Q}_N,$$

we have that, for every $p, q \in \mathcal{Q}_N$,

$$\langle p, q \rangle_N = \hat{\mathbf{q}}^T \, W \, \hat{\mathbf{p}} \quad \text{and} \quad \langle \partial_{t_1} p, \partial_{t_2} q \rangle_N = (S_{t_2} \hat{\mathbf{q}})^T \, W \, (S_{t_1} \hat{\mathbf{p}}),$$

where t_1 and t_2 are x or y. Then we obtain

$$\begin{aligned}
&\langle \mathbf{u} - \nabla p, \, \mathbf{v} - \nabla q \rangle_N = (\hat{\mathbf{v}}_1 - S_x \hat{\mathbf{q}})^T \, W \, (\hat{\mathbf{u}}_1 - S_x \hat{\mathbf{p}}) + (\hat{\mathbf{v}}_2 - S_y \hat{\mathbf{q}})^T \, W \, (\hat{\mathbf{u}}_2 - S_y \hat{\mathbf{p}}), \\
&\langle \nabla \cdot \mathbf{u} - \mathbf{b} \cdot \mathbf{u} - c_0 \, p, \, \nabla \cdot \mathbf{v} - \mathbf{b} \cdot \mathbf{v} - c_0 \, q \rangle_N \\
&\quad = (S_x \hat{\mathbf{v}}_1 + S_y \hat{\mathbf{v}}_2 - b_1 \hat{\mathbf{v}}_1 - b_2 \hat{\mathbf{v}}_2 - c_0 \hat{\mathbf{q}})^T \, W \, (S_x \hat{\mathbf{u}}_1 + S_y \hat{\mathbf{u}}_2 - b_1 \hat{\mathbf{u}}_1 - b_2 \hat{\mathbf{u}}_2 - c_0 \hat{\mathbf{p}}), \\
&\langle \nabla \times \mathbf{u}, \, \nabla \times \mathbf{v} \rangle_N = (-S_y \hat{\mathbf{v}}_1 + S_x \hat{\mathbf{v}}_2)^T \, W \, (-S_y \hat{\mathbf{u}}_1 + S_x \hat{\mathbf{u}}_2),
\end{aligned}$$

and

$$\langle f, \, \nabla \cdot \mathbf{v} - \mathbf{b} \cdot \mathbf{v} - c_0 \, q \rangle_N = (S_x \hat{\mathbf{v}}_1 + S_y \hat{\mathbf{v}}_2 - b_1 \hat{\mathbf{v}}_1 - b_2 \hat{\mathbf{v}}_2 - c_0 \hat{\mathbf{q}})^T \, W \, \hat{\mathbf{f}},$$

where $\mathbf{u} = (u_1, u_2)^T$, $\mathbf{v} = (v_1, v_2)^T$, and $\mathbf{b} = (b_1, b_2)^T$. Now, the symmetric matrix $\mathcal{A}_N^* = \big(A_{ij}\big)_{i,j=1,2,3}$ corresponding to $a_N(\cdot;\cdot)$ consists of

$$\begin{aligned}
A_{11} &= S_x^T \, W \, S_x + S_y^T \, W \, S_y + W - b_1 (S_x^T \, W + W \, S_x) + b_1^2 \, W, \\
A_{12} &= S_x^T \, W \, S_y - S_y^T \, W \, S_x - b_2 S_x^T \, W + b_1 b_2 \, W - b_1 \, W \, S_y, \\
A_{13} &= -W \, S_x - c_0 \, S_x^T \, W + b_1 c_0 \, W, \\
A_{22} &= S_x^T \, W \, S_x + S_y^T \, W \, S_y + W - b_2 (S_y^T \, W + W \, S_y) + b_2^2 \, W, \\
A_{23} &= -W \, S_y - c_0 \, S_y^T \, W + b_2 c_0 \, W, \\
A_{33} &= S_x^T \, W \, S_x + S_y^T \, W \, S_y + c_0^2 \, W.
\end{aligned}$$

Also, the vector F_N^* corresponding to $f_N(\mathbf{v}, q)$ consists of

$$F_N^* = \begin{pmatrix} -S_x^T \, W + b_1 \, W \\ -S_y^T \, W + b_2 \, W \\ c_0 \, W \end{pmatrix} \hat{\mathbf{f}}.$$

Let $\mathcal{A}_N$ and F_N be the matrix and vector eliminated rows and columns from $\mathcal{A}_N^*$ and F_N^* relative to the nodes on the boundary where the solution is required to be 0. Now, we are led to the matrix problem associated with (3.9):

$$\mathcal{A}_N \, X = F_N, \tag{5.1}$$

where $X = (\hat{\mathbf{u}}_1, \hat{\mathbf{u}}_2, \hat{\mathbf{p}})^T$.

5.2. Preconditioning. The spectral system (5.1) is generally handled with iterative methods due to the topological structure of $\mathcal{A}_N$, which makes direct methods unsuitable. One may employ a preconditioning conjugate gradient method for an efficient implementation of a linear system (5.1). In this section we use a finite element preconditioner generated by bilinear elements based on the Gauss–Lobatto nodes, which can be found in [7], [8], [19], [20], and [24].

Let B_N be the space of continuous piecewise bilinear functions with respect to the grid induced by the Gauss–Lobatto nodes $\{\mathrm{x}_i\}_{i=0}^{(N+1)^2-1}$, and let $\{\varphi_i\}_{i=0}^{(N+1)^2-1}$ be its nodal basis functions of B_N. Define the interpolation operator $J_N : \mathcal{Q}_N \to B_N$ by

$$J_N\, q = \sum_{j=0}^{(N+1)^2-1} q(\mathrm{x}_j)\varphi_j \in B_N \quad \text{for given } q = \sum_{j=0}^{(N+1)^2-1} q(\mathrm{x}_j)\psi_j \in Q_N.$$

Let $\mathcal{J}_N : \mathbf{W}_N \times V_N \to B_N^3$ be the interpolation operator given by

$$\mathcal{J}_N(v_1, v_2, q) = (J_N\, v_1, J_N\, v_2, J_N\, q) \qquad \forall\, (v_1, v_2, q) \in \mathbf{W}_N \times V_N.$$

Then, for the Legendre approximation we have from [20] that

$$\frac{1}{C}\, \|(\mathbf{v}, q)\|_1 \le \|\mathcal{J}_N(\mathbf{v}, q)\|_1 \le C\, \|(\mathbf{v}, q)\|_1 \qquad \forall\, (\mathbf{v}, q) \in \mathbf{W}_N \times V_N,$$

and for the Chebyshev approximation we have from [19] that

$$\frac{1}{C}\, \|(\mathbf{v}, q)\|_{1,w} \le \|\mathcal{J}_N(\mathbf{v}, q)\|_{1,w} \le C\, \|(\mathbf{v}, q)\|_{1,w} \qquad \forall\, (\mathbf{v}, q) \in \mathbf{W}_N \times V_N.$$

For the Legendre approximation, from (3.12) we have the following equivalent relation:

$$\frac{1}{C}\, \|\mathcal{J}_N(\mathbf{v}, q)\|_1 \le a_N(\mathbf{v}, q;\, \mathbf{v}, q) \le C\, \|\mathcal{J}_N(\mathbf{v}, q)\|_1 \quad \forall\, (\mathbf{v}, q) \in \mathbf{W}_N \times V_N. \tag{5.2}$$

Let $\mathcal{B}_{\mathcal{N}} := \mathcal{J}_N(\mathbf{W}_N \times V_N)$ be the subspace of B_N^3. Let the linear operator L be defined by $Lp = -\Delta p + x\mathbf{b} \cdot \nabla p + c_0 p$. The idea of optimal preconditioning in [16] is to consider preconditioning by the leading term $Bp = -\Delta p + \beta p$ for L, in which the choice of a nonnegative β is discussed so that B remains an optimal preconditioner. In this paper, using such ideas, we propose a finite element preconditioner. Let $\mathcal{R}_N :=$ diag(R_1, R_2, R_3) be the block diagonal stiffness matrix, where R_i is the stiffness matrix based on the continuous piecewise bilinear element space with respect to the operator $Bp := -\Delta p + \beta p$, with the boundary conditions shared with the ith component space of $\mathcal{B}_{\mathcal{N}}$, where β is a nonnegative constant. By (5.2) the matrix $\mathcal{R}_N$ is spectrally equivalent to $\mathcal{A}_N$ for any $\beta \ge 0$, and thus the spectral condition number of the preconditioned matrix $\mathcal{R}_N^{-1}\mathcal{A}_N$ is $O(1)$ in comparison with the spectral condition number $O(N^3)$ of $\mathcal{A}_N$ (see [1], [23]). In the next section, we perform some numerical experiments to find an optimal β which leads to the smallest spectral condition number of the preconditioned matrix $\mathcal{R}_N^{-1}\mathcal{A}_N$.

For the Chebyshev approximation, we cannot guarantee the norm equivalence

$$\frac{1}{C}\, \|(\mathbf{v}, q)\|_{1,w}^2 \le \|\mathbf{v}\|_W^2 + \|q\|_{1,w}^2 \le C\, \|(\mathbf{v}, q)\|_{1,w}^2 \qquad \forall\, (\mathbf{v}, q) \in \mathbf{W}_N \times V_N, \tag{5.3}$$

but we will propose a block diagonal finite element preconditioner based on the Chebyshev-weighted inner product. In the following section, we give some numerical evidence that the block diagonal finite element preconditioner $\mathcal{R}_{w,N}^{-1}$ associated

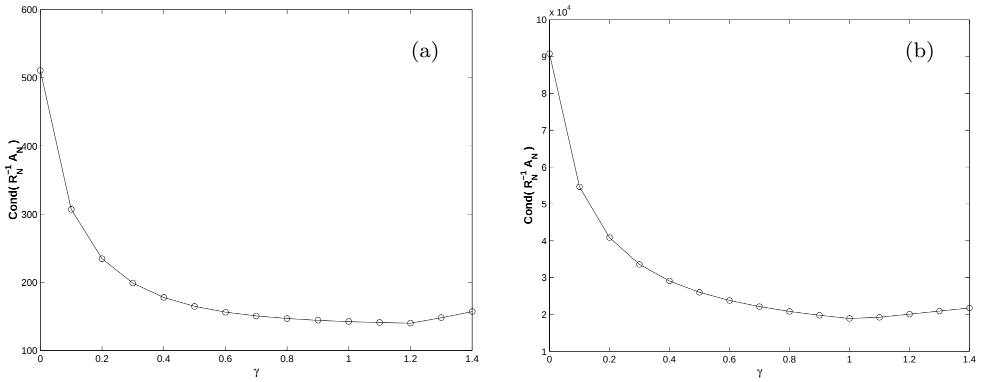


FIG. 1. *Spectral condition numbers of* $\mathcal{L}_{16} = \mathcal{R}_{16}^{-1}\mathcal{A}_{16}$ *for the Legendre cases* (a) $\mathbf{b} = (6,9)^t$, $c_0 = 0$, $\beta = \gamma\sqrt{b_1^2 + b_2^2}$; (b) $\mathbf{b} = (0,0)^t$, $c_0 = -10$, $\beta = \gamma\,|c_0|$.

with $Bp := -\Delta p + \beta p$ can be a good preconditioner, where the optimal choice of β is also discussed numerically. The spectral condition number of the preconditioned matrix $\mathcal{R}_{w,N}^{-1}\,\mathcal{A}_{w,N}$ seems to be bounded independently of N in comparison with the spectral condition number $O(N^4)$ of $\mathcal{A}_{w,N}$ (see [1], [23]). Also, one may consider a preconditioner $\widetilde{\mathcal{R}}_{w,N}^{-1}$ with respect to the norm $\|\mathbf{v}\|_W^2 + \|q\|_{1,w}^2$. However, $\widetilde{\mathcal{R}}_{w,N}$ is not a block diagonal matrix, and it is not easy to invert $\widetilde{\mathcal{R}}_{w,N}$, though it is spectrally equivalent to $\mathcal{A}_{w,N}$.

6. Numerical results. In this section, we present numerical experiments for the first-order system (3.1) associated with the elliptic partial differential equation

$$\left\{\begin{array}{rcll} -\nabla\cdot\nabla p + \mathbf{b}\cdot\nabla p + c_0\, p &=& f & \text{in } \Omega,\\ p &=& 0 & \text{on } \Gamma_D,\\ \mathbf{n}\cdot\nabla p &=& 0 & \text{on } \Gamma_N. \end{array}\right.$$

Let $\Omega = (-1,1)^2$. We take $\Gamma_D := (\{-1\}\times[-1,1])\cup([-1,1]\times\{-1\})$ and $\Gamma_N = \partial\Omega\backslash\Gamma_D$ for the Legendre approximation, and $\Gamma_D = \partial\Omega$ and $\Gamma_N = \emptyset$ for the Chebyshev approximation. Denote by $(\mathbf{u}_N, p_N)$ the discrete solution to (3.9) or (4.15), and by $e_{\mathbf{u}} = \mathbf{u} - \mathbf{u}_N$ and $e_p = p - p_N$ the errors.

6.1. Legendre pseudospectral approximation. We give the numerical experiments by Legendre pseudospectral least-squares approximation (3.9). Using the idea in [16], we first study the performance of the preconditioner $\mathcal{R}_N^{-1}$ for the system (5.1), where $\mathcal{R}_N$ is the block diagonal finite element stiffness matrix based on the continuous piecewise bilinear element space associated with the operator $Bp = -\Delta p + \beta p$. The constant β will be used to reduce the condition number of the preconditioned system for our examples. Define $\beta = \gamma\sqrt{b_1^2 + b_2^2}$ if $\mathbf{b} = (b_1, b_2)^t \neq \mathbf{0}$, and $\beta = \gamma|c_0|$ if $\mathbf{b} = (b_1, b_2)^t = \mathbf{0}$. Let $\mathcal{L}_N$ be the preconditioned matrix, i.e., $\mathcal{L}_N =: \mathcal{R}_N^{-1}\,\mathcal{A}_N$.

In Figure 1, we plot the spectral condition numbers of $\mathcal{L}_{16}$ along with γ for two different cases: (a) $\mathbf{b} = (6,9)^t$, $c_0 = 0$ and (b) $\mathbf{b} = (0,0)^t$, $c_0 = -10$. Numerical experiments indicate that the spectrum deforms smoothly as γ increases from zero to one for both cases, in which the best choice of γ is near 1. With $\gamma = 1$, we report the spectral condition numbers of $\mathcal{L}_N$ and $\mathcal{A}_N$ in Table 1 for several convection coefficients $\mathbf{b} = (0,0)^t, (2,3)^t, (4,6)^t, (6,9)^t$ and $c_0 = 0$, and in Table 2 for $c_0 = -1, -10$ and

Table 1
Condition numbers for $c_0 = 0$ *and* $\beta = \sqrt{b_1^2 + b_2^2}$.

	$\mathbf{b}^t = (0,0)$		$\mathbf{b}^t = (2,3)$		$\mathbf{b}^t = (4,6)$		$\mathbf{b}^t = (6,9)$	
N	$\mathcal{L}_N$	$\mathcal{A}_N$	$\mathcal{L}_N$	$\mathcal{A}_N$	$\mathcal{L}_N$	$\mathcal{A}_N$	$\mathcal{L}_N$	$\mathcal{A}_N$
4	11	177	28	129	87	246	199	448
8	14	710	31	393	73	503	147	689
12	15	1753	32	906	70	977	141	1172
16	15	3695	33	1864	69	1890	142	1986
20	16	6818	33	3419	69	3433	142	3541

Table 2
Condition numbers for $\mathbf{b}^t = (0,0)$ *and* $\beta = |c_0|$.

	$c_0 = -1$		$c_0 = -10$	
N	$\mathcal{L}_N$	$\mathcal{A}_N$	$\mathcal{L}_N$	$\mathcal{A}_N$
4	911	7552	4080	7817
8	1094	29587	20595	69107
12	1173	71838	19165	120819
16	1215	150524	18885	202646
20	1241	277620	19107	357455

$\mathbf{b} = (0,0)^t$. Both tables show that the spectral condition numbers of $\mathcal{A}_N$ behave like $O(N^3)$ for all cases, but those of $\mathcal{L}_N$ are bounded regardless of the degree N of polynomials, and they are increasing as the size $\sqrt{b_1^2 + b_2^2}$ of convection coefficient $\mathbf{b}$ or the absolute value $|c_0|$ of negative reaction coefficient c_0 increases.

We now present the discretization errors along with several coefficients $\mathbf{b}$ and c_0. The exact solutions p and $\mathbf{u} = \nabla p$ that we take are

$$p = \sin\left(\frac{7\pi}{4}(x+1)\right) \sin\left(\frac{7\pi}{4}(y+1)\right),$$

$$\mathbf{u} = \begin{pmatrix} \dfrac{7\pi}{4} \cos\left(\dfrac{7\pi}{4}(x+1)\right) \sin\left(\dfrac{7\pi}{4}(y+1)\right) \\ \dfrac{7\pi}{4} \sin\left(\dfrac{7\pi}{4}(x+1)\right) \cos\left(\dfrac{7\pi}{4}(y+1)\right) \end{pmatrix}.$$

The present solutions satisfy the given boundary conditions and, by substituting the solutions into (3.1), we have the right-hand side f along with various coefficients $\mathbf{b}$ and c_0. Table 3 shows that the spectral errors decay exponentially with respect to N, independent of the coefficients $\mathbf{b}$ and c_0.

6.2. Chebyshev pseudospectral approximation. We give the numerical experiments obtained by Chebyshev pseudospectral least-squares approximation (4.15). Let $\mathcal{A}_{w,N}$ be the matrix associated with the bilinear form $a_{w,N}(\cdot,\cdot)$. As a preconditioner for $\mathcal{A}_{w,N}$, we similarly take the inverse of the block diagonal finite element Chebyshev-weighted stiffness matrix $\mathcal{R}_{w,N}$ associated with the operator $Bp = -\Delta p + \beta p$ and based on the continuous piecewise bilinear element space with two-dimensional CGL grid points. Define $\beta = \gamma\sqrt{b_1^2 + b_2^2}$ if $\mathbf{b} = (b_1, b_2)^t \neq \mathbf{0}$, and $\beta = \gamma|c_0|$ if $\mathbf{b} = (b_1, b_2)^t = \mathbf{0}$. Let $\mathcal{L}_{w,N}$ be the preconditioned matrix, i.e., $\mathcal{L}_{w,N} =: \mathcal{R}_{w,N}^{-1} \mathcal{A}_{w,N}$.

We first give the performance of the preconditioner $\mathcal{R}_{w,N}^{-1}$. In Figure 2, we also plot the spectral condition numbers of $\mathcal{L}_{w,16}$ along with γ for two different cases: (a)

TABLE 3
Discretization errors for the Legendre approximation.

$\mathbf{b}^t$	c_0	N	$\|e_p\|_N$	$\|\nabla e_p\|_N$	$\|e_{\mathbf{u}}\|_N$	$\|\nabla e_{\mathbf{u}}\|_N$
(0,0)	0	4	1.2200e+00	4.7629e+00	9.7112e+00	5.5725e+01
		8	1.3455e−02	1.5819e−01	2.2849e−01	2.3819e+00
		12	5.0563e−05	7.7413e−04	7.1000e−04	1.0296e−02
		16	4.8412e−08	9.3810e−07	6.2503e−07	1.1763e−05
		20	7.0796e−11	7.7338e−10	4.9643e−10	7.0516e−09
(6, 9)	0	4	1.1591e+00	6.4543e+00	7.2067e+00	4.2593e+01
		8	2.4715e−02	2.4829e−01	2.9515e−01	3.0774e+00
		12	6.6759e−05	9.7210e−04	9.8871e−04	1.4567e−02
		16	5.7087e−08	1.0802e−06	9.4188e−07	1.7996e−05
		20	8.7126e−11	8.3847e−10	6.0020e−10	9.7925e−09
(0, 0)	−10	4	4.3143e+00	2.6675e+01	7.8625e+00	4.8069e+01
		8	1.4433e−01	1.5830e+00	2.6797e−01	2.6751e+00
		12	5.4262e−04	8.0872e−03	7.9288e−04	1.1163e−02
		16	5.2702e−07	1.0072e−05	6.7266e−07	1.2412e−05
		20	2.2602e−10	4.9411e−09	5.8513e−10	7.7477e−09

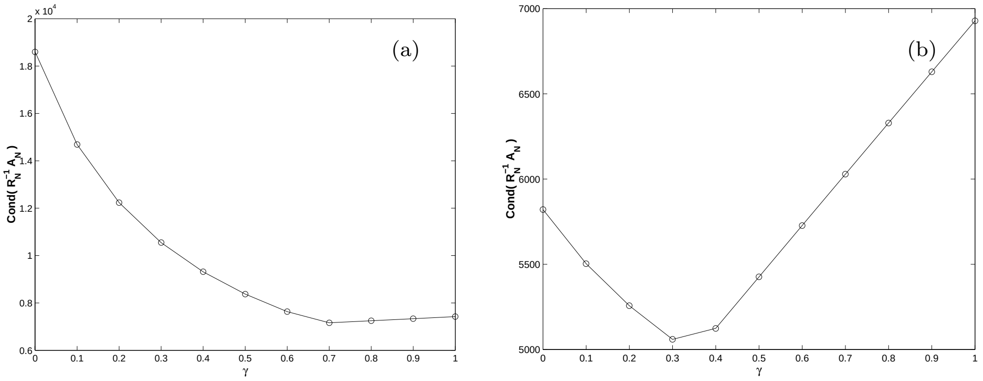


FIG. 2. *Spectral condition numbers of* $\mathcal{L}_{w,16} = \mathcal{R}_{w,16}^{-1}\mathcal{A}_{w,16}$ *for the Chebyshev cases* (a) $\mathbf{b} = (6,9)$, $c_0 = 0$, $\beta = \gamma\sqrt{b_1^2 + b_2^2}$; (b) $\mathbf{b} = (0,0)$, $c_0 = -10$, $\beta = \gamma\,|c_0|$.

$\mathbf{b} = (6,9)^t$, $c_0 = 0$ and (b) $\mathbf{b} = (0,0)^t$, $c_0 = -10$. We have the similar type of spectral condition numbers as in the Legendre case. The figure shows that the spectrum smoothly deforms as γ increases from zero to 0.7 for case (a) and to 0.3 for case (b); i.e., the best choice of γ is 0.7 for case (a) and 0.3 for case (b). We report the spectral condition numbers of $\mathcal{L}_{w,N}$ and $\mathcal{A}_{w,N}$ in Table 4 for $\mathbf{b} = (0,0)^t, (2,3)^t, (4,6)^t, (6,9)^t$ and $c_0 = 0$ with $\gamma = 0.7$, and in Table 5 for $c_0 = -1, -10$ and $\mathbf{b} = (0,0)^t$ with $\gamma = 0.3$. Both tables show that the spectral condition numbers of $\mathcal{A}_{w,N}$ behave like $O(N^3)$ for all cases, but those of $\mathcal{L}_{w,N}$ are bounded regardless of the degree N of polynomials, and they are increasing as the size $\sqrt{b_1^2 + b_2^2}$ of convection coefficient $\mathbf{b}$ or the absolute value $|c_0|$ of negative reaction coefficient c_0 increases. Tables 4 and 5 show that the condition numbers of $\mathcal{A}_{w,N}$ behave like $O(N^4)$ for all cases, but those of $\mathcal{L}_{w,N}$ are like $O(1)$. These numerical experiments demonstrate that $\mathcal{R}_{w,N}^{-1}$ can be a good preconditioner even though we could not prove the spectral equivalence (5.3).

TABLE 4
Condition numbers for $c_0 = 0$ and $\beta = \gamma\sqrt{b_1^2 + b_2^2}$ with $\gamma = 0.7$.

	$\mathbf{b}^t = (0,0)$		$\mathbf{b}^t = (2,3)$		$\mathbf{b}^t = (4,6)$		$\mathbf{b}^t = (6,9)$	
N	$\mathcal{L}_{w,N}$	$\mathcal{A}_{w,N}$	$\mathcal{L}_{w,N}$	$\mathcal{A}_{w,N}$	$\mathcal{L}_{w,N}$	$\mathcal{A}_{w,N}$	$\mathcal{L}_{w,N}$	$\mathcal{A}_{w,N}$
4	18	50	135	218	103	113	109	87
8	27	541	246	2473	1684	7432	4288	9955
12	31	2508	268	10882	1805	31049	7147	63531
16	33	7645	277	32440	1851	90316	7160	180232
20	35	18305	282	76622	1877	210555	7232	415418

TABLE 5
Condition numbers for $\mathbf{b}^t = (0,0)$ and $\beta = \gamma\,|c_0|$ with $\gamma = 0.3$.

	$c_0 = -1$		$c_0 = -10$	
N	$\mathcal{L}_{w,N}$	$\mathcal{A}_{w,N}$	$\mathcal{L}_{w,N}$	$\mathcal{A}_{w,N}$
4	36	96	752	1315
8	51	1054	5405	39955
12	55	4904	5012	175905
16	57	14985	5059	532050
20	58	35931	5088	1269432

TABLE 6
Discretization errors for the Chebyshev approximation.

$\mathbf{b}^t$	c_0	N	$\|e_p\|_{w,N}$	$\|\nabla e_p\|_{w,N}$	$\|e_{\mathbf{u}}\|_{w,N}$	$\|\nabla e_{\mathbf{u}}\|_{w,N}$
(0,0)	0	4	2.7814e+00	1.8449e+01	3.1952e+01	1.8336e+02
		8	7.1988e−02	1.0248e+00	1.5477e+00	1.9066e+01
		12	3.3955e−04	9.6570e−03	9.9796e−03	1.8252e−01
		16	4.8428e−07	2.1596e−05	1.6699e−05	4.1045e−04
		20	2.8875e−10	1.8534e−08	1.0792e−08	3.3829e−07
(6,9)	0	4	4.5616e+00	2.2512e+01	2.8557e+01	1.8355e+02
		8	2.2940e−01	1.6986e+00	2.1323e+00	2.3336e+01
		12	3.4315e−04	9.3054e−03	9.4352e−03	1.7511e−01
		16	4.6830e−07	2.1125e−05	1.5771e−05	3.9683e−04
		20	2.9720e−10	1.8699e−08	1.0337e−08	3.3173e−07
(0,0)	−10	4	2.2412e+00	1.4867e+01	2.5448e+01	1.5450e+02
		8	3.6606e−01	4.1592e+00	1.9013e+00	2.1507e+01
		12	1.9093e−03	3.4635e−02	1.1251e−02	1.9454e−01
		16	2.6919e−06	6.5936e−05	1.8115e−05	4.2733e−04
		20	1.4600e−09	4.6310e−08	1.1407e−08	3.4721e−07

Finally, we present the discretization errors with the following exact solutions p and $\mathbf{u} = \nabla p$:

$$p = \sin 2\pi x \, \sin 2\pi y, \qquad \mathbf{u} = \begin{pmatrix} 2\pi \cos 2\pi x \, \sin 2\pi y \\ 2\pi \sin 2\pi x \, \cos 2\pi y \end{pmatrix}.$$

Table 6 shows that the spectral errors decay exponentially with respect to N, independent of the coefficients $\mathbf{b}$ and c_0.

7. Conclusion. The analysis and computations for combining least-squares techniques and collocation methods using high-order elements have been shown for elliptic boundary value problems like (1.1). We saw that the continuous and discrete least-squares functionals are equivalent to a product H^1-norm, and we proved the spectral convergence consequently for the Legendre case. Hence the block finite element preconditioner corresponding to $-\Delta p + \beta p$ was demonstrated to be numerically optimal by emphasizing the importance of the choice of β, adopting the discussion in [16] for a finite difference case. For the Chebyshev case, convergence analysis was provided and the finite element preconditioner, which seems to be numerically optimal, was developed even though the continuous and discrete least-squares functionals were not shown to be equivalent to a product H^1_w-norm but to a product *div-curl* L^2_w-norm for the flux variable and $H^1_{0,w}$-norm for a primitive variable. In the near future we believe that it may be possible to analyze such an equivalence. The success of finite element least-squares (or fosls) can be fused over the Legendre (Chebyshev) pseudospectral method and the spectral method with a staggered grid for various kinds of partial differential equations. This fusion approach opens many possible applications. In particular, a practical application of these developed theories to transport-dominated problems by setting $c_0 = 0$ and making $|\mathbf{b}|$ large needs to be studied (see [22]). Incidentally, the neutron transported problem was analyzed for a finite element least-squares method in [17] and [18].

Acknowledgements. We would like to thank the referees, whose many valuable comments and corrections improved the paper.

REFERENCES

[1] C. Bernardi and Y. Maday, *Approximation Spectrales de Problémes aux Limites Elliptiques*, Springer-Verlag, Paris, 1992.

[2] P. B. Bochev and M. D. Gunzburger, *Analysis of least-squares finite element methods for the Stokes equations*, Math. Comp., 63 (1994), pp. 479–506.

[3] J. H. Bramble, R. D. Lazarov, and J. E. Pasciak, *A least-squares approach based on a discrete minus one inner product for first order system*, Math. Comp., 66 (1997), pp. 935–955.

[4] Z. Cai, R. Lazarov, T. A. Manteuffel, and S. F. McCormick, *First-order system least squares for second-order partial differential equations: Part* I, SIAM J. Numer. Anal., 31 (1994), pp. 1785–1799.

[5] Z. Cai, T. A. Manteuffel, and S. F. McCormick, *First-order system least squares for second-order partial differential equations: Part* II, SIAM J. Numer. Anal., 34 (1997), pp. 425–454.

[6] C. Canuto, M. Y. Hussaini, A. Quarteroni, and T. A. Zang, *Spectral Methods in Fluid Dynamics*, Springer-Verlag, New York, 1988.

[7] C. Canuto and P. Pietra, *Boundary and interface conditions within a finite element preconditioner for spectral methods*, J. Comput. Phys., 91 (1991), pp. 310–343.

[8] M. O. Deville and E. H. Mund, *Finite-element preconditioning for pseudospectral solutions of elliptic problems*, SIAM J. Sci. Stat. Comput., 11 (1990), pp. 311–342.

[9] D. Funaro, *A variational formulation for the Chebyshev pseudospectral approximation of Neumann problems*, SIAM J. Numer. Anal., 27 (1990), pp. 695–703.

[10] D. Funaro, *Polynomial Approximation of Differential Equations*, Lecture Notes in Phys. 8, Springer-Verlag, New York, 1992.

[11] D. Funaro, *Spectral Element for Transport-Dominated Equations*, Lecture Notes in Comput. Sci. Engrg. 1, Springer-Verlag, New York, 1997.

[12] G. J. Fix, M. Gunzburger, and R. Nicolaides, *On finite element methods of least-squares type*, Comput. Math. Appl., 5 (1979), pp. 87–98.

[13] G. J. Fix and E. Stephen, *On the finite element least-squares approximation to higher order elliptic systems*, Arch. Raton. Mech. Anal., 91–2 (1986), pp. 137–151.

[14] D. Jesperson, *A least-squares decomposition method for solving elliptic equations*, Math. Comp., 31 (1977), pp. 873–880.

[15] Bo-nan Jiang, *The Least-Squares Finite Element Method*, Springer-Verlag, New York, Berlin, 1998.

[16] T. Manteuffel and J. Otto, *Optimal equivalent preconditioners*, SIAM J. Numer. Anal., 30 (1993), pp. 790–812.

[17] T. A. Manteuffel and K. J. Ressel, *Least-squares finite-element solution of the neutron transport equation in diffusive regimes*, SIAM J. Numer. Anal., 35 (1998), pp. 806–835.

[18] T. A. Manteuffel, K. J. Ressel, and G. Starke, *A boundary functional for the least-squares finite-element solution of neutron transport problems*, SIAM J. Numer. Anal., 37 (2000), pp. 556–586.

[19] S. D. Kim and S. V. Parter, *Preconditioning Chebyshev spectral collocation by finite-difference operators*, SIAM J. Numer. Anal., 34 (1997), pp. 939–958.

[20] S. V. Parter and E. E. Rothman, *Preconditioning Legendre spectral collocation approximations to elliptic problems*, SIAM J. Numer. Anal., 32 (1995), pp. 333–385.

[21] A. I. Pehlivavov, G. F. Carey, and R. D. Lazarov, *Least-squares mixed finite elements for second-order elliptic problems*, SIAM J. Numer. Anal., 31 (1994), pp. 1368–1377.

[22] A. Pinelli, W. Couzy, M. O. Deville, and C. Benocci, *An efficient iteratrative solution method for the Chebyshev collocation of advection-dominated transport problems*, SIAM J. Sci. Comput., 17 (1996), pp. 647–657.

[23] A. Quarteroni and A. Valli, *Numerical Approximation of Partial Differential Equations*, Springer-Verlag, Berlin, Heidelberg, 1994.

[24] A. Quarteroni and E. Zampieri, *Finite element preconditioning for Legendre spectral collocation approximations to elliptic equations and systems*, SIAM J. Numer. Anal., 29 (1992), pp. 917–936.

SIAM J. NUMER. ANAL.
Vol. 41, No. 4, pp. 1388–1405

CONVERGENCE ANALYSIS OF THE GAUSS–SEIDEL PRECONDITIONER FOR DISCRETIZED ONE DIMENSIONAL EULER EQUATIONS[*]

ARNOLD REUSKEN[†]

Abstract. We consider the nonlinear system of equations that results from the Van Leer flux vector-splitting discretization of the one dimensional Euler equations. This nonlinear system is linearized at the discrete solution. The main topic of this paper is a convergence analysis of block-Gauss–Seidel methods applied to this linear system of equations. Both the lexicographic and the symmetric block-Gauss–Seidel method are considered. We derive results which quantify the quality of these methods as preconditioners. These results show, for example, that for the subsonic case the symmetric Gauss–Seidel method can be expected to be a much better preconditioner than the lexicographic variant. Sharp bounds for the condition number of the preconditioned matrix are derived.

AMS subject classifications. 65F10, 65N22, 65N06

Key words. Gauss–Seidel method, Euler equations, convergence analysis

DOI. 10.1137/S0036142902407393

1. Introduction.

In this paper we consider iterative methods for discrete stationary Euler equations. Two important solution approaches known from the literature are the following. First, one can use some "simple" explicit iterative method, like a block nonlinear Gauss–Seidel method or a Runge–Kutta method (obtained by introducing an artificial time variable), which then is accelerated by multigrid techniques (e.g., [12, 13, 16, 20, 25, 27]). The second approach is based on linearization combined with fast iterative solvers for large sparse linear systems, such as multigrid solvers or (preconditioned) Krylov-subspace methods. A typical example of this is the Newton–Krylov technique from [5, 14, 15, 18, 19, 24]. In the literature one can find many studies in which different iterative solution techniques for solving stationary (or instationary) discrete Euler equations are compared (e.g., [17, 26]). There are, however, as far as we know, no rigorous theoretical results available which yield any insight into convergence properties of certain iterative methods applied to (linearized) discrete Euler equations. In this paper a first step towards such theoretical results is made.

In this paper, as a model problem we consider the stationary Euler equations that model one dimensional subsonic and transonic flows through a nozzle [11, 21], and use the Van Leer flux vector-splitting method for discretization. The discrete nonlinear problem is linearized at the discrete solution. We apply a GMRES method with block-Gauss–Seidel preconditioning to this Jacobian linear problem. In the Gauss–Seidel preconditioner the three unknowns at each grid point are collected in a block and updated simultaneously. (This is also often called a collective Gauss–Seidel method.) Both a lexicographic (LGS) and a symmetric (SGS) Gauss–Seidel

[*] Received by the editors May 10, 2002, accepted for publication (in revised form) February 26, 2003; published electronically September 9, 2003. This work was supported by the Deutsche Forschungsgemeinschaft in the Collaborative Research Centre SFB 401 "Flow Modulation and Fluid-Structure Interaction at Airplane Wings" of the RWTH Aachen, University of Technology, Aachen, Germany.

http://www.siam.org/journals/sinum/41-4/40739.html

[†] Institut für Geometrie und Praktische Mathematik, Rheinisch-Westfälische Technische Hochschule Aachen, D-52056 Aachen, Germany (reusken@igpm.rwth-aachen.de).

method are used. We emphasize that we do not recommend using such an iterative method for these one dimensional linearized Euler equations, because the Jacobian matrix has a block-tridiagonal structure with 3 × 3 blocks, and thus a direct solver is efficient for this problem. Our main interest, however, is not the efficient solution of these one dimensional Euler equations, but a better understanding of convergence properties of the block-Gauss–Seidel method applied to discrete Euler equations.

As is well known, direction of flow essentially influences not only the discretization of Euler equations, but also the convergence of iterative methods. If the flow is subsonic, then even in the one dimensional case the LGS method cannot be consistent with the flow direction. In one dimensional flow one has only two directions, and thus the SGS method can be expected to be a fast iterative solver. These elementary observations are part of common knowledge. However, even for one dimensional flows many questions related to Gauss–Seidel preconditioning are still unanswered. As will be illustrated by numerical experiments, for GMRES with block-Gauss–Seidel preconditioning there are some interesting dependencies of the rate of convergence on the Mach number and the mesh size. As far as we know, there is no analysis available which explains these dependencies. The main topic of the paper is a theoretical analysis in which we try to explain some of the convergence phenomena that are observed in the numerical experiments. In this analysis we use the technique of "frozen coefficients"; i.e., we linearize the discrete Euler equations at a function triple (ρ, u, p) (density, velocity, pressure) which is constant as a function of the space variable and is such that the solution is subsonic. We consider the LGS and SGS methods applied to this problem and derive results which quantify the quality of these methods as a preconditioner. Our results show, for example, that the SGS method can be expected to be a (much) better preconditioner than the LGS method. Sharp bounds for the condition number of the preconditioned matrix are derived, which show that in case of the SGS preconditioner for a large range of Mach numbers $M \in (M_0, 1)$ this condition number increases only (very) slowly if the grid size decreases.

We realize that although some first theoretical results are given in this paper, we are still far from a complete theoretical convergence analysis of Gauss–Seidel methods applied to linearized discrete one dimensional Euler equations. The theoretical analysis presented supports the numerical observation that for many subsonic and transonic one dimensional linearized Euler equations the SGS method is a (very) effective preconditioner. However, as already noted above, in the one dimensional case a direct solver is the best choice. In two and three dimensional problems, however, block-Gauss–Seidel techniques or other basic iterative methods (ILU) combined with Krylov subspace methods can result in very efficient solvers [3, 4, 6, 19]. Clearly, in higher dimensions flow has many directions to go and the relation between a Gauss–Seidel-type splitting and direction of flow becomes much more complicated. This then makes the analysis of this class of iterative methods much more difficult, as in the one dimensional case. We do not claim that our analysis can easily be applied to the much more interesting higher dimensional case. Nevertheless, starting from the results presented in this paper we do see some possibilities for the analysis of a two dimensional problem. These are briefly discussed in Remark 3 at the end of the paper.

2. The one dimensional nozzle flow and its discretization. We consider the stationary quasi–one dimensional Euler flow in a channel of varying cross section

$S(x)$ $(x \in \mathbb{R})$. This problem can be modeled by the equations (cf. [11, 21])

$$\begin{cases} \frac{d(\rho u S)}{dx} = 0, \\ \frac{d(\rho u^2 S + pS)}{dx} = p\frac{dS}{dx}, \\ \frac{d(\rho u H S)}{dx} = 0, \end{cases}$$

with density ρ, velocity u, pressure p, and stagnation enthalpy $H = E + \frac{p}{\rho}$. Further relations are

$$E = e + \frac{1}{2}u^2, \quad p = (\gamma - 1)\rho e.$$

Here e denotes the internal energy and γ is a gas parameter (ratio of specific heats; $\gamma = 1.4$ for air). As unknowns one can take the primitive variables $V := (\rho, u, p)^T$. We introduce the conservative variables U, the source term Q_S, and the flux function f:

$$U = \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} := S \begin{pmatrix} \rho \\ \rho u \\ \rho E \end{pmatrix}, \quad Q_S(U) := \begin{pmatrix} 0 \\ \frac{dS}{dx} \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ (\gamma - 1)(u_3 - \frac{1}{2}\frac{u_2^2}{u_1})\frac{d \ln S}{dx} \\ 0 \end{pmatrix},$$

$$f(U) := S \begin{pmatrix} \rho u \\ \rho u^2 + p \\ \rho u H \end{pmatrix} = \begin{pmatrix} u_2 \\ \frac{1}{2}(3 - \gamma)\frac{u_2^2}{u_1} + (\gamma - 1)u_3 \\ \gamma \frac{u_2 u_3}{u_1} - \frac{1}{2}(\gamma - 1)\frac{u_2^3}{u_1^2} \end{pmatrix}.$$

In compact form the problem can be represented as

$$f(U)_x = Q_S(U). \tag{2.1}$$

Note that for $S(x) \equiv 1$ we obtain the homogeneous one dimensional Euler equations. Formulas for the transformation between the primitive variables V and the conservative variables U are known (cf. [11]). Important quantities are the speed of sound $c = (\gamma p \rho^{-1})^{\frac{1}{2}}$ and the Mach number $M = uc^{-1}$. In our experiments we take the following nozzle with throat at $x = 1$:

$$S(x) = \begin{cases} 1 + 1\frac{1}{2}\left(1 - \frac{1}{5}(x + 4)\right)^2 & \text{for } 0 \leq x \leq 1, \\ 1 + \frac{1}{2}\left(1 - \frac{1}{5}(x + 4)\right)^2 & \text{for } 1 \leq x \leq 4. \end{cases} \tag{2.2}$$

Nozzle flows are well-known test cases for steady-state computations (cf. [11, 13]). By specifying certain problem parameters (inflow Mach number and critical throat section), the problem (2.1) can have several types of solutions: a smooth subsonic flow, a smooth hypersonic flow, a transonic flow without shocks, or a transonic flow with shocks. Moreover, these solutions depend on only *one* parameter (for example, the Mach number $M = M(x)$), and a simple procedure for computing the exact solution of the continuous problem is available (cf. [11, section 16.6.4]). For two cases the function $x \to M(x)$ corresponding to the exact solution of the problem (2.1), (2.2) is shown in Figures 2.1 and 2.2. In Figure 2.1 we have a smooth subsonic flow with critical throat section $S^* = 0.5$. The solution in Figure 2.2 corresponds to a transonic flow with a critical throat value $S^* = 1$, which equals the throat value $S(1)$, and a shock at $x = 3$.

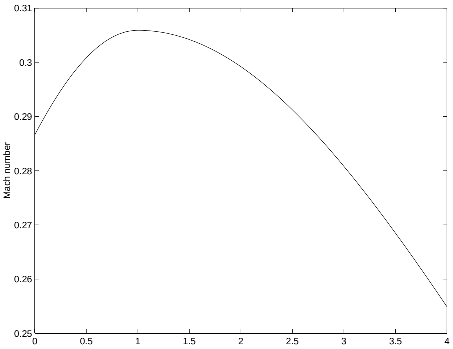


FIG. 2.1. $x \to M(x)$ *for a smooth subsonic flow.*

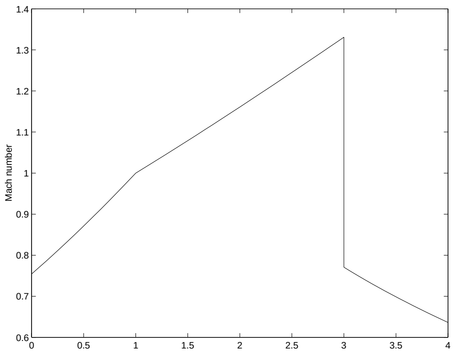


FIG. 2.2. $x \to M(x)$ *for a transonic flow with shock.*

We now outline the numerical solution method for this test problem (for which the exact solution is available). We consider only problems with subsonic inflow and outflow conditions ($0 < M(0) < 1$ and $0 < M(4) < 1$). For the boundary conditions we prescribe values for ρ and u at the inflow boundary $x = 0$, and for p at the outflow boundary $x = 4$. We use a uniform grid $x_i = ih$, $0 \le i \le n+1$, with a mesh size $h = 4/(n+1)$. We introduce the discrete unknowns

$$U_i := \begin{pmatrix} u_1(x_i) \\ u_2(x_i) \\ u_3(x_i) \end{pmatrix}, \quad \mathbf{U} := \big(U_i\big)_{0 \le i \le n+1}.$$

For the discretization at the boundaries we use compatibility relations as discussed in [11, section 19.1.2]; i.e., at the inflow boundary we discretize with one-sided differences the equation $(u-c)\big(\frac{du}{dx} - \frac{1}{\rho c}\frac{dp}{dx}\big) = uc\frac{d\ln S}{dx}$ that corresponds to the left-going characteristic. Similarly, the two right-going characteristic equations at $x = 4$ are discretized using one-sided differences. Together with the prescribed boundary values this yields equations

$$F_0 : \mathbb{R}^6 \to \mathbb{R}^3, \qquad F_0(U_0, U_1) = 0, \tag{2.3}$$

$$F_{n+1} : \mathbb{R}^6 \to \mathbb{R}^3, \qquad F_{n+1}(U_n, U_{n+1}) = 0. \tag{2.4}$$

For the discretization in the interior grid points we use an upwind method based on the Van Leer flux vector-splitting (see [11, 28]):

$$\begin{aligned} &f(V) = f^+(V) + f^-(V), \\ &f^+(V) := \frac{\rho}{4c}(u+c)^2 \begin{pmatrix} 1 \\ \frac{(\gamma-1)u+2c}{\gamma} \\ \frac{((\gamma-1)u+2c)^2}{2(\gamma^2-1)} \end{pmatrix} \quad \text{if} \quad -1 \le M \le 1, \\ &f^+ := 0 \quad \text{if} \quad M \le -1, \quad f^+ := f \quad \text{if} \quad M \ge 1. \end{aligned} \tag{2.5}$$

We use backward differences for the approximation of $f^+(U)_x$, and forward differences for the approximation of $f^-(U)_x$. This yields the equations

$$F_i(U_{i-1}, U_i, U_{i+1}) := -f^+(U_{i-1}) + f^+(U_i) - f^-(U_i) + f^-(U_{i+1}) - hQ_S(U_i) = 0 \tag{2.6}$$

for $i = 1, \dots, n$. The equations (2.3), (2.4), and (2.6) yield a nonlinear system of equations

$$F : \mathbb{R}^{3(n+2)} \to \mathbb{R}^{3(n+2)}, \quad F(\mathbf{U}) = 0. \tag{2.7}$$

For the iterative solution of this problem we apply the Newton method. The Jacobian matrices $DF(\mathbf{U}) \in \mathbb{R}^{3(n+2)\times 3(n+2)}$ have a block-tridiagonal structure. Hence, the linear systems in the Newton iteration can be solved efficiently using a direct method. The main topic of this paper is the analysis of block-Gauss–Seidel iterative methods applied to these linear systems. We emphasize that we do *not* suggest using such a Gauss–Seidel method as an efficient solver in this one dimensional setting. The analysis for the one dimensional case is a first step towards a better theoretical understanding of basic iterative methods applied to two or three dimensional linearized Euler equations.

3. Numerical experiments. In this section we show results of a few numerical experiments which illustrate some interesting phenomena related to the rate of convergence of block-Gauss–Seidel methods. Let $\mathbf{U}_h^*$ be the solution of the discrete problem (2.7). We consider the linear system

$$DF(\mathbf{U}_h^*)\mathbf{v} = \mathbf{b}. \tag{3.1}$$

In the experiments we take $\mathbf{b} = (1, \dots, 1)^T$, and for the starting vector in the iterative method we use $\mathbf{v}^0 = 0$. It turns out that in many cases (often due to the treatment of the boundary conditions) the block-Gauss–Seidel method does not converge. It turns out, however, that the method is a (very) good preconditioner. Hence, we use the block-Gauss–Seidel method in combination with a Krylov subspace method. We choose the GMRES(m) iterative method. Experiments with BiCGSTAB yielded similar results.

We use the LGS and SGS methods. In the GMRES method we make a restart after $m = 20$ iterations. The choice $m = 20$ is rather arbitrary, however; for other values of $m \in [10, 40]$ we observe similar phenomena. We use the GMRES(m) implementation in MATLAB. In a first experiment, as a comparison for other results, we consider a standard very simple model problem. We take the one dimensional diffusion equation $-u_{xx} = g$ discretized by second order differences. This results in an $n \times n$ tridiagonal matrix tridiag$(-1, 2, -1)$. For different n-values the convergence history of the SGS-GMRES(20) iterative solver applied to this problem is shown in Figure 3.1. For the linearized compressible Euler equations (3.1) we show results for the following problems.

Problem 1. We consider a problem with a smooth subsonic solution, as shown in Figure 2.1. The convergence history of the SGS-GMRES(20) method is shown in Figure 3.2.

Problem 2. We take a smooth subsonic flow with larger Mach numbers than in Problem 1. The solution is shown in Figure 3.3 (with critical throat value $S^* = 0.85$). The corresponding convergence history is presented in Figure 3.4.

Problem 3. We consider a transonic flow with a shock, as shown in Figure 2.2. The convergence behavior of the SGS-GMRES(20) solver is shown in Figure 3.5. If instead of SGS we use the LGS preconditioner, we obtain the results in Figure 3.6.

From these experiments we observe that in all three problems the rate of convergence of the SGS-GMRES(20) method is (much) higher than for the one dimensional discrete Poisson equation. We also see that in Problem 2 (subsonic flow with relatively high Mach numbers) the rate of convergence is much higher than in Problem

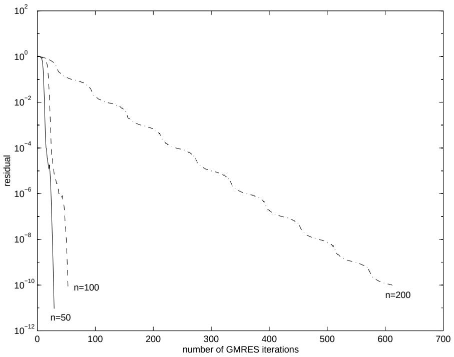


Fig. 3.1. *SGS-GMRES*(20) *method applied to a one dimensional Poisson equation.*

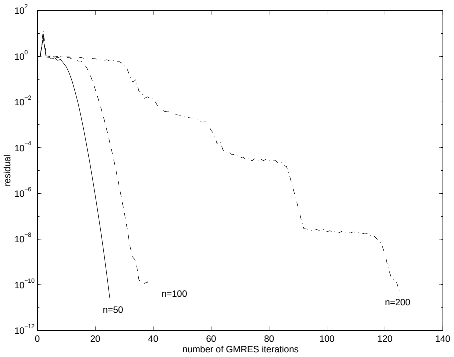


Fig. 3.2. *Problem* 1: *SGS-GMRES*(20) *method for the subsonic flow in Figure* 2.1.

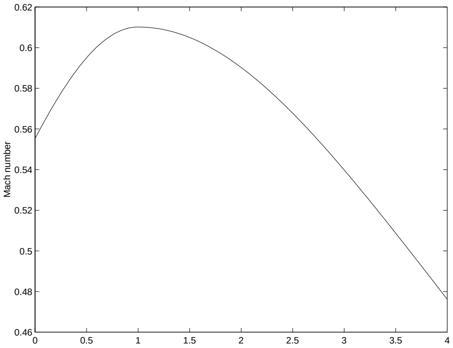


Fig. 3.3. *Problem* 2: $x \to M(x)$ *for a smooth subsonic flow.*

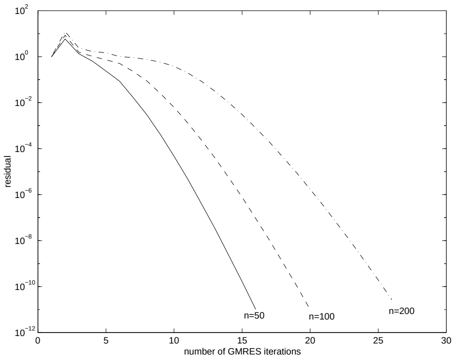


Fig. 3.4. *Problem* 2: *SGS-GMRES*(20) *method for the subsonic flow in Figure* 3.3.

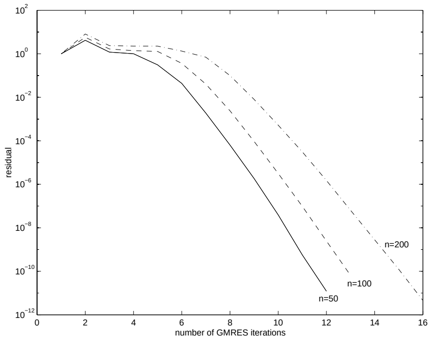


Fig. 3.5. *Problem* 3: *SGS-GMRES*(20) *method for the transonic flow in Figure* 2.2.

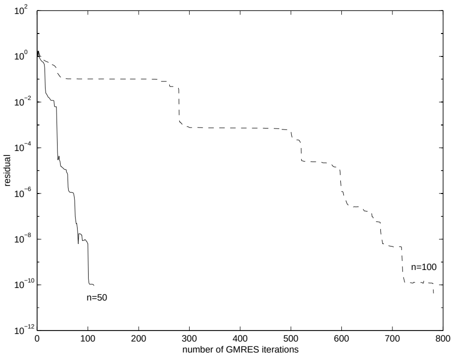


Fig. 3.6. *Problem* 3: *LGS-GMRES*(20) *method for the transonic flow in Figure* 2.2.

1. In the case of the transonic flow in Problem 3 the rate of convergence of the SGS-GMRES(20) method is even higher. We also note that the results presented in Figures 3.4 and 3.5 show a weak dependence of the rate of convergence on the mesh size h. Finally note that in Problem 3 the LGS-GMRES(20) method is much slower than the SGS-GMRES(20) method.

In the next section we present an analysis which yields some theoretical results on the quality of the block-Gauss–Seidel preconditioner. These theoretical results yield a better understanding of the convergence phenomena that are observed in the numerical experiments above.

4. Convergence analysis of the block-Gauss–Seidel method. For the (block) Gauss–Seidel method many convergence results are known in the literature (e.g., see [1, 2, 7, 8, 23]). These results apply to certain classes of matrices, like, for example, symmetric positive definite matrices or M-matrices. We did not find a convergence analysis which yields a satisfactory result when applied to the linearized discrete one dimensional Euler equations. In this section we present an analysis that partly fills this gap.

For the theoretical analysis we consider the homogeneous Euler equations $f(U)_x = 0$ with a *constant* solution $(\rho(x), u(x), p(x)) = (\rho, u, p) =: \bar{V}$ for all x. We consider only data with

$$\rho > 0, \quad p > 0, \quad M \in (0,1), \quad \gamma := 1.4. \tag{4.1}$$

The corresponding solution vector in conservative variables is denoted by $\bar{U}^*$. The Van Leer discretization method as described in section 2 results in a nonlinear system as in (2.3), (2.4), (2.6) with $Q_S = 0$. The treatment of the boundary conditions (first order accurate) is such that

$$F_0(\bar{U}_0^*, \bar{U}_1^*) = 0, \quad F_{n+1}(\bar{U}_n^*, \bar{U}_{n+1}^*) = 0$$

holds. Hence, the discrete problem has the constant solution $\bar{U}_h^*(x_i) := \bar{U}^*(x_i),\ i = 0, \ldots, n+1$. To avoid technical complications related to the specific treatment of the boundary conditions we consider the nonlinear system in the interior points only; i.e., as unknowns we take $\mathbf{U} = (U_1, \ldots, U_n)^T \in \mathbb{R}^{3n}$, and the system of nonlinear equations is given by

$$\begin{aligned} F_1(U_1, U_2) &:= f^+(U_1) - f^-(U_1) + f^-(U_2) = f^+(\bar{U}_0^*), \\ F_i(U_{i-1}, U_i, U_{i+1}) &:= -f^+(U_{i-1}) + f^+(U_i) - f^-(U_i) + f^-(U_{i+1}) = 0, \quad 2 \le i \le n-1, \\ F_n(U_{n-1}, U_n) &:= -f^+(U_{n-1}) + f^+(U_n) - f^-(U_n) = -f^-(\bar{U}_{n+1}^*). \end{aligned} \tag{4.2}$$

The vector $\bar{U}_h^*(x_i) = \bar{U}^*(x_i),\ i = 1, \ldots, n$, is a solution of this nonlinear system of equations. The Jacobian system

$$\mathbf{A}\mathbf{v} = \mathbf{b}, \quad \mathbf{A} := DF(\bar{\mathbf{U}}_h^*) \in \mathbb{R}^{3n \times 3n} \tag{4.3}$$

has a block-tridiagonal matrix

$$\begin{aligned} &\mathbf{A} = \text{blocktridiag}(-A^+, A^+ - A^-, A^-)_{1 \le i \le n}, \\ &A^+ := Df^+(\bar{U}_h^*) \in \mathbb{R}^{3\times 3}, \quad A^- := Df^-(\bar{U}_h^*) \in \mathbb{R}^{3\times 3}. \end{aligned} \tag{4.4}$$

The eigenvalues of $A^{\pm}$ are denoted by $\lambda_i^{\pm},\ i = 1, 2, 3$. The Van Leer splitting has been constructed in such a way that both A^+ and A^- have one zero eigenvalue: $\lambda_1^+ = \lambda_1^- = 0$. The other eigenvalues λ_2^+, λ_3^+ of A^+ and λ_2^-, λ_3^- of $A^-$are strictly positive and strictly negative, respectively. For these eigenvalues explicit formulas in terms of c and M are known [11, 28].

Using MAPLE, one obtains

$$\det(A^+ - A^-) = \frac{c^3}{24}(M^6 - 15M^4 + 3M^2 + 11).$$

The polynomial in M on the right-hand side has no zeros for $M \in (-1, 1)$. Hence (cf. (4.1)) the matrix $A^+ - A^-$ is nonsingular. The matrix

$$B = B(\bar{U}_h^*) := -(A^+ - A^-)^{-1} A^- \tag{4.5}$$

plays an important role in the analysis. From $\ker(B) = \ker(A^-)$ and $\ker(I - B) = \ker(A^+)$ it follows that

$$\sigma(B) = \{\, 1,\, 0,\, \mu(\rho, c, M) \,\}. \tag{4.6}$$

Using MAPLE, an explicit representation for B can be obtained. The resulting formulas are rather long and not relevant here. We note only that from these formulas it immediately follows that B can be factorized as

$$B = E\tilde{B}(M)E^{-1}, \quad E = \mathrm{diag}(1, c, c^2), \tag{4.7}$$

with a matrix $\tilde{B}(M)$ which depends only on M. Hence, the eigenvalue μ of B in (4.6) depends only on M. A further MAPLE computation yields a representation of an eigenvector basis of the matrix $\tilde{B}$:

$$\tilde{B}X = X\mathrm{diag}(1, 0, \mu(M)),$$

$$X = \begin{pmatrix} 1 & 1 & 1 \\ \frac{M^2+4M-5}{M+9} & \frac{M^2-4M-5}{M-9} & \frac{6M}{11} \\ \frac{7M^3-7M^2+5M+275}{14(M+9)} & \frac{7M^3+7M^2+5M-275}{14(M-9)} & \frac{16M^2}{77} \end{pmatrix}, \tag{4.8}$$

$$\mu(M) = \frac{1}{2}\frac{M^4 - 14M^2 + 24M - 11}{M^4 - 14M^2 - 11}. \tag{4.9}$$

The function $M \to \mu(M)$ is shown in Figure 4.1. An important observation is that for a large range of Mach numbers $M \in [M_0, 1]$ the eigenvalue $\mu(M)$ is small (e.g., $\mu(M) \in [0, 0.1]$ for $M \in [0.5, 1]$). The condition number of the matrix X is bounded uniformly in $M \in [0, 1]$. The function $M \to \|X\|_2 \|X^{-1}\|_2$ is given in Figure 4.2.

In the remainder of this section we analyze block-Gauss–Seidel methods applied to the system (4.3). For any block-tridiagonal matrix $\mathbf{C} = \mathrm{blocktridiag}(C_l, C_d, C_u)$ we introduce the decomposition $\mathbf{C} = \mathbf{D} - \mathbf{L} - \mathbf{U}$ with $\mathbf{D} := \mathrm{blockdiag}(C_d)$ and strictly lower and upper triangular matrices $\mathbf{L}$ and $\mathbf{U}$, respectively. We assume that the matrix $\mathbf{D}$ is nonsingular and define the lexicographic and symmetric Gauss–Seidel preconditioners:

$$\mathbf{W}_C^{LGS} := \mathbf{D} - \mathbf{L}, \quad \mathbf{W}_C^{SGS} := (\mathbf{D} - \mathbf{L})\mathbf{D}^{-1}(\mathbf{D} - \mathbf{U}).$$

Below, the symbol $\mathbf{W}_C$ is used to denote both $\mathbf{W}_C^{LGS}$ and $\mathbf{W}_C^{SGS}$; i.e., statements involving $\mathbf{W}_C$ hold both for the lexicographic and the symmetric block-Gauss–Seidel preconditioner. We apply these preconditioners to the matrix $\mathbf{A}$ in (4.4). The block-Gauss–Seidel methods are invariant under block-diagonal scaling, and thus the following result holds.

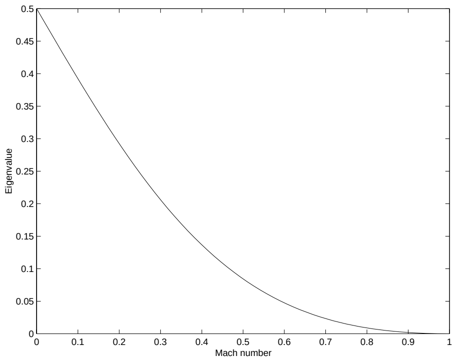


FIG. 4.1. *Function $M \to \mu(M)$.*

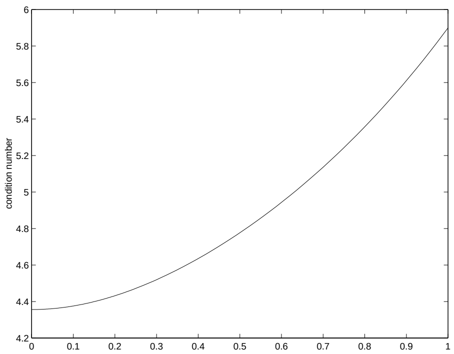


FIG. 4.2. *Function $M \to \|X\|_2\|X^{-1}\|_2$.*

LEMMA 4.1. *Define*

$$\tilde{\mathbf{A}} := \text{blocktridiag}\big(-(I-B), I, -B\big)_{1\le i\le n}, \quad \textit{with } B \textit{ as in } (4.5).$$

Then for the block-Gauss–Seidel preconditioner we have

$$\mathbf{W}_A^{-1}\mathbf{A} = \mathbf{W}_{\tilde{A}}^{-1}\tilde{\mathbf{A}}.$$

We apply a further transformation with the well-conditioned eigenvector basis X of the matrix $\tilde{B}$. For this we introduce

$$\mathbf{X} := \text{blockdiag}(X)_{1\le i\le n}, \quad \mathbf{E} := \text{blockdiag}\begin{pmatrix} 1 & 0 & 0 \\ 0 & c & 0 \\ 0 & 0 & c^2 \end{pmatrix}_{1\le i\le n}, \quad c := (\gamma p \rho^{-1})^{\frac{1}{2}}.$$

LEMMA 4.2. *Define*

$$\hat{\mathbf{A}} := \text{blocktridiag}\left(-\begin{pmatrix} 0 & & \emptyset \\ & 1 & \\ \emptyset & & 1-\mu \end{pmatrix}, \begin{pmatrix} 1 & & \emptyset \\ & 1 & \\ \emptyset & & 1 \end{pmatrix}, -\begin{pmatrix} 1 & & \emptyset \\ & 0 & \\ \emptyset & & \mu \end{pmatrix}\right) \in \mathbb{R}^{3n\times 3n},$$

with $\mu = \mu(M)$ as in (4.9). *Then*

$$\mathbf{W}_A^{-1}\mathbf{A} = \mathbf{E}\mathbf{X}\mathbf{W}_{\hat{A}}^{-1}\hat{\mathbf{A}}\mathbf{X}^{-1}\mathbf{E}^{-1}$$

holds.

Proof. This follows from

$$\tilde{\mathbf{A}} = \mathbf{E}\mathbf{X}\hat{\mathbf{A}}\mathbf{X}^{-1}\mathbf{E}^{-1}, \quad \mathbf{W}_{\tilde{A}} = \mathbf{E}\mathbf{X}\mathbf{W}_{\hat{A}}\mathbf{X}^{-1}\mathbf{E}^{-1},$$

and the result in Lemma 4.1. □

From Lemma 4.2 it follows that $\sigma(\mathbf{W}_A^{-1}\mathbf{A}) = \sigma(\mathbf{W}_{\hat{A}}^{-1}\hat{\mathbf{A}})$. However, it is well known that in a setting with strongly nonnormal matrices the eigenvalues (spectral radius) are in general not a good measure for the rate of convergence of an iterative method (cf. [8, 23]). Because the blocks in the block-tridiagonal matrix $\hat{\mathbf{A}}$ are diagonal, this matrix represents three decoupled systems of dimension n, and a block-Gauss–Seidel method applied to $\hat{\mathbf{A}}$ is the same as a *point* Gauss–Seidel method. To make this more precise we introduce for $\mathbf{H} = \mathbf{D} - \mathbf{L} - \mathbf{U}$ with $\mathbf{D} = \mathrm{diag}(\mathbf{H})$, $\mathbf{L}$ and $\mathbf{U}$ strictly lower and strictly upper triangular matrices, respectively, the *point* Gauss–Seidel splittings

$$\mathbf{G}_H^{LGS} := \mathbf{D} - \mathbf{L}, \quad \mathbf{G}_H^{SGS} := (\mathbf{D} - \mathbf{L})\mathbf{D}^{-1}(\mathbf{D} - \mathbf{U}).$$

The symbol $\mathbf{G}_H$ is used to denote both $\mathbf{G}_H^{LGS}$ and $\mathbf{G}_H^{SGS}$. Let $\mathbf{P} \in \mathbb{R}^{3n\times 3n}$ be the permutation matrix given by

$$(\mathbf{Px})_{k+3(i-1)} = x_{(k-1)n+i}, \quad k = 1, 2, 3, \quad i = 1, \dots, n.$$

We introduce the tridiagonal $n \times n$-matrices

(4.10)
$$\mathbf{L} := \begin{pmatrix} 1 & & & \\ -1 & 1 & & \\ & \ddots & \ddots & \\ & & -1 & 1 \end{pmatrix}, \quad \mathbf{T} = \mathbf{T}_\mu := \begin{pmatrix} 1 & -\mu & & \\ -(1-\mu) & 1 & \ddots & \\ & \ddots & \ddots & -\mu \\ & & -(1-\mu) & 1 \end{pmatrix}.$$

From the result in Lemma 4.2 one obtains the following.

Lemma 4.3. *The following holds:*

$$\mathbf{E}^{-1}\mathbf{W}_A^{-1}\mathbf{A}\mathbf{E} = \mathbf{X}\mathbf{P}\mathbf{Q}\mathbf{P}^{-1}\mathbf{X}^{-1}$$
$$\textit{with} \;\; \mathbf{Q} = \begin{pmatrix} \mathbf{Q}_1 & & \emptyset \\ & \mathbf{Q}_2 & \\ \emptyset & & \mathbf{Q}_3 \end{pmatrix} := \begin{pmatrix} \mathbf{G}_{L^T}^{-1}\mathbf{L}^T & & \emptyset \\ & \mathbf{G}_L^{-1}\mathbf{L} & \\ \emptyset & & \mathbf{G}_T^{-1}\mathbf{T} \end{pmatrix}.$$

We now consider a Krylov subspace method applied to the matrix $\mathbf{W}_A^{-1}\mathbf{A}$. Let $\mathcal{P}_k$ be the space of polynomials of degree less than or equal to k and $\mathcal{P}_k^* := \{\, p \in \mathcal{P}_k \,|\, p(0) = 1 \,\}$. A Krylov subspace method can be described by a corresponding polynomial $p_k \in \mathcal{P}_k^*$. Based on the result in Lemma 4.3 we use the problem dependent scaled Euclidean norm

$$\|\mathbf{y}\|_E := \|\mathbf{E}^{-1}\mathbf{y}\|_2, \quad \mathbf{y} \in \mathbb{R}^{3n}.$$

Let $\kappa_2(\mathbf{C}) := \|\mathbf{C}^{-1}\|_2\|\mathbf{C}\|_2$ be the spectral condition number. From Lemma 4.3 it follows that

$$\kappa_2(X)^{-1}\|p_k(\mathbf{Q})\|_2 \le \|p_k(\mathbf{W}_A^{-1}\mathbf{A})\|_E \le \kappa_2(X)\|p_k(\mathbf{Q})\|_2.$$

Since $\kappa_2(X)$ is independent of n and uniformly (w.r.t. M) bounded, the quantity $\|p_k(\mathbf{Q})\|_2$ is a reasonable measure for the rate of convergence of the Krylov subspace method applied to $\mathbf{W}_A^{-1}\mathbf{A}$. We therefore consider

$$\|p_k(\mathbf{Q})\|_2 = \max_{1\le i\le 3} \|p_k(\mathbf{Q}_i)\|_2 \,. \tag{4.11}$$

In order to derive bounds for $\|p_k(\mathbf{C})\|_2$, $\mathbf{C} \in \mathbb{R}^n$, one usually makes the natural assumption that the symmetric part of the matrix $\mathbf{C}$ is positive definite. This assumption is satisfied in our case.

Lemma 4.4. *The following holds:*

$$\frac{1}{2}\lambda_{\min}(\mathbf{Q}_i + \mathbf{Q}_i^T) := \min\{\, \mathbf{y}^T\mathbf{Q}_i\mathbf{y} \,|\, \mathbf{y} \in \mathbb{R}^n,\ \|\mathbf{y}\|_2 = 1 \,\} > 0 \quad \textit{for} \;\; i = 1, 2, 3.$$

Proof. Note that for the LGS and the SGS methods we have

$$\|\mathbf{I} - \mathbf{G}_T^{-1}\mathbf{T}\|_\infty < 1, \quad \|\mathbf{I} - \mathbf{G}_T^{-1}\mathbf{T}\|_1 < 1.$$

From this it follows that

$$\rho\left(\mathbf{I} - \frac{1}{2}(\mathbf{G}_T^{-1}\mathbf{T} + (\mathbf{G}_T^{-1}\mathbf{T})^T)\right) \leq \frac{1}{2}\|\mathbf{I} - \mathbf{G}_T^{-1}\mathbf{T}\|_\infty + \frac{1}{2}\|\mathbf{I} - \mathbf{G}_T^{-1}\mathbf{T}\|_1 < 1,$$

and thus

$$\lambda_{\min}\left(\frac{1}{2}(\mathbf{Q}_3 + \mathbf{Q}_3^T)\right) > 0.$$

Similar arguments can be used to prove the results for $i = 1$ and $i = 2$. □

In the literature one can find analyses in which for several classes of Krylov subspace methods, under the assumption that the symmetric part of $\mathbf{C}$ is positive definite, bounds for $\|p_k(\mathbf{C})\|_2$ in terms of the quantity

$$\xi(\mathbf{C}) := \frac{\|\mathbf{C}\|_2}{\frac{1}{2}\lambda_{\min}(\mathbf{C} + \mathbf{C}^T)} \tag{4.12}$$

are derived (cf. [9, 10, 22]). These bounds are in general very pessimistic but indicate that if $\xi(\mathbf{C})$ is "small" (i.e., close to one), one can expect fast convergence of the Krylov subspace method applied to $\mathbf{C}$. Another interesting quantity related to the rate of convergence is the spectral condition number $\kappa_2(\mathbf{C})$. Note that

$$1 \leq \kappa_2(\mathbf{C}) \leq \xi(\mathbf{C})$$

holds. Based on this and on the result in (4.11) we take

$$\xi_{\max} := \max_{1 \leq i \leq 3} \xi(\mathbf{Q}_i), \quad \kappa_{\max} := \max_{1 \leq i \leq 3} \kappa_2(\mathbf{Q}_i)$$

as measures for the quality of the block-Gauss–Seidel preconditioner.

We now distinguish between the LGS and SGS methods.

THEOREM 4.5. *For the lexicographic block-Gauss–Seidel method we have*

$$\mathbf{G} = \mathbf{G}^{LGS}, \quad \xi_{\max} = \max\{\xi(\mathbf{Q}_1), \xi(\mathbf{Q}_3)\}, \quad \kappa_{\max} = \max\{\kappa_2(\mathbf{Q}_1), \kappa_2(\mathbf{Q}_3)\}. \tag{4.13}$$

For the symmetric block-Gauss–Seidel method we have

$$\mathbf{G} = \mathbf{G}^{SGS}, \quad \xi_{\max} = \xi(\mathbf{Q}_3), \quad \kappa_{\max} = \kappa_2(\mathbf{Q}_3). \tag{4.14}$$

Proof. For the LGS method we have

$$\mathbf{G}_{L^T} = \mathbf{I}, \quad \mathbf{G}_L = \mathbf{L},$$

and for the SGS method

$$\mathbf{G}_{L^T} = \mathbf{L}^T, \quad \mathbf{G}_L = \mathbf{L}.$$

Hence $\mathbf{Q}_2 = \mathbf{I}$ for the LGS and for the SGS methods, and $\mathbf{Q}_1 = \mathbf{I}$ for the SGS method. □

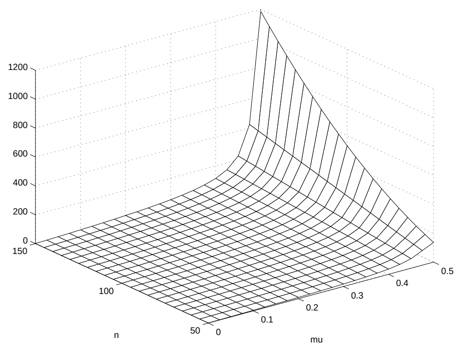


FIG. 4.3. $\kappa_2(\mathbf{G}_T^{-1}\mathbf{T})$ *for* $\mathbf{G}_T = \mathbf{G}_T^{SGS}$.

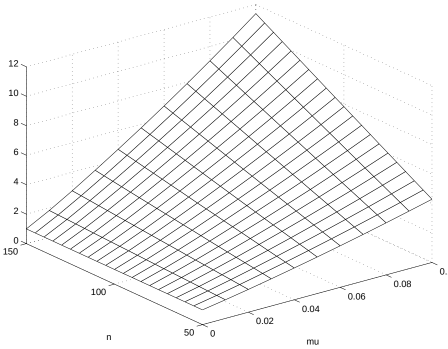


FIG. 4.4. $\kappa_2(\mathbf{G}_T^{-1}\mathbf{T})$ *for* $\mathbf{G}_T = \mathbf{G}_T^{SGS}$.

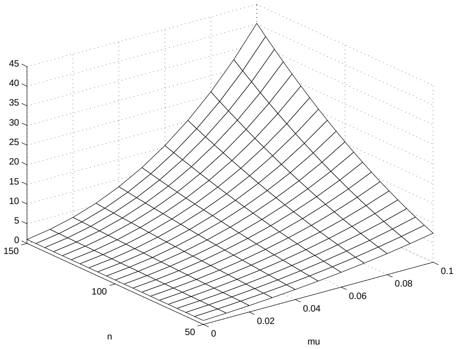


FIG. 4.5. $\xi(\mathbf{G}_T^{-1}\mathbf{T})$ *for* $\mathbf{G}_T = \mathbf{G}_T^{SGS}$.

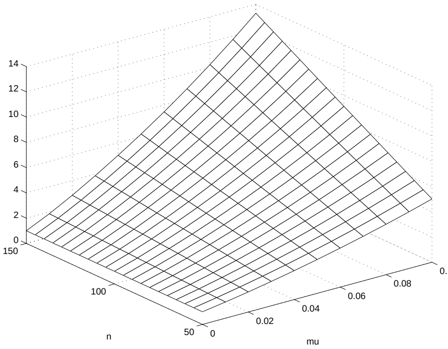


FIG. 4.6. $\kappa_2(\mathbf{G}_T^{-1}\mathbf{T})$ *for* $\mathbf{G}_T = \mathbf{G}_T^{LGS}$.

In Figures 4.3 and 4.4 for the SGS method the dependence of $\kappa_2(\mathbf{Q}_3) = \kappa_2(\mathbf{G}_T^{-1}\mathbf{T})$ on μ and n is shown.

From these figures and the result in (4.14) it follows that for $\mu \in (0, \mu_0)$ with $\mu_0 \ll \frac{1}{2}$ the function $n \to \kappa_{\max}(n)$ increases only slowly. Hence, for "small" μ-values the SGS-preconditioned matrix has a corresponding $\kappa_{\max}$-value which is small, even for "large" n-values. Now note that the dependence of μ on the Mach number M is as in (4.9) (Figure 4.1), and thus for a large range of Mach numbers $M \in [M_0, 1]$ the corresponding $\mu(M)$-values are (very) small, and thus the condition number $\kappa_{\max}$ is small, too. In Figure 4.5 for the SGS method we show, for small μ-values, the dependence of $\xi_{max} = \xi(\mathbf{G}_T^{-1}\mathbf{T})$ on μ and n. Note that for small μ-values the function $n \to \xi_{max}(n)$ increases slowly, too. These observations yield some theoretical explanation of the fast convergence of the SGS-GMRES(20) method in Problems 1 and 2 as compared to the diffusion problem (cf. Figures 3.1, 3.2, 3.4), and of the fact that in Problem 2 (Figure 3.4) the rate of convergence is much higher than in Problem 1 (Figure 3.2).

For the LGS method the term $\xi(\mathbf{Q}_1)$ in (4.13) has to be taken into account. For this term we have

$$\xi(\mathbf{Q}_1) = \frac{\|\mathbf{L}^T\|_2}{\frac{1}{2}\lambda_{\min}(\mathbf{L} + \mathbf{L}^T)} \approx 4\left(\frac{n}{\pi}\right)^2,$$

which is, independently of μ, large if n is large. This gives a theoretical justification of the intuitive conjecture that for a subsonic or transonic one dimensional flow problem with characteristics going in both directions the SGS method should perform (much) better than the LGS method (cf. also the large difference in the rates of convergence in Figures 3.5, 3.6).

The result in (4.14) relates the quality measure $\kappa_{\max}$ of the SGS-method to the condition number $\kappa_2(\mathbf{G}_T^{-1}\mathbf{T})$. The behavior of the function $(\mu, n) \rightarrow \kappa_2(\mathbf{G}_T^{-1}\mathbf{T})$ is shown in Figures 4.3 and 4.4. An important observation is that for "small" μ-values these condition numbers are small. The same holds for the LGS method (cf. Figure 4.6). One can derive (fairly sharp) bounds for $\kappa_2(\mathbf{G}_T^{-1}\mathbf{T})$, which show the dependence of this condition number on n and μ. Here we present such a result for the simplest case, namely for the LGS method. For completeness a proof is given in the appendix. A similar result can be shown to hold for the SGS method.

THEOREM 4.6. *Let $\mathbf{G} = \mathbf{G}_T^{LGS}$ be the LGS preconditioner for the matrix $\mathbf{T} = \mathbf{T}_\mu \in \mathbb{R}^{n\times n}$. For the condition number of the preconditioned matrix the following holds for $\mu \in [0, \frac{1}{2}]$:*

$$\|\mathbf{G}^{-1}\mathbf{T}\|_2\|\mathbf{T}^{-1}\mathbf{G}\|_2 \leq \left(1+\min\left\{\frac{\mu}{h},1\right\}\right)\frac{2\delta_\mu}{1-2\mu}\left(\frac{\mu}{h}+1+\frac{\mu\delta_\mu}{1-2\mu\delta_\mu}\frac{1}{\sqrt{h}}\right),$$
$$\textit{with}\ \ h=\frac{1}{n+1}, \quad \delta_\mu = \min\left\{1, \frac{1-2\mu}{8\mu}\frac{1}{h}\right\}.$$

Remark 1. In our model problem we are interested in the case $\mu \ll \frac{1}{2}$ (e.g., $\mu \in (0, 0.1)$) and $h \ll 1$. For this case we have $\delta_\mu = 1$, and we obtain the following bound for the condition number:

$$\|\mathbf{G}^{-1}\mathbf{T}\|_2\|\mathbf{T}^{-1}\mathbf{G}\|_2 \lesssim \begin{cases} 2(1+\frac{\mu}{h})^2 & \text{if } \frac{\mu}{h}<1, \\ 4(1+\frac{\mu}{h}) & \text{if } \frac{\mu}{h}\geq 1. \end{cases}$$

This bound clearly shows that for small μ there is (at worst) only a slow growth in the condition number as a function of $n = h^{-1} - 1$.

Remark 2. We briefly comment on the very high rate of convergence of the SGS-GMRES(20) method for the transonic flow problem in section 3 (Figures 2.2 and 3.5). In part of the domain the flow is supersonic ($M > 1$), and in another part of the domain the flow is subsonic with Mach numbers $M \in (0.6, 1)$. The upwind discretization in the supersonic part of the domain results in a block lower triangular matrix. Hence in this part of the domain the information is propagated exactly by the symmetric block-Gauss–Seidel method. In the subsonic part of the domain the Mach numbers are ≥ 0.6, and thus the corresponding $\mu(M)$-values lie in the interval $[0, 0.05]$. The analysis in this section shows that in such a case if we freeze the coefficients, the SGS method can be expected to be a very effective preconditioner. At the "critical" points $x = 1$ and $x = 3$ we do not have a smooth behavior, and this results in a low dimensional subspace in which the Gauss–Seidel preconditioner may perform relatively poorly. Due to its very low dimension the error components in this subspace can be reduced effectively by the GMRES method. These arguments give some heuristic explanation of the convergence behavior shown in Figure 3.5. A rigorous analysis for the transonic case is still lacking.

Remark 3. We briefly comment on possible topics for further research towards two dimensional problems. Consider a stationary two dimensional Euler equation that is discretized on a uniform square grid using the Van Leer flux vector-splitting method. The resulting nonlinear problem is linearized at the discrete solution. For the analysis we assume that $V = (\rho, u, v, p)$ is constant as a function of the space variable. In stencil notation the discrete problem has the structure (cf. (4.4))

$$\begin{bmatrix} & B^- & \\ -A^+ & A^+ - A^- + B^+ - B^- & A^- \\ & -B^+ & \end{bmatrix} \tag{4.15}$$

with $A^\pm, B^\pm \in \mathbb{R}^{4\times 4}$. For the Mach numbers in one direction we use the notation $M_u := \frac{u}{c}$, $M_v := \frac{v}{c}$. We consider only $M_u \geq 0, M_v \geq 0$.

In the supersonic case $M_u \geq 1, M_v \geq 1$ we have $B^- = A^- = 0$, and thus the matrix is block lower triangular. Hence the block-Gauss–Seidel method is a direct solver.

For the supersonic case $M_u \in (0,1), M_v \geq 1$ we have that $B^- = 0$, and thus the x-line block-Gauss–Seidel method is a direct solver. To analyze convergence properties of the symmetric block-Gauss–Seidel method (which is not a direct solver) one has to investigate the SGS method applied to the matrix

$$\text{blocktridiag}(-A^+ \ ,\ A^+ - A^- + B^+ \ ,\ A^-).$$

This corresponds to a one dimensional problem, and thus for the analysis one can try to use the same approach used in section 4. Note, however, that the matrix $A^+ \in \mathbb{R}^{4\times 4}$ differs from the one in (4.4).

For the subsonic case $M_u = M_v \in (0,1)$ one has nice symmetry properties. We have $PB^\pm = A^\pm$ with a simple permutation matrix P. Hence properties of the matrix corresponding to (4.15) essentially depend only on those of A^+ and A^-. Suitable transformations (as in section 4) based on the known eigenvector bases of $A^\pm$ may help to determine some of these properties.

Appendix. Proof of Theorem 4.6. In this appendix we give a proof of the result in Theorem 4.6. We consider the tridiagonal matrix $\mathbf{T} = \mathbf{T}_\mu$ as (4.10) with $\mu \in (0, \frac{1}{2})$, and for the preconditioner we take the LGS method:

$$\mathbf{G} = \text{tridiag}(-(1-\mu), 1, 0) \in \mathbb{R}^{n\times n}.$$

In Figure 4.6 we showed the numerically computed values of the function $(\mu, n) \to \kappa_2(\mathbf{G}^{-1}\mathbf{T})$. In this section we derive a rigorous (sharp) bound for this condition number, which shows its dependence on μ and $h = 1/(n+1)$.

We use the notation

$$\mathbf{S} = \begin{pmatrix} 0 & 1 & & \\ & 0 & \ddots & \\ & & \ddots & 1 \\ & & & 0 \end{pmatrix} \in \mathbb{R}^{n\times n}, \qquad \mathbf{W} = \mathbf{I} - \mathbf{S}^T.$$

LEMMA A.1. *The following holds:*

$$\|\mathbf{G}^{-1}\mathbf{T}\|_2 \leq 1 + \min\left\{\frac{\mu}{h}, 1\right\}.$$

Proof. Using $\mathbf{T} = \mathbf{G} - \mu\mathbf{S}$, we obtain

$$\begin{aligned} \|\mathbf{G}^{-1}\mathbf{T}\|_2 = \|\mathbf{I} - \mu\mathbf{G}^{-1}\mathbf{S}\|_2 &\leq 1 + \mu\big(\|\mathbf{G}^{-1}\mathbf{S}\|_\infty \|\mathbf{G}^{-1}\mathbf{S}\|_1\big)^{\frac{1}{2}} \\ &\leq 1 + \mu \sum_{k=0}^{n-1} (1-\mu)^k \leq 1 + \min\{\mu n, 1\}, \end{aligned}$$

and thus the result of this lemma holds. □

We now derive a bound for $\|\mathbf{T}^{-1}\mathbf{G}\|_2$. First we note that $\mathbf{T} = \mathbf{G} - \mu\mathbf{S}$ is a weakly regular splitting; i.e., $\mathbf{G}^{-1} \geq 0$ and $\mu\mathbf{G}^{-1}\mathbf{S} \geq 0$ hold. Moreover, $\mathbf{T}^{-1} \geq 0$ holds, and thus $\mu\rho(\mathbf{G}^{-1}\mathbf{S}) < 1$. From this we obtain that $\mathbf{T}^{-1}\mathbf{G}$ is a positive matrix:

$$\mathbf{T}^{-1}\mathbf{G} = (\mathbf{I} - \mu\mathbf{G}^{-1}\mathbf{S})^{-1} = \sum_{k=0}^{\infty}(\mu\mathbf{G}^{-1}\mathbf{S})^k \geq 0.$$

In our analysis we use the numerical radius

$$r(\mathbf{A}) := \max\{\, |\mathbf{x}^H\mathbf{A}\mathbf{x}| \mid \mathbf{x} \in \mathbb{C}^n, \|\mathbf{x}\|_2 = 1\,\}.$$

We also use the following properties:

$$\begin{aligned} &\|\mathbf{A}\|_2 \leq 2r(\mathbf{A}), \\ &r(\mathbf{A}) = \frac{1}{2}\rho(\mathbf{A} + \mathbf{A}^T) \quad \text{if } \ \mathbf{A} \geq 0. \end{aligned}$$

Using $\mathbf{G} = \mathbf{I} - (1-\mu)\mathbf{S}^T = \mathbf{W} + \mu\mathbf{S}^T$, we get

$$\begin{aligned} \|\mathbf{T}^{-1}\mathbf{G}\|_2 &\leq 2r(\mathbf{T}^{-1}\mathbf{G}) = \rho(\mathbf{T}^{-1}\mathbf{G} + \mathbf{G}^T\mathbf{T}^{-T}) \\ &= \rho\big((\mathbf{T}^{-1}\mathbf{W} + \mathbf{W}^T\mathbf{T}^{-T}) + \mu(\mathbf{T}^{-1}\mathbf{S}^T + \mathbf{S}\mathbf{T}^{-T})\big) \\ &\leq \rho(\mathbf{T}^{-1}\mathbf{W} + \mathbf{W}^T\mathbf{T}^{-T}) + \mu\rho(\mathbf{T}^{-1}\mathbf{S}^T + \mathbf{S}\mathbf{T}^{-T}). \end{aligned} \tag{A.1}$$

In the following two lemmas we derive bounds for the two terms in (A.1).

Lemma A.2. *The following holds:*

$$\mu\rho(\mathbf{T}^{-1}\mathbf{S}^T + \mathbf{S}\mathbf{T}^{-T}) \leq \frac{2\delta_\mu}{1-2\mu}\frac{\mu}{h},$$
$$\textit{with} \quad \delta_\mu := \min\left\{1, \frac{1-2\mu}{8\mu}\frac{1}{h}\right\}.$$

Proof. Note that

$$\rho(\mathbf{T}^{-1}\mathbf{S}^T + \mathbf{S}\mathbf{T}^{-T}) \leq \|\mathbf{T}^{-1}\mathbf{S}^T + \mathbf{S}\mathbf{T}^{-T}\|_\infty \leq \|\mathbf{T}^{-1}\|_\infty + \|\mathbf{T}^{-T}\|_\infty. \tag{A.2}$$

We derive a bound on $\|\mathbf{T}^{-1}\|_\infty$ using $\mathbf{T}^{-1} \geq 0$ and an appropriate barrier function. The difference operator corresponding to $\mathbf{T}$ is given by

$$\begin{aligned} [T]_{x_i} &= \mu[-1 \;\; 2 \;\; -1]_{x_i} + (1-2\mu)[-1 \;\; 1 \;\; 0]_{x_i} \\ &= (1-2\mu)h\left(\frac{\varepsilon}{h^2}[-1 \;\; 2 \;\; -1]_{x_i} + \frac{1}{h}[-1 \;\; 1 \;\; 0]_{x_i}\right), \end{aligned}$$

with $x_i = ih$, $0 \leq i \leq n+1$, and $\varepsilon = \frac{\mu h}{1-2\mu} \in (0, \infty)$. To obtain a suitable barrier function we consider the boundary value problem

$$-\varepsilon u''(x) + u'(x) = 1, \quad x \in (0,1), \quad u(0) = u(1) = 0,$$

with solution given by

$$\bar{u}(x) = x - \frac{\exp(\frac{x}{\varepsilon}) - 1}{\exp(\frac{1}{\varepsilon}) - 1} \ \in [0,1].$$

For $x \in (0,1)$ and $m \geq 2$, $\bar{u}^{(m)}(x) \leq 0$ holds. Using this, it follows from a Taylor expansion that

$$[T]_{x_i}\bar{u} \geq (1-2\mu)h\big(-\varepsilon\bar{u}''(x_i) + \bar{u}'(x_i)\big) = (1-2\mu)h.$$

From this and the fact that $\mathbf{T}$ is inverse positive we obtain

$$\|\mathbf{T}^{-1}\|_\infty \leq \frac{\|\bar{u}\|_{\infty,[0,1]}}{(1-2\mu)h}.$$

We introduce the notation $z_\varepsilon := \varepsilon(\exp(\frac{1}{\varepsilon}) - 1)$. A simple computation yields that on $[0,1]$ the function $\bar{u}$ attains its maximum at $x = \varepsilon \ln z_\varepsilon$, and this maximum is given by

$$\|\bar{u}\|_{\infty,[0,1]} = \varepsilon(\ln z_\varepsilon + z_\varepsilon^{-1} - 1) =: m(\varepsilon).$$

On $(0,\infty)$ the function $\varepsilon \to m(\varepsilon)$ has the following properties:

$$\begin{aligned}
&\lim_{\varepsilon\downarrow 0} m(\varepsilon) = 1, \quad m'(\varepsilon) < 0, \quad \lim_{\varepsilon\to\infty} m(\varepsilon) = 0,\\
&\lim_{\varepsilon\downarrow 0} \varepsilon m(\varepsilon) = 0, \quad (\varepsilon m(\varepsilon))' > 0, \quad \lim_{\varepsilon\to\infty} \varepsilon m(\varepsilon) = \frac{1}{8}.
\end{aligned}$$

It follows that

$$\begin{aligned}
\|\mathbf{T}^{-1}\|_\infty &\leq \frac{1}{(1-2\mu)h}\, m(\varepsilon) \leq \frac{1}{(1-2\mu)h},\\
\|\mathbf{T}^{-1}\|_\infty &\leq \frac{\varepsilon^{-1}}{(1-2\mu)h}\, \varepsilon m(\varepsilon) \leq \frac{1}{\mu h^2}\frac{1}{8},
\end{aligned}$$

and thus

$$\|\mathbf{T}^{-1}\|_\infty \leq \frac{1}{(1-2\mu)h}\delta_\mu.$$

The same bound can be derived for $\|\mathbf{T}^{-T}\|_\infty$ if one uses the (adjoint) equation $-\varepsilon u'' - u' = 1$. These bounds in combination with (A.2) prove the result. □

LEMMA A.3. *The following holds, with δ_μ as in Lemma* A.2*:*

$$\rho(\mathbf{T}^{-1}\mathbf{W} + \mathbf{W}^T\mathbf{T}^{-T}) \leq \frac{2\delta_\mu}{1-2\mu}\left(1 + \frac{\mu\delta_\mu}{1-2\mu+\mu\delta_\mu}\, h^{-\frac{1}{2}}\right).$$

Proof. We use the notation

$$\begin{aligned}
&\xi = \frac{\mu}{1-\mu}, \quad \mathbf{1} = (1,\dots,1)^T \in \mathbb{R}^n, \quad \mathbf{e}_1 = (1,0,\dots,0)^T \in \mathbb{R}^n,\\
&\mathbf{x} = (\mathbf{I} - \xi\mathbf{S})^{-1}\mathbf{1}, \quad \mathbf{y} = (\mathbf{I} - \xi\mathbf{S}^T)^{-1}\mathbf{e}_1 = (1,\xi,\xi^2,\dots,\xi^{n-1})^T,\\
&\beta = \|\mathbf{y}\|_1 = \sum_{k=0}^{n-1}\xi^k\ , \quad \tau = \frac{\xi}{1+\xi\beta}.
\end{aligned}$$

Note that

$$\begin{aligned}\mathbf{T}^{-1}\mathbf{W} &= \left(\mathbf{W}-\mu(\mathbf{S}-\mathbf{S}^T)\right)^{-1}\mathbf{W} = \left(\mathbf{I}-\mu\mathbf{W}^{-1}(\mathbf{S}-\mathbf{S}^T)\right)^{-1}\\ &= \left(\mathbf{I}-\mu(\mathbf{I}+\mathbf{S}-\mathbf{1}\mathbf{e}_1^T)\right)^{-1} = \frac{1}{1-\mu}(\mathbf{I}-\xi\mathbf{S}+\xi\mathbf{1}\mathbf{e}_1^T)^{-1}\\ &= \frac{1}{1-\mu}(\mathbf{I}+\xi\mathbf{x}\mathbf{e}_1^T)^{-1}(\mathbf{I}-\xi\mathbf{S})^{-1} = \frac{1}{1-\mu}(\mathbf{I}-\tau\mathbf{x}\mathbf{e}_1^T)(\mathbf{I}-\xi\mathbf{S})^{-1}\\ &= \frac{1}{1-\mu}\left((\mathbf{I}-\xi\mathbf{S})^{-1}-\tau\mathbf{x}\mathbf{y}^T\right).\end{aligned}$$

Using

$$\begin{aligned}\|(\mathbf{I}-\xi\mathbf{S})^{-1}\|_2 &\le \left(\|(\mathbf{I}-\xi\mathbf{S})^{-1}\|_\infty\|(\mathbf{I}-\xi\mathbf{S})^{-1}\|_1\right)^{\frac{1}{2}} = \beta,\\ \|\mathbf{x}\|_2 &\le \|(\mathbf{I}-\xi\mathbf{S})^{-1}\|_2\|\mathbf{1}\|_2 \le \beta\sqrt{n} \le \beta h^{-\frac{1}{2}},\\ \|\mathbf{y}\|_2 &\le \|(\mathbf{I}-\xi\mathbf{S})^{-1}\|_2\|\mathbf{e}_1\|_2 \le \beta,\end{aligned}$$

we obtain

$$\begin{aligned}\rho(\mathbf{T}^{-1}\mathbf{W}+\mathbf{W}^T\mathbf{T}^{-T}) &= \frac{1}{1-\mu}\rho\left((\mathbf{I}-\xi\mathbf{S})^{-1}+(\mathbf{I}-\xi\mathbf{S}^T)^{-1}-\tau(\mathbf{x}\mathbf{y}^T+\mathbf{y}\mathbf{x}^T)\right)\\ &\le \frac{2}{1-\mu}\left(\|(\mathbf{I}-\xi\mathbf{S})^{-1}\|_2+\tau\|\mathbf{x}\|_2\|\mathbf{y}\|_2\right)\\ &\le \frac{2\beta}{1-\mu}\left(1+\frac{\xi\beta}{1+\xi\beta}h^{-\frac{1}{2}}\right). \qquad \text{(A.3)}\end{aligned}$$

We use

$$\begin{aligned}\beta &\le \min\left\{\frac{1}{1-\xi}, n\right\} \le \frac{1-\mu}{1-2\mu}\min\left\{1, \frac{1-2\mu}{1-\mu}h^{-1}\right\}\\ &\le \frac{1-\mu}{1-2\mu}\min\left\{1, \frac{1-2\mu}{8\mu}h^{-1}\right\} = \frac{1-\mu}{1-2\mu}\delta_\mu.\end{aligned}$$

Hence

$$\frac{2\beta}{1-\mu} \le \frac{2\delta_\mu}{1-2\mu} \qquad \text{(A.4)}$$

holds. Finally, note that

$$\frac{\xi\beta}{1+\xi\beta} \le \frac{\frac{\mu}{1-2\mu}\delta_\mu}{1+\frac{\mu}{1-2\mu}\delta_\mu} = \frac{\mu\delta_\mu}{1-2\mu+\mu\delta_\mu}. \qquad \text{(A.5)}$$

Combination of (A.3), (A.4), and (A.5) yields the result. □

Substitution of the results of Lemmas A.2 and A.3 into (A.1) yields

$$\|\mathbf{T}^{-1}\mathbf{G}\|_2 \le \frac{2\delta_\mu}{1-2\mu}\left(\frac{\mu}{h}+1+\frac{\mu\delta_\mu}{1-2\mu+\mu\delta_\mu}\frac{1}{\sqrt{h}}\right).$$

Combination of this result with the result of Lemma A.1 shows that the inequality in Theorem 4.6 holds.

REFERENCES

[1] O. Axelsson, *Iterative Solution Methods*, Cambridge University Press, New York, 1994.

[2] J. Bey and A. Reusken, *On the convergence of basic iterative methods for convection-diffusion equations*, Numer. Linear Algebra Appl., 6 (1999), pp. 329–352.

[3] Ph. Birken, *Preconditioning GMRES for Steady Compressible Inviscid Flows*, Institute für Geometric und Praktische Mathematik Report 212, Rheinisch-Westfälische Technische Hochschule-Aachen, Aachen, Germany, 2002.

[4] F. Bramkamp, J. Ballmann, and S. Müller, *Development of a flow solver employing local adaptation based on multiscale analysis on B-spline grids*, in Proceedings of the 8th Annual Conference of the CFD Society of Canada, Montreal, 2000, pp. 113–118.

[5] P.N. Brown, and Y. Saad, *Hybrid Krylov methods for nonlinear systems of equations*, SIAM J. Sci. Statist. Comput., 11 (1990), pp. 450–481.

[6] R.F. Chen and Z.J. Wang, *Fast block lower-upper symmetric Gauß-Seidel scheme for arbitrary grids*, AIAA J., 38 (2000), pp. 2238–2245.

[7] W. Hackbusch, *Iterative Solution of Large Sparse Systems*, Springer-Verlag, Berlin, 1994.

[8] M. Eiermann, *Fields of values and iterative methods*, Linear Algebra Appl., 180 (1993), pp. 167–197.

[9] S.C. Eisenstat, H.C. Elman, and M.H. Schultz, *Variational iterative methods for nonsymmetric systems of linear equations*, SIAM J. Numer. Anal., 20 (1983), pp. 345–357.

[10] A. Greenbaum, *Iterative Methods for Solving Linear Systems*, Frontiers Appl. Math. 17, SIAM, Philadelphia, 1997.

[11] C. Hirsch, *Numerical Computation of Internal and External Flows: Computational Methods for Inviscid and Viscous Flows*, Vol. 2, Wiley, Chichester, UK, 1988.

[12] A. Jameson, *Solution of the Euler equations for two-dimensional transonic flow by a multigrid method*, Appl. Math. Comp., 13 (1983), pp. 327–356.

[13] A. Jameson and D.A. Caughey, *How many steps are required to solve the Euler equations of steady, compressible flow: In search of a fast solution algorithm*, AIAA paper 2001–2673, in Proceedings of the 15th AIAA Computational Fluid Dynamics Conference, Anaheim, CA, 2001.

[14] D.A. Knoll, P. McHugh, and D. Keyes, *Newton–Krylov methods for low Mach number compressible combustion*, AIAA J., 34 (1996), pp. 961–967.

[15] D.A. Knoll and W.J. Rider, *A multigrid preconditioned Newton–Krylov method*, SIAM J. Sci. Comput., 21 (1999), pp. 691–710.

[16] B. Koren, *Defect correction and multigrid for an efficient and accurate computation of airfoil flows*, J. Comput. Phys., 183 (1988), pp. 193–206.

[17] A. Meister, *Comparison of different Krylov subspace methods embedded in an implicit finite volume scheme for the computation of viscous and inviscid flow fields on unstructured grids*, J. Comput. Phys., 140 (1998), pp. 311–345.

[18] A. Meister and Th. Sonar, *Finite-volume schemes for compressible fluid flow*, Surveys Math. Indust., 8 (1998), pp. 1–36.

[19] A. Meister and C. Vömel, *Efficient preconditioning of linear systems arising from the discretization of hyperbolic conservation laws*, Adv. Comput. Math., 14 (2001), pp. 49–73.

[20] N.A. Pierce and M.B. Giles, *Preconditioned multigrid methods for compressible flow computations on stretched meshes*, J. Comput. Phys., 136 (1997), pp. 425–445.

[21] A.H. Shapiro, *The Dynamics and Thermodynamics of Compressible Fluid Flow*, Ronald Press, New York, 1953.

[22] G. Starke, *Field-of-values analysis of preconditioned iterative methods for nonsymmetric elliptic problems*, Numer. Math., 78 (1997), pp. 103–117.

[23] R.S. Varga, *Matrix Iterative Analysis*, Prentice–Hall, Englewood Cliffs, NJ, 1962.

[24] V. Venkatakrishnan and D.J. Mavripilis, *Implicit solvers for unstructured meshes*, J. Comput. Phys., 105 (1993), pp. 83–91.

[25] V. Venkatakrishnan and D.J. Mavripilis, *Implicit method for the computation of unsteady flows on unstructured grids*, J. Comput. Phys., 127 (1996), pp. 380–397.

[26] V. Venkatakrishnan, *Implicit schemes and parallel computing in unstructured grid CFD*, ICASE report 95-28, ICASE, Langley Research Center, Hampton, VA, 1995.

[27] P. Wesseling, *An Introduction to Multigrid Methods*, Wiley, Chichester, UK, 1992.

[28] P. Wesseling, *Principles of Computational Fluid Dynamics*, Springer-Verlag, Berlin, 2001.

SIAM J. NUMER. ANAL.
Vol. 41, No. 4, pp. 1406–1426

AN EXPLICIT UNCONDITIONALLY STABLE NUMERICAL METHOD FOR SOLVING DAMPED NONLINEAR SCHRÖDINGER EQUATIONS WITH A FOCUSING NONLINEARITY*

WEIZHU BAO† AND DIETER JAKSCH‡

Abstract. This paper introduces an extension of the time-splitting sine-spectral (TSSP) method for solving damped focusing nonlinear Schrödinger equations (NLSs). The method is explicit, unconditionally stable, and time transversal invariant. Moreover, it preserves the exact decay rate for the normalization of the wave function if linear damping terms are added to the NLS. Extensive numerical tests are presented for cubic focusing NLSs in two dimensions with a linear, cubic, or quintic damping term. Our numerical results show that quintic or cubic damping always arrests blowup, while linear damping can arrest blowup only when the damping parameter δ is larger than a threshold value $\delta_{\rm th}$. We note that our method can also be applied to solve the three-dimensional Gross–Pitaevskii equation with a quintic damping term to model the dynamics of a collapsing and exploding Bose–Einstein condensate (BEC).

Key words. damped nonlinear Schrödinger equation (DNLS), time-splitting sine-spectral (TSSP) method, Gross–Pitaevskii equation (GPE), Bose–Einstein condensate (BEC), complex Ginzburg–Landau (CGL)

AMS subject classifications. 35Q55, 65T40, 65N12, 65N35, 81-08

DOI. 10.1137/S0036142902413391

1. Introduction.

Since the first experimental realization of Bose–Einstein condensation (BEC) in dilute weakly interacting gases, the nonlinear Schrödinger equation (NLS) has been used extensively to describe the single particle properties of BECs. The results obtained by solving the NLS showed excellent agreement with most of the experiments (for a review, see [2, 3, 11, 10]). In fact, up to now there have been very few experiments in ultracold dilute bosonic gases which could not be described properly by using theoretical methods based on the NLS [20, 23].

Recent experiments by Donley et al. [12] provide new experimental results for checking the validity of describing a BEC by using the NLS in the case of attractive interactions (focusing nonlinearity) in three dimensions. Since the particle density might become very large in the case of attractive interactions, inelastic collisions become important and cannot be neglected. These inelastic collisions are assumed to be accounted for by adding damping terms to the NLS. Two particle inelastic processes are taken into account by a cubic damping term, while three particle inelastic collisions are described by a quintic damping term. Collisions with the background gas and feeding of the condensate can be studied by adding linear damping terms. One of the major theoretical challenges in comparing results obtained in the experiment with theoretical results is to find reliable methods for solving the NLS with a focusing

*Received by the editors August 25, 2002; accepted for publication (in revised form) January 31, 2003; published electronically September 9, 2003. This work was supported by the WITTGENSTEIN-AWARD of P. Markowich and P. Zoller which is funded by the Austrian National Science Foundation FWF.

http://www.siam.org/journals/sinum/41-4/41339.html

†Department of Computational Science, National University of Singapore, Singapore 117543 (bao@cz3.nus.edu.sg). The research of this author was supported by National University of Singapore grant R-151-000-027-112.

‡Institut für Theoretische Physik, Universität Innsbruck, A-6020 Innsbruck, Austria (dieter.jaksch@physics.oxford.ac.uk).

nonlinearity and damping terms in the parameter regime where the experiments are performed.

The aim of this paper is to extend the time-splitting sine-spectral (TSSP) method for solving the focusing NLS with additional damping terms and to present extensive numerical tests. The comparison of our numerical results with the experimental results obtained for a collapsing BEC [12] will be presented elsewhere [8].

We consider the NLS [7, 36]

$$i\,\psi_t = -\frac{1}{2}\,\Delta\psi + V(\mathbf{x})\,\psi - \beta|\psi|^{2\sigma}\psi, \qquad t>0, \qquad \mathbf{x}\in\mathbb{R}^d, \tag{1.1}$$

$$\psi(\mathbf{x},t=0) = \psi_0(\mathbf{x}), \qquad \mathbf{x}\in\mathbb{R}^d, \tag{1.2}$$

with $\sigma > 0$ a positive constant, where $\sigma = 1$ corresponds to a cubic nonlinearity, $\sigma = 2$ corresponds to a quintic nonlinearity, $V(\mathbf{x})$ is a real-valued potential whose shape is determined by the type of system under investigation, and β positive/negative corresponds to the focusing/defocusing NLS. In BEC, where (1.1) is also known as the Gross–Pitaevskii equation (GPE) [21, 26, 33], ψ is the macroscopic wave function of the condensate, t is time, $\mathbf{x}$ is the spatial coordinate, and $V(\mathbf{x})$ is a trapping potential which usually is harmonic and can thus be written as $V(\mathbf{x}) = \frac{1}{2}\left(\gamma_1^2x_1^2 + \cdots + \gamma_d^2x_d^2\right)$ with $\gamma_1,\ldots,\gamma_d \geq 0$. Two important invariants of (1.1) are the *normalization of the wave function*

$$N(t) = \int_{\mathbb{R}^d} |\psi(\mathbf{x},t)|^2\,d\mathbf{x}, \qquad t\geq 0, \tag{1.3}$$

and the *energy*

$$E(t) = \int_{\mathbb{R}^d} \left[\frac{1}{2}|\nabla\psi(\mathbf{x},t)|^2 + V(\mathbf{x})|\psi(\mathbf{x},t)|^2 - \frac{\beta}{\sigma+1}|\psi(\mathbf{x},t)|^{2\sigma+2}\right] d\mathbf{x}, \quad t\geq 0. \tag{1.4}$$

From the theory for the local existence of the solution of (1.1), it is well known that if $\|\psi(\cdot,t)\|_{H^1}$ is bounded, the solution exists for all t [36]. As a result, when the NLS is defocusing ($\beta < 0$), conservation of energy implies that $\int_{\mathbb{R}^d} |\nabla\psi(\mathbf{x},t)|^2\,d\mathbf{x}$ is bounded and the solution exists globally. On the other hand, if the NLS is focusing ($\beta > 0$) at critical ($\sigma d = 2$) or supercritical ($\sigma d > 2$) dimensions and for an initial energy $E(0) < 0$, then the solutions of (1.1) can self-focus and become singular in finite time; i.e., there exists a time $t_* < \infty$ such that (see [36])

$$\lim_{t\to t_*} |\nabla\psi|_{L^2} = \infty \qquad \text{and} \qquad \lim_{t\to t_*} |\psi|_{L^\infty} = \infty.$$

However, the physical quantities modeled by ψ do not become infinite, which implies that the validity of (1.1) breaks down near the singularity. Additional physical mechanisms, which were initially small, become important near the singular point and prevent the formation of the singularity. In BEC, the particle density $|\psi|^2$ becomes large close to the critical point, and inelastic collisions between particles which are negligible for small densities become important. Therefore, a small damping (absorption) term is introduced into the NLS (1.1) which describes inelastic processes. We are interested in the cases where these damping mechanisms are important and therefore restrict ourselves to the case of focusing nonlinearities $\beta > 0$, where β may also be time dependent. We consider the damped nonlinear Schrödinger equation (DNLS)

$$i\,\psi_t = -\frac{1}{2}\,\Delta\psi + V(\mathbf{x})\,\psi - \beta|\psi|^{2\sigma}\psi - i\,g(|\psi|^2)\psi, \qquad t>0, \quad \mathbf{x}\in\mathbb{R}^d, \tag{1.5}$$

$$\psi(\mathbf{x},t=0) = \psi_0(\mathbf{x}), \qquad \mathbf{x}\in\mathbb{R}^d, \tag{1.6}$$

where $g(\rho) \ge 0$ for $\rho = |\psi|^2 \ge 0$ is a real-valued monotonically increasing function.

The general form of (1.5) covers many DNLSs arising in various different applications. In BEC, for example, when $g(\rho) \equiv 0$, (1.5) reduces to the usual GPE (1.1); a linear damping term $g(\rho) \equiv \delta$ with $\delta > 0$ describes inelastic collisions with the background gas; cubic damping $g(\rho) = \delta_1 \beta \rho$ with $\delta_1 > 0$ corresponds to two-body loss [13, 35, 34]; and a quintic damping term of the form $g(\rho) = \delta_2 \beta^2 \rho^2$ with $\delta_2 > 0$ adds three-body loss to the GPE (1.1) [1, 35, 34]. It is easy to see that the decay of the normalization according to (1.5) due to damping is given by

$$(1.7)\quad N'(t) = \frac{d}{dt}\int_{\mathbb{R}^d} |\psi(\mathbf{x},t)|^2 \, d\mathbf{x} = -2\int_{\mathbb{R}^d} g(|\psi(\mathbf{x},t)|^2)|\psi(\mathbf{x},t)|^2 \, d\mathbf{x} \le 0, \quad t > 0.$$

In particular, if $g(\rho) \equiv \delta$ with $\delta > 0$, the normalization is given by

$$(1.8)\quad N(t) = \int_{\mathbb{R}^d} |\psi(\mathbf{x},t)|^2 \, d\mathbf{x} = e^{-2\delta\, t} N(0) = e^{-2\delta\, t} \int_{\mathbb{R}^d} |\psi_0(\mathbf{x})|^2 \, d\mathbf{x}, \quad t \ge 0.$$

There has been a series of recent studies which deals with the analysis and numerical solution of the DNLS. Fibich [14] analyzed the effect of linear damping (absorption) on the critical self-focusing NLS, Tsutsumi [37, 38] studied the global solutions of the NLS with linear damping, and the regularity of attractors and approximate inertial manifolds for a weakly damped NLS were given in Goubet [17, 19, 18] and by Jolly, Temam, and Xiong [24]. For numerically solving the linearly damped NLS, Peranich [32] proposed a finite difference scheme, and this method was revisited recently by Ciegis and Pakalnyte [9] and Zhang and Lu [39]. Moebs and Temam [30] presented a multilevel method for weakly damped NLS, and Moebs applied it to solve a stochastic weakly damped NLS in [29]. Variable mesh difference schemes for the NLS with a linear damping term were used by Iyengar, Jayaraman, and Balasubramanian [22].

Also, the TSSP method, which we will use in this paper to solve the DNLS, was already successfully used for solving the Schrödinger equation in the semiclassical regime and for describing BEC using the GPE by Bao et al. [4, 5, 7]. The TSSP method is explicit, unconditionally stable, and time transversal invariant. Moreover, it gives the exact decay rate of the normalization when linear damping is applied to the NLS (i.e., $g(\rho) \equiv \delta$ with $\delta > 0$ in (1.5)) and yields spectral accuracy for spatial derivatives and second-order accuracy for the time derivative. Thus this method is a very good candidate for solving the DNLS, especially in two or three dimensions. We test the novel numerical method extensively in two dimensions.

Finally, we want to emphasize that the NLS is also used in nonlinear optics, e.g., to describe the propagation of an intense laser beam through a medium with a Kerr nonlinearity [16, 36]. In nonlinear optics, $\psi = \psi(\mathbf{x}, t)$ describes the electrical field amplitude, t is the spatial coordinate in the direction of propagation, $\mathbf{x} = (x_1, \ldots, x_d)^T$ is the transverse spatial coordinate, and $V(\mathbf{x})$ is determined by the index of refraction. Nonlinear damping terms of the form $g(\rho) = \delta \beta^q \rho^q$ with $\delta, q > 0$ correspond to multiphoton absorption processes [14].

The paper is organized as follows. In section 2, we present the TSSP approximation for the damped nonlinear Schrödinger equation. In section 3, numerical tests are presented for the cubic focusing NLS in two dimensions with a linear, cubic, or quintic damping term. In section 4, some conclusions are drawn.

2. Time-splitting sine-spectral method.

In this section we present a time-splitting sine-spectral (TSSP) method for solving the problem (1.5), (1.6) with homogeneous periodic boundary conditions. For simplicity of notation, we shall introduce the method for the case of one spatial dimension ($d = 1$). Generalizations to $d > 1$ are straightforward for tensor product grids, and the results remain valid without modifications. For $d = 1$, the problem becomes

$$i\,\psi_t = -\frac{1}{2}\psi_{xx} + V(x)\psi - \beta|\psi|^{2\sigma}\psi - i\,g(|\psi|^2)\psi, \qquad a < x < b, \quad t > 0, \tag{2.1}$$

$$\psi(x, t = 0) = \psi_0(x), \quad a \le x \le b, \qquad \psi(a, t) = \psi(b, t) = \mathbf{0}, \quad t \ge 0. \tag{2.2}$$

2.1. General damping term. We choose the spatial mesh size $h = \Delta x > 0$ with $h = (b - a)/M$ and M an even positive integer. The time step is given by $k = \Delta t > 0$, and we define grid points and time steps by

$$x_j := a + j\,h, \qquad t_n := n\,k, \qquad j = 0, 1, \ldots, M, \qquad n = 0, 1, 2, \ldots.$$

Let ψ_j^n be the numerical approximation of $\psi(x_j, t_n)$ and ψ^n the solution vector at time $t = t_n = nk$ with components ψ_j^n.

From time $t = t_n$ to time $t = t_{n+1}$, the DNLS (2.1) is solved in two steps. One solves

$$i\,\psi_t = -\frac{1}{2}\psi_{xx} \tag{2.3}$$

for one time step, followed by solving

$$i\,\psi_t(x, t) = V(x)\psi(x, t) - \beta|\psi(x, t)|^{2\sigma}\psi(x, t) - i\,g(|\psi(x, t)|^2)\psi(x, t), \tag{2.4}$$

again for the same time step. Equation (2.3) is discretized in space by the sine-spectral method and integrated in time *exactly.* For $t \in [t_n, t_{n+1}]$, multiplying the ODE (2.4) by $\overline{\psi(x, t)}$, the conjugate of $\psi(x, t)$, one obtains

$$i\,\psi_t(x, t)\overline{\psi(x, t)} = V(x)|\psi(x, t)|^2 - \beta|\psi(x, t)|^{2\sigma+2} - i\,g(|\psi(x, t)|^2)|\psi(x, t)|^2. \tag{2.5}$$

Subtracting the conjugate of (2.5) from (2.5) and multiplying by $-i$, one obtains

$$\frac{d}{dt}|\psi(x, t)|^2 = \overline{\psi_t(x, t)}\psi(x, t) + \psi_t(x, t)\overline{\psi(x, t)} = -2g(|\psi(x, t)|^2)|\psi(x, t)|^2. \tag{2.6}$$

Let

$$f(s) = \int \frac{1}{s\,g(s)}\,ds, \qquad h(s, \tau) = \begin{cases} f^{-1}\left(f(s) - 2\tau\right), & s > 0,\ \tau \ge 0, \\ 0, & s = 0,\ \tau \ge 0. \end{cases} \tag{2.7}$$

Then, if $g(s) \ge 0$ for $s \ge 0$, we find

$$0 \le h(s, \tau) \le s \qquad \text{for} \quad s \ge 0, \quad \tau \ge 0, \tag{2.8}$$

and the solution of the ODE (2.6) can be expressed as (with $\tau = t - t_n$)

$$\begin{aligned} 0 \le \rho(t) = \rho(t_n + \tau) &:= |\psi(x, t)|^2 = h\left(|\psi(x, t_n)|^2, t - t_n\right) := h\left(\rho(t_n), \tau\right) \\ &\le \rho(t_n) = |\psi(x, t_n)|^2, \qquad t_n \le t \le t_{n+1}. \end{aligned} \tag{2.9}$$

Combining (2.9) and (2.4), we obtain

$$\begin{aligned} i\,\psi_t(x, t) = V(x)\psi(x, t) &- \beta\left[h\left(|\psi(x, t_n)|^2, t - t_n\right)\right]^{\sigma}\psi(x, t) \\ &- i\,g\left(h\left(|\psi(x, t_n)|^2, t - t_n\right)\right)\psi(x, t), \qquad t_n \le t \le t_{n+1}. \end{aligned} \tag{2.10}$$

Integrating (2.10) from t_n to t, we find

$$\begin{aligned}\psi(x,t) &= \exp\left\{i\left[-V(x)(t-t_n) + G\left(|\psi(x,t_n)|^2, t-t_n\right)\right] - F\left(|\psi(x,t_n)|^2, t-t_n\right)\right\}\\ &\quad\times \psi(x,t_n), \qquad t_n \le t \le t_{n+1}, \end{aligned} \tag{2.11}$$

where we have defined

$$F(s,r) = \int_0^r g(h(s,\tau))\, d\tau \ge 0, \quad G(s,r) = \int_0^r \beta\ [h(s,\tau)]^\sigma\ d\tau. \tag{2.12}$$

To find the time evolution between $t = t_n$ and $t = t_{n+1}$, we combine the splitting steps via the standard second-order Strang splitting TSSP method for solving the DNLS (2.1). In detail, the steps for obtaining ψ_j^{n+1} from ψ_j^n are given by

$$\begin{aligned}\psi_j^* &= \exp\left\{-F\left(|\psi_j^n|^2, k/2\right) + i\left[-V(x_j)k/2 + G\left(|\psi_j^n|^2, k/2\right)\right]\right\}\ \psi_j^n,\\ \psi_j^{**} &= \sum_{l=1}^{M-1} e^{-ik\mu_l^2/2}\ \widehat{\psi}_l^*\ \sin(\mu_l(x_j-a)), \qquad j = 1,2,\ldots,M-1,\\ \psi_j^{n+1} &= \exp\left\{-F\left(|\psi_j^{**}|^2, k/2\right) + i\left[-V(x_j)k/2 + G\left(|\psi_j^{**}|^2, k/2\right)\right]\right\}\ \psi_j^{**},\end{aligned} \tag{2.13}$$

where $\widehat{U}_l$ are the sine-transform coefficients of a complex vector $U = (U_0, U_1, \ldots, U_M)$ with $U_0 = U_M = \mathbf{0}$ which are defined as

$$\mu_l = \frac{\pi l}{b-a}, \quad \widehat{U}_l = \frac{2}{M}\sum_{j=1}^{M-1} U_j\ \sin(\mu_l(x_j-a)),\ l = 1,2,\ldots,M-1, \tag{2.14}$$

where

$$\psi_j^0 = \psi(x_j, 0) = \psi_0(x_j), \qquad j = 0,1,2,\ldots,M. \tag{2.15}$$

Note that the only time discretization error of the TSSP method is the splitting error, which is second-order in k if the integrals in (2.7) and (2.12) can be evaluated analytically.

2.2. Most frequently used damping terms. In this subsection we present explicit formulae for using the TSSP method when solving the NLS with those damping terms most frequently appearing in BEC and nonlinear optics.

Case I. *NLS with a linear damping term.* We choose $g(\rho) \equiv \delta$ with $\delta > 0$ in (1.5). In BEC, this damping term describes inelastic collisions of condensate particles with the background gas. From (2.7), we find

$$f(s) = \int \frac{1}{\delta s} ds = \frac{1}{\delta}\ln s \qquad \text{and} \qquad h(s,\tau) = e^{-2\delta\tau}\ s. \tag{2.16}$$

Substituting (2.16) into (2.9) and (2.12), we obtain

$$\rho(t) = e^{-2\delta(t-t_n)}\ |\psi(x,t_n)|^2, \quad t_n \le t \le t_{n+1}, \tag{2.17}$$

$$F(s,r) = \delta r, \tag{2.18}$$

$$G(s,r) = \frac{\beta s^\sigma}{2\delta\sigma}\left(1 - e^{-2\delta\sigma r}\right). \tag{2.19}$$

Substituting (2.18) and (2.19) into (2.13), we get the following second-order TSSP steps for the NLS with a linear damping term:

$$\begin{aligned}
&\psi_j^* = \exp\left\{-k\delta/2 + i\left[-V(x_j)k/2 + \beta|\psi_j^n|^{2\sigma}\left(1-e^{-\delta\sigma k}\right)/(2\delta\sigma)\right]\right\}\ \psi_j^n,\\
&\psi_j^{**} = \sum_{l=1}^{M-1} e^{-ik\mu_l^2/2}\ \widehat{\psi}_l^*\ \sin(\mu_l(x_j-a)), \qquad j=1,2,\dots,M-1, \qquad (2.20)\\
&\psi_j^{n+1} = \exp\left\{-k\delta/2 + i\left[-V(x_j)k/2 + \beta|\psi_j^{**}|^{2\sigma}\left(1-e^{-\delta\sigma k}\right)/(2\delta\sigma)\right]\right\}\ \psi_j^{**}.
\end{aligned}$$

Case II. *NLS with a damping term of the form $g(\rho) = \delta\beta^q\rho^q$, where δ, $q > 0$ in* (1.5). For $q = 1$ ($q = 2$), we obtain the damping term describing two (three) particle inelastic collisions in BEC. From (2.7) we get

$$f(s) = \int \frac{1}{\delta\beta^q s^{q+1}}ds = -\frac{1}{q\delta\beta^q s^q} \qquad \text{and} \qquad h(s,\tau) = \frac{s}{(1+2q\delta\tau\beta^q s^q)^{1/q}}. \qquad (2.21)$$

Substituting (2.21) into (2.9) and (2.12), we obtain

$$\rho(t) = \frac{|\psi(x,t_n)|^2}{[1+2q\delta\beta^q(t-t_n)|\psi(x,t_n)|^{2q}]^{1/q}}, \quad t_n \le t \le t_{n+1}, \qquad (2.22)$$

$$F(s,r) = \frac{1}{2q}\ln\left(1+2q\delta r\beta^q s^q\right), \qquad (2.23)$$

$$G(s,r) = \begin{cases} \dfrac{\beta^{1-q}}{2\delta q}\ln\left(1+2q\delta r\beta^q s^q\right), & q=\sigma,\\ \dfrac{\beta^{1-q}s^{\sigma-q}\left[-1+(1+2q\delta r\beta^q s^q)^{(q-\sigma)/q}\right]}{2\delta(q-\sigma)}, & \sigma\neq q. \end{cases} \qquad (2.24)$$

Substituting (2.23) and (2.24) into (2.13), we get the following second-order TSSP method for the NLS:

(2.25)

$$\psi_j^* = \begin{cases} \dfrac{\exp\left\{i\left[-V(x_j)k/2 + \beta^{1-q}\ln\left(1+\delta qk\beta^q|\psi_j^n|^{2q}\right)/(2\delta q)\right]\right\}}{\left(1+q\delta k\beta^q|\psi_j^n|^{2q}\right)^{1/2q}}\ \psi_j^n, & \sigma = q,\\ \dfrac{\exp\left\{i\left[-\frac{V(x_j)k}{2} + \frac{\beta^{1-q}|\psi_j^n|^{2\sigma-2q}}{2\delta(q-\sigma)}\left(-1+\left(1+\delta qk\beta^q|\psi_j^n|^{2q}\right)^{\frac{q-\sigma}{q}}\right)\right]\right\}}{\left(1+q\delta k\beta^q|\psi_j^n|^{2q}\right)^{1/2q}}\ \psi_j^n, & \sigma\neq q, \end{cases}$$

$$\psi_j^{**} = \sum_{l=1}^{M-1} e^{-ik\mu_l^2/2}\ \widehat{\psi}_l^*\ \sin(\mu_l(x_j-a)), \qquad j=1,2,\dots,M-1,$$

$$\psi_j^{n+1} = \begin{cases} \dfrac{\exp\left\{i\left[-V(x_j)k/2 + \beta^{1-q}\ln\left(1+\delta qk\beta^q|\psi_j^{**}|^{2q}\right)/(2\delta q)\right]\right\}}{\left(1+q\delta k\beta^q|\psi_j^{**}|^{2q}\right)^{1/2q}}\ \psi_j^{**}, & \sigma = q,\\ \dfrac{\exp\left\{i\left[-\frac{V(x_j)k}{2} + \frac{\beta^{1-q}|\psi_j^{**}|^{2\sigma-2q}}{2\delta(q-\sigma)}\left(-1+\left(1+\delta qk\beta^q|\psi_j^{**}|^{2q}\right)^{\frac{q-\sigma}{q}}\right)\right]\right\}}{\left(1+q\delta k\beta^q|\psi_j^{**}|^{2q}\right)^{1/2q}}\ \psi_j^{**}, & \sigma\neq q. \end{cases}$$

Case III. *Focusing cubic NLS with a damping term that accounts for two-body and three-body losses in a BEC* [35]. We choose $\sigma = 1$, $g(\rho) = \delta_1\beta\rho + \delta_2\beta^2\rho^2$ with

$\delta_1,\ \delta_2 > 0$, in (1.5). Using (2.7), we get

$$f(s) = \begin{cases} -\frac{1}{\delta_1\beta s} + \frac{\delta_2}{\delta_1^2}\ln\left(\delta_2\beta + \delta_1/s\right), & s > 0, \\ 0, & s = 0. \end{cases} \tag{2.26}$$

Substituting (2.7) into (2.12) and changing the variable of integration, we obtain

$$\begin{aligned} F(s,r) &= \int_0^r g\left(f^{-1}(f(s)-2\tau)\right)\,d\tau \overset{\tau=(f(s)-f(h))/2}{=} \int_s^{h(s,r)} -\frac{1}{2}g(h)f'(h)\,dh \\ &= \int_s^{h(s,r)} -\frac{1}{2h}\,dh = \begin{cases} -\frac{1}{2}\ln\left(h(s,r)/s\right), & s > 0, \\ 0, & s = 0, \end{cases} \end{aligned} \tag{2.27}$$

where $h(s,r)$ is the solution of

$$f(s) - f(h(s,r)) = 2r \qquad \text{for any } r > 0, \tag{2.28}$$

with f given in (2.26). Similarly we find

$$G(s,r) = \int_s^{h(s,r)} -\frac{\beta}{2g(h)}\,dh = \begin{cases} -\frac{1}{2\delta_1}\ln\frac{h(s,r)(\delta_1+\delta_2\beta s)}{s(\delta_1+\delta_2\beta h(s,r))}, & s > 0, \\ 0, & s = 0. \end{cases} \tag{2.29}$$

Substituting (2.27) and (2.29) into (2.13), we get the following second-order TSSP steps for the NLS with a combination of cubic and quintic damping terms:

(2.30)

$$\begin{aligned} \psi_j^* &= \begin{cases} \frac{\sqrt{h(|\psi_j^n|^2,k/2)}}{|\psi_j^n|}\exp\left\{i\left[-\frac{V(x_j)k}{2} - \frac{1}{2\delta_1}\ln\frac{h(|\psi_j^n|^2,k/2)(\delta_1+\delta_2\beta|\psi_j^n|^2)}{|\psi_j^n|^2(\delta_1+\delta_2\beta h(|\psi_j^n|^2,k/2))}\right]\right\}\psi_j^n, & \psi_j^n \neq 0, \\ 0, & \psi_j^n = 0, \end{cases} \\ \psi_j^{**} &= \sum_{l=1}^{M-1} e^{-ik\mu_l^2/2}\,\widehat{\psi}_l^*\ \sin(\mu_l(x_j-a)), \qquad j = 1,2,\dots,M-1, \\ \psi_j^{n+1} &= \begin{cases} \frac{\sqrt{h(|\psi_j^{**}|^2,k/2)}}{|\psi_j^{**}|}\exp\left\{i\left[-\frac{V(x_j)k}{2} - \frac{1}{2\delta_1}\ln\frac{h(|\psi_j^{**}|^2,k/2)(\delta_1+\delta_2\beta|\psi_j^{**}|^2)}{|\psi_j^{**}|^2(\delta_1+\delta_2\beta h(|\psi_j^{**}|^2,k/2))}\right]\right\}\psi_j^{**}, & \psi_j^{**} \neq 0, \\ 0, & \psi_j^{**} = 0. \end{cases} \end{aligned}$$

Remark 2.1. As demonstrated in this subsection, the integrals in (2.7) and (2.12) can be evaluated *analytically* for the damping terms which most frequently appear in physical applications. If the integrals in (2.7) or (2.12) cannot be evaluated analytically or the inverse of f in (2.7) cannot be expressed explicitly, e.g., if $g(\rho)$ in (1.5) is not a polynomial, one can solve the following ODE numerically by either the second- or fourth-order Runge–Kutta method

$$\begin{aligned} &\frac{dh(t)}{dt} = -2g(h(t))\,h(t), \qquad 0 \le t \le k/2, \\ &h(0) = s, \end{aligned}$$

to get $h(s,k/2)$ for any given $s > 0$ and set $h(s,k/2) = 0$ for $s = 0$. By changing the variable of integration in (2.12) (see detail in (2.27) and (2.29)), the first integral in (2.12), i.e., $F(s,k/2)$, can be evaluated exactly (see detail in (2.27)), and the second integral in (2.12), i.e., $G(s,k/2)=\int_s^{h(s,k/2)} -\frac{\beta h^{\sigma-1}}{2g(h)}\,dh$, can be evaluated numerically by using a numerical quadrature, e.g., the trapezoidal rule or Simpson's rule.

The TSSP scheme is explicit and is unconditionally stable as we will demonstrate in the next subsection. Another main advantage of the time-splitting method is its time transversal invariance, which also holds for the NLS and the DNLS themselves. If a constant α is added to the potential V, then the discrete wave functions $\psi_j^{\varepsilon,n+1}$ obtained from the TSSP method get multiplied by the phase factor $e^{-i\alpha(n+1)k}$, which leaves the discrete normalization unchanged. This property does not hold for finite difference schemes.

Remark 2.2. For the focusing cubic NLS with a quintic damping term describing three-body recombination loss and an additional feeding term for the BEC [25], we choose $\sigma=1$, $g(\rho)=-\delta_1+\delta_2\beta^2\rho^2$ with $\delta_1,\delta_2>0$ in (1.5). The idea of constructing the TSSP method is also applicable to this case, although we could not prove that it is unconditionally stable due to the feeding term. Inserting the above feeding term into (2.7), we get

$$f(s)=\begin{cases}\frac{1}{2\delta_1}\ln\left|\delta_2\beta^2-\delta_1/s^2\right|, & s>0,\\ 0, & s=0.\end{cases} \tag{2.31}$$

Inserting (2.31) into (2.9), we find

$$h(s,\tau)=\frac{s\sqrt{\delta_1}}{\sqrt{\delta_1 e^{-4\tau\delta_1}+(1-e^{-4\tau\delta_1})\delta_2\beta^2 s^2}}, \tag{2.32}$$

and substituting (2.32) into (2.9) and (2.12), we obtain

$$\rho(t)=\frac{|\psi(x,t_n)|^2\sqrt{\delta_1}}{\sqrt{\delta_1 e^{-4\tau\delta_1}+(1-e^{-4\tau\delta_1})\delta_2\beta^2|\psi(x,t_n)|^4}},\ t_n\le t\le t_{n+1}, \tag{2.33}$$

$$F(s,r)=-\delta_1 r+\frac{1}{4}\ln\left[1+\delta_2\beta^2 s^2(e^{4\delta_1 r}-1)/\delta_1\right], \tag{2.34}$$

$$G(s,r)=\frac{1}{2\sqrt{\delta_1\delta_2}}\ln\frac{\beta s\sqrt{\delta_2}e^{2r\delta_1}+\sqrt{\delta_1+\delta_2\beta^2 s^2\left(e^{4r\delta_1}-1\right)}}{\sqrt{\delta_1}+\beta s\sqrt{\delta_2}}. \tag{2.35}$$

Inserting (2.34) and (2.35) into (2.13), we get the following second-order TSSP steps for the NLS with a quintic damping term and a feeding term:

(2.36)

$$\psi_j^*=\frac{e^{k\delta_1/2}\exp\left[i\left(-\frac{V(x_j)k}{2}+\frac{1}{2\sqrt{\delta_1\delta_2}}\ln\frac{\beta|\psi_j^n|^2\sqrt{\delta_2}e^{k\delta_1}+\sqrt{\delta_1+\delta_2\beta^2|\psi_j^n|^4\left(e^{2k\delta_1}-1\right)}}{\sqrt{\delta_1}+\beta|\psi_j^n|^2\sqrt{\delta_2}}\right)\right]}{\left[1+\delta_2\beta^2|\psi_j^n|^4(e^{2k\delta_1}-1)/\delta_1\right]^{1/4}}\,\psi_j^n,$$

$$\psi_j^{**}=\sum_{l=1}^{M-1} e^{-ik\mu_l^2/2}\,\widehat{\psi}_l^*\,\sin(\mu_l(x_j-a)),\qquad j=1,2,\ldots,M-1,$$

$$\psi_j^{n+1}=\frac{e^{k\delta_1/2}\exp\left[i\left(-\frac{V(x_j)k}{2}+\frac{1}{2\sqrt{\delta_1\delta_2}}\ln\frac{\beta|\psi_j^{**}|^2\sqrt{\delta_2}e^{k\delta_1}+\sqrt{\delta_1+\delta_2\beta^2|\psi_j^{**}|^4\left(e^{2k\delta_1}-1\right)}}{\sqrt{\delta_1}+\beta|\psi_j^{**}|^2\sqrt{\delta_2}}\right)\right]}{\left[1+\delta_2\beta^2|\psi_j^{**}|^4(e^{2k\delta_1}-1)/\delta_1\right]^{1/4}}\,\psi_j^{**}.$$

Remark 2.3. The TSSP scheme (2.13) can easily be extended for solving the complex Ginzburg–Landau (CGL) equation [15, 28]

$$i\,\psi_t=-(1-i\,\varepsilon)\,\Delta\psi-|\psi|^2\psi-i\left(\delta_2|\psi|^2-\delta_1\right)\psi, \tag{2.37}$$

where ε, δ_1, and δ_2 are positive constants. The idea of constructing the TSSP method for the DNLS is also applicable to the CGL equation provided that we solve

$$i\,\psi_t = -\,(1 - i\,\varepsilon)\,\Delta\psi \tag{2.38}$$

in the first step instead of (2.3). Inserting $\sigma = 1$, $\beta = 1$, and $g(\rho) = \delta_2\rho - \delta_1$ with δ_1, $\delta_2 > 0$ into (1.5) and using (2.7), we get

$$f(s) = \begin{cases} \frac{1}{\delta_1}\ln|\delta_2 - \delta_1/s|\,, & s > 0, \\ 0, & s = 0. \end{cases} \tag{2.39}$$

Inserting (2.39) into (2.7), we find

$$h(s,\tau) = \frac{s\delta_1}{s\delta_2\left(1 - e^{-2\tau\delta_1}\right) + \delta_1 e^{-2\tau\delta_1}}, \tag{2.40}$$

and substituting (2.40) into (2.9) and (2.12), we obtain

$$\rho(t) = \frac{\delta_1\,|\psi(x,t_n)|^2}{\delta_2\,|\psi(x,t_n)|^2\left(1 - e^{-2\tau\delta_1}\right) + \delta_1 e^{-2\tau\delta_1}}, \quad t_n \le t \le t_{n+1}, \tag{2.41}$$

$$F(s,r) = -\frac{1}{2}\ln\frac{\delta_1}{s\delta_2 + (\delta_1 - s\delta_2)\,e^{-2r\delta_1}}, \tag{2.42}$$

$$G(s,r) = \frac{1}{2\delta_2}\ln\frac{\delta_1 - s\delta_2 + s\delta_2 e^{2r\delta_1}}{\delta_1}. \tag{2.43}$$

Inserting (2.42) and (2.43) into (2.13), we get the following second-order TSSP steps for the CGL equation (2.37):

(2.44)

$$\psi_j^* = \sqrt{\frac{\delta_1}{\delta_2|\psi_j^n|^2 + \left(\delta_1 - \delta_2|\psi_j^n|^2\right)e^{-k\delta_1}}}\exp\left[\frac{i}{2\delta_2}\ln\frac{\delta_1 - \delta_2|\psi_j^n|^2 + \delta_2|\psi_j^n|^2 e^{k\delta_1}}{\delta_1}\right]\,\psi_j^n,$$

$$\psi_j^{**} = \sum_{l=1}^{M-1} e^{-(\varepsilon+i)k\mu_l^2}\,\widehat{\psi}_l^*\,\sin(\mu_l(x_j - a)), \qquad j = 1, 2, \ldots, M-1,$$

$$\psi_j^{n+1} = \sqrt{\frac{\delta_1}{\delta_2|\psi_j^{**}|^2 + \left(\delta_1 - \delta_2|\psi_j^{**}|^2\right)e^{-k\delta_1}}}\exp\left[\frac{i}{2\delta_2}\ln\frac{\delta_1 - \delta_2|\psi_j^{**}|^2 + \delta_2|\psi_j^{**}|^2 e^{k\delta_1}}{\delta_1}\right]\,\psi_j^{**}.$$

Remark 2.4. If the homogeneous periodic boundary conditions in (2.2) are replaced by the periodic boundary conditions

$$\psi(a,t) = \psi(b,t), \qquad \psi_x(a,t) = \psi_x(b,t), \qquad t \ge 0, \tag{2.45}$$

the TSSP scheme (2.13) still works provided that one replaces the sine-series in (2.13) by a Fourier series [4, 5, 7].

2.3. Stability and decay rate. Let $U = (U_0, U_1, \ldots, U_M)^T$ with $U_0 = U_M = \mathbf{0}$ and $\|\cdot\|_{l^2}$ be the usual discrete l^2-norm on the interval (a,b), i.e.,

$$\|U\|_{l^2} = \sqrt{\frac{b-a}{M}\sum_{j=1}^{M-1}|U_j|^2}. \tag{2.46}$$

For the *stability* of the TSSP approximations (2.13), we have the following lemma, which shows that the total normalization does not increase.

LEMMA 2.1. *The TSSP schemes* (2.13) *are unconditionally stable if* $g(s) \geq 0$ *for* $s \geq 0$. *In fact, for every mesh size* $h > 0$ *and time step* $k > 0$,

$$\|\psi^{n+1}\|_{l^2} \leq \|\psi^n\|_{l^2} \leq \|\psi^0\|_{l^2} = \|\psi_0\|_{l^2}, \qquad n = 0, 1, 2, \ldots. \tag{2.47}$$

Furthermore, when a linear damping term is used in (1.5), *i.e., when we choose* $g(\rho) \equiv \delta$ *with* $\delta > 0$, *the decay rate of the normalization satisfies*

$$\|\psi^n\|_{l^2} = e^{-2\delta t_n}\|\psi^0\|_{l^2} = e^{-2\delta t_n}\|\psi_0\|_{l^2}, \qquad n = 1, 2, \ldots. \tag{2.48}$$

In fact, (2.48) *is a discretized version of the decay rate of the normalization* $N(t)$ *in* (1.8).

Proof. We combine (2.13), (2.14), and (2.46) and note that $F(s,\tau) \geq 0$ for $s \geq 0$ and $\tau \geq 0$ to obtain

$$\begin{aligned}
\frac{1}{b-a}\|\psi^{n+1}\|_{l^2}^2 &= \frac{1}{M}\sum_{j=1}^{M-1}\left|\psi_j^{n+1}\right|^2 \\
&= \frac{1}{M}\sum_{j=1}^{M-1}\exp\left[-2F\left(|\psi_j^{**}|^2, k/2\right)\right]\ \left|\psi_j^{**}\right|^2 \leq \frac{1}{M}\sum_{j=1}^{M-1}\left|\psi_j^{**}\right|^2 \\
&= \frac{1}{M}\sum_{j=1}^{M-1}\left|\sum_{l=1}^{M-1} e^{-ik\mu_l^2/2}\ \widehat{\psi}_l^*\ \sin(\mu_l(x_j-a))\right|^2 = \frac{1}{2}\sum_{l=1}^{M-1}\left|e^{-ik\mu_l^2/2}\ \hat{\psi}_l^*\right|^2 = \frac{1}{2}\sum_{l=1}^{M-1}\left|\hat{\psi}_l^*\right|^2 \\
&= \frac{1}{2}\sum_{l=1}^{M-1}\left|\frac{2}{M}\sum_{j=1}^{M-1}\psi_j^*\ \sin(\mu_l(x_j-a))\right|^2 = \frac{1}{M}\sum_{j=1}^{M-1}\left|\psi_j^*\right|^2 \\
&= \frac{1}{M}\sum_{j=1}^{M-1}\exp\left[-2F\left(|\psi_j^n|^2, k/2\right)\right]\left|\psi_j^n\right|^2 \leq \frac{1}{M}\sum_{j=1}^{M-1}\left|\psi_j^n\right|^2 \\
&= \frac{1}{b-a}\|\psi^n\|_{l^2}^2.
\end{aligned} \tag{2.49}$$

Here, we used the identity

$$\sum_{j=1}^{M-1}\sin\left(\frac{\pi r\, j}{M}\right)\ \sin\left(\frac{\pi s\, j}{M}\right) = \left\{\begin{array}{ll} 0, & r-s \neq 2mM, \\ M/2, & r-s = 2mM, r \neq 2nM, \end{array}\right. \qquad m, n \text{ integer.} \tag{2.50}$$

When a linear damping term is added to the NLS (1.5), the equality (2.48) follows from the above proof, (2.18), and

$$\sum_{j=1}^{M-1}\exp\left[-2F\left(|\psi_j^n|^2, k/2\right)\right]\left|\psi_j^n\right|^2 = \sum_{j=1}^{M-1}e^{-\delta k}\left|\psi_j^n\right|^2 = e^{-\delta k}\sum_{j=1}^{M-1}\left|\psi_j^n\right|^2.$$

3. Numerical examples. In this section, we present numerical tests of the TSSP scheme (2.13) for solving a focusing cubic NLS appearing in nonlinear optics [16, 36] and for the GPE in BEC [7] in two dimensions with a linear, a cubic, or a quintic

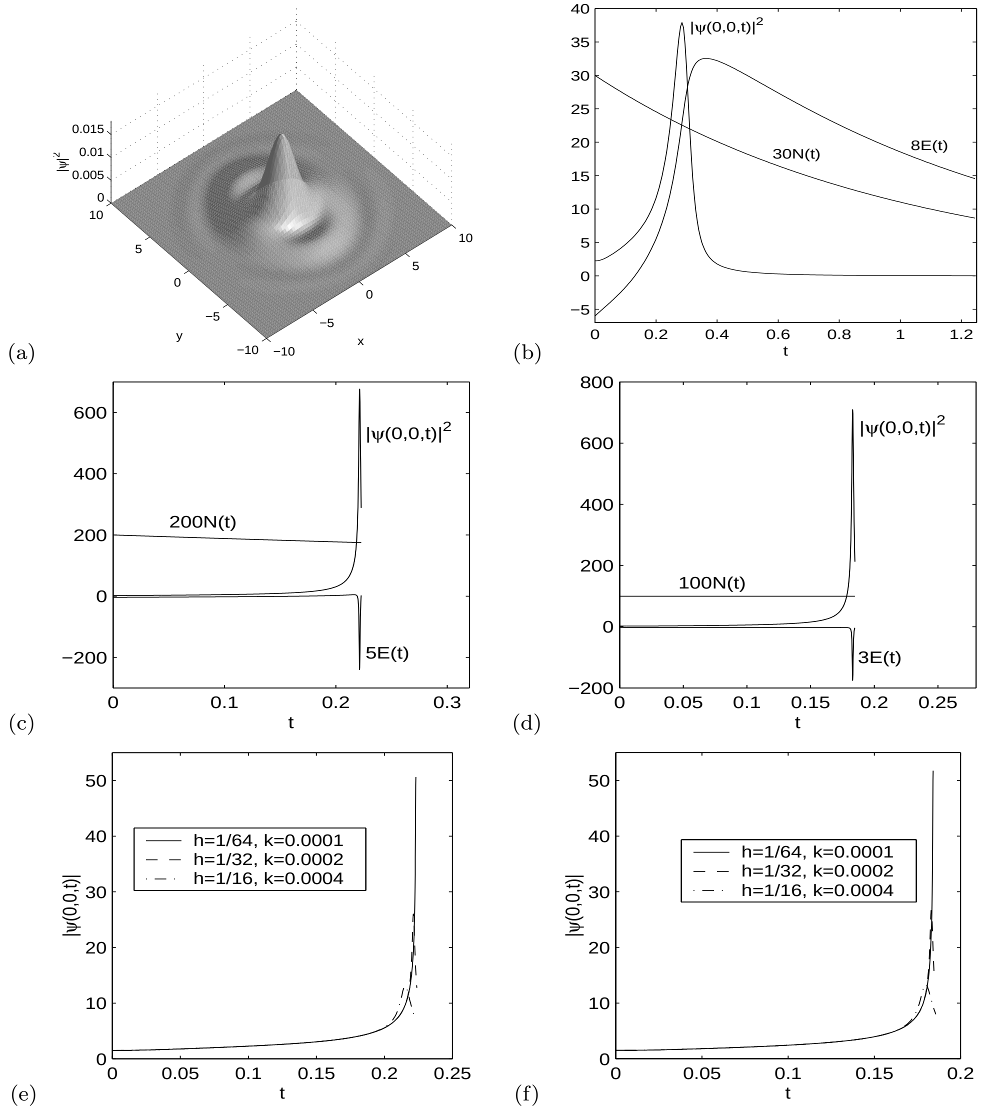


FIG. 1. *Numerical results in Example* 1, *case* I. (a) *Surface plot of the density* $|\psi|^2$ *at time* $t = 1.25$ *with* $\delta = 0.5$. *Normalization, energy, and central density* $|\psi(0,0,t)|^2$ *as functions of time:* (b) *with* $\delta = 0.5$, (c) $\delta = 0.3$, (d) $\delta = 0$ *(no damping). Blowup study:* (e) $\delta = 0.3$, (f) $\delta = 0$ *(no damping).*

damping term. In our computations, the initial condition (1.2) is always chosen such that $|\psi_0(\mathbf{x})|$ decays to zero sufficiently fast as $|\mathbf{x}| \to \infty$. We choose an appropriately large rectangle $[a, b] \times [c, d]$ in two dimensions to prevent the homogeneous periodic boundary condition (2.2) from introducing a significant (aliasing) error relative to the whole space problem. To quantify the numerical results of the GPE for a BEC, we define the condensate widths along the x, y, and z axes by

$$\sigma_\alpha^2 = \langle \alpha^2 \rangle = \frac{1}{N(t)} \int_{\mathbb{R}^d} \alpha^2 |\psi(\mathbf{x}, t)|^2 \, d\mathbf{x}, \qquad \text{with} \quad \alpha = x,\ y,\ \text{or } z.$$

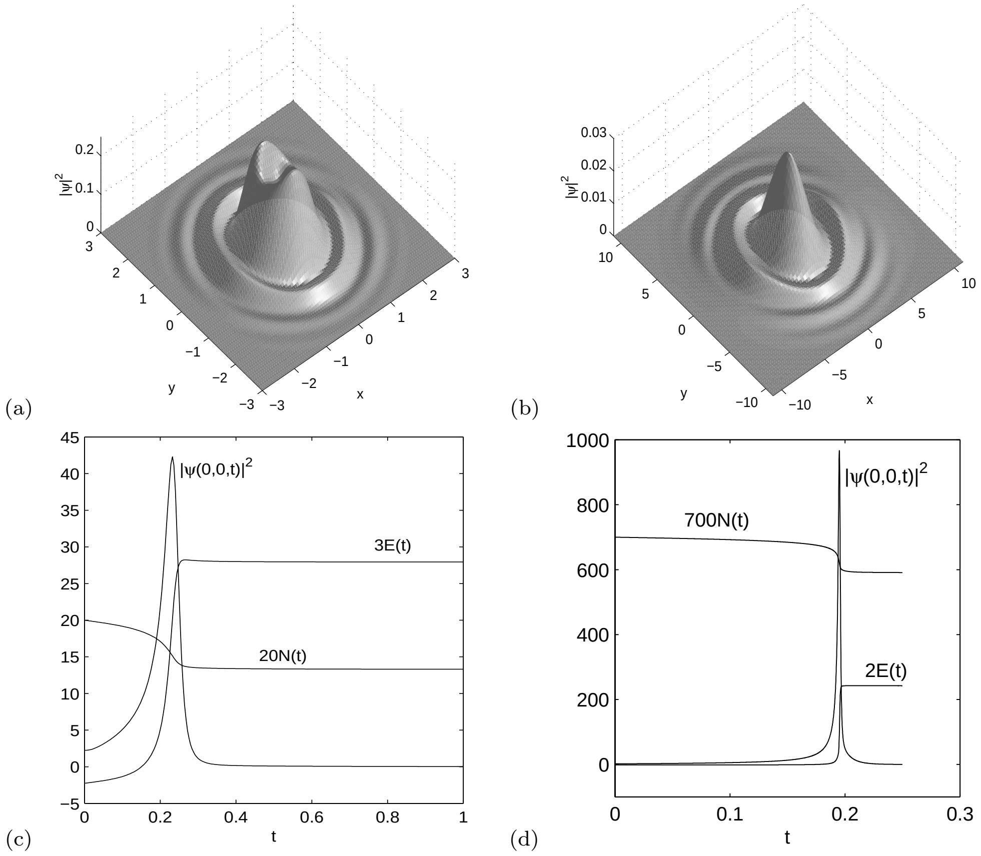


FIG. 2. *Numerical results in Example* 1*, case* II*. Surface plot of the density* $|\psi|^2$ *with* $\delta = 0.02$*:* (a) *At time* $t = 0.4$*,* (b) $t = 1.0$*. Normalization, energy, and central density* $|\psi(0,0,t)|^2$ *as functions of time:* (c) *with* $\delta = 0.02$*,* (d) $\delta = 0.005$ *(with* $h = 1/128$*,* $k = 0.00002$*).*

Example 1. *Solution of the two-dimensional damped focusing cubic NLS*. We choose $d = 2$, $\sigma = 1$, and $V(x, y) \equiv 0$ in (1.5) and present computations for three different damping terms ($\delta > 0$):

I. A linear damping term; i.e., we choose $g(\rho) \equiv \delta$.

II. A cubic damping term; i.e., we choose $g(\rho) \equiv \delta\beta\rho$.

III. A quintic damping term; i.e., we choose $g(\rho) \equiv \delta\beta^2\rho^2$.

The initial condition (1.6) is taken to be

$$(3.1) \qquad \psi(x, y, 0) = \psi_0(x, y) = \frac{\gamma_y^{1/4}}{\sqrt{\pi\varepsilon}}\, e^{-(x^2+\gamma_y y^2)/2\varepsilon}, \qquad (x, y) \in \mathbb{R}^2.$$

We assume $\gamma_y = 2$, $\varepsilon = 0.2$, and $\beta = 8$ in (1.5) such that $E(0) = -0.751582 < 0$ in (1.4). We solve the NLS on the square $[-16, 16]^2$; i.e., $a = c = -16$ and $b = d = 16$ with mesh size $h = \frac{1}{32}$, time step $k = 0.0002$, and homogeneous periodic boundary conditions along the boundary of the square. We compare the effect of changing the damping parameter δ in the three different cases I, II, and III.

Figure 1 shows the surface plot of the density $|\psi(x, y, t)|^2$ at time $t = 1.25$ with $\delta = 0.5$; plots of the normalization, energy, and central density $|\psi(0, 0, t)|^2$ are shown

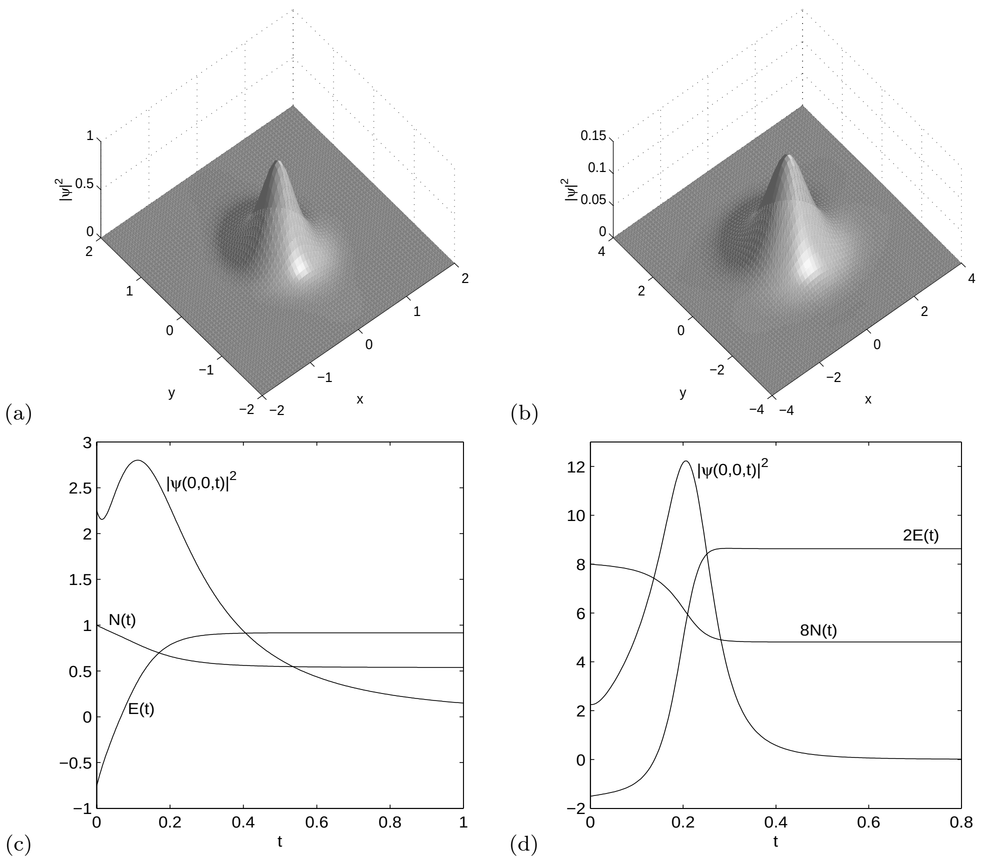


Fig. 3. *Numerical results in Example* 1, *case* III. *Surface plot of the density $|\psi|^2$ with $\delta = 0.01$:* (a) *At time $t = 0.4$,* (b) *$t = 1.0$. Normalization, energy, and central density $|\psi(0, 0, t)|^2$ as functions of time:* (c) *with $\delta = 0.01$,* (d) *$\delta = 0.001$.*

as functions of time with $\delta = 0.5,\ 0.3$, and $\delta = 0$ (no damping) for case I. Figure 2 shows similar results for case II and Figure 3 for case III. Furthermore, Figure 4 shows contour plots of the density $|\psi|^2$ at different times for case III with $\delta = 0.01$.

In the numerical computations, a blowup is detected either from the plot of the central density $|\psi(0, 0, t)|^2$, which at the blowup shows a very sharp spike with a peak value that increases when the mesh size h decreases, or from the plot of the energy $E(t)$, which has a very sharp spike with negative values at the blowup. In fact, the TSSP method (2.13) aims to capture the solution of the DNLS without blowup, i.e., physical relevant solution. If one wants to capture the blowup rate of the NLS, we refer to [27, 31].

From the numerical results we find the following conditions for arresting a blowup of the wave function with initial energy $E(0) < 0$. (1) For linear damping, the blowup is arrested if the damping parameter is bigger than a certain threshold value which we find to be $\delta_{\text{th}} \approx 0.461$ by numerical experiments. As shown in Figure 1(b), blowup is arrested for $\delta = 0.5 > \delta_{\text{th}}$, while the wave function blows up for $\delta < \delta_{\text{th}}$, as can be seen from Figure 1(c),(d), where we have chosen $\delta = 0.3 < \delta_{\text{th}}$ and $\delta = 0 < \delta_{\text{th}}$, respectively. The time at which the blowup of the wave function happens, however, increases with increasing δ (cf. Figure 1(c)(d)). (2) For a cubic damping term with

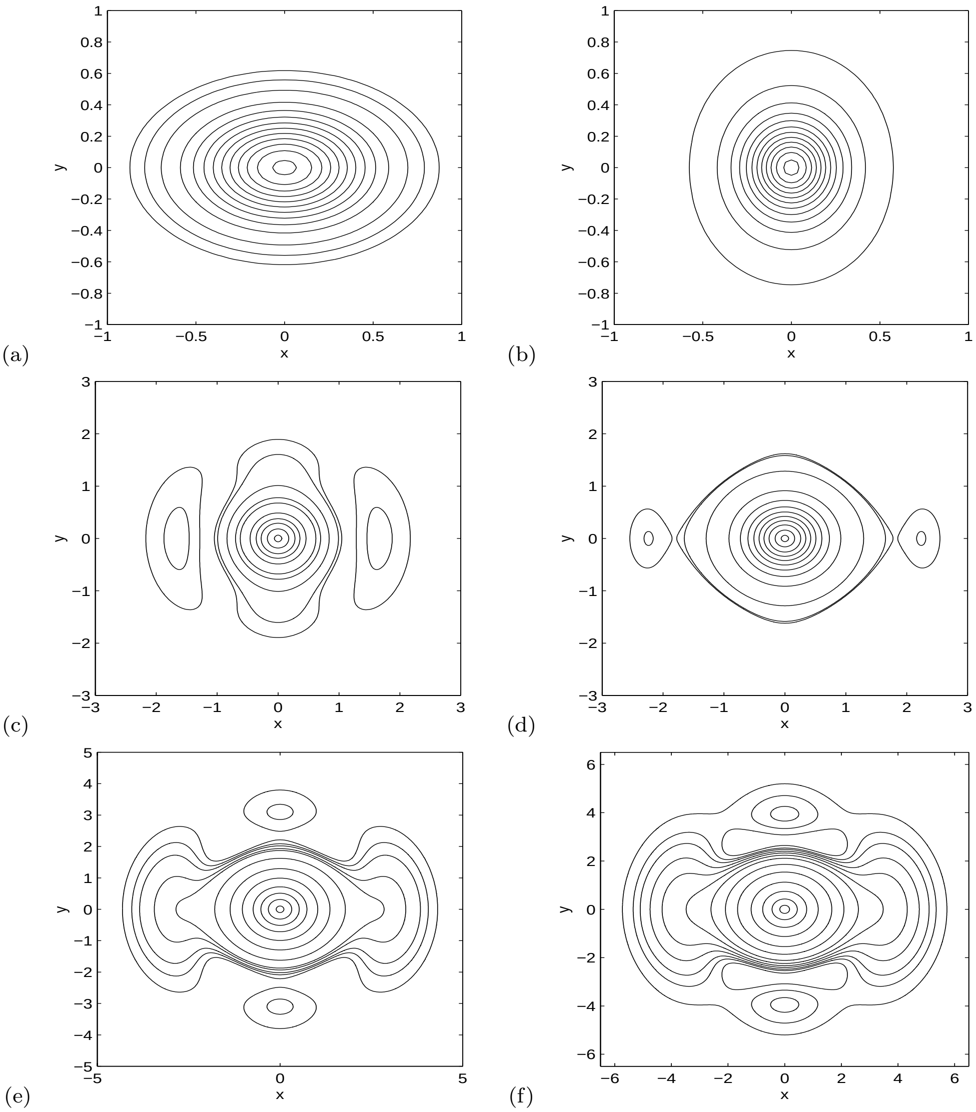


FIG. 4. *Contour plots of the density* $|\psi|^2$ *at different times in Example* 1, *case* III, *with* $\delta = 0.01$. (a) $t = 0$, (b) $t = 0.2$, (c) $t = 0.4$, (d) $t = 0.6$, (e) $t = 0.8$, (f) $t = 1$.

$\delta > 0$, the blowup of the wave function is always arrested (cf. Figure 2). (3) The above observation (2) also holds for a quintic damping term (cf. Figure 3).

For linear damping, we also test the dependence of the threshold value of the damping parameter δ_{th} on β and the initial data. First we take $\gamma_y = 2$ and $\varepsilon = 0.2$ in (3.1). Table 1 shows the threshold values δ_{th} for different β in (1.5), and $E(0)$ represents the initial energy. Then we choose $\beta = 16$ in (1.5) and $\gamma_y = 2$ in (3.1). Table 2 displays the threshold values δ_{th} for different values of ε in (3.1).

From Table 1, we find by a least square fitting

$$\delta_{\text{th}} = -0.6930E(0) \qquad \text{or} \qquad \delta_{\text{th}} = 0.3872\beta - 2.4627.$$

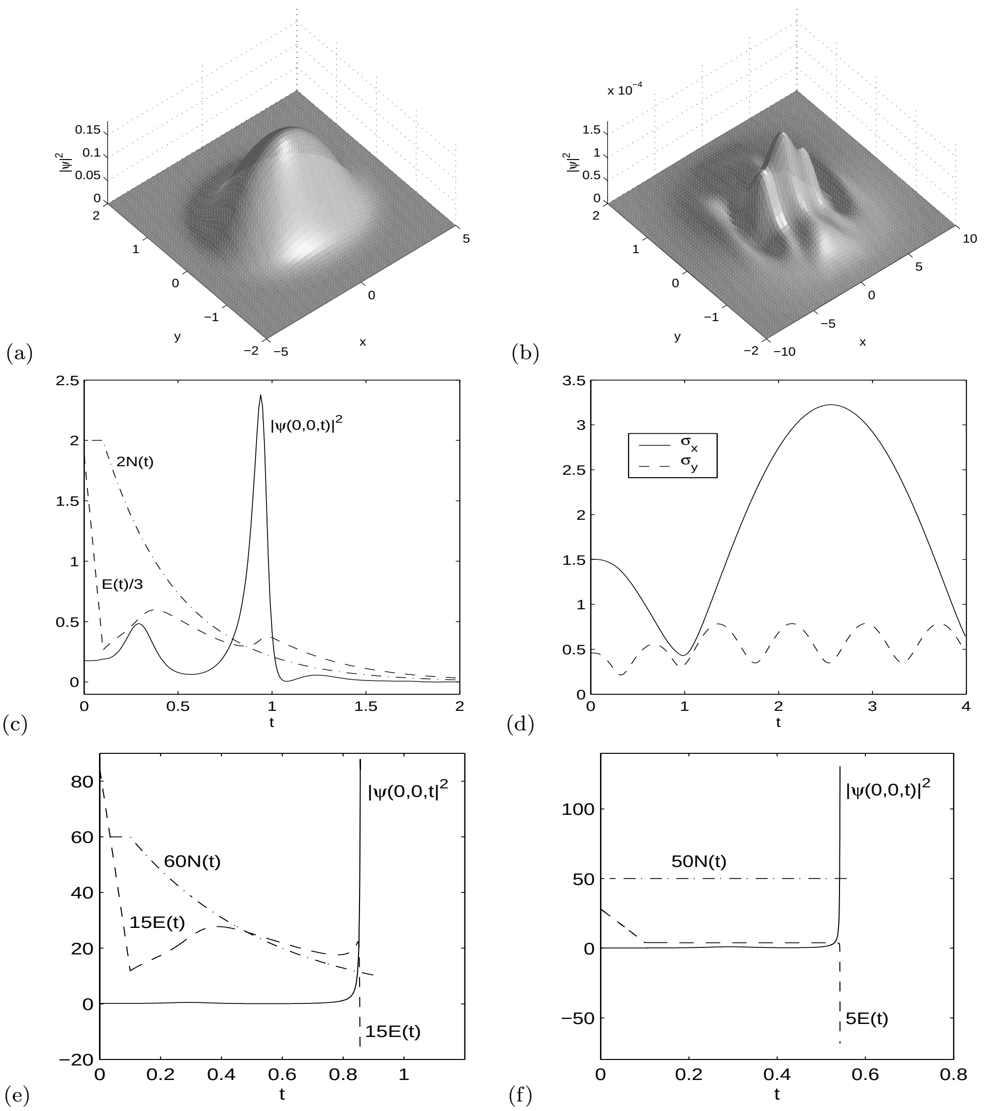


Fig. 5. *Numerical results in Example* 2, *case* I. *Surface plot of the density* $|\psi|^2$ *with* $\delta = 1.25$*:* (a) *At time* $t = 0$ *(ground-state solution),* (b) $t = 2.8$*. Normalization, energy, and central density* $|\psi(0,0,t)|^2$ *as functions of time:* (c) *with* $\delta = 1.25$*,* (e) $\delta = 1.1$*,* (f) $\delta = 0$ *(no damping).* (d) *Condensate widths with* $\delta = 1.25$*.*

Table 1

Dependence of $\delta_{\rm th}$ *on* β *for* $\gamma_y = 2$ *and* $\varepsilon = 0.2$ *in* (3.1).

	$\beta = 8$	$\beta = 16$	$\beta = 32$	$\beta = 64$	$\beta = 128$
$E(0)$	-0.7516	-5.253	-14.256	-32.263	-68.275
$\delta_{\rm th}$	0.461	3.655	10.35	22.15	40.05

Table 2

Dependence of $\delta_{\rm th}$ *on* ε *in* (3.1) *for* $\beta = 16$ *in* (1.5) *and* $\gamma_y = 2$ *in* (3.1).

	$\varepsilon = 0.8$	$\varepsilon = 0.4$	$\varepsilon = 0.2$	$\varepsilon = 0.1$	$\varepsilon = 0.05$
$E(0)$	-1.3133	-2.6266	-5.2532	-10.506	-21.013
$\delta_{\rm th}$	0.895	1.845	3.655	7.25	14.55

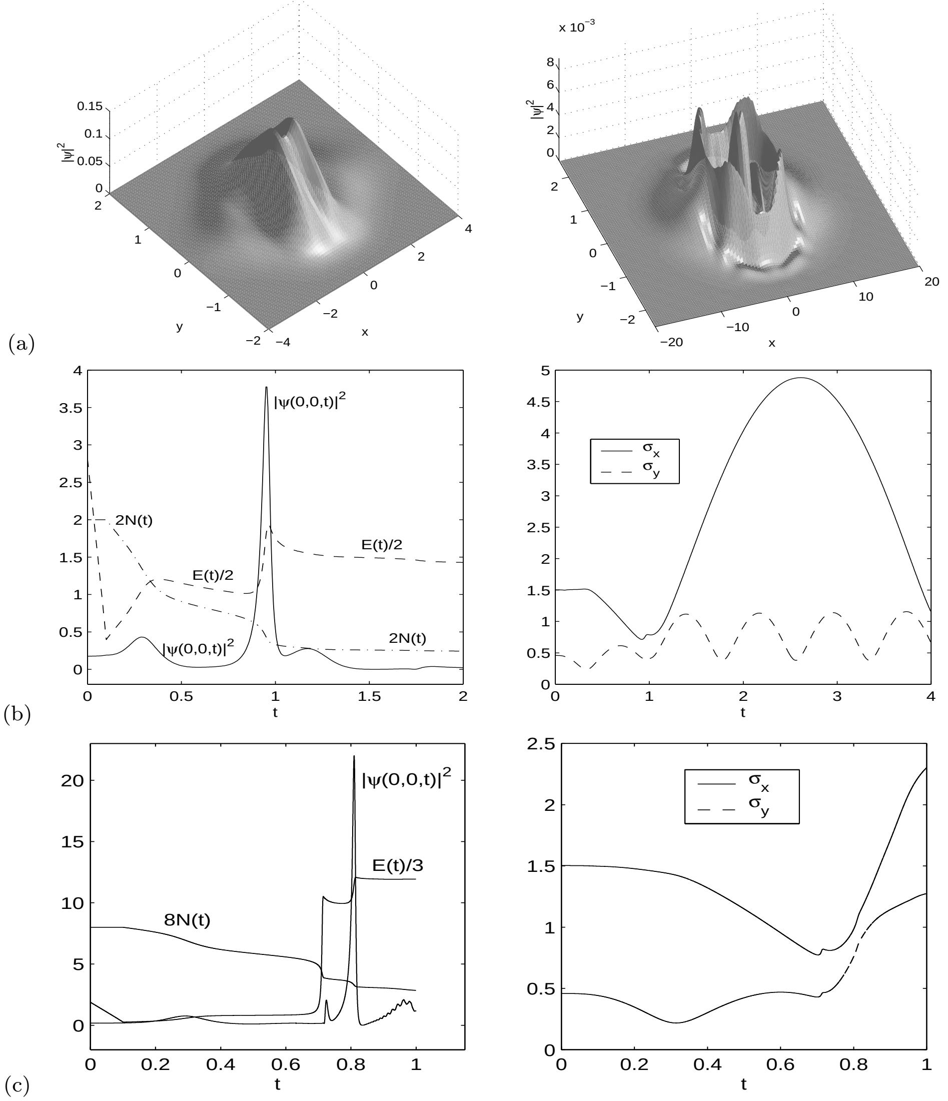


FIG. 6. *Numerical results in Example* 2*, case* II. (a) *Surface plot of the density* $|\psi|^2$ *with* $\delta = 0.15$*: At time* $t = 0.8$ *(left column) and* $t = 2.4$ *(right column). Normalization, energy, and central density* $|\psi(0,0,t)|^2$ *(left column) and condensate widths (right column) as functions of time:* (b) *With* $\delta = 0.15$*;* (c) $\delta = 0.04$ *(under* $h = 1/128, k = 0.00002$ *for* (c)*).*

Similarly, from Table 2, we obtain

$$\delta_{\mathrm{th}} = -0.6922E(0).$$

Based on this observation, we conclude that the threshold value of the linear damping parameter δ_{th} depends linearly on the initial energy $E(0)$.

Example 2. *Solution of the two-dimensional damped GPE with focusing nonlinearity.* We choose $d = 2$, $\sigma = 1$, and $V(x,y) = \frac{1}{2}(\gamma_x^2 x^2 + \gamma_y^2 y^2)$ to be a harmonic oscillator potential with $\gamma_x, \gamma_y > 0$ in (1.5). Again, we present computations for the same three different damping terms in (1.5) that we studied in Example 1.

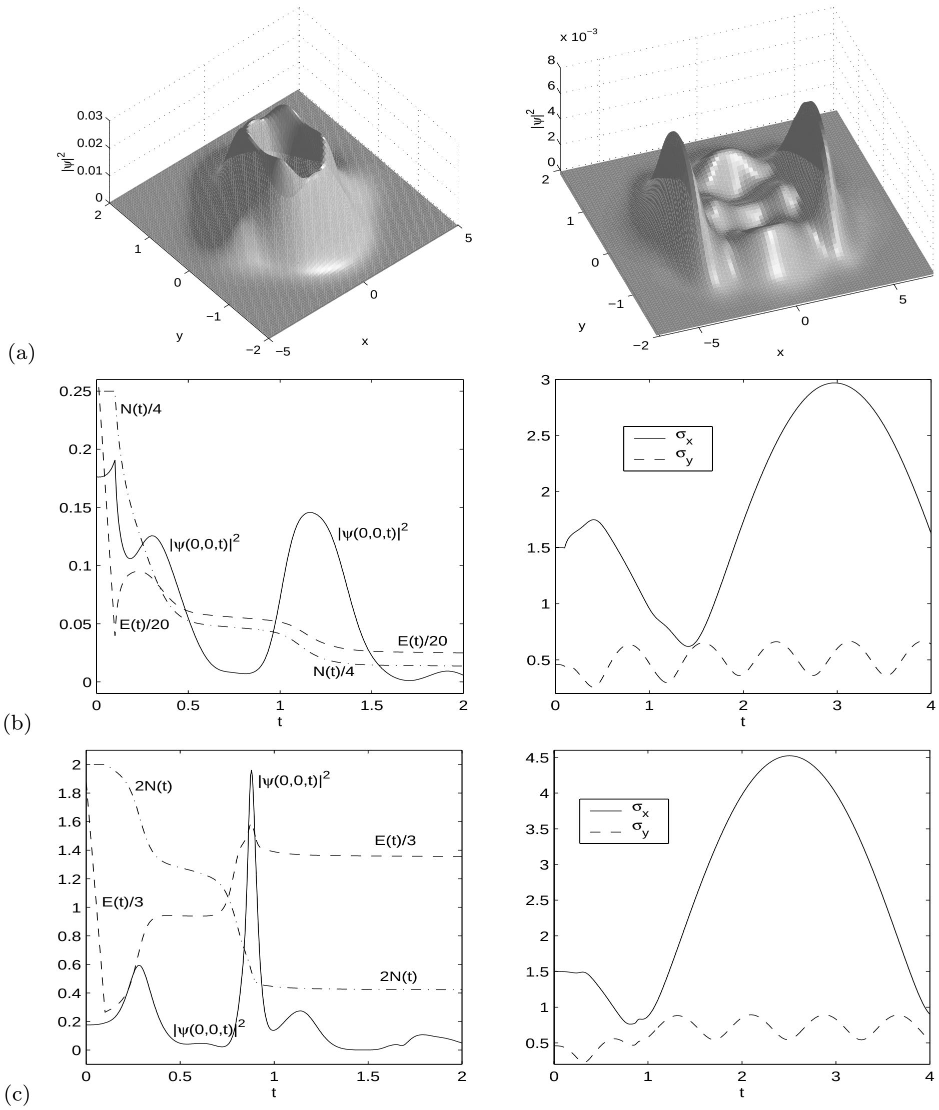


FIG. 7. *Numerical results in Example* 2*, case* III. (a) *Surface plot of the density* $|\psi|^2$ *with* $\delta = 0.15$*: At time* $t = 0.8$ *(left column) and* $t = 3.2$ *(right column). Normalization, energy, and central density* $|\psi(0,0,t)|^2$ *(left column) and condensate widths (right column) as functions of time:* (b) *With* $\delta = 0.15$*;* (c) $\delta = 0.005$.

We take $\gamma_x = 1$ and $\gamma_y = 4$. The initial condition (1.6) is assumed to be the ground-state solution of (1.5) with $g(\rho) \equiv 0$ (i.e., undamped case) and $\beta = -40$ [6]. The cubic nonlinearity is ramped linearly from $\beta = -40$ (defocusing) to $\beta = 50$ (focusing) during the time interval $[0, 0.1]$ and afterward kept constant. The absorption parameter was set to $\delta = 0$ during the time interval $[0, 0.1]$ and increased to a positive value $\delta > 0$ afterward.

We solve the GPE on the rectangle $[-24, 24] \times [-6, 6]$, i.e., for $a = -24$, $b = 24$, $c = -6$, and $d = 6$ with mesh size $h_x = \frac{3}{64}$, $h_y = \frac{3}{128}$, time step $k = 0.0005$, and homogeneous periodic boundary conditions along the boundary of the rectangle.

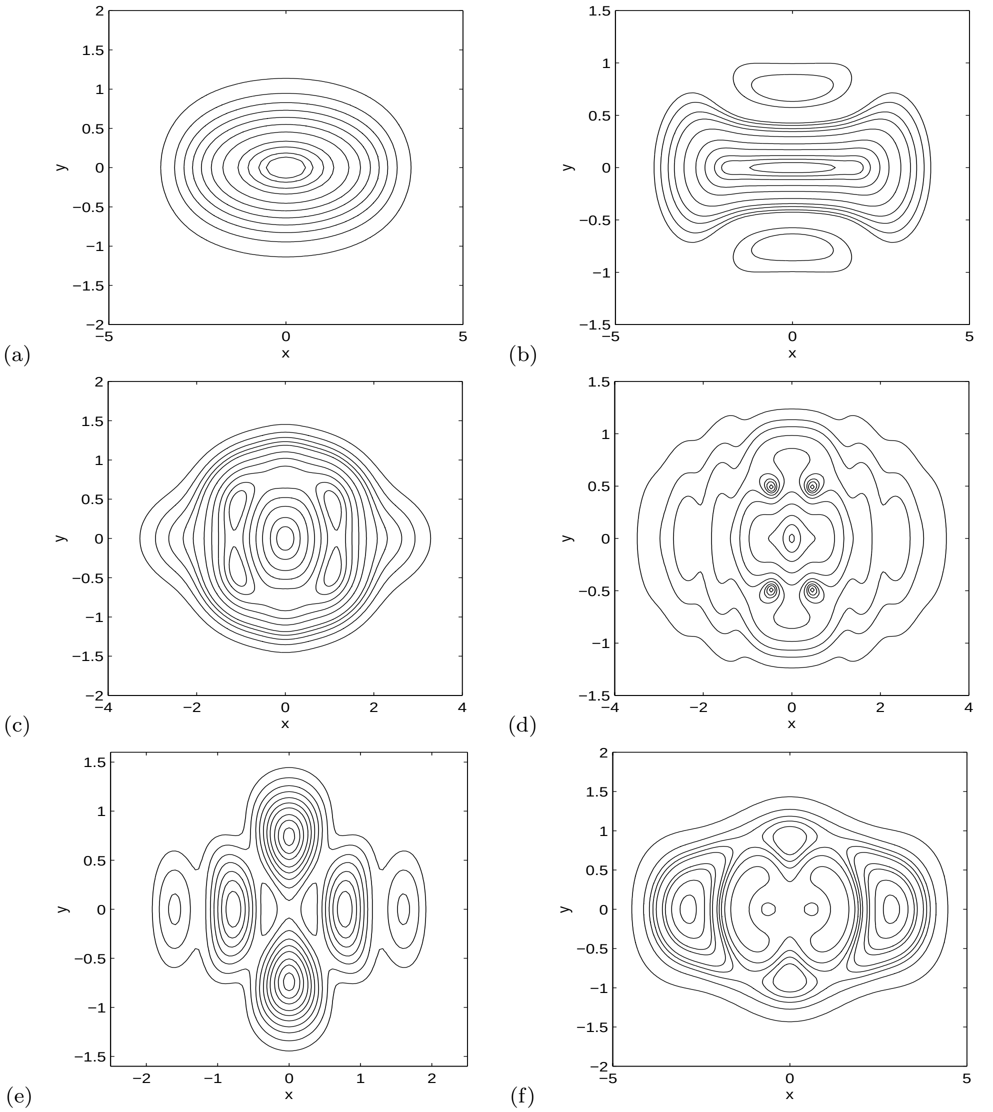


FIG. 8. *Contour plots of the density* $|\psi|^2$ *at different times in Example* 2, *case* III, *with* $\delta = 0.15$. (a) $t = 0$, (b) $t = 0.4$, (c) $t = 0.8$, (d) $t = 1.2$, (e) $t = 1.6$, (f) $t = 2.4$.

Again, we compare the effect of changing the damping parameter δ in the three different cases I, II, and III.

Figure 5 shows a surface plot of the density $|\psi(x, y, t)|^2$ at times $t = 0$ (ground-state solution) and $t = 2.8$ with $\delta = 1.25$; normalization, energy, and central density $|\psi(0, 0, t)|^2$ are shown as functions of time with $\delta = 1.25$, 1.1, and 0 (no damping) for case I. Figure 6 shows similar results for case II and Figure 7 for case III. Furthermore, Figure 8 shows contour plots of the density $|\psi|^2$ at different times for case III with $\delta = 0.15$.

From our numerical results we find that the observations (1)–(3) made for Example 1 are still valid with the additional trapping potential. However, the value of δ_{th} depends on β (or initial energy $E(0)$), and we find $\delta_{\text{th}} \approx 1.185$ for linear damping (cf. Figure 5).

3.1. Discussion. In this subsection, we discuss our numerical results in terms of physical properties of a BEC described by the GPE. We concentrate on those cases where a collapse of the wave function is arrested since this collapse leads to unphysical processes like the negative peaks in the energy $E(t)$ shown in Figures 1(c), (d) and 5(e), (f).

The general form of the time evolution in Example 1 is similar for all three cases. Initially the cloud of atoms contracts due to the attractive interaction between the particles. This contraction is accompanied by an increase in the energy due to particle loss which is most efficient in regions of high particle density. These regions are characterized by a negative local energy density leading to an increase in energy for each particle lost there. After the central particle density has reached a maximum, the cloud starts to expand due to the kinetic energy gained by the particles during the contraction. Particles are emitted from the cloud in burst-like pulses which can be seen in Figures 4 and 8. Such bursts have also been seen in BEC experiments [12]. The main differences between the three cases are the behavior of the energy and the number of particles as a function of time. In case I, where we assumed a linear damping term, the loss rate of particles from the condensate is independent of the shape of the condensate wave function. The energy decrease during the condensate expansion is determined by the loss of particles (cf. Figure 1(b)). In the cases of cubic and quintic damping, the loss term has a significant effect only on the time evolution of the condensate during the contraction. When the condensate expands, the density of particles is so low that the loss terms have only a very small effect and the energy $E(t)$ and the number of particles $N(t)$ remain almost constant (see Figures 2(c) and 3(c), (d)).

In Example 2, we add an additional trap potential which confines the BEC, and we assume a realistic scenario (described above) to prepare the condensate in the trap (cf. the experiments by Donley et al. [12]). We find that the initial process of turning on the attractive interactions between the particles leads to oscillations in the widths of the condensate [7] as can be seen from Figures 5, 6, and 7. However, neither the additional trap potential nor these oscillations significantly alter the behavior of the system compared to Example 1, when the condensate is strongly contracted. Before and after this contraction, some differences can be seen. By looking at Figures 5 and 6 we find that the first minimum in σ_y due to the oscillations of the condensate causes an increase in the central density and in the energy. For cubic and quintic damping, this is accompanied by an increased particle loss. However, an arrested collapse of the wave function happens only when both σ_x and σ_y attain a minimum value due to the attractive interactions (cf. Figures 5(d) and 6(b)). We also note that the frequency of the oscillations after an arrested collapse has happened is not significantly influenced by the damping terms. The amplitude of these oscillations is, however, strongly dependent on δ and decreases with increasing δ. Finally, we want to mention that a series of contractions and expansions of the condensate is possible. In Figure 7(b), we find three contractions of the condensate where only the first one reaches a sufficiently high particle density to lead to an increase in energy while the next two contractions show a rather smooth decrease in energy and particle number. For a smaller quintic damping term, we obtain two contractions of the condensate which increase the energy (see Figure 7(c)).

4. Conclusions. We extended the explicit unconditionally stable second-order TSSP method for solving damped focusing NLSs. We showed that this method is time transversal invariant and preserves the exact decay rate of the normalization for a linear damping of the NLS. Extensive numerical tests were presented for the cubic focusing NLS in two dimensions with linear, cubic, and quintic damping terms. Our numerical results show that quintic and cubic damping always arrest blowup, whereas linear damping can arrest blowup only when the damping parameter δ is bigger than a certain threshold value δ_{th}. We will apply this novel method to solve the three-dimensional GPE with a quintic damping term and will compare the numerical results with the experimental dynamics [12] of collapsing and exploding BECs [8].

Acknowledgment. The authors acknowledge the hospitality of the International Erwin Schrödinger Institute in Vienna, where this work was initiated.

REFERENCES

[1] S. K. Adhikari, *Mean-field description for collapsing and exploding Bose-Einstein condensates*, Phys. Rev. A, 66 (2002), pp. 3611–3615.

[2] M. H. Anderson, J. R. Ensher, M. R. Matthews, C. E. Wieman, and E. A. Cornell, *Observation of Bose-Einstein condensation in a dilute atomic vapor*, Science, 269 (1995), pp. 198–201.

[3] J. R. Anglin and W. Ketterle, *Bose-Einstein condensation of atomic gases*, Nature, 416 (2002), pp. 211–218.

[4] W. Bao, S. Jin, and P. A. Markowich, *On time-splitting spectral approximations for the Schrödinger equation in the semiclassical regime*, J. Comput. Phys., 175 (2002), pp. 487–524.

[5] W. Bao, S. Jin, and P. A. Markowich, *Numerical study of time-splitting spectral discretizations of nonlinear Schrödinger equations in the semiclassical regimes*, SIAM J. Sci. Comput., 25 (2003), pp. 27–64.

[6] W. Bao and W. Tang, *Ground state solution of trapped interacting Bose-Einstein condensate by minimizing a functional*, J. Comput. Phys., 187 (2003), pp. 230–254.

[7] W. Bao, D. Jaksch, and P. A. Markowich, *Numerical solution of the Gross-Pitaevskii equation for Bose-Einstein condensation*, J. Comput. Phys., 187 (2003), pp. 318–342.

[8] W. Bao, D. Jaksch, and P. A. Markowich, *Three Dimensional Simulation of Jet Formation in Collapsing Condensates*, preprint, 2003. Available online from http://arXiv.org/abs/cond-mat/0307344.

[9] R. Ciegis and V. Pakalnyte, *The finite difference scheme for the solution of weakly damped nonlinear Schrödinger equation*, Internat. J. Appl. Sci. Comput., 8 (2001), pp. 127–134.

[10] E. Cornell, *Very cold indeed: The nanokelvin physics of Bose-Einstein condensation*, J. Res. Natl. Inst. Stan., 101 (1996), pp. 419–434.

[11] F. Dalfovo, S. Giorgini, L. P. Pitaevskii, and S. Stringari, *Theory of Bose-Einstein condensation in trapped gases*, Rev. Modern Phys., 71 (1999), pp. 463–512.

[12] E. A. Donley, N. R. Claussen, S. L. Cornish, J. L. Roberts, E. A. Cornell, and C. E. Wieman, *Dynamics of collapsing and exploding Bose-Einstein condensates*, Nature, 412 (2001), pp. 295–299.

[13] R. A. Duine and H. T. C. Stoof, *Explosion of a collapsing Bose-Einstein condensate*, Phys. Rev. Lett., 86 (2001), pp. 2204–2207.

[14] G. Fibich, *Self-focusing in the damped nonlinear Schrödinger equation*, SIAM J. Appl. Math., 61 (2001), pp. 1680–1705.

[15] G. Fibich and D. Levy, *Self-focusing in the complex Ginzburg-Landau limit of the critical nonlinear Schrödinger equation*, Phys. Lett. A, 249 (1998), pp. 286–294.

[16] G. Fibich and G. Papanicolaou, *Self-focusing in the perturbed and unperturbed nonlinear Schrödinger equation in critical dimension*, SIAM J. Appl. Math., 60 (1999), pp. 183–240.

[17] O. Goubet, *Asymptotic smoothing effect for a weakly damped nonlinear Schrödinger equation in T^2*, J. Differential Equations, 165 (2000), pp. 96–122.

[18] O. Goubet, *Regularity of the attractor for a weakly damped nonlinear Schrödinger equation in $R^2$2*, Adv. Differential Equations, 3 (1998), pp. 337–360.

[19] O. Goubet, *Approximate inertial manifolds for a weakly damped nonlinear Schrödinger equation*, Discrete Contin. Dynam. Systems, 3 (1997), pp. 503–530.

[20] M. Greiner, O. Mandel, T. Esslinger, T. W. Hänsch, and I. Bloch, *Quantum phase transition from a superfluid to a mott insulator in a gas of ultracold atoms*, Nature, 415 (2002), pp. 39–45.

[21] E. P. Gross, *Structure of a quantized vortex in boson systems*, Nuovo Cimento (10), 20 (1961), pp. 454–477.

[22] S. R. K. Iyengar, G. Jayaraman, and V. Balasubramanian, *Variable mesh difference schemes for solving a nonlinear Schrödinger equation with a linear damping term. Advances in partial differential equations*, III, Comput. Math. Appl., 40 (2000), pp. 1375–1385.

[23] D. Jaksch, C. Bruder, J. I. Cirac, C. W. Gardiner, and P. Zoller, *Cold bosonic atoms in optical lattices*, Phys. Rev. Lett., 81 (1998), pp. 3108–3111.

[24] M. S. Jolly, R. Temam, and C. Xiong, *An application of approximate inertial manifolds to a weakly damped nonlinear Schrödinger equation*, Numer. Funct. Anal. Optim., 16 (1995), pp. 923–937.

[25] Y. Kagan, A. E. Muryshev, and G. V. Shlyapnikov, *Collapse and Bose-Einstein condensation in a trapped Bose gas with negative scattering length*, Phys. Rev. Lett., 81 (1998), pp. 933–937.

[26] L. Landau and E. Lifschitz, *Quantum Mechanics: Non-Relativistic Theory*, Pergamon Press, New York, 1977.

[27] M. J. Landman, G. C. Papanicolaou, C. Sulem, P. L. Sulem, and X. P. Wang, *Stability of isotropic singularities for the nonlinear Schrödinger equation*, Phys. D, 47 (1991), pp. 393–415.

[28] A. Mielke, *The complex Ginzburg-Landau equation on large and unbounded domains: Sharper bounds and attractors*, Nonlinearity, 10 (1997), pp. 199–222.

[29] G. Moebs, *Guy A multilevel method for the resolution of a stochastic weakly damped nonlinear Schrödinger equation*, Appl. Numer. Math., 26 (1998), pp. 353–375.

[30] G. Moebs and R. Temam, *Resolution of a stochastic weakly damped nonlinear Schrödinger equation by a multilevel numerical method*, J. Opt. Soc. Amer. A, 17 (2000), pp. 1870–1879.

[31] G. C. Papanicolaou, C. Sulem, P. L. Sulem, and X. P. Wang, *Singular solutions of the Zakharov equations for Langmuir turbulence*, Phys. Fluids B, 3 (1991), pp. 969–980.

[32] L. S. Peranich, *A finite difference scheme for solving a non-linear Schrödinger equation with a linear damping term*, J. Comput. Phys., 68 (1987), pp. 501–505.

[33] L. P. Pitaevskii, *Vortex lines in an imperfect Bose gaze*, Z. Èksper. Teoret. Fiz., 40 (1961), pp. 646–651 (in Russian); Soviet Physics JETP, 13 (1961), pp. 451–454.

[34] J. L. Roberts, N. R. Claussen, S. L. Cornish, and C. E. Wieman, *Magnetic field dependence of ultracold inelastic collisions near a Feshbach resonance*, Phys. Rev. Lett., 85 (2000), pp. 728–731.

[35] H. Saito and M. Ueda, *Intermittent implosion and pattern formation of trapped Bose-Einstein condensates with an attractive interaction*, Phys. Rev. Lett., 86 (2001), pp. 1406–1409.

[36] C. Sulem and P. L. Sulem, *The Nonlinear Schrödinger Equation: Self-Focusing and Wave Collapse*, Springer-Verlag, New York, 1999.

[37] M. Tsutsumi, *Nonexistence of global solutions to the Cauchy problem for the damped nonlinear Schrödinger equations*, SIAM J. Math. Anal., 15 (1984), pp. 357–366.

[38] M. Tsutsumi, *On global solutions to the initial-boundary value problem for the damped nonlinear Schrödinger equations*, J. Math. Anal. Appl., 145 (1990), pp. 328–341.

[39] F.-Y. Zhang and S.-J. Lu, *Long-time behavior of finite difference solutions of a nonlinear Schrödinger equation with weakly damped*, J. Comput. Math., 19 (2001), pp. 393–406.

SIAM J. NUMER. ANAL.
Vol. 41, No. 4, pp. 1427–1464

FINITE ELEMENT APPROXIMATION OF SURFACTANT SPREADING ON A THIN FILM*

JOHN W. BARRETT[†][§], HARALD GARCKE[‡], AND ROBERT NÜRNBERG[†][¶]

Abstract. We consider a fully practical finite element approximation of the following system of nonlinear degenerate parabolic equations:

$$\frac{\partial u}{\partial t} + \tfrac{1}{2}\,\nabla.(u^2\,\nabla[\sigma(v)]) - \tfrac{1}{3}\,\nabla.(u^3\,\nabla w) = 0, \qquad w = -c\,\Delta u + a\,u^{-3} - \delta\,u^{-\nu},$$

$$\frac{\partial v}{\partial t} + \nabla.(u\,v\,\nabla[\sigma(v)]) - \rho\,\Delta v - \tfrac{1}{2}\,\nabla.(u^2\,v\,\nabla w) = 0.$$

The above models a surfactant-driven thin film flow in the presence of both attractive, $a > 0$, and repulsive, $\delta > 0$ with $\nu > 3$, van der Waals forces, where u is the height of the film, v is the concentration of the insoluble surfactant monolayer, and $\sigma(v) := 1 - v$ is the typical surface tension. Here $\rho \geq 0$ and $c > 0$ are the inverses of the surface Peclet number and the modified capillary number. In addition to showing stability bounds for our approximation, we prove convergence in one space dimension when $\rho > 0$ and either $a = \delta = 0$ or $\delta > 0$. Furthermore, iterative schemes for solving the resulting nonlinear discrete system are discussed. Finally, some numerical experiments are presented.

Key words. thin film flow, surfactant, fourth order degenerate parabolic system, finite elements, convergence analysis

AMS subject classifications. 65M60, 65M12, 35K55, 35K65, 35K35, 76A20, 76D08

DOI. 10.1137/S003614290139799X

1. Introduction.

The study of the motion of surfactants placed on a thin layer of a viscous fluid is motivated by applications ranging from the medical treatment of premature infants to industrial coating and drying processes (cf. [13, 18, 24, 26, 27]). We are interested in situations in which a free surface of a thin film contains a monolayer of a surfactant, which is a chemical that lowers the surface tension. Surface tension gradients then lead to shear stresses which force the liquid to flow toward regions of higher surface tension (Marangoni effect). In total, the liquid flow is driven by capillarity and surfactant gradient induced convection (Marangoni convection).

We consider a situation in which the thin layer of a viscous fluid spreads on a horizontal planar surface. The evolution then can be described by a free boundary problem for the Navier–Stokes equations coupled to a convection-diffusion equation for the surfactant, where the latter equation has to be solved only on the free surface. Starting from this complicated free boundary problem, it is possible, under appropriate assumptions, to use lubrication theory to derive a coupled set of nonlinear partial differential equations for the two unknowns: film thickness and surfactant concentration. It is the goal of this paper to develop and analyze a finite element method for this set of equations.

*Received by the editors November 13, 2001; accepted for publication (in revised form) February 17, 2003; published electronically September 9, 2003.

http://www.siam.org/journals/sinum/41-4/39799.html

[†]Department of Mathematics, Imperial College, London, SW7 2AZ, UK (j.barrett@ic.ac.uk).

[‡]Institut für Angewandte Mathematik, Wegelerstraße 6, 53115 Bonn, Germany. Present address: NWF I – Mathematik, Universität Regensburg, 93040 Regensburg, Germany (harald.garcke@mathematik.uni-regensburg.de). The research of this author was partially supported by the RTN Programme Fronts–Singularities.

[§]The research of this author was partially supported by the DFG through the SFB256 Nichtlineare Partielle DGL.

[¶]The research of this author was supported by the EPSRC, and by the DAAD through a Doktorandenstipendium.

Denoting by u the height of the film, by w the pressure, and by v the concentration of the insoluble surfactant, the governing equations that one derives from lubrication theory (cf. [13, 18, 31]) are

$$\tag{1.1a} \tfrac{\partial u}{\partial t} + \tfrac{1}{2}\,\nabla.(u^2\,\nabla[\sigma(v)]) - \tfrac{1}{3}\,\nabla.(u^3\,\nabla w) = 0, \qquad w = -c\,\Delta u\,,$$

$$\tag{1.1b} \tfrac{\partial v}{\partial t} + \nabla.(u\,v\,\nabla[\sigma(v)]) - \rho\,\Delta v - \tfrac{1}{2}\,\nabla.(u^2\,v\,\nabla w) = 0.$$

In the following, we will denote by the vector x the horizontal variables and by $y \in \mathbb{R}$ the vertical variable. The functions u, v, and w are functions of x and the time t, and all spatial operators like $\nabla, \nabla.$, and Δ in this paper act on the horizontal variables only. The given data are $\rho \in \mathbb{R}_{\geq 0}$, the inverse of the surface Peclet number, and $c \in \mathbb{R}_{>0}$, the inverse of a modified capillary number. In addition, $\sigma : [0,1] \to [0,1]$ is the constitutive equation of state relating the surface tension σ to v, e.g.,

$$\tag{1.2} \sigma(s) := (\beta+1)\,[1+\theta(\beta)\,s]^{-3} - \beta, \quad \text{where} \quad \theta(\beta) := \big(\tfrac{\beta+1}{\beta}\big)^{\frac{1}{3}} - 1,$$

and $\beta \in \mathbb{R}_{>0}$ relates to the activity of the surfactant (cf. [24, p. 262]). It is reasonable to assume throughout that σ is a monotonically decreasing function of v, as the surfactant lowers surface tension.

Let us now discuss some properties of the system which are important for later developments. First, we want to show how one can recover the pressure and velocity in the fluid if one knows a solution $\{u, v, w\}$ of (1.1a,b). In lubrication theory (see [19]), it turns out that the pressure in the fluid is independent of the vertical variable y, and we obtain that $p(x,y,t) \equiv w(x,t) = -c\,\Delta u(x,t)$, where the right-hand side is an approximation to the mean curvature of the air/liquid interface. We remark that one obtains this identity from the leading order equation in lubrication theory. Another important quantity in lubrication theory is the horizontal component of the velocity, $\vec{\mathbf{V}}_H$, which can be computed from $\{u(x,t), v(x,t), w(x,t)\}$ as follows:

$$\tag{1.3} \vec{\mathbf{V}}_H(x,y,t) = y\,\nabla[\sigma(v)] + \big(\tfrac{1}{2}\,y^2 - y\,u\big)\,\nabla w,$$

where a no-slip condition has been assumed at $y = 0$. One notices that $\vec{\mathbf{V}}_H$ is quadratic in the y-direction. Furthermore, the fluid is driven by two effects: namely, by pressure gradients due to capillarity effects, $-c\,\nabla(\Delta u)$, and by surface tension gradients, $\nabla[\sigma(v)]$. Equation (1.3) evaluated for $y = u(x,t)$ yields the horizontal velocity on the free surface, and hence the equation for the surfactant concentration, (1.1b), can be rewritten as

$$\tfrac{\partial v}{\partial t} + \nabla.(v\,\vec{\mathbf{V}}_H(x, u(x,t), t)) = \rho\,\Delta v\,,$$

which shows that it can be interpreted as a convection-diffusion equation, where the surfactant is transported with the velocity of the fluid. In addition, the equation for the film height can be expressed with the help of the fluid velocity. A straightforward computation starting from (1.1a) shows that the change of height of the film is given in terms of the horizontal component of the velocity as follows:

$$\tfrac{\partial u}{\partial t} = -\nabla.\left(\int_0^{u(x,t)} \vec{\mathbf{V}}_H(x,y,t)\;\mathrm{d}y\right)\,.$$

A basic ingredient of our approach is an energy estimate for surfactant-driven flows. To derive an energy estimate involving a density function $F(v)$, we use some ideas from thermodynamics. First, we relate F to σ by the Gibbs identity

$$\sigma(v) = F(v) - v\,F'(v) \qquad \Rightarrow \qquad \sigma'(v) = -v\,F''(v)\,. \tag{1.4}$$

Knowing σ, the above identity determines F up to a linear term. Assuming appropriate boundary conditions, which will be specified later on, one can derive an energy estimate for the surfactant-driven thin film system as

$$\tfrac{\mathrm{d}}{\mathrm{d}t}\int_\Omega \left[\tfrac{c}{2}\,|\nabla u|^2 + F(v)\right]\,\mathrm{d}x + \int_\Omega\int_0^u |\partial_y \vec{\mathbf{V}}_H|^2\,\mathrm{d}y\,\mathrm{d}x + \rho\int_\Omega F''(v)\,|\nabla v|^2\,\mathrm{d}x = 0\,, \tag{1.5}$$

where Ω is a bounded domain in $\mathbb{R}^d$, $d = 1$ or 2. To derive the above, we have used the identity

$$\int_0^u |\partial_y \vec{\mathbf{V}}_H|^2\,\mathrm{d}y = u\,|\nabla[\sigma(v)]|^2 - u^2\,\nabla[\sigma(v)]\,.\,\nabla w + \tfrac{1}{3}\,u^3\,|\nabla w|^2\,. \tag{1.6}$$

The identity (1.5) directly corresponds to the energy estimate for the free boundary value problem for the Navier–Stokes equations. The term $\frac{\mathrm{d}}{\mathrm{d}t}\int_\Omega [\frac{c}{2}\,|\nabla u|^2 + F(v)]\,\mathrm{d}x$ describes the rate of change of energy in time, and, since one neglects inertia effects, only capillarity terms appear in the energy. The two remaining terms in (1.5) represent energy dissipation due to friction in the fluid and diffusion of the surfactant, respectively.

On recalling that σ is monotonically decreasing, we deduce from (1.4) that F'' is nonnegative, and hence the identity (1.5) shows that the energy decreases in time. This energy estimate will lead to important a priori estimates. In particular, the identity (1.6) together with the inequality

$$u^2\,\nabla[\sigma(v)]\,.\,\nabla w \le \tfrac{\gamma}{2}\,u\,|\nabla[\sigma(v)]|^2 + \tfrac{1}{2\gamma}\,u^3\,|\nabla w|^2, \qquad \gamma \in (\tfrac{3}{2}, 2), \tag{1.7}$$

then shows that we can control $u\,|\nabla[\sigma(v)]|^2$ and $u^3\,|\nabla w|^2$ with the help of the energy estimate (1.5).

It is the goal of this paper to derive a finite element method that is consistent with the energy estimate (1.5); i.e., we want to derive a method that satisfies a discrete analogue.

To conclude the system, we need to specify initial and boundary conditions for (1.1a–c). One possibility that leads to the above energy estimate would be to describe periodic boundary conditions. Instead we prescribe the following conditions at the horizontal boundary: a no-penetration condition for the velocity, a 90° angle condition for the film height, and a no-flux condition for the surfactant concentration. This implies the following conditions for $x \in \partial\Omega$ and $0 \le y \le u(x,t)$:

$$\nu_{\partial\Omega}\,.\,\vec{\mathbf{V}}_H(x,y,t) \equiv \nu_{\partial\Omega}\,.\,\big(y\,\nabla[\sigma(v)] + \big(\tfrac{1}{2}\,y^2 - y\,u\big)\,\nabla w\big) = 0, \qquad \tfrac{\partial u}{\partial \nu_{\partial\Omega}} = 0, \tag{1.8a}$$

$$\nu_{\partial\Omega}.\,(v\,\vec{\mathbf{V}}_H(x,u(x,t),t) - \rho\,\nabla v) \equiv \nu_{\partial\Omega}.(v\,(u\,\nabla[\sigma(v)] - \tfrac{1}{2}\,u^2\,\nabla w) - \rho\,\nabla v) = 0, \tag{1.8b}$$

where $\nu_{\partial\Omega}$ is normal to $\partial\Omega$, the Lipschitz boundary of Ω. Integrating the first equation in (1.8a) with respect to y yields that

$$\tfrac{1}{2}\,u^2\,\tfrac{\partial[\sigma(v)]}{\partial\nu_{\partial\Omega}} - \tfrac{1}{3}\,u^3\,\tfrac{\partial w}{\partial\nu_{\partial\Omega}} = 0 \qquad \text{on } \partial\Omega\,, \tag{1.9}$$

which means that the height averaged normal component of the horizontal velocity is zero on the boundary. Note that in the case that either $\rho > 0$ or $v\,\sigma'(v) \neq 0$, it can be seen that the first boundary condition in (1.8a) and (1.8b) are equivalent to (1.9) and (1.8b) (observe that (1.8a) holds for all $y \in [0, u(x,t)]$, $x \in \partial\Omega$).

In what follows, we will therefore specify the boundary conditions

$$\tfrac{1}{2}\, u^2\, \tfrac{\partial[\sigma(v)]}{\partial\nu_{\partial\Omega}} - \tfrac{1}{3}\, u^3\, \tfrac{\partial w}{\partial\nu_{\partial\Omega}} = \tfrac{\partial u}{\partial\nu_{\partial\Omega}} = u\, v\, \tfrac{\partial[\sigma(v)]}{\partial\nu_{\partial\Omega}} - \tfrac{1}{2}\, u^2\, v\, \tfrac{\partial w}{\partial\nu_{\partial\Omega}} - \rho\, \tfrac{\partial v}{\partial\nu_{\partial\Omega}} = 0 \quad \text{on } \partial\Omega\,.$$

We remark that if either $u\,\rho > 0$ or $-u\,v\,\sigma'(v) > 0$ on $\partial\Omega$, these boundary conditions are equivalent to $\frac{\partial u}{\partial\nu_{\partial\Omega}} = \frac{\partial w}{\partial\nu_{\partial\Omega}} = \frac{\partial v}{\partial\nu_{\partial\Omega}} = 0$ on $\partial\Omega$.

When surfactant is placed on the film, a thinning effect can be observed (see the numerical results in section 5). If the film thickness is in the range of a few hundred Angstroms, then molecular effects due to van der Waals forces become important. If van der Waals forces are included, an additional conservative body force enters the Navier–Stokes equations (see, e.g., [28]). In the thin film equations and in all the formulae above, the pressure w, related to the height u previously by (1.1a), is then replaced by the reduced pressure

$$w = -c\,\Delta u + \phi(u)\,, \qquad \text{where} \quad \phi(u) := a\,u^{-3} - \delta\,u^{-\nu}, \quad \nu > 3\,. \tag{1.10}$$

Here $a \in \mathbb{R}_{\geq 0}$ is the scaled dimensionless Hamaker constant and $\delta \in \mathbb{R}_{\geq 0}$ represents the effect of repulsive van der Waals forces; see, e.g., [28]. When van der Waals forces are included, the energy estimate (1.5) is replaced by

(1.11)
$$\frac{\mathrm{d}}{\mathrm{d}t}\int_\Omega \left[\tfrac{c}{2}\,|\nabla u|^2 + F(v) + \Phi(u)\right] \mathrm{d}x + \int_\Omega \int_0^u |\partial_y \vec{\mathbf{V}}_H|^2 \,\mathrm{d}y\,\mathrm{d}x + \rho \int_\Omega F''(v)\,|\nabla v|^2 \,\mathrm{d}x = 0,$$

where Φ is an antiderivative of ϕ, i.e., $\Phi' = \phi$, and the horizontal velocity $\vec{\mathbf{V}}_H$ is still given by (1.3) but with w now defined by (1.10).

Altogether, in this paper we consider the following initial boundary value problem.

(P) Find functions u, w, $v : \Omega \times [0,T] \to \mathbb{R}$ such that

$$\tfrac{\partial u}{\partial t} + \tfrac{1}{2}\,\nabla.(u^2\,\nabla[\sigma(v)]) - \tfrac{1}{3}\,\nabla.(u^3\,\nabla\,w) = 0 \quad \text{in } \Omega_T, \tag{1.12a}$$

$$w = -c\,\Delta u + \phi(u) \quad \text{in } \Omega_T, \text{ where } u > 0, \tag{1.12b}$$

$$\tfrac{\partial v}{\partial t} + \nabla.(u\,\lambda(v)\,\nabla[\sigma(v)]) - \rho\,\Delta v - \tfrac{1}{2}\,\nabla.(u^2\,\lambda(v)\,\nabla w) = 0 \quad \text{in } \Omega_T, \tag{1.12c}$$

$$u(x,0) = u^0(x), \qquad v(x,0) = v^0(x) \quad \forall\, x \in \Omega, \tag{1.12d}$$

$$\begin{aligned} &\tfrac{1}{2}\, u^2\, \tfrac{\partial[\sigma(v)]}{\partial\nu_{\partial\Omega}} - \tfrac{1}{3}\, u^3\, \tfrac{\partial w}{\partial\nu_{\partial\Omega}} = \tfrac{\partial u}{\partial\nu_{\partial\Omega}} = u\,\lambda(v)\, \tfrac{\partial[\sigma(v)]}{\partial\nu_{\partial\Omega}} - \tfrac{1}{2}\, u^2\,\lambda(v)\, \tfrac{\partial w}{\partial\nu_{\partial\Omega}} \\ &-\rho\,\tfrac{\partial v}{\partial\nu_{\partial\Omega}} = 0 \quad \text{on } \partial\Omega \times (0,T), \end{aligned} \tag{1.12e}$$

where $\Omega_T := \Omega \times (0,T]$ and $T > 0$ is a fixed positive time. In the above, $c \in \mathbb{R}_{>0}$ and $\rho \in \mathbb{R}_{\geq 0}$ are given constants, while $\phi : \mathbb{R}_{>0} \to \mathbb{R}$ is defined as in (1.10); u^0 and v^0 are given nonnegative initial profiles (e.g., $u^0 \equiv 1$ for a film of uniform thickness and u^0 having support $\subset\subset \Omega$ for a drop). Throughout this paper, we will restrict ourselves to the model case $\sigma(v) := 1 - v$, the $\beta \to \infty$ limit of (1.2). We remark that physically relevant values of $v(x,t)$ lie in the interval $[0,1]$. Noting this, it is convenient for the analysis in this paper to replace the terms $u^i\,v$, $i = 1 \to 2$, in (1.1b) by $u^i\,\lambda(v)$, where $\lambda : \mathbb{R} \to (-\infty, 1]$ is defined as

$$\lambda(s) := [s-1]_- + 1, \qquad \text{with} \qquad [s]_- := \min\{s,\, 0\}. \tag{1.13}$$

As u and $\lambda(v)$ can take on zero values, (P) is a degenerate parabolic system, which is fourth order in u. This degeneracy makes the analysis/numerical analysis of (P) particularly difficult. Although we have assumed a no-slip condition at $y = 0$ in the derivation of (P) (see (1.3)) the results in this paper can easily be generalized to models allowing slip (see [31]).

Let us mention some work on problems that also lead to degenerate parabolic equations of fourth order. In particular, we would like to mention work on the following topics: thin film flow (cf. [9, 19, 10, 8, 11]), the Cahn–Hilliard equation with a degenerate mobility (cf. [14, 16, 17]), and models that describe dislocation densities in plasticity (cf. [20]). An existence result for the system (P) studied in this paper has been given by Wieland [31] in the case of one space dimension.

Problem (P) with $v^0 \equiv 0$ and $\phi \equiv 0$ collapses to the thin film equation, i.e., a degenerate parabolic equation of fourth order. Degenerate parabolic equations of higher order exhibit some new characteristic features which are fundamentally different from those for second order degenerate parabolic equations such as the porous medium equation $\frac{\partial u}{\partial t} - \nabla.(|u|^\alpha \nabla u) = 0$ for a given $\alpha \in \mathbb{R}_{>0}$. The key point is that there is no maximum or comparison principle for parabolic equations of higher order. This drastically complicates the analysis since many results which are known for second order equations are proven with the help of comparison techniques. Related to this is the fact that there is no uniqueness result known for the thin film equation. Although there is no comparison principle, one of the main features of the thin film equation is the fact that one can show existence of nonnegative solutions if given nonnegative initial data. This is in contrast to linear parabolic equations of fourth order, where solutions which are initially positive may become negative in certain regions.

There is very little work on the numerical analysis of degenerate parabolic equations of fourth order; for work on thin film flows in the absence of surfactants, see [4, 32, 22, 21] and for work on degenerate Cahn–Hilliard systems, see [5, 6, 3].

This paper is organized as follows. In section 2 we formulate a fully practical finite element approximation of problem (P). On the discrete level, the nonnegativity of the approximation to u is not guaranteed when we discretize the system (1.12a–e) in a naive way. Following [4], we impose the nonnegativity of the discrete approximation to u as a constraint and require (1.12b) only where u is positive. In addition, in order to derive a discrete analogue of the energy estimate (1.11), we adapt a technique introduced in [32] and [22] for deriving a discrete entropy bound for the thin film equation.

We can derive stability bounds in space dimensions $d = 1$ and 2, but we are only able to show convergence in one space dimension. This is due to the fact that the a priori bounds we derive guarantee in one space dimension only that the discrete approximation to u is uniformly bounded and equicontinuous, which is necessary to be able to pass to the limit in the discrete problem. For similar reasons, the results in [9, 4, 22, 5, 6, 3] were restricted to one space dimension. This convergence is carried out in section 3. A convergence result for a finite element method of the thin film equation in two dimensions has been given recently by Grün [21]. Unfortunately a generalization of Grün's result to the problem presented in this paper does not seem to be possible in a straightforward manner. In section 4, we introduce and analyze algorithms to solve the nonlinear algebraic systems at each time level. Finally, in section 5, we present some numerical computations in one and two space dimensions. We compare the computed discrete solutions with results published in [24, 26] and other papers.

Notation and auxiliary results. We adopt the standard notation for Sobolev spaces, denoting the norm of $W^{m,q}(\Omega)$ ($m \in \mathbb{N}$, $q \in [1,\infty]$) by $\|\cdot\|_{m,q}$ and the seminorm by $|\cdot|_{m,q}$. For $q = 2$, $W^{m,2}(\Omega)$ will be denoted by $H^m(\Omega)$ with the associated norm and seminorm written as, respectively, $\|\cdot\|_m$ and $|\cdot|_m$. Throughout $(\cdot,\cdot)$ denotes the standard L^2 inner product over Ω and $\langle\cdot,\cdot\rangle$ denotes the duality pairing between $\big(H^1(\Omega)\big)'$ and $H^1(\Omega)$. In addition we define

$$(1.14)\qquad ⨍\eta := \tfrac{1}{\underline{m}(\Omega)}\,(\eta,1) \qquad \forall\ \eta \in L^1(\Omega), \qquad \text{where } \underline{m}(\Omega) \text{ is the measure of } \Omega\,.$$

We require also the Hölder space $C^{p_1,p_2}_{x,t}(\overline{\Omega}_T)$ for $p_i \in (0,1]$, which denotes those functions whose time (spatial) derivative(s) is (are) Hölder continuous over $\overline{\Omega}_T$ with exponent $p_1(p_2)$.

For later purposes, we recall the following well-known Sobolev interpolation result (see, e.g., [1]): Let $m \geq 1$, $q \in (\frac{d}{m},\infty]$, $r \in [q,\infty]$, and $\mu := \frac{d}{m}\left(\frac{1}{q} - \frac{1}{r}\right)$. Then there is a constant C depending only on Ω, m, q, r such that

$$(1.15)\qquad |z|_{0,r} \leq C\,|z|_{0,q}^{1-\mu}\,\|z\|_{m,q}^{\mu} \quad \forall\ z \in W^{m,q}(\Omega)\,.$$

It is convenient to introduce the "inverse Laplacian" operator $\mathcal{G} : \mathcal{F} \to Z$ such that

$$(1.16)\qquad (\nabla\mathcal{G}z, \nabla\eta) = \langle z, \eta\rangle \quad \forall\ \eta \in H^1(\Omega),$$

where $\mathcal{F} := \big\{z \in (H^1(\Omega))' : \langle z, 1\rangle = 0\big\}$ and $Z := \{z \in H^1(\Omega) : (z,1) = 0\}$. The well-posedness of $\mathcal{G}$ follows from the Lax–Milgram theorem and the Poincaré inequality

$$(1.17)\qquad |\eta|_{0,q} \leq C\,(\,|\eta|_{1,q} + |(\eta,1)|\,) \quad \forall\ \eta \in W^{1,q}(\Omega) \quad \text{and} \quad q \in [1,\infty].$$

One can define a norm on $\mathcal{F}$ by

$$(1.18)\qquad \|z\|_{-1} := |\mathcal{G}z|_1 = \langle z, \mathcal{G}z\rangle^{\frac{1}{2}} \quad \forall\ z \in \mathcal{F}.$$

We note also for future reference that using Young's inequality

$$(1.19)\qquad r\,s \leq \tfrac{\gamma}{2}\,r^2 + \tfrac{1}{2\gamma}\,s^2 \quad \forall\ r,s \in \mathbb{R},\ \gamma \in \mathbb{R}_{>0}$$

yields for all $\gamma \in \mathbb{R}_{>0}$ that

$$(1.20)\qquad \langle z, \eta\rangle = (\nabla\mathcal{G}z, \nabla\eta) \leq \|z\|_{-1}|\eta|_1 \leq \tfrac{\gamma}{2}\,|\eta|_1^2 + \tfrac{1}{2\gamma}\,\|z\|_{-1}^2 \quad \forall\ z \in \mathcal{F},\ \eta \in H^1(\Omega).$$

Throughout C denotes a generic constant independent of h, τ, and ε, the mesh and temporal discretization parameters and the regularization parameter. In addition, $C(a_1,\ldots,a_I)$ denotes a constant depending on the arguments $\{a_i\}_{i=1}^I$.

2. Finite element approximation. We consider the finite element approximation of (P) at first under the following assumptions on the mesh:

(A) Let Ω be a polygonal domain if $d = 2$. Let $\{\mathcal{T}^h\}_{h>0}$ be a quasi-uniform family of partitionings of Ω into disjoint open simplices κ with $h_\kappa := \text{diam}(\kappa)$ and $h := \max_{\kappa\in\mathcal{T}^h} h_\kappa$ so that $\overline{\Omega} = \cup_{\kappa\in\mathcal{T}^h}\overline{\kappa}$. In addition, it is assumed for $d = 2$ that all simplices $\kappa \in \mathcal{T}^h$ are right-angled.

We note that the quasi uniformity assumption can be avoided at the expense of a mild constraint on the minimum time step; see Remark 3.5 below. Furthermore we note that the right-angled simplices assumption is not a severe constraint, as there exist adaptive finite element codes that satisfy this requirement; see, e.g., [30].

Associated with $\mathcal{T}^h$ is the finite element space $S^h := \{\chi \in C(\overline{\Omega}) : \chi\,|_\kappa$ is linear for all $\kappa \in \mathcal{T}^h\} \subset H^1(\Omega)$. We introduce also $K^h := \{\chi \in S^h : \chi \geq 0 \text{ in } \Omega\} \subset K$, where $K := \{\eta \in H^1(\Omega) : \eta \geq 0 \text{ a.e. in } \Omega\}$. Let J be the set of nodes of $\mathcal{T}^h$ and $\{p_j\}_{j\in J}$ the coordinates of these nodes. Let $\{\chi_j\}_{j\in J}$ be the standard basis functions for S^h; that is, $\chi_j \in K^h$ and $\chi_j(p_i) = \delta_{ij}$ for all $i, j \in J$. We introduce $\pi^h : C(\overline{\Omega}) \to S^h$, the interpolation operator, such that $(\pi^h\eta)(p_j) = \eta(p_j)$ for all $j \in J$. A discrete semi-inner product on $C(\overline{\Omega})$ is then defined by

$$(\eta_1, \eta_2)^h := \int_\Omega \pi^h(\eta_1(x)\,\eta_2(x))\,\mathrm{d}x = \sum_{j\in J} m_j\,\eta_1(p_j)\,\eta_2(p_j), \tag{2.1}$$

where $m_j := (1, \chi_j) > 0$. The induced discrete seminorm is then $|\eta|_h := [\,(\eta,\eta)^h\,]^{\frac{1}{2}}$, where $\eta \in C(\overline{\Omega})$. We introduce also the L^2 projection $Q^h : L^2(\Omega) \to S^h$ defined by

$$(Q^h\eta, \chi)^h = (\eta, \chi) \qquad \forall\ \chi \in S^h. \tag{2.2}$$

In this paper, for simplicity, we consider only the model case when the surface tension is given by $\sigma(s) := 1-s$, which is the limit as $\beta \to \infty$ in (1.2) and is commonly used in the physics/engineering literature (cf. [27]). On recalling (1.4), we then define a function F such that $v\,\nabla[F'(v)] = -\nabla[\sigma(v)]$; that is, $F''(s) = -s^{-1}\,\sigma'(s) = s^{-1} \Rightarrow F(s) = s\,(\ln s - 1) + 1$. For computational purposes, we replace $F \in C^\infty(\mathbb{R}_{>0})$ for any $\varepsilon \in (0,1)$ by the regularized function $F_\varepsilon : \mathbb{R} \to \mathbb{R}_{\geq 0}$ such that

$$F_\varepsilon(s) := \begin{cases} \frac{s^2-\varepsilon^2}{2\,\varepsilon} + (\ln\varepsilon - 1)\,s + 1, & s \leq \varepsilon, \\ s\,(\ln s - 1) + 1, & \varepsilon \leq s \leq 1, \\ \frac{1}{2}\,(s-1)^2, & 1 \leq s\,. \end{cases}$$

Hence $F_\varepsilon \in C^{2,1}(\mathbb{R})$ with the first two derivatives of F_ε given by

$$F_\varepsilon'(s) := \begin{cases} \varepsilon^{-1}\,s + \ln\varepsilon - 1, & s \leq \varepsilon, \\ \ln s, & \varepsilon \leq s \leq 1, \\ s - 1, & 1 \leq s, \end{cases} \quad \text{and} \quad F_\varepsilon''(s) := \begin{cases} \varepsilon^{-1}, & s \leq \varepsilon, \\ s^{-1}, & \varepsilon \leq s \leq 1, \\ 1, & 1 \leq s\,, \end{cases} \tag{2.3}$$

respectively. For later purposes, we note that

$$F_\varepsilon(s) \geq \tfrac{s^2}{4} - \tfrac{1}{2} \quad \forall\ s \geq 0 \qquad \text{and} \qquad F_\varepsilon(s) \geq \tfrac{s^2}{2\,\varepsilon} \quad \forall\ s \leq 0\,. \tag{2.4}$$

This holds since

$$\begin{aligned} F_\varepsilon(s) := \tfrac{1}{2}(s-1)^2 \geq \tfrac{1}{2}(s-1)^2 - (\tfrac{1}{2}s - 1)^2 = \tfrac{s^2}{4} - \tfrac{1}{2} \qquad & \text{if } s \geq 1\,, \\ F_\varepsilon'(s) \leq 0 \quad \Rightarrow \quad F_\varepsilon(s) \geq F_\varepsilon(1) = 0 \geq \tfrac{s^2}{4} - \tfrac{1}{2} \qquad & \text{if } s \in [0,1]\,, \\ F_\varepsilon(s) := \tfrac{s^2-\varepsilon^2}{2\,\varepsilon} + (\ln\varepsilon - 1)\,s + 1 \geq \tfrac{s^2}{2\,\varepsilon} + (1 - \tfrac{\varepsilon}{2}) \geq \tfrac{s^2}{2\,\varepsilon} \qquad & \text{if } s \leq 0\,. \end{aligned}$$

Similarly to the approach in [32] and [22], we introduce $\Lambda_\varepsilon : S^h \to [L^\infty(\Omega)]^{d\times d}$ such that for all $z^h \in S^h$ and a.e. in Ω,

$$\Lambda_\varepsilon(z^h) \text{ is symmetric and positive semidefinite and } \Lambda_\varepsilon(z^h)\,\nabla\pi^h[F_\varepsilon'(z^h)] = \nabla z^h. \tag{2.5}$$

Following [22], we now give the construction of Λ_ε. Let $\{e_i\}_{i=1}^d$ be the orthonormal vectors in $\mathbb{R}^d$ such that the jth component of e_i is δ_{ij}, i, $j = 1 \to d$. Given nonzero constants α_i, $i = 1 \to d$, let $\widehat{\kappa}(\{\alpha_i\}_{i=1}^d)$ be the reference open simplex in $\mathbb{R}^d$ with vertices $\{\widehat{p}_i\}_{i=0}^d$, where $\widehat{p}_0$ is the origin and $\widehat{p}_i = \alpha_i\, e_i$, $i = 1 \to d$. Given a $\kappa \in \mathcal{T}^h$ with vertices $\{p_{j_i}\}_{i=0}^d$ such that p_{j_0} is the right-angled vertex, there exist a rotation matrix R_κ and nonzero constants $\{\alpha_i\}_{i=1}^d$ such that the mapping $\mathcal{R}_\kappa : \widehat{x} \in \mathbb{R}^d \to p_{j_0} + R_\kappa \widehat{x} \in \mathbb{R}^d$ maps the vertex $\widehat{p}_i$ to p_{j_i}, $i = 0 \to d$, and hence $\widehat{\kappa} \equiv \widehat{\kappa}(\{\alpha_i\}_{i=1}^d)$ to κ. For any $z^h \in S^h$, we then set

$$\Lambda_\varepsilon(z^h)\,|_\kappa := R_\kappa\, \widehat{\Lambda}_\varepsilon(\widehat{z}^h)\,|_{\widehat{\kappa}}\, R_\kappa^T, \tag{2.6}$$

where $\widehat{z}^h(\widehat{x}) \equiv z^h(\mathcal{R}_\kappa \widehat{x})$ for all $\widehat{x} \in \overline{\widehat{\kappa}}$ and $\widehat{\Lambda}_\varepsilon(\widehat{z}^h)\,|_{\widehat{\kappa}}$ is the $d \times d$ diagonal matrix with diagonal entries, $k = 1 \to d$,

$$[\widehat{\Lambda}_\varepsilon(\widehat{z}^h)]_{kk}\,|_{\widehat{\kappa}} := \begin{cases} \frac{\widehat{z}^h(\widehat{p}_k) - \widehat{z}^h(\widehat{p}_0)}{F'_\varepsilon(\widehat{z}^h(\widehat{p}_k)) - F'_\varepsilon(\widehat{z}^h(\widehat{p}_0))} \equiv \frac{z^h(p_{j_k}) - z^h(p_{j_0})}{F'_\varepsilon(z^h(p_{j_k})) - F'_\varepsilon(z^h(p_{j_0}))} & \text{if } z^h(p_{j_k}) \neq z^h(p_{j_0}), \\ \frac{1}{F''_\varepsilon(\widehat{z}^h(\widehat{p}_0))} \equiv \frac{1}{F''_\varepsilon(z^h(p_{j_0}))} & \text{if } z^h(p_{j_k}) = z^h(p_{j_0}). \end{cases} \tag{2.7}$$

As $R_\kappa^T \equiv R_\kappa^{-1}$, $\nabla z^h \equiv R_\kappa\, \widehat{\nabla} \widehat{z}^h$, where $x \equiv (x_1, \ldots, x_d)^T$, $\nabla \equiv (\frac{\partial}{\partial x_1}, \ldots, \frac{\partial}{\partial x_d})^T$, $\widehat{x} \equiv (\widehat{x}_1, \ldots, \widehat{x}_d)^T$, and $\widehat{\nabla} \equiv (\frac{\partial}{\partial \widehat{x}_1}, \ldots, \frac{\partial}{\partial \widehat{x}_d})^T$, it easily follows that $\Lambda_\varepsilon(z^h)$ constructed in (2.6) and (2.7) satisfies (2.5). It is this construction that requires the right angle constraint on the partitioning $\mathcal{T}^h$. Furthermore, we note from (2.3) that for all $\kappa \in \mathcal{T}^h$,

$$\begin{aligned} \nabla z^h . \nabla \pi^h[F'_\varepsilon(z^h)]\,|_\kappa &\equiv R_\kappa \widehat{\nabla} \widehat{z}^h . R_\kappa \widehat{\nabla} \widehat{\pi}^h[F'_\varepsilon(\widehat{z}^h)]\,|_{\widehat{\kappa}} \\ &\equiv \widehat{\nabla} \widehat{z}^h . \widehat{\nabla} \widehat{\pi}^h[F'_\varepsilon(\widehat{z}^h)]\,|_{\widehat{\kappa}} \geq |\widehat{\nabla} \widehat{z}^h|^2\,|_{\widehat{\kappa}} \equiv |\nabla z^h|^2\,|_\kappa \\ \Longrightarrow \quad (\nabla z^h, \nabla \pi^h[F'_\varepsilon(z^h)]) &\geq |z^h|_1^2 \qquad \forall\ z^h \in S^h, \end{aligned} \tag{2.8}$$

where $(\widehat{\pi}^h \widehat{\eta})(\widehat{x}) \equiv (\pi^h \eta)(\mathcal{R}_\kappa \widehat{x})$ and $\widehat{\eta}(\widehat{x}) \equiv \eta(\mathcal{R}_\kappa \widehat{x})$ for all $\widehat{x} \in \overline{\widehat{\kappa}}$.

To define our approximation of (P), it is convenient to split Φ (recall (1.11)) into its convex and concave parts. We have for given a, $\delta \in \mathbb{R}_{\geq 0}$, and $\nu > 3$ that for all $s \in \mathbb{R}_{>0}$

$$\begin{aligned} \Phi(s) &= \Phi^+(s) + \Phi^-(s), && \text{where} \quad \Phi^+(s) := \tfrac{\delta}{\nu - 1}\, s^{1-\nu}, \quad \Phi^-(s) := -\tfrac{a}{2}\, s^{-2}, \\ \phi(s) &= \Phi'(s) = \phi^+(s) + \phi^-(s), && \text{where} \quad \phi^+(s) := (\Phi^+)'(s) = -\delta\, s^{-\nu}, \\ & && \phantom{\text{where}} \quad \phi^-(s) := (\Phi^-)'(s) = a\, s^{-3}. \end{aligned} \tag{2.9}$$

For future reference, we note that the following hold for all $s \in \mathbb{R}_{>0}$:

$$\Phi(s) \geq \Phi(\,(\tfrac{\delta}{a})^{\frac{1}{\nu-3}}) = \tfrac{a\,(3-\nu)}{2\,(\nu-1)} (\tfrac{a}{\delta})^{\frac{2}{\nu-3}} \quad \text{and} \quad |\Phi^-(s)| \leq \tfrac{a\,(\nu-3)}{2\,(\nu-1)} (\tfrac{2\,a}{\delta})^{\frac{2}{\nu-3}} + \tfrac{1}{2}\, \Phi^+(s). \tag{2.10}$$

In addition to $\mathcal{T}^h$, let $0 = t_0 < t_1 < \cdots < t_{N-1} < t_N = T$ be a partitioning of $[0, T]$ into possibly variable time steps $\tau_n := t_n - t_{n-1}$, $n = 1 \to N$. We set $\tau := \max_{n=1 \to N} \tau_n$. For any given $\varepsilon \in (0, 1)$, we then consider the following fully practical finite element approximation of (P) with $\sigma(v) := 1 - v$ and $\delta = 0\,(\phi^+ \equiv 0)$.

$(\mathrm{P}_{\varepsilon}^{h,\tau})$ For $n \geq 1$ find $\{U_{\varepsilon}^{n}, W_{\varepsilon}^{n}, V_{\varepsilon}^{n}\} \in K^h \times [S^h]^2$ such that for all $\chi \in S^h$, $z^h \in K^h$,

$$\left(\frac{U_{\varepsilon}^{n}-U_{\varepsilon}^{n-1}}{\tau_n}, \chi\right)^h + \tfrac{1}{3}\left(\pi^h[(U_{\varepsilon}^{n-1})^3]\,\nabla W_{\varepsilon}^{n}, \nabla\chi\right) = -\tfrac{1}{2}\,(\pi^h[(U_{\varepsilon}^{n-1})^2]\nabla V_{\varepsilon}^{n-1}, \nabla\chi), \tag{2.11a}$$

$$c\,(\nabla U_{\varepsilon}^{n}, \nabla(z^h - U_{\varepsilon}^{n})) + (\phi^-(U_{\varepsilon}^{n-1}+\varepsilon), z^h - U_{\varepsilon}^{n})^h \geq (W_{\varepsilon}^{n}, z^h - U_{\varepsilon}^{n})^h, \tag{2.11b}$$

$$\begin{aligned}\left(\frac{V_{\varepsilon}^{n}-V_{\varepsilon}^{n-1}}{\tau_n}, \chi\right)^h &+ \rho\,(\nabla V_{\varepsilon}^{n}, \nabla\chi) + (U_{\varepsilon}^{n}\,\Lambda_{\varepsilon}(V_{\varepsilon}^{n})\,\nabla V_{\varepsilon}^{n}, \nabla\chi)\\ &= -\tfrac{1}{2}\,(\pi^h[(U_{\varepsilon}^{n})^{\frac{1}{2}}\,(U_{\varepsilon}^{n-1})^{\frac{3}{2}}]\,\Lambda_{\varepsilon}(V_{\varepsilon}^{n})\,\nabla W_{\varepsilon}^{n}, \nabla\chi),\end{aligned} \tag{2.11c}$$

where $U_{\varepsilon}^{0} \in K^h$ and $V_{\varepsilon}^{0} \in S^h$ are approximations of u^0 and v^0, respectively; e.g., $U_{\varepsilon}^{0} \equiv \pi^h u^0$ or $Q^h u^0$ and similarly V_{ε}^{0}.

If $a = 0\,(\phi^- \equiv 0)$, then setting $V_{\varepsilon}^{n} \equiv 0$, $n = 0 \to N$, (2.11a,b) collapses to the approximation of the thin film equation, (1.12a–e) with $v \equiv 0$, studied in [4], except that there $\pi^h[(U_{\varepsilon}^{n-1})^3]$ in (2.11a) is replaced by $(U_{\varepsilon}^{n-1})^3$ and so is less practical than (2.11a). If $\delta > 0\,(\phi^+ \not\equiv 0)$, then $(\mathrm{P}_{\varepsilon}^{h,\tau})$ above is modified as follows.

$(\mathrm{P}_{\delta,\varepsilon}^{h,\tau})$ For $n \geq 1$ find $\{U_{\varepsilon}^{n}, W_{\varepsilon}^{n}, V_{\varepsilon}^{n}\} \in [S^h]^3$ such that (2.11a,c) hold with (2.11b) replaced by

$$c\,(\nabla U_{\varepsilon}^{n}, \nabla\chi) + (\phi^+(U_{\varepsilon}^{n}) + \phi^-(U_{\varepsilon}^{n-1}), \chi)^h = (W_{\varepsilon}^{n}, \chi)^h \quad \forall\, \chi \in S^h\,, \tag{2.12}$$

where in addition it is assumed that $U_{\varepsilon}^{0} > 0$.

Note that the convex (concave) terms in Φ are approximated implicitly (explicitly) in (2.11b) and (2.12). If $\delta = 0$, we can guarantee only that $U_{\varepsilon}^{n-1} \geq 0$ and hence the choice of $\phi^-(U_{\varepsilon}^{n-1}+\varepsilon)$ instead of $\phi^-(U_{\varepsilon}^{n-1})$ on the left-hand side of (2.11b). Whereas if $\delta > 0$ and $U_{\varepsilon}^{0} > 0$, one can ensure that $\phi^-(U_{\varepsilon}^{n-1})$ is well defined for $n \geq 1$; see Theorem 2.6 below.

Below we recall some well-known results concerning S^h for any $\kappa \in \mathcal{T}^h$, $\chi, z^h \in S^h$, $\eta_1, \eta_2 \in C(\overline{\Omega})$ and for $m = 0$ or 1:

$$\lim_{h\to 0} |(I - \pi^h)\eta_1|_{0,\infty} = 0\,, \tag{2.13}$$

$$\textstyle\int_{\kappa} \chi^2\,\mathrm{d}x \leq \int_{\kappa} \pi^h[\chi^2]\,\mathrm{d}x \leq (d+2)\,\int_{\kappa} \chi^2\,\mathrm{d}x\,, \tag{2.14}$$

$$\textstyle\int_{\kappa} \pi^h[\eta_1\eta_2]\,\nabla\chi.\nabla z^h\,\mathrm{d}x \leq \left[\int_{\kappa} \pi^h[\eta_1^2]\,|\nabla\chi|^2\,\mathrm{d}x\right]^{\frac{1}{2}}\,\left[\int_{\kappa} \pi^h[\eta_2^2]\,|\nabla z^h|^2\,\mathrm{d}x\right]^{\frac{1}{2}}\,, \tag{2.15}$$

$$|\pi^h[\eta_1\eta_2](x)\,|^2 \leq |\pi^h\eta_1|_{0,\infty}^2\,\pi^h[\eta_2^2](x) \quad \forall\, x \in \overline{\Omega}; \tag{2.16}$$

$$|(\chi, z^h) - (\chi, z^h)^h| \leq |(I - \pi^h)(\chi\, z^h)|_{0,1} \;\leq\; C\,h^{1+m}\,|\chi|_m\,|z^h|_1\,. \tag{2.17}$$

If $d = 1$, then we have for $m = 0$ or 1 that

$$|(I - \pi^h)\eta|_{m,r} \leq C\,h^{1-m}\,|\eta|_{1,r} \quad \forall\, \eta \in W^{1,r}(\Omega), \quad \text{for any } r \in [1,\infty]; \tag{2.18}$$

$$\lim_{h\to 0} \|(I - \pi^h)\eta\|_1 = 0 \qquad \forall\, \eta \in H^1(\Omega). \tag{2.19}$$

It follows from (2.2) that

$$(Q^h\eta)(p_j) = \tfrac{(\eta,\chi_j)}{(1,\chi_j)} \quad \forall\, j \in J \quad\Longrightarrow\quad |Q^h\eta|_{0,\infty} \leq |\eta|_{0,\infty} \quad \forall\, \eta \in L^{\infty}(\Omega). \tag{2.20}$$

Finally, as we have a quasi-uniform family of partitionings, it holds for $m = 0$ or 1 that

$$|(I - Q^h)\eta|_m \leq C\,h^{1-m}\,\|\eta\|_1 \quad \forall\, \eta \in H^1(\Omega). \tag{2.21}$$

We define $Z^h := \{z^h \in S^h : (z^h, 1) = 0\} \subset \mathcal{F}^h := \{z \in C(\overline{\Omega}) : (z, 1)^h = 0\}$. Then, similarly to (1.16), we introduce $\mathcal{G}^h : \mathcal{F}^h \to Z^h$ such that

$$(\nabla \mathcal{G}^h z, \nabla \chi) = (z, \chi)^h \quad \forall\, \chi \in S^h\,. \tag{2.22}$$

It is easily established, as we have a quasi-uniform family of partitionings, that

$$|z^h|_0 \leq C\, h^{-1}\, |\mathcal{G} z^h|_1 \quad \forall\, z^h \in Z^h. \tag{2.23}$$

We now adapt and extend the approach in [4] to establish the existence of a solution $\{U_\varepsilon^n, W_\varepsilon^n, V_\varepsilon^n\}_{n=1}^N$ to $(\mathrm{P}_\varepsilon^{h,\tau})$. First, we need to introduce some notation. In particular, we define sets $Z^h(U_\varepsilon^{n-1})$ in which we seek the update $U_\varepsilon^n - U_\varepsilon^{n-1}$. Given $q^h \in K^h$, we set $J_0(q^h) := \{j \in J : (\pi^h[(q^h)^3], \chi_j) = 0\}$. All other nodes we call active nodes, and they can be uniquely partitioned so that $J_+(q^h) := J \setminus J_0(q^h) = \bigcup_{m=1}^M I_m(q^h)$, $M \geq 1$, where $I_m(q^h)$, $m = 1 \to M$, are mutually disjoint and maximally connected in the following sense: $I_m(q^h)$ is said to be connected if for all j, $k \in I_m(q^h)$, there exist $\{\kappa_\ell\}_{\ell=1}^L \subseteq \mathcal{T}^h$, not necessarily distinct, such that

$$p_j \in \overline{\kappa}_1,\ p_k \in \overline{\kappa}_L; \qquad \overline{\kappa}_\ell \cap \overline{\kappa}_{\ell+1} \neq \emptyset,\ \ell = 1 \to L-1; \qquad q^h \not\equiv 0 \text{ on } \kappa_\ell,\ \ell = 1 \to L.$$

$I_m(q^h)$ is said to be maximally connected if there is no other connected subset of $J_+(q^h)$, which contains $I_m(q^h)$. We then set

$$\begin{aligned} Z^h(q^h) := \{z^h \in S^h : z^h(p_j) = 0 \ \ \forall\, j \in J_0(q^h) \quad \text{and} \\ (z^h, \Xi_m(q^h))^h = 0,\ m = 1 \to M\,\}, \end{aligned} \tag{2.24}$$

where $\Xi_m(q^h) := \sum_{j \in I_m(q^h)} \chi_j$ for $m = 1 \to M$. An immediate consequence of the above definitions is that on any $\kappa \in \mathcal{T}^h$ either

$$q^h \equiv 0 \qquad \text{or} \qquad \Xi_{m_\star}(q^h) \equiv 1 \text{ for some } m_\star \text{ and } \Xi_m(q^h) \equiv 0 \text{ for } m \neq m_\star. \tag{2.25}$$

This follows since if $q^h \not\equiv 0$ on κ, then all vertices of κ belong to the set of active nodes $J_+(q^h)$. Using the fact that $I_m(q^h)$, $m = 1 \to M$, are maximally connected, we can conclude that there exists an $m_\star$ such that all vertices of κ belong to $I_{m_\star}(q^h)$ and therefore, $\Xi_{m_\star}(q^h) \equiv 1$ on κ. The desired result now follows since $I_m(q^h)$, $m = 1 \to M$, are mutually disjoint.

For later reference, we state that any $z^h \in S^h$ can be written as

$$z^h \equiv \sum_{j \in J} z^h(p_j)\, \chi_j \ \equiv\ \overline{z}^h + \sum_{j \in J_0(q^h)} z^h(p_j)\, \chi_j + \sum_{m=1}^M \frac{(z^h, \Xi_m(q^h))^h}{(1, \Xi_m(q^h))}\, \Xi_m(q^h), \tag{2.26}$$

where $\overline{z}^h := \sum_{m=1}^M \sum_{j \in I_m(q^h)} [z^h(p_j) - \frac{(z^h, \Xi_m(q^h))^h}{(1, \Xi_m(q^h))}] \chi_j \in Z^h(q^h)$ is the projection with respect to the $(\cdot,\cdot)^h$ scalar product of z^h onto $Z^h(q^h)$. In order to express W_ε^n in terms of U_ε^n and U_ε^{n-1}, we introduce for all $q^h \in K^h$ the discrete anisotropic Green's operator $\mathcal{G}_{q^h}^h : Z^h(q^h) \to Z^h(q^h)$ such that

$$(\pi^h[(q^h)^3]\, \nabla \mathcal{G}_{q^h}^h z^h, \nabla \chi) = (z^h, \chi)^h \qquad \forall\, \chi \in S^h. \tag{2.27}$$

To show the well-posedness of $\mathcal{G}_{q^h}^h$, we first note that on choosing $\chi \equiv \chi_j$, $j \in J_0(q^h)$, in (2.27) leads to both sides vanishing on noting (2.24). Similarly, choosing $\chi \equiv \Xi_m(q^h)$,

$m = 1 \to M$, in (2.27) leads to both sides vanishing on noting (2.25) and (2.24). Therefore, for well-posedness, it remains to prove uniqueness as $Z^h(q^h)$ has finite dimension. If there exist two solutions $Z^{(i)}$, $i = 1, 2$, with $(\pi^h[(q^h)^3]\,\nabla Z^{(i)}, \nabla\chi) = (z^h, \chi)^h$ for all $\chi \in S^h$, then $Z := Z^{(2)} - Z^{(1)} \in Z^h(q^h)$ satisfies, on noting (2.25),

$$C(q^h, h) \sum_{m=1}^{M} \int_{\Omega_m} |\nabla Z|^2 \,\mathrm{d}x \le \sum_{m=1}^{M} \int_{\Omega_m} \pi^h[(q^h)^3]\,|\nabla Z|^2 \,\mathrm{d}x = \int_{\Omega} \pi^h[(q^h)^3]\,|\nabla Z|^2 \,\mathrm{d}x = 0$$

for some positive constant $C(q^h, h)$, where $\Omega_m := \{\cup_{\kappa \in \mathcal{T}^h} \overline{\kappa} : \Xi_m(q^h)|_\kappa \equiv 1\}$. Hence it follows that Z is constant on each Ω_m. However, as $Z \in Z^h(q^h)$, it follows that $Z \equiv 0$. Finally, note that $Z^h(q^h) \subseteq Z^h$ for all $q^h \in K^h$ and in addition that $Z^h(q^h)$ defined in (2.24) is equal to Z^h if q^h is strictly positive.

LEMMA 2.1. *Let the assumptions* (A) *hold, and let* $\|\cdot\|$ *denote the spectral norm on* $\mathbb{R}^{d\times d}$. *Then for any given* $\varepsilon \in (0, 1)$, *the function* $\Lambda_\varepsilon : S^h \to [L^\infty(\Omega)]^{d\times d}$ *satisfies*

$$\varepsilon\,\xi^T\xi \le \xi^T \Lambda_\varepsilon(z^h)\xi \le \xi^T\xi \qquad \forall\ \xi \in \mathbb{R}^d, \quad \forall\ z^h \in S^h \tag{2.28}$$

and is continuous. In particular, it holds for all z_1^h, $z_2^h \in S^h$, $\kappa \in \mathcal{T}^h$ *that*

$$\begin{aligned}
\|[\Lambda_\varepsilon(z_1^h) - \Lambda_\varepsilon(z_2^h)]\,|_\kappa\,\| &= \|[\widehat{\Lambda}_\varepsilon(\widehat{z}_1^h) - \widehat{\Lambda}_\varepsilon(\widehat{z}_2^h)]\,|_{\widehat{\kappa}}\,\| \\
&\le \max_{s\in\mathbb{R}} F_\varepsilon''(s)\ \max_{s\in\mathbb{R}} [F_\varepsilon''(s)]^{-1} \max_{k=1\to d} \left[\,|z_1^h(p_{j_k}) - z_2^h(p_{j_k})| + |z_1^h(p_{j_0}) - z_2^h(p_{j_0})|\,\right] \\
&\le \varepsilon^{-1} \max_{k=1\to d} \left[\,|z_1^h(p_{j_k}) - z_2^h(p_{j_k})| + |z_1^h(p_{j_0}) - z_2^h(p_{j_0})|\,\right],
\end{aligned} \tag{2.29}$$

where we have adopted the notation of (2.6) *and* (2.7).

Proof. It follows immediately from (2.6), (2.7), and (2.3) that (2.28) holds. The proof of (2.29) is a straightforward adaptation of the proof of Lemma 2.1 in [7], where a similar inequality is shown for a slightly modified F_ε. □

THEOREM 2.2. *Let the assumptions* (A) *hold and* $U_\varepsilon^{n-1} \in K^h$, $V_\varepsilon^{n-1} \in S^h$. *Then for all* $\varepsilon \in (0, 1)$ *and for all* h, $\tau_n > 0$, *there exists a solution* $\{U_\varepsilon^n, W_\varepsilon^n, V_\varepsilon^n\}$ *to the* n*th step of* $(\mathrm{P}_\varepsilon^{h,\tau})$ *with* $⨍ U_\varepsilon^n = ⨍ U_\varepsilon^{n-1}$ *and* $⨍ V_\varepsilon^n = ⨍ V_\varepsilon^{n-1}$. *Moreover,* U_ε^n *is unique. In addition,* $W^n(p_j)$ *is unique if* $(\pi^h[(U_\varepsilon^{n-1})^3], \chi_j) > 0$ *for all* $j \in J$.

Proof. For $n \ge 1$, given $U_\varepsilon^{n-1} \in K^h$, $V_\varepsilon^{n-1} \in S^h$, we define $X_\varepsilon^{n-1} \in Z^h(U_\varepsilon^{n-1})$ such that

$$(X_\varepsilon^{n-1}, \chi)^h := \tfrac{1}{2}\,(\pi^h[(U_\varepsilon^{n-1})^2]\,\nabla V_\varepsilon^{n-1}, \nabla\chi) \qquad \forall\ \chi \in S^h\,. \tag{2.30}$$

It follows from (2.11a), (2.27), and (2.30) that we seek $U_\varepsilon^n \in K^h(U_\varepsilon^{n-1})$, where for all $q^h \in K^h$

$$S^h(q^h) := \{\,\chi \in S^h : \chi - q^h \in Z^h(q^h)\,\} \qquad \text{and} \qquad K^h(q^h) := S^h(q^h) \cap K^h. \tag{2.31}$$

In addition, we have that (cf. (2.26))

$$W_\varepsilon^n \equiv -3\,\mathcal{G}^h_{U_\varepsilon^{n-1}} \left[\tfrac{U_\varepsilon^n - U_\varepsilon^{n-1}}{\tau_n} + X_\varepsilon^{n-1}\right] + \sum_{j\in J_0(U_\varepsilon^{n-1})} \alpha_j^n\,\chi_j + \sum_{m=1}^{M} \beta_m^n\,\Xi_m(U_\varepsilon^{n-1}), \tag{2.32}$$

where $\{\alpha_j^n\}_{j\in J_0(U_\varepsilon^{n-1})}$ and $\{\beta_m^n\}_{m=1}^M$ are arbitrary constants. Hence (2.11a) and (2.11b) can be restated as follows.

For $n \geq 1$, find $U_\varepsilon^n \in K^h(U_\varepsilon^{n-1})$ and constant Lagrange multipliers $\{\alpha_j^n\}_{j\in J_0(U_\varepsilon^{n-1})}$, $\{\beta_m^n\}_{m=1}^M$ such that

$$\begin{aligned} & c\,(\nabla U_\varepsilon^n, \nabla(\chi - U_\varepsilon^n)) + 3\,\left(\mathcal{G}^h_{U_\varepsilon^{n-1}}\left[\tfrac{U_\varepsilon^n - U_\varepsilon^{n-1}}{\tau_n}\right], \chi - U_\varepsilon^n\right)^h \\ & \qquad \geq \left(\sum_{j\in J_0(U_\varepsilon^{n-1})} \alpha_j^n\,\chi_j + \sum_{m=1}^M \beta_m^n\,\Xi_m(U_\varepsilon^{n-1}) + \overline{X}_\varepsilon^{n-1}, \chi - U_\varepsilon^n\right)^h \quad \forall\, \chi \in K^h, \end{aligned} \tag{2.33}$$

where $\overline{X}_\varepsilon^{n-1} \in S^h$ is such that

$$(\overline{X}_\varepsilon^{n-1}, \chi)^h := -(\phi^-(U_\varepsilon^{n-1} + \varepsilon) + 3\,\mathcal{G}^h_{U_\varepsilon^{n-1}}\,X_\varepsilon^{n-1}, \chi)^h \qquad \forall\, \chi \in S^h\,.$$

It follows from (2.33), (2.31), and (2.24) that $U_\varepsilon^n \in K^h(U_\varepsilon^{n-1})$ is such that

$$A_{U_\varepsilon^{n-1}}(U_\varepsilon^n, \widetilde{z}^h - U_\varepsilon^n) \geq (\overline{X}_\varepsilon^{n-1}, \widetilde{z}^h - U_\varepsilon^n)^h \qquad \forall\, \widetilde{z}^h \in K^h(U_\varepsilon^{n-1}), \tag{2.34}$$

where $A_{U_\varepsilon^{n-1}} : S^h(U_\varepsilon^{n-1}) \times S^h \to \mathbb{R}$ is defined by

$$A_{U_\varepsilon^{n-1}}(z^h, \chi) := c\,(\nabla z^h, \nabla \chi) + 3\,(\mathcal{G}^h_{U_\varepsilon^{n-1}}[\tfrac{z^h - U_\varepsilon^{n-1}}{\tau_n}], \chi)^h \quad \forall\, z^h \in S^h(U_\varepsilon^{n-1}),\ \chi \in S^h. \tag{2.35}$$

There exists $U_\varepsilon^n \in K^h(U_\varepsilon^{n-1})$ satisfying (2.34) since, on noting (2.27), this is the Euler–Lagrange variational inequality of the convex minimization problem

$$\min_{\widetilde{z}^h \in K^h(U_\varepsilon^{n-1})} \left\{ \tfrac{c}{2}\,|\widetilde{z}^h|_1^2 + \tfrac{3}{2\tau_n}\,|[\pi^h[(U_\varepsilon^{n-1})^3]]^{\frac{1}{2}}\,\nabla\mathcal{G}^h_{U_\varepsilon^{n-1}}(\widetilde{z}^h - U_\varepsilon^{n-1})|_0^2 - (\overline{X}_\varepsilon^{n-1}, \widetilde{z}^h)^h \right\}. \tag{2.36}$$

Furthermore, given any $z^h \in K^h$, similarly to (2.26) there exists a $\zeta \in \mathbb{R}_{>0}$ such that

$$\begin{aligned} \widetilde{z}^h &:= U_\varepsilon^n + \zeta\,\left\{ (z^h - U_\varepsilon^n) - \sum_{j\in J_0(U_\varepsilon^{n-1})} (z^h - U_\varepsilon^n)(p_j)\,\chi_j \right. \\ & \qquad\qquad \left. - \sum_{m=1}^M \frac{(z^h - U_\varepsilon^n, \Xi_m(U_\varepsilon^{n-1}))^h}{(U_\varepsilon^n, \Xi_m(U_\varepsilon^{n-1}))^h}\,\pi^h[U_\varepsilon^n\,\Xi_m(U_\varepsilon^{n-1})] \right\} \\ & \equiv \pi^h\left[\left(1 - \zeta\left(1 + \sum_{m=1}^M \frac{(z^h - U_\varepsilon^n, \Xi_m(U_\varepsilon^{n-1}))^h}{(U_\varepsilon^n, \Xi_m(U_\varepsilon^{n-1}))^h}\,\Xi_m(U_\varepsilon^{n-1})\right)\right)\,U_\varepsilon^n\right] \\ & \qquad\qquad + \ \zeta\,\left(z^h - \sum_{j\in J_0(U_\varepsilon^{n-1})} z^h(p_j)\,\chi_j\right) \in K^h(U_\varepsilon^{n-1}). \end{aligned} \tag{2.37}$$

Here we have used that $\Xi_m(U_\varepsilon^{n-1})(p_j) = U_\varepsilon^n(p_j) = 0$ for all $j \in J_0(U_\varepsilon^{n-1})$, and $(\pi^h[U_\varepsilon^n\,\Xi_m(U_\varepsilon^{n-1})], \Xi_m(U_\varepsilon^{n-1}))^h = (U_\varepsilon^n, \Xi_m(U_\varepsilon^{n-1}))^h = (U_\varepsilon^{n-1}, \Xi_m(U_\varepsilon^{n-1}))^h > 0$ for

$m = 1 \to M$. For all $z^h \in K^h$, choosing $\widetilde{z}^h \in K^h(U_\varepsilon^{n-1})$ (as constructed in (2.37)) in (2.34) yields the existence of a solution to (2.33) with

$$\alpha_j^n = \frac{A_{U_\varepsilon^{n-1}}(U_\varepsilon^n, \chi_j) - (\overline{X}_\varepsilon^{n-1}, \chi_j)^h}{(1, \chi_j)} \qquad \forall\ j \in J_0(U_\varepsilon^{n-1})$$

and

$$\beta_m^n = \frac{A_{U_\varepsilon^{n-1}}(U_\varepsilon^n, \pi^h[U_\varepsilon^n\, \Xi_m(U_\varepsilon^{n-1})]) - (\overline{X}_\varepsilon^{n-1}, U_\varepsilon^n\, \Xi_m(U_\varepsilon^{n-1}))^h}{(U_\varepsilon^n, \Xi_m(U_\varepsilon^{n-1}))^h} \quad m = 1 \to M.$$

Therefore, on noting (2.32), we have the existence of a solution $\{U_\varepsilon^n, W_\varepsilon^n\}$ to $(\mathrm{P}_\varepsilon^{h,\tau})$ with $⨍ U_\varepsilon^n = ⨍ U_\varepsilon^{n-1}$.

To prove the existence of V_ε^n, we will make use of the *Brouwer fixed point theorem* (see, e.g., [29, Theorem 9.36, p. 357]). Let $\mathcal{J} := \#J$, and let $g : \mathbb{R}^{\mathcal{J}} \to \mathbb{R}^{\mathcal{J}}$ be defined by

$$\begin{aligned} g_j(\underline{V}) :=& (V, \chi_j)^h + \rho\, \tau_n\, (\nabla V, \nabla \chi_j) + \tau_n\, (U_\varepsilon^n\, \Lambda_\varepsilon(V)\, \nabla V, \nabla \chi_j) \\ &+ \tfrac{\tau_n}{2}\, (\pi^h[(U_\varepsilon^n)^{\frac{1}{2}}\, (U_\varepsilon^{n-1})^{\frac{3}{2}}]\, \Lambda_\varepsilon(V)\, \nabla W_\varepsilon^n, \nabla \chi_j) \qquad \forall\ j \in J\,, \end{aligned}$$

where $V \equiv \sum_{j \in J} V_j\, \chi_j$ and $\underline{V} := (V_1, \ldots, V_{\mathcal{J}})^T \in \mathbb{R}^{\mathcal{J}}$. Noting Lemma 2.1, we have that g is continuous, and hence it is sufficient to show that g is coercive. We have that

$$\begin{aligned} \sum_{j \in J} g_j(\underline{V})\, V_j =& |V|_h^2 + \rho\, \tau_n\, |V|_1^2 + \tau_n\, (U_\varepsilon^n\, \Lambda_\varepsilon(V)\, \nabla V, \nabla V) \\ &+ \tfrac{\tau_n}{2}\, (\pi^h[(U_\varepsilon^n)^{\frac{1}{2}}\, (U_\varepsilon^{n-1})^{\frac{3}{2}}]\, \Lambda_\varepsilon(V)\, \nabla W_\varepsilon^n, \nabla V) \quad \forall\ V \in S^h\,. \end{aligned} \tag{2.38}$$

From (1.19), (2.15), and (2.28), we have

$$\begin{aligned} \tfrac{\tau_n}{2} & \left| (\pi^h[(U_\varepsilon^n)^{\frac{1}{2}}\, (U_\varepsilon^{n-1})^{\frac{3}{2}}]\, \Lambda_\varepsilon(V)\, \nabla W_\varepsilon^n, \nabla V) \right| \\ &\leq \tfrac{\tau_n}{2}\, (U_\varepsilon^n\, \Lambda_\varepsilon(V)\, \nabla V, \nabla V) + \tfrac{\tau_n}{8}\, (\pi^h[(U_\varepsilon^{n-1})^3]\, \Lambda_\varepsilon(V)\, \nabla W_\varepsilon^n, \nabla W_\varepsilon^n) \\ &\leq \tfrac{\tau_n}{2}\, (U_\varepsilon^n\, \Lambda_\varepsilon(V)\, \nabla V, \nabla V) + C(\tau_n,\, U_\varepsilon^{n-1},\, W_\varepsilon^n)\,. \end{aligned} \tag{2.39}$$

It follows from (2.38), (2.39), and (2.28) that

$$\sum_{j \in J} g_j(\underline{V})\, V_j \geq |V|_h^2 - C(\tau_n,\, U_\varepsilon^{n-1},\, W_\varepsilon^n) \qquad \forall\ V \in S^h\,. \tag{2.40}$$

Hence the coerciveness of g follows from (2.40) and (2.1). Therefore, on noting the aforementioned theorem, we have the existence of V_ε^n to (2.11c) and hence the existence of a solution $\{U_\varepsilon^n, W_\varepsilon^n, V_\varepsilon^n\}$ to $(\mathrm{P}_\varepsilon^{h,\tau})$. Choosing $\chi \equiv 1$ in (2.11c) yields that $⨍ V_\varepsilon^n = ⨍ V_\varepsilon^{n-1}$.

If (2.33) has two solutions $\{\, U_\varepsilon^{n,i},\ \{\alpha_j^{n,i}\}_{j \in J_0(U_\varepsilon^{n-1})},\ \{\beta_m^{n,i}\}_{m=1}^M \,\}$, $i = 1, 2$, then it follows from (2.34) and (2.27) that $\widetilde{U}_\varepsilon^n := U_\varepsilon^{n,1} - U_\varepsilon^{n,2} \in Z^h(U_\varepsilon^{n-1})$ satisfies

$$c\, |\widetilde{U}_\varepsilon^n|_1^2 + 3\, \tau_n^{-1}\, |[\pi^h[(U_\varepsilon^{n-1})^3]\,]^{\frac{1}{2}}\, \nabla(\mathcal{G}_{U_\varepsilon^{n-1}}^h \widetilde{U}_\varepsilon^n)|_0^2 \leq 0.$$

Therefore, the uniqueness of U^n_ε follows from (1.17). For any $\zeta \in (0,1)$, choosing $\chi \equiv U^n_\varepsilon \pm \zeta\, \pi^h[U^n_\varepsilon\, \Xi_m(U^{n-1}_\varepsilon)] \equiv \pi^h[\,(1 \pm \zeta\, \Xi_m(U^{n-1}_\varepsilon))\, U^n_\varepsilon\,]$ in (2.33) for $m = 1 \to M$ yields the uniqueness of the Lagrange multipliers $\{\beta^n_m\}^M_{m=1}$. Hence the desired uniqueness result on W^n_ε follows from noting (2.32). □

LEMMA 2.3. *Let the assumptions of Theorem* 2.2 *hold. Then for all* $\varepsilon \in (0,1)$ *and for all* h, $\tau_n > 0$ *a solution* $\{U^n_\varepsilon, W^n_\varepsilon, V^n_\varepsilon\}$ *to the nth step of* $(\mathrm{P}^{h,\tau}_\varepsilon)$ *is such that*

$$\begin{aligned}\mathcal{E}(U^n_\varepsilon, V^n_\varepsilon) + \tfrac{c}{2}\, |U^n_\varepsilon - U^{n-1}_\varepsilon|^2_1 + \tfrac{1}{2}\, |V^n_\varepsilon - V^{n-1}_\varepsilon|^2_h + \rho\, \tau_n\, (\nabla V^n_\varepsilon, \nabla \pi^h[F'_\varepsilon(V^n_\varepsilon)]) \\ + \tfrac{\tau_n}{24}\, (\pi^h[(U^{n-1}_\varepsilon)^3]\, \nabla W^n_\varepsilon, \nabla W^n_\varepsilon) + \tfrac{5}{8}\, \tau_n\, (U^n_\varepsilon\, \nabla V^n_\varepsilon, \nabla V^n_\varepsilon) \\ \leq \mathcal{E}(U^{n-1}_\varepsilon, V^{n-1}_\varepsilon) + \tfrac{\tau_n}{2}\, (U^{n-1}_\varepsilon\, \nabla V^{n-1}_\varepsilon, \nabla V^{n-1}_\varepsilon)\,, \end{aligned} \tag{2.41}$$

where

$$\mathcal{E}(U^n_\varepsilon, V^n_\varepsilon) := \tfrac{c}{2}\, |U^n_\varepsilon|^2_1 + (F_\varepsilon(V^n_\varepsilon) + \Phi^-(U^n_\varepsilon + \varepsilon), 1)^h\,. \tag{2.42}$$

Proof. Choosing $\chi \equiv W^n_\varepsilon$ in (2.11a), $z^h \equiv U^{n-1}_\varepsilon$ in (2.11b), and $\chi \equiv \pi^h[F'_\varepsilon(V^n_\varepsilon)]$ in (2.11c) and noting (2.5) yield that

$$\begin{aligned}(U^n_\varepsilon - U^{n-1}_\varepsilon, W^n_\varepsilon)^h + \tfrac{\tau_n}{3}\, (\pi^h[(U^{n-1}_\varepsilon)^3]\, \nabla W^n_\varepsilon, \nabla W^n_\varepsilon) \\ = -\tfrac{\tau_n}{2}\, (\pi^h[(U^{n-1}_\varepsilon)^2]\, \nabla V^{n-1}_\varepsilon, \nabla W^n_\varepsilon)\,,\end{aligned} \tag{2.43a}$$

$$\begin{aligned}c\, (\nabla U^n_\varepsilon, \nabla (U^n_\varepsilon - U^{n-1}_\varepsilon)) + (\phi^-(U^{n-1}_\varepsilon + \varepsilon), U^n_\varepsilon - U^{n-1}_\varepsilon)^h \\ \leq (W^n_\varepsilon, U^n_\varepsilon - U^{n-1}_\varepsilon)^h\,,\end{aligned} \tag{2.43b}$$

$$\begin{aligned}(V^n_\varepsilon - V^{n-1}_\varepsilon, F'_\varepsilon(V^n_\varepsilon))^h + \rho\, \tau_n\, (\nabla V^n_\varepsilon, \nabla \pi^h[F'_\varepsilon(V^n_\varepsilon)]) \\ + \tau_n\, (U^n_\varepsilon\, \nabla V^n_\varepsilon, \nabla V^n_\varepsilon) \\ = -\tfrac{\tau_n}{2}\, (\pi^h[(U^n_\varepsilon)^{\frac{1}{2}}\, (U^{n-1}_\varepsilon)^{\frac{3}{2}}]\, \nabla W^n_\varepsilon, \nabla V^n_\varepsilon)\,.\end{aligned} \tag{2.43c}$$

On noting the elementary identity

$$2\, r\, (r - s) = (r^2 - s^2) + (r - s)^2 \quad \forall\ r, s \in \mathbb{R} \tag{2.44}$$

and the concavity of Φ^-, it follows from (2.43b) that

$$\begin{aligned}\tfrac{c}{2}\, |U^n_\varepsilon|^2_1 + \tfrac{c}{2}\, |U^n_\varepsilon - U^{n-1}_\varepsilon|^2_1 + (\Phi^-(U^n_\varepsilon + \varepsilon), 1)^h \\ \leq \tfrac{c}{2}\, |U^{n-1}_\varepsilon|^2_1 + (\Phi^-(U^{n-1}_\varepsilon + \varepsilon), 1)^h + (W^n_\varepsilon, U^n_\varepsilon - U^{n-1}_\varepsilon)^h\,.\end{aligned} \tag{2.45}$$

Combining (2.43a) and (2.45) yields that

$$\begin{aligned}\tfrac{c}{2}\, |U^n_\varepsilon|^2_1 + (\Phi^-(U^n_\varepsilon + \varepsilon), 1)^h + \tfrac{\tau_n}{3}\, (\pi^h[(U^{n-1}_\varepsilon)^3]\, \nabla W^n_\varepsilon, \nabla W^n_\varepsilon) + \tfrac{c}{2}\, |U^n_\varepsilon - U^{n-1}_\varepsilon|^2_1 \\ \leq \tfrac{c}{2}\, |U^{n-1}_\varepsilon|^2_1 + (\Phi^-(U^{n-1}_\varepsilon + \varepsilon), 1)^h - \tfrac{\tau_n}{2}\, (\pi^h[(U^{n-1}_\varepsilon)^2]\, \nabla V^{n-1}_\varepsilon, \nabla W^n_\varepsilon)\,.\end{aligned} \tag{2.46}$$

Now $F''_\varepsilon \geq 1$ implies that

$$(V^n_\varepsilon - V^{n-1}_\varepsilon, F'_\varepsilon(V^n_\varepsilon))^h \geq (F_\varepsilon(V^n_\varepsilon) - F_\varepsilon(V^{n-1}_\varepsilon), 1)^h + \tfrac{1}{2}\, |V^{n-1}_\varepsilon - V^n_\varepsilon|^2_h\,. \tag{2.47}$$

Combining (2.43c), (2.46), and (2.47) and noting (1.19), (2.15), and (2.42) yield that

$$\begin{aligned}\mathcal{E}(U^n_\varepsilon, V^n_\varepsilon) + \tfrac{c}{2}\, |U^n_\varepsilon - U^{n-1}_\varepsilon|^2_1 + \tfrac{1}{2}\, |V^n_\varepsilon - V^{n-1}_\varepsilon|^2_h + \rho\, \tau_n\, (\nabla V^n_\varepsilon, \nabla \pi^h[F'_\varepsilon(V^n_\varepsilon)]) \\ + \tfrac{\tau_n}{3}\, (\pi^h[(U^{n-1}_\varepsilon)^3]\, \nabla W^n_\varepsilon, \nabla W^n_\varepsilon) + \tau_n\, (U^n_\varepsilon\, \nabla V^n_\varepsilon, \nabla V^n_\varepsilon) \\ \leq \mathcal{E}(U^{n-1}_\varepsilon, V^{n-1}_\varepsilon) - \tfrac{\tau_n}{2}\, (\pi^h[(U^n_\varepsilon)^{\frac{1}{2}}\, (U^{n-1}_\varepsilon)^{\frac{3}{2}}]\, \nabla W^n_\varepsilon, \nabla V^n_\varepsilon) \\ - \tfrac{\tau_n}{2}\, (\pi^h[(U^{n-1}_\varepsilon)^2]\, \nabla W^n_\varepsilon, \nabla V^{n-1}_\varepsilon) \\ \leq \mathcal{E}(U^{n-1}_\varepsilon, V^{n-1}_\varepsilon) + \tfrac{\zeta + \gamma}{4}\, \tau_n\, (\pi^h[(U^{n-1}_\varepsilon)^3]\, \nabla W^n_\varepsilon, \nabla W^n_\varepsilon) \\ + \tfrac{\tau_n}{4\zeta}\, (U^{n-1}_\varepsilon\, \nabla V^{n-1}_\varepsilon, \nabla V^{n-1}_\varepsilon) + \tfrac{\tau_n}{4\gamma}\, (U^n_\varepsilon\, \nabla V^n_\varepsilon, \nabla V^n_\varepsilon)\end{aligned}$$

for arbitrary $\zeta, \gamma \in \mathbb{R}_{>0}$. Choosing $\zeta = \frac{1}{2}$ and $\gamma = \frac{2}{3}$ leads to the desired result (2.41). □

THEOREM 2.4. *Let the assumptions* (A) *hold and $U^0_\varepsilon \in K^h$, $V^0_\varepsilon \in S^h$. Then for all $\varepsilon \in (0,1)$, $h > 0$ a solution $\{U^n_\varepsilon, W^n_\varepsilon, V^n_\varepsilon\}_{n=1}^N$ to* $(\mathrm{P}^{h,\tau}_\varepsilon)$ *with $a = 0\,(\phi^- \equiv 0)$ is such that $⨍ U^n_\varepsilon = ⨍ U^0_\varepsilon$ and $⨍ V^n_\varepsilon = ⨍ V^0_\varepsilon$, and if $\tau_n \leq \frac{5}{4}\,\omega\,\tau_{n-1}$, $n = 2 \to N$, for an $\omega \in (0,1)$, then*

$$
\begin{aligned}
&c \max_{1\leq n\leq N} \|U^n_\varepsilon\|_1^2 + \max_{1\leq n\leq N} (F_\varepsilon(V^n_\varepsilon), 1)^h + \max_{1\leq n\leq N} |V^n_\varepsilon|_0^2 + \varepsilon^{-1} \max_{1\leq n\leq N} |\pi^h[V^n_\varepsilon]_-|_0^2 \\
&\quad + c\sum_{n=1}^N \|U^n_\varepsilon - U^{n-1}_\varepsilon\|_1^2 + \sum_{n=1}^N |V^n_\varepsilon - V^{n-1}_\varepsilon|_0^2 + \rho \sum_{n=1}^N \tau_n (\nabla V^n_\varepsilon, \nabla \pi^h[F'_\varepsilon(V^n_\varepsilon)]) \\
&\quad + \sum_{n=1}^N \tau_n\,(\pi^h[(U^{n-1}_\varepsilon)^3]\,\nabla W^n_\varepsilon, \nabla W^n_\varepsilon) + (1-\omega) \sum_{n=1}^N \tau_n\,(U^n_\varepsilon\,\nabla V^n_\varepsilon, \nabla V^n_\varepsilon) \\
&\quad + \rho \sum_{n=1}^N \tau_n \|V^n_\varepsilon\|_1^2 \;\leq C\left[\,1 + \|U^0_\varepsilon\|_1^2 + (U^0_\varepsilon\,\nabla V^0_\varepsilon, \nabla V^0_\varepsilon) + (F_\varepsilon(V^0_\varepsilon), 1)^h\,\right],
\end{aligned}
\tag{2.48a}
$$

$$
\begin{aligned}
&\sum_{n=1}^N \tau_n \left|\mathcal{G}[\tfrac{U^n_\varepsilon - U^{n-1}_\varepsilon}{\tau_n}]\right|_1^2 + \sum_{n=1}^N \tau_n \left|\mathcal{G}[\tfrac{V^n_\varepsilon - V^{n-1}_\varepsilon}{\tau_n}]\right|_1^2 \\
&\quad \leq C(\max_{n=0\to N} \|U^n_\varepsilon\|_{0,\infty})\left[\,1 + \|U^0_\varepsilon\|_1^2 + (U^0_\varepsilon\,\nabla V^0_\varepsilon, \nabla V^0_\varepsilon) + (F_\varepsilon(V^0_\varepsilon), 1)^h\,\right].
\end{aligned}
\tag{2.48b}
$$

Proof. Summing (2.41) from $n = 1 \to k$ and observing that $\tau_n \leq \frac{5}{4}\,\omega\,\tau_{n-1}$, $n = 2 \to k$, yield for any $k \leq N$ that

$$
\begin{aligned}
&\mathcal{E}(U^k_\varepsilon, V^k_\varepsilon) + \tfrac{1}{2} \sum_{n=1}^k \left[c\,|U^n_\varepsilon - U^{n-1}_\varepsilon|_1^2 + |V^n_\varepsilon - V^{n-1}_\varepsilon|_h^2\right] + \rho \sum_{n=1}^k \tau_n (\nabla V^n_\varepsilon, \nabla \pi^h[F'_\varepsilon(V^n_\varepsilon)]) \\
&\quad + \tfrac{1}{24} \sum_{n=1}^k \tau_n\,(\pi^h[(U^{n-1}_\varepsilon)^3]\,\nabla W^n_\varepsilon, \nabla W^n_\varepsilon) + \tfrac{5}{8}\,(1-\omega) \sum_{n=1}^k \tau_n\,(U^n_\varepsilon\,\nabla V^n_\varepsilon, \nabla V^n_\varepsilon) \\
&\quad \leq \mathcal{E}(U^0_\varepsilon, V^0_\varepsilon) + \tfrac{\tau_1}{2}\,(U^0_\varepsilon\,\nabla V^0_\varepsilon, \nabla V^0_\varepsilon)\,.
\end{aligned}
\tag{2.49}
$$

As $a = 0$, we have that

$$
\mathcal{E}(U^n_\varepsilon, V^n_\varepsilon) = \tfrac{c}{2}\,|U^n_\varepsilon|_1^2 + (F_\varepsilon(V^n_\varepsilon), 1)^h \geq 0\,. \tag{2.50}
$$

Therefore, the bounds $1 \to 2$ and $5 \to 9$ in (2.48a) follow from (2.49), (2.50), $U^n_\varepsilon - U^{n-1}_\varepsilon \in Z^h$, (1.17), (2.1), and (2.14). Combining the bound on $F_\varepsilon(V^n_\varepsilon)$ in (2.48a) and (2.4) yields the bounds $3 \to 4$ in (2.48a). Bounds 3 and 7 in (2.48a) yield, on noting (2.8), bound 10 in (2.48a).

From (1.16), (2.2), (2.11a), (2.21), and (1.17), we obtain that

$$\begin{aligned}\left|\mathcal{G}[\tfrac{U_\varepsilon^n-U_\varepsilon^{n-1}}{\tau_n}]\right|_1^2 &= \left(\tfrac{U_\varepsilon^n-U_\varepsilon^{n-1}}{\tau_n}, \mathcal{G}[\tfrac{U_\varepsilon^n-U_\varepsilon^{n-1}}{\tau_n}]\right) = \left(\tfrac{U_\varepsilon^n-U_\varepsilon^{n-1}}{\tau_n}, Q^h\,\mathcal{G}[\tfrac{U_\varepsilon^n-U_\varepsilon^{n-1}}{\tau_n}]\right)^h \\ &= -\left(\tfrac{1}{3}\,\pi^h[(U_\varepsilon^{n-1})^3]\,\nabla W_\varepsilon^n + \tfrac{1}{2}\,\pi^h[(U_\varepsilon^{n-1})^2]\,\nabla V_\varepsilon^{n-1}, \nabla\left[Q^h\,\mathcal{G}[\tfrac{U_\varepsilon^n-U_\varepsilon^{n-1}}{\tau_n}]\right]\right) \\ &\leq C\,|U_\varepsilon^{n-1}|_{0,\infty}^{\frac{3}{2}}\left[\,|\,[\pi^h[(U_\varepsilon^{n-1})^3]\,]^{\frac{1}{2}}\,\nabla W_\varepsilon^n|_0 + |(U_\varepsilon^{n-1})^{\frac{1}{2}}\,\nabla V_\varepsilon^{n-1}|_0\right]\left|Q^h\,\mathcal{G}[\tfrac{U_\varepsilon^n-U_\varepsilon^{n-1}}{\tau_n}]\right|_1 \\ &\leq C\,|U_\varepsilon^{n-1}|_{0,\infty}^{3}\left[\,|\,[\pi^h[(U_\varepsilon^{n-1})^3]\,]^{\frac{1}{2}}\,\nabla W_\varepsilon^n|_0^2 + |(U_\varepsilon^{n-1})^{\frac{1}{2}}\,\nabla V_\varepsilon^{n-1}|_0^2\right].\end{aligned}\tag{2.51}$$

Similarly to (2.51), from (1.16), (2.2), (2.11c), (2.28), (2.16), (2.21), and (1.17), we obtain that

$$\begin{aligned}\left|\mathcal{G}[\tfrac{V_\varepsilon^n-V_\varepsilon^{n-1}}{\tau_n}]\right|_1^2 &= \left(\tfrac{V_\varepsilon^n-V_\varepsilon^{n-1}}{\tau_n}, Q^h\,\mathcal{G}[\tfrac{V_\varepsilon^n-V_\varepsilon^{n-1}}{\tau_n}]\right)^h = -\rho\left(\nabla V_\varepsilon^n, \nabla\left[Q^h\,\mathcal{G}[\tfrac{V_\varepsilon^n-V_\varepsilon^{n-1}}{\tau_n}]\right]\right) \\ &\quad - \left(U_\varepsilon^n\,\Lambda_\varepsilon(V_\varepsilon^n)\,\nabla V_\varepsilon^n + \tfrac{1}{2}\,\pi^h[(U_\varepsilon^n)^{\frac{1}{2}}\,(U_\varepsilon^{n-1})^{\frac{3}{2}}]\,\Lambda_\varepsilon(V_\varepsilon^n)\,\nabla W_\varepsilon^n, \nabla\left[Q^h\,\mathcal{G}[\tfrac{V_\varepsilon^n-V_\varepsilon^{n-1}}{\tau_n}]\right]\right) \\ &\leq C\left[\rho^2\,|\nabla V_\varepsilon^n|_0^2 + |U_\varepsilon^n|_{0,\infty}\left(|(U_\varepsilon^n)^{\frac{1}{2}}\,\nabla V_\varepsilon^n|_0^2 + |\,[\pi^h[(U_\varepsilon^{n-1})^3]\,]^{\frac{1}{2}}\,\nabla W_\varepsilon^n|_0^2\right)\right].\end{aligned}\tag{2.52}$$

□

Combining (2.51), (2.52), the assumptions on τ_n, and the bounds 8 → 10 in (2.48a) yields the bounds (2.48b).

LEMMA 2.5. *Let u^0, $v^0 \in K$, and the assumptions* (A) *hold. On choosing $U_\varepsilon^0 \equiv Q^h u^0$ and $V_\varepsilon^0 \equiv Q^h v^0$, or $U_\varepsilon^0 \equiv \pi^h u^0$ and $V_\varepsilon^0 \equiv \pi^h v^0$ in the case $d = 1$, it follows that U_ε^0, $V_\varepsilon^0 \in K^h$ are such that for all $h > 0$*

$$\|U_\varepsilon^0\|_1^2 + (U_\varepsilon^0\,\nabla V_\varepsilon^0, \nabla V_\varepsilon^0) + (F_\varepsilon(V_\varepsilon^0), 1)^h \leq C\,.\tag{2.53}$$

Proof. The desired result (2.53) follows from (2.21), (2.20), and (2.18). □

2.1. Inclusion of repulsive van der Waals forces. We end this section by extending Theorems 2.2 and 2.4 and Lemmas 2.3 and 2.5 to the approximation $(\mathrm{P}_{\delta,\,\varepsilon}^{h,\tau})$. In order to prove the existence of a solution to $(\mathrm{P}_{\delta,\,\varepsilon}^{h,\tau})$, we need to go through a regularization procedure which is similar to that used for the logarithmic Cahn–Hilliard equation; see, e.g., [5, 3]. For this purpose we introduce, for any $\mu \in \mathbb{R}_{>0}$, the $C^{2,1}$ convex function $\Phi_\mu^+ : \mathbb{R} \to \mathbb{R}_{\geq 0}$ such that

$$\Phi_\mu^+(s) := \begin{cases} \Phi^+(\mu) + \phi^+(\mu)\,(s-\mu) + \frac{(s-\mu)^2}{2}\,(\phi^+)'(\mu), & s \leq \mu, \\ \Phi^+(s), & \mu \leq s\,. \end{cases}\tag{2.54}$$

We set $\phi_\mu^+(\cdot) := (\Phi_\mu^+)'(\cdot)$ and note that $\Phi^+(s) \geq \Phi_\mu^+(s) \geq 0$ for all $s \in \mathbb{R}_{>0}$.

A consequence of the monotonicity of ϕ_μ^+ and our mesh assumption (A) is that for all $\mu \in \mathbb{R}_{>0}$

$$|\pi^h[\phi_\mu^+(\chi)]\,|_1^2 \leq (\phi^+)'(\mu)\,(\nabla\chi, \nabla\pi^h[\phi_\mu^+(\chi)]\,) \qquad \forall\ \chi \in S^h;\tag{2.55}$$

see, for example, [15].

THEOREM 2.6. *Let the assumptions* (A) *hold and U_ε^{n-1}, $V_\varepsilon^{n-1} \in S^h$ with $U_\varepsilon^{n-1} > 0$. Then for all $\varepsilon \in (0,1)$ and for all h, $\tau_n > 0$ there exists a solution $\{U_\varepsilon^n, W_\varepsilon^n, V_\varepsilon^n\}$ to the nth step of $(\mathrm{P}_{\delta,\,\varepsilon}^{h,\tau})$ with $U_\varepsilon^n > 0$, $\fint U_\varepsilon^n = \fint U_\varepsilon^{n-1}$, and $\fint V_\varepsilon^n = \fint V_\varepsilon^{n-1}$. Moreover, U_ε^n and W_ε^n are unique.*

Proof. As $\mathcal{U}_\varepsilon^{n-1} := \min_{x\in\overline{\Omega}} U_\varepsilon^{n-1}(x) > 0$, we have in place of (2.32), on noting (2.12) for $\chi \equiv 1$, that

$$W_\varepsilon^n \equiv -3\,\mathcal{G}^h_{U_\varepsilon^{n-1}}[\tfrac{U_\varepsilon^n - U_\varepsilon^{n-1}}{\tau_n} + X_\varepsilon^{n-1}] + \tfrac{1}{\underline{m}(\Omega)}\,(\phi^+(U_\varepsilon^n) + \phi^-(U_\varepsilon^{n-1}), 1)^h\,, \tag{2.56}$$

where $X_\varepsilon^{n-1} \in Z^h$ is defined by (2.30). Hence (2.11a) and (2.12) can be restated as follows.

Find $U_\varepsilon^n \in S^h(U_\varepsilon^{n-1})$ such that

$$A_{U_\varepsilon^{n-1}}(U_\varepsilon^n, \chi) + (\phi^+(U_\varepsilon^n), (I - \fint)\chi)^h = (\overline{X}_{\delta,\,\varepsilon}^{n-1}, \chi)^h \qquad \forall\, \chi \in S^h\,, \tag{2.57}$$

where $A_{U_\varepsilon^{n-1}}(\cdot,\cdot)$ is defined as in (2.35) and $\overline{X}_{\delta,\,\varepsilon}^{n-1} \in Z^h$ is such that

$$(\overline{X}_{\delta,\,\varepsilon}^{n-1}, \chi)^h := -(\phi^-(U_\varepsilon^{n-1}) + 3\,\mathcal{G}^h_{U_\varepsilon^{n-1}}\, X_\varepsilon^{n-1}, (I - \fint)\chi)^h \qquad \forall\, \chi \in S^h\,.$$

Due to the singular nature of the nonlinearity $\phi^+(s)$, we have to go through a regularization procedure in order to prove the existence of a solution to (2.57). For any $\mu \in \mathbb{R}_{>0}$, we introduce the regularized version of (2.57): Find $U_{\varepsilon,\,\mu}^n \in S^h(U_\varepsilon^{n-1})$ such that

$$A_{U_\varepsilon^{n-1}}(U_{\varepsilon,\,\mu}^n, \chi) + (\phi_\mu^+(U_{\varepsilon,\,\mu}^n), (I - \fint)\chi)^h = (\overline{X}_{\delta,\,\varepsilon}^{n-1}, \chi)^h \qquad \forall\, \chi \in S^h, \tag{2.58}$$

where ϕ_μ^+ is defined via (2.54). Similarly to (2.36), there exists a unique $U_{\varepsilon,\,\mu}^n$ satisfying (2.58) since this is the Euler–Lagrange equation of the convex minimization problem

$$\min_{\chi\in S^h(U_\varepsilon^{n-1})} \big\{\tfrac{c}{2}\,|\chi|_1^2 + (\Phi_\mu^+(\chi), 1)^h + \tfrac{3}{2\tau_n}\,|[\pi^h[(U_\varepsilon^{n-1})^3]\,]^{\frac{1}{2}}\,\nabla\mathcal{G}^h_{U_\varepsilon^{n-1}}(\chi - U_\varepsilon^{n-1})|_0^2 - (\overline{X}_{\delta,\,\varepsilon}^{n-1}, \chi)^h\big\}.$$

Choosing $\chi \equiv U_{\varepsilon,\,\mu}^n - U_\varepsilon^{n-1} \in Z^h$ in (2.58) and rearranging using (2.44), (2.27), (1.17), and the convexity of $\Phi_\mu^+ \leq \Phi^+$ yield that

$$\begin{aligned} &c\,\|U_{\varepsilon,\,\mu}^n\|_1^2 + \tau_n\,|\,[\pi^h[(U_\varepsilon^{n-1})^3]\,]^{\frac{1}{2}}\,\nabla\mathcal{G}^h_{U_\varepsilon^{n-1}}[\tfrac{U_{\varepsilon,\,\mu}^n - U_\varepsilon^{n-1}}{\tau_n}]\,|_0^2 \\ &\qquad \leq C\,[\,(\Phi^+(U_\varepsilon^{n-1}), 1)^h + |\overline{X}_{\delta,\,\varepsilon}^{n-1}|_h^2 + \|U_\varepsilon^{n-1}\|_1^2\,] \leq C, \end{aligned} \tag{2.59}$$

where, in the above and below, $C \in \mathbb{R}_{>0}$ is also independent of μ. Choosing $\chi \equiv \pi^h[\phi_\mu^+(U_{\varepsilon,\,\mu}^n)]$ in (2.58) and noting (2.55), $\overline{X}_{\delta,\,\varepsilon}^{n-1} \in Z^h$, (2.27), (2.14), (1.17), and (2.59) yield that

$$\begin{aligned} &\tau_n\,|(I - \fint)\,\pi^h[\phi_\mu^+(U_{\varepsilon,\,\mu}^n)]\,|_h^2 \leq C\,\tau_n\,[\,|\overline{X}_{\delta,\,\varepsilon}^{n-1}|_h^2 + |\mathcal{G}^h_{U_\varepsilon^{n-1}}[\tfrac{U_{\varepsilon,\,\mu}^n - U_\varepsilon^{n-1}}{\tau_n}]\,|_h^2\,] \\ &\quad \leq C(\mathcal{U}_\varepsilon^{n-1})\,\tau_n\,[\,|\overline{X}_{\delta,\,\varepsilon}^{n-1}|_h^2 + |\,[\pi^h[(U_\varepsilon^{n-1})^3]\,]^{\frac{1}{2}}\,\nabla\mathcal{G}^h_{U_\varepsilon^{n-1}}[\tfrac{U_{\varepsilon,\,\mu}^n - U_\varepsilon^{n-1}}{\tau_n}]\,|_0^2\,] \;\leq\; C(\mathcal{U}_\varepsilon^{n-1})\,. \end{aligned} \tag{2.60}$$

Choosing $\chi \equiv U_{\varepsilon,\,\mu}^n$ in (2.58) and noting the convexity of Φ_μ^+, it follows for any constant $\zeta \in \mathbb{R}_{>0}$ that

$$\begin{aligned} &(\phi_\mu^+(U_{\varepsilon,\,\mu}^n), \zeta - \fint U_{\varepsilon,\,\mu}^n)^h \\ &\qquad \leq (\Phi_\mu^+(\zeta) - \Phi_\mu^+(U_{\varepsilon,\,\mu}^n), 1)^h + (\overline{X}_{\delta,\,\varepsilon}^{n-1} - 3\,\mathcal{G}^h_{U_\varepsilon^{n-1}}[\tfrac{U_{\varepsilon,\,\mu}^n - U_\varepsilon^{n-1}}{\tau_n}], U_{\varepsilon,\,\mu}^n)^h \\ &\qquad \leq (\Phi^+(\zeta), 1)^h + (\overline{X}_{\delta,\,\varepsilon}^{n-1} - 3\,\mathcal{G}^h_{U_\varepsilon^{n-1}}[\tfrac{U_{\varepsilon,\,\mu}^n - U_\varepsilon^{n-1}}{\tau_n}], U_{\varepsilon,\,\mu}^n)^h\,. \end{aligned} \tag{2.61}$$

Choosing $\zeta = (⨍ U^n_{\varepsilon,\,\mu}) \pm \frac{1}{2}\,\mathcal{U}^{n-1}_\varepsilon = (⨍ U^{n-1}_\varepsilon) \pm \frac{1}{2}\,\mathcal{U}^{n-1}_\varepsilon \geq \frac{1}{2}\,\mathcal{U}^{n-1}_\varepsilon > 0$ in (2.61) and noting (2.59) and (2.60) yield that

$$\tau_n\,|\,⨍(\pi^h[\phi^+_\mu(U^n_{\varepsilon,\,\mu})]\,)|^2_h \leq C(\mathcal{U}^{n-1}_\varepsilon). \tag{2.62}$$

It follows from (2.59), (2.60), and (2.62) that there exist $U^n_\varepsilon \in S^h(U^{n-1}_\varepsilon)$, $\phi^+_h \in S^h$, and a subsequence $\{U^n_{\varepsilon,\,\mu'},\, \pi^h[\phi^+_{\mu'}(U^n_{\varepsilon,\,\mu'})]\,\}_{\mu'}$ such that $U^n_{\varepsilon,\,\mu'} \to U^n_\varepsilon$ and $\pi^h[\phi^+_{\mu'}(U^n_{\varepsilon,\,\mu'})] \to \phi^+_h$ as $\mu' \to 0$. Noting that for all $s \in \mathbb{R}$, $[\phi^+_\mu]^{-1}(s) \to [\phi^+]^{-1}(s)$ as $\mu \to 0$, we have that $(U^n_\varepsilon(p_j) - [\phi^+]^{-1}(s))\,(\phi^+_h(p_j) - s) \geq 0$ for all $s \in \mathbb{R}$, $j \in J$ and hence that $\phi^+_h \equiv \pi^h[\phi^+(U^n_\varepsilon)]$. Therefore, we may pass to the limit $\mu' \to 0$ in (2.58) to prove the existence of a solution $U^n_\varepsilon > 0$ to (2.57). Uniqueness of this solution follows from the monotonicity of ϕ^+. Hence noting (2.56), we have existence and uniqueness of a solution $\{U^n_\varepsilon,\, W^n_\varepsilon\}$ to (2.11a) and (2.12). Finally, existence of a solution V^n_ε to (2.11c) follows as in the proof of Theorem 2.2. □

LEMMA 2.7. *Let the assumptions of Theorem* 2.6 *hold. Then for all* $\varepsilon \in (0,1)$ *and for all* $h,\, \tau_n > 0$ *a solution* $\{U^n_\varepsilon, W^n_\varepsilon, V^n_\varepsilon\}$ *to the nth step of* $(\mathrm{P}^{h,\tau}_{\delta,\,\varepsilon})$ *is such that* (2.41) *holds with* $\mathcal{E}(\cdot,\cdot)$ *replaced by* $\mathcal{E}_\delta(U^n_\varepsilon, V^n_\varepsilon) := \frac{c}{2}\,|U^n_\varepsilon|^2_1 + (F_\varepsilon(V^n_\varepsilon) + \Phi(U^n_\varepsilon), 1)^h$.

Proof. The proof is a straightforward adaptation of the proof of Lemma 2.3 on noting (2.9) and the convexity of Φ^+. □

THEOREM 2.8. *Let the assumptions* (A) *hold and* $U^0_\varepsilon,\, V^0_\varepsilon \in S^h$ *with* $U^0_\varepsilon > 0$. *Then for all* $\varepsilon \in (0,1)$, $h > 0$, *a solution* $\{U^n_\varepsilon, W^n_\varepsilon, V^n_\varepsilon\}^N_{n=1}$ *to* $(\mathrm{P}^{h,\tau}_{\delta,\,\varepsilon})$ *is such that* $⨍ U^n_\varepsilon = ⨍ U^0_\varepsilon$ *and* $⨍ V^n_\varepsilon = ⨍ V^0_\varepsilon$, *and if* $\tau_n \leq \frac{5}{4}\,\omega\,\tau_{n-1}$, $n = 2 \to N$, *for an* $\omega \in (0,1)$, *then* (2.48a,b) *hold with the additional terms* $\max_{1\leq n\leq N}(\Phi(U^n_\varepsilon), 1)^h$ *on the left-hand side of* (2.48a) *and* $(\Phi(U^0_\varepsilon), 1)^h$ *inside the square brackets on the right-hand sides of* (2.48a,b).

Proof. A straightforward adaptation of the proof of Theorem 2.4 on noting (2.10) yields the desired result. □

LEMMA 2.9. *Let* u^0, $v^0 \in K$, *with* $u^0 \in L^\infty(\Omega)$ *and* $u^0(x) \geq \zeta > 0$ *for a.e.* $x \in \Omega$, *and let the assumptions* (A) *hold. On choosing* $U^0_\varepsilon \equiv Q^h u^0$ *and* $V^0_\varepsilon \equiv Q^h v^0$, *or* $U^0_\varepsilon \equiv \pi^h u^0$ *and* $V^0_\varepsilon \equiv \pi^h v^0$ *in the case* $d = 1$, *it follows that* U^0_ε, $V^0_\varepsilon \in K^h$ *with* $U^0_\varepsilon \geq \zeta$ *are such that for all* $h > 0$

$$\|U^0_\varepsilon\|^2_1 + (U^0_\varepsilon\,\nabla V^0_\varepsilon, \nabla V^0_\varepsilon) + (F_\varepsilon(V^0_\varepsilon) + \Phi(U^0_\varepsilon), 1)^h \leq C\,. \tag{2.63}$$

Proof. The desired result (2.63) follows from (2.21), (2.20), and (2.18). □

Remark 2.10. We note that Lemmas 2.3 and 2.7 are the discrete analogues of the energy estimates (1.5) and (1.11), respectively, on recalling (1.6), (1.7), and that $\sigma(s) := 1 - s$.

3. Convergence in one space dimension. Let

$$U_\varepsilon(t) := \tfrac{t-t_{n-1}}{\tau_n}\,U^n_\varepsilon + \tfrac{t_n-t}{\tau_n}\,U^{n-1}_\varepsilon, \qquad t \in [t_{n-1}, t_n], \quad n \geq 1, \tag{3.1a}$$

$$U^+_\varepsilon(t) := U^n_\varepsilon, \qquad U^-_\varepsilon(t) := U^{n-1}_\varepsilon, \qquad t \in (t_{n-1}, t_n], \quad n \geq 1. \tag{3.1b}$$

We note for future reference that

$$U_\varepsilon - U^\pm_\varepsilon = (t - t^\pm_n)\,\tfrac{\partial U_\varepsilon}{\partial t}, \qquad t \in (t_{n-1}, t_n) \quad n \geq 1, \tag{3.2}$$

where $t^+_n := t_n$ and $t^-_n := t_{n-1}$. We introduce also $\bar\tau(t) := \tau_n$ for $t \in (t_{n-1}, t_n]$ and $n \geq 1$. Using the above notation and introducing analogous notation for W_ε and V_ε, (2.11a–c) can be restated as follows.

Find $\{U_\varepsilon,\, W_\varepsilon,\, V_\varepsilon\} \in H^1(0,T;S^h)\times L^2(0,T;S^h)\times H^1(0,T;S^h)$ such that $U_\varepsilon(\cdot,t)\in K^h$ and for all $\chi \in L^2(0,T;S^h)$, $z^h \in L^2(0,T;K^h)$,

(3.3a)
$$\int_0^T \Big[\big(\tfrac{\partial U_\varepsilon}{\partial t},\chi\big)^h + \tfrac13\,\big(\pi^h[(U_\varepsilon^-)^3]\,\nabla W_\varepsilon^+,\nabla\chi\big)\Big]\,\mathrm{d}t = -\tfrac12\int_0^T (\pi^h[(U_\varepsilon^-)^2]\,\nabla V_\varepsilon^-,\nabla\chi)\,\mathrm{d}t,$$

(3.3b)
$$\int_0^T \big[\,c\,(\nabla U_\varepsilon^+,\nabla(z^h-U_\varepsilon^+)) + (\phi^-(U_\varepsilon^-+\varepsilon)-W_\varepsilon^+,z^h-U_\varepsilon^+)^h\big]\,\mathrm{d}t \geq 0,$$

$$\int_0^T \Big[\big(\tfrac{\partial V_\varepsilon}{\partial t},\chi\big)^h + \rho\,\big(\nabla V_\varepsilon^+,\nabla\chi\big) + \big(U_\varepsilon^+\,\Lambda_\varepsilon(V_\varepsilon^+)\,\nabla V_\varepsilon^+,\nabla\chi\big)\Big]\,\mathrm{d}t$$

(3.3c)
$$= -\tfrac12\int_0^T (\pi^h[(U_\varepsilon^+)^{\frac12}\,(U_\varepsilon^-)^{\frac32}]\,\Lambda_\varepsilon(V_\varepsilon^+)\,\nabla W_\varepsilon^+,\nabla\chi)\,\mathrm{d}t.$$

LEMMA 3.1. *Let $d=1$, $a=0\,(\phi^-\equiv 0)$, $\rho>0$, and u^0, $v^0\in K$ with $u^0\not\equiv 0$. Let $\{\mathcal{T}^h, U_\varepsilon^0, V_\varepsilon^0, \{\tau_n\}_{n=1}^N, \varepsilon\}_{h>0}$ be such that*

(i) *$U_\varepsilon^0 \equiv \pi^h u^0$, $V_\varepsilon^0\equiv \pi^h v^0$;*

(ii) *Ω and $\{\mathcal{T}^h\}_{h>0}$ fulfil assumption* (A), *$\varepsilon\in(0,1)$, and $\tau_n \leq \frac54\,\omega\,\tau_{n-1}$, $n=2\rightarrow N$, for an $\omega\in(0,1)$;*

(iii) *ε, $\tau\rightarrow 0$ as $h\rightarrow 0$.*

Then there exist a subsequence of $\{U_\varepsilon, V_\varepsilon\}_h$, where $\{U_\varepsilon,\, W_\varepsilon,\, V_\varepsilon\}$ solve $(\mathrm{P}_\varepsilon^{h,\tau})$, *and functions*

$$u \in L^\infty(0,T;K)\cap H^1(0,T;(H^1(\Omega))')\cap C^{\frac12,\frac18}_{x,t}(\overline{\Omega}_T), \tag{3.4a}$$

$$v \in L^\infty(0,T;L^2(\Omega))\cap L^2(0,T;K)\cap H^1(0,T;(H^1(\Omega))') \tag{3.4b}$$

with $u(x,0)=u^0(x)$ for all $x\in\overline{\Omega}$, $v(\cdot,0)=v^0(\cdot)$ in $(H^1(\Omega))'$, $⨍ u(\cdot,t)=⨍ u^0>0$ for all $t\in[0,T]$ and $⨍ v(\cdot,t)=⨍ v^0$ for a.e. $t\in[0,T]$ such that as $h\rightarrow 0$

(3.5a)	$U_\varepsilon,\, U_\varepsilon^\pm \rightarrow u$			*uniformly on* $\overline{\Omega}_T$,
(3.5b)	$U_\varepsilon,\, U_\varepsilon^\pm \rightarrow u$	*and*	$\mathcal{G}\frac{\partial U_\varepsilon}{\partial t}\rightarrow \mathcal{G}\frac{\partial u}{\partial t}$	*weakly in* $L^2(0,T;H^1(\Omega))$,
(3.6a)	$V_\varepsilon,\, V_\varepsilon^\pm \rightarrow v$	*and*	$\mathcal{G}\frac{\partial V_\varepsilon}{\partial t}\rightarrow \mathcal{G}\frac{\partial v}{\partial t}$	*weakly in* $L^2(0,T;H^1(\Omega))$,
(3.6b)	$V_\varepsilon,\, V_\varepsilon^\pm \rightarrow v$	*and*	$\Lambda_\varepsilon(V_\varepsilon^+)\rightarrow\lambda(v)$	*strongly in* $L^2(\Omega_T)$.

Proof. From (2.48a), (2.53), and (1.15), we have for $d=1$

$$\max_{1\leq n\leq N}\|U_\varepsilon^n\|_1 \leq C \qquad\Longrightarrow\qquad \max_{1\leq n\leq N}|U_\varepsilon^n|_{0,\infty}\leq C\,. \tag{3.7}$$

Noting the definitions (3.1a,b) and (3.7), the bounds in (2.48a,b) together with (1.17), (2.53), and the time step control in (ii) imply that

$$\begin{aligned}&\|U_\varepsilon\|^2_{L^\infty(0,T;H^1(\Omega))} + \|V_\varepsilon\|^2_{L^\infty(0,T;L^2(\Omega))} + \rho\,\|V_\varepsilon\|^2_{L^2(0,T;H^1(\Omega))} + \rho\,\|V_\varepsilon^\pm\|^2_{L^2(0,T;H^1(\Omega))}\\ &\quad + \varepsilon^{-1}\,\|\pi^h[V_\varepsilon^+]_-\|^2_{L^\infty(0,T;L^2(\Omega))} + \|\bar\tau^{\frac12}\,\tfrac{\partial U_\varepsilon}{\partial t}\|^2_{L^2(0,T;H^1(\Omega))} + \|\bar\tau^{\frac12}\,\tfrac{\partial V_\varepsilon}{\partial t}\|^2_{L^2(\Omega_T)}\\ &\quad + \|\,[\pi^h[(U_\varepsilon^-)^3]\,]^{\frac12}\,\nabla W_\varepsilon^+\|^2_{L^2(\Omega_T)} + \|\mathcal{G}\tfrac{\partial U_\varepsilon}{\partial t}\|^2_{L^2(0,T;H^1(\Omega))} + \|\mathcal{G}\tfrac{\partial V_\varepsilon}{\partial t}\|^2_{L^2(0,T;H^1(\Omega))} \leq C.\end{aligned} \tag{3.8}$$

In addition, we deduce from (3.2) and (3.8) that

$$\|U_\varepsilon - U_\varepsilon^\pm\|^2_{L^2(0,T;H^1(\Omega))} + \|V_\varepsilon - V_\varepsilon^\pm\|^2_{L^2(\Omega_T)} \leq \|\bar\tau\,\tfrac{\partial U_\varepsilon}{\partial t}\|^2_{L^2(0,T;H^1(\Omega))} + \|\bar\tau\,\tfrac{\partial V_\varepsilon}{\partial t}\|^2_{L^2(\Omega_T)} \leq C\,\tau. \tag{3.9}$$

Moreover, the first and ninth bound in (3.8) imply that the $C^{\frac{1}{2},\frac{1}{8}}_{x,t}(\overline{\Omega}_T)$ norm of U_ε is bounded independently of h, τ, ε, and T; see, e.g., [4, Theorem 2.2]. Therefore, by the Arzelà–Ascoli theorem there exists a subsequence $\{U_\varepsilon, V_\varepsilon\}_h$ and a $u \geq 0$, as $U_\varepsilon(\cdot,t) \in K^h$, such that

$$U_\varepsilon,\ U_\varepsilon^\pm \to u \in C^{\frac{1}{2},\frac{1}{8}}_{x,t}(\overline{\Omega}_T) \qquad \text{uniformly on } \overline{\Omega}_T \text{ as } h \to 0. \tag{3.10}$$

Furthermore, (3.10), (3.8), (2.19), (3.1a,b), (3.9), $⨍ U_\varepsilon^n = ⨍ U_\varepsilon^0$, $⨍ V_\varepsilon^n = ⨍ V_\varepsilon^0$, and our assumptions (iii) imply, as $\rho > 0$, that this same subsequence $\{U_\varepsilon, V_\varepsilon\}_h$ can be chosen such that (3.4a), (3.4b) with K replaced by $H^1(\Omega)$, (3.5a,b), and (3.6a) hold. The strong convergence result for $V_\varepsilon^{(\pm)}$ in (3.6b) follows immediately from (3.6a) and a standard embedding result. Furthermore, it follows from (3.8), (3.6a), (1.18), and a standard compactness argument that

$$V_\varepsilon \to v \quad \text{in} \quad C([0,T];(H^1(\Omega))')\,. \tag{3.11}$$

Noting the assumptions (i) and (2.19), we have that

$$U_\varepsilon^0 \to u^0 \quad \text{and} \quad V_\varepsilon^0 \to v^0 \quad \text{strongly in} \quad H^1(\Omega)\,. \tag{3.12}$$

Combining (3.12), (3.11), and (3.10) yields that $u(\cdot,0) = u^0(\cdot)$ in $C(\overline{\Omega})$ and $v(\cdot,0) = v^0(\cdot)$ in $(H^1(\Omega))'$.

We now prove the remaining result in (3.6b). For this we introduce for all $\varepsilon \in (0,1)$, $\lambda_\varepsilon : \mathbb{R} \to [\varepsilon, 1]$ defined, on recalling (1.13), by

$$\lambda_\varepsilon(s) := [\lambda(s) - \varepsilon]_+ + \varepsilon, \qquad \text{where} \qquad [s]_+ := \max\{s,\,0\}. \tag{3.13}$$

Then we have that

$$\begin{aligned}\|\lambda(v) - \Lambda_\varepsilon(V_\varepsilon^+)\|_{L^2(\Omega_T)} &\leq \|\lambda(v) - \lambda(V_\varepsilon^+)\|_{L^2(\Omega_T)} + \|(I - \pi^h)\lambda(V_\varepsilon^+)\|_{L^2(\Omega_T)} \\ &\quad + \|\pi^h[\,\lambda(V_\varepsilon^+) - \lambda_\varepsilon(V_\varepsilon^+)\,]\|_{L^2(\Omega_T)} + \|\pi^h[\lambda_\varepsilon(V_\varepsilon^+)] - \Lambda_\varepsilon(V_\varepsilon^+)\|_{L^2(\Omega_T)}.\end{aligned} \tag{3.14}$$

Noting the global Lipschitz continuity of λ, (2.18), and (3.8), we have that

$$\begin{aligned}&\|\lambda(v) - \lambda(V_\varepsilon^+)\|_{L^2(\Omega_T)} + \|(I - \pi^h)\lambda(V_\varepsilon^+)\|_{L^2(\Omega_T)} \\ &\quad \leq \|v - V_\varepsilon^+\|_{L^2(\Omega_T)} + C\,h\,\|\nabla[\lambda(V_\varepsilon^+)]\|_{L^2(\Omega_T)} \\ &\quad \leq \|v - V_\varepsilon^+\|_{L^2(\Omega_T)} + C\,h\,\|\nabla V_\varepsilon^+\|_{L^2(\Omega_T)} \leq \|v - V_\varepsilon^+\|_{L^2(\Omega_T)} + C\,h\,.\end{aligned} \tag{3.15}$$

It follows from (2.1), (2.14), (3.1b), (1.13), (3.13), and (3.8) that

$$\begin{aligned}\|\pi^h[\,\lambda(V_\varepsilon^+) - \lambda_\varepsilon(V_\varepsilon^+)\,]\,\|^2_{L^2(\Omega_T)} &\leq \sum_{n=1}^N \tau_n\,|\pi^h[\,\lambda(V_\varepsilon^n) - \lambda_\varepsilon(V_\varepsilon^n)\,]|_h^2 \\ &\leq \sum_{n=1}^N \tau_n\,|\pi^h[\varepsilon - [V_\varepsilon^n]_-]\,|_h^2 \leq C\left[\varepsilon^2 + \sum_{n=1}^N \tau_n\,|\pi^h[V_\varepsilon^n]_-\,|_0^2\right] \leq C\,\varepsilon^2.\end{aligned} \tag{3.16}$$

From (2.7), (2.3), and (3.13), we have that $\Lambda_\varepsilon(V_\varepsilon^n)\,|_{(p_{j-1},p_j)}$ lies between $\lambda_\varepsilon(V_\varepsilon^n(p_{j-1}))$ and $\lambda_\varepsilon(V_\varepsilon^n(p_j))$ for $j = 1 \to J$ and $n = 1 \to N$. This together with (2.18), the global Lipschitz continuity of λ_ε, and (3.8) implies that

$$\begin{aligned}\|\pi^h[\lambda_\varepsilon(V_\varepsilon^+)] - \Lambda_\varepsilon(V_\varepsilon^+)\|_{L^2(\Omega_T)} &\leq C\,h\,\|\nabla\pi^h[\lambda_\varepsilon(V_\varepsilon^+)]\|_{L^2(\Omega_T)}\\ &\leq C\,h\,\|\nabla[\lambda_\varepsilon(V_\varepsilon^+)]\|_{L^2(\Omega_T)} \leq C\,h\,\|\nabla V_\varepsilon^+\|_{L^2(\Omega_T)} \leq C\,h\,.\end{aligned}\tag{3.17}$$

Combining (3.14), (3.15), (3.16), and (3.17) and noting the result on V_ε^+ in (3.6b) and our assumption (iii) on ε yield the desired result on $\Lambda_\varepsilon(V_\varepsilon^+)$ in (3.6b). Finally, we note that $\Lambda_\varepsilon(V_\varepsilon^+) \geq 0$ and (3.6b) $\Rightarrow$ $\lambda(v) \geq 0$ a.e. $\Rightarrow$ $v \geq 0$ a.e. $\Rightarrow$ K in (3.4b). □

For any $\alpha > 0$, we set

$$B_\alpha := \{\,(x,t) \in \overline{\Omega}_T : u(x,t) > \alpha\,\} \quad\text{and}\quad B_\alpha(t) := \{\,x \in \overline{\Omega} : u(x,t) > \alpha\,\}.\tag{3.18}$$

From (3.4a), we have that there exist positive constants C_x and C_t such that

$$|u(y_2,t) - u(y_1,t)| \leq C_x\,|y_2 - y_1|^{\frac{1}{2}} \qquad \forall\; y_1,\, y_2 \in \overline{\Omega}, \qquad \forall\; t \in [0,T];\tag{3.19a}$$

$$|u(x,t_b) - u(x,t_a)| \leq C_t\,|t_b - t_a|^{\frac{1}{8}} \qquad \forall\; t_a,\, t_b \in [0,T], \quad \forall\; x \in \overline{\Omega}.\tag{3.19b}$$

As $⨍ u(\cdot,t) = ⨍ u^0 > 0$ for all $t \in [0,T]$, it follows that there exists an $\alpha_0 \in (0, ⨍ u^0)$ such that $B_{\alpha_0}(t) \neq \emptyset$ for all $t \in [0,T]$. It immediately follows from (3.18) and (3.19a,b) for any t_a, $t_b \in [0,T]$ and for any α_1, $\alpha_2 \in (0,\alpha_0)$ with $\alpha_1 > \alpha_2$ that

$$\begin{aligned}&y_1 \in B_{\alpha_1}(t_a) \text{ and } y_2 \in \partial B_{\alpha_2}(t_b) \text{ with } y_2 \notin \partial\Omega \quad \Longrightarrow\\ &C_x\,|y_2 - y_1|^{\frac{1}{2}} + C_t\,|t_b - t_a|^{\frac{1}{8}} \geq u(y_1,t_a) - u(y_2,t_b) > (\alpha_1 - \alpha_2),\end{aligned}\tag{3.20}$$

where $\partial B_\alpha(t)$ is the boundary of $B_\alpha(t)$. Therefore, (3.20) implies that for any $\alpha \in (0,\alpha_0)$, there exists an $h_0(\alpha)$ such that for all $h \leq h_0(\alpha)$ and $t \in [0,T]$ there exists a collection of simplices $\mathcal{T}_\alpha^h(t) \subset \mathcal{T}^h$ such that

$$B_\alpha(t) \subset B_\alpha^h(t) := \cup_{\kappa \in \mathcal{T}_\alpha^h(t)}\,\overline{\kappa} \subset B_{\frac{\alpha}{2}}(t) \qquad \forall\; t \in [0,T].\tag{3.21}$$

Similarly, it follows from (3.20) that for any $\alpha \in (0,\alpha_0)$, there exists a $\tau_0(\alpha)$ such that for all $\tau \leq \tau_0(\alpha)$

$$B_\alpha(t) \subset B_{\frac{\alpha}{2}}(t_n) \subset B_{\frac{\alpha}{4}}(t) \qquad \forall\; t \in (t_{n-1},t_n], \quad n = 1 \to N.\tag{3.22}$$

Clearly, we have from (3.21) and (3.22) that $\alpha_2 < \alpha_1 < \alpha_0$ implies that $h_0(\alpha_2) \leq h_0(\alpha_1)$ and $\tau_0(\alpha_2) \leq \tau_0(\alpha_1)$. For a fixed $\alpha \in (0,\alpha_0)$, it follows from (3.18), (3.5a), and our assumption (iii) of Lemma 3.1 that there exists an $\widehat{h}_0(\alpha) \leq h_0(\alpha)$ such that for $h \leq \widehat{h}_0(\alpha)$

$$\begin{array}{lll} 0 \;\leq U_\varepsilon^\pm(x,t) \leq 2\,\alpha & \forall\;(x,t) \notin B_\alpha, & \\ \frac{1}{2}\,\alpha \;\leq U_\varepsilon^\pm(x,t) & \forall\;(x,t) \in B_\alpha, & \end{array} \quad \text{and} \quad \tau \leq \tau_0(\alpha).\tag{3.23}$$

THEOREM 3.2. *Let the assumptions of Lemma* 3.1 *hold. Then there exist a subsequence of* $\{U_\varepsilon, W_\varepsilon, V_\varepsilon\}_h$*, where* $\{U_\varepsilon,\, W_\varepsilon,\, V_\varepsilon\}$ *solve* $(\mathrm{P}_\varepsilon^{h,\tau})$*, and functions* $\{u,w,v\}$ *satisfying* (3.4a,b) *and*

$$w \in L^2_{loc}(\{u > 0\}) \qquad \textit{with} \qquad \nabla w \in L^2_{loc}(\{u > 0\}),\tag{3.24}$$

where $\{u > 0\} := \{(x,t) \in \Omega_T \,:\, u(x,t) > 0\,\}$ *such that as* $h \to 0$ (3.5a,b), (3.6a–c) *hold and* $W_\varepsilon^+ \to w$, $\nabla W_\varepsilon^+ \to \nabla w$ *weakly in* $L^2_{loc}(\{u > 0\})$. *Furthermore, we* *have that* u*,* v*, and* w *fulfil* $u(\cdot,0) = u^0(\cdot)$*,* $v(\cdot,0) = v^0(\cdot)$ *and are such that for all* $\eta, z \in L^2(0,T;H^1(\Omega))$*, with* $\mathrm{supp}(z) \subset \{u > 0\}$*,*

$$\int_0^T \langle \tfrac{\partial u}{\partial t}, \eta\rangle \,\mathrm{d}t + \tfrac{1}{3}\int_{\{u>0\}} u^3 \,\nabla w \,.\, \nabla \eta \,\mathrm{d}x\,\mathrm{d}t + \tfrac{1}{2}\int_{\Omega_T} u^2\,\nabla v \,.\, \nabla\eta\,\mathrm{d}x\,\mathrm{d}t = 0, \tag{3.25a}$$

$$\int_{\{u>0\}} [\, c\,\nabla u \,.\, \nabla z - w\,z\,]\,\mathrm{d}x\,\mathrm{d}t = 0, \tag{3.25b}$$

$$\begin{aligned}\int_0^T \langle \tfrac{\partial v}{\partial t}, \eta\rangle \,\mathrm{d}t + \int_{\Omega_T} [\,\rho\,\nabla v \,.\, \nabla \eta + u\,\lambda(v)\,\nabla v \,.\, \nabla\eta\,]\,\mathrm{d}x\,\mathrm{d}t \\ + \tfrac{1}{2}\int_{\{u>0\}} u^2\,\lambda(v)\,\nabla w \,.\, \nabla\eta\,\mathrm{d}x\,\mathrm{d}t = 0.\end{aligned} \tag{3.25c}$$

Proof. For any $\eta \in L^2(0,T;H^1(\Omega))$, we choose $\chi \equiv \pi^h\eta$ in (3.3a,c) and now analyze the subsequent terms. First, (2.17), (2.23), (2.18), (1.15) in time, and (3.8) yield for $Z \equiv U_\varepsilon$ and V_ε, respectively, and for all $\widetilde{\eta} \in H^1(0,T;H^1(\Omega))$ that

$$\begin{aligned}\left|\int_0^T \Big[\big(\tfrac{\partial Z}{\partial t}, \pi^h\eta\big)^h - \big(\tfrac{\partial Z}{\partial t}, \pi^h\eta\big)\Big]\,\mathrm{d}t\right| &\leq \left|\int_0^T \Big[\big(\tfrac{\partial Z}{\partial t}, \pi^h[\eta-\widetilde{\eta}]\big)^h - \big(\tfrac{\partial Z}{\partial t}, \pi^h[\eta-\widetilde{\eta}]\big)\Big]\,\mathrm{d}t\right| \\ &+ \left| -\int_0^T \Big(Z, \tfrac{\partial(\pi^h\widetilde{\eta})}{\partial t}\Big)^h \,\mathrm{d}t + (\,Z(\cdot,T), \pi^h\widetilde{\eta}(\cdot,T)\,)^h - (\,Z(\cdot,0), \pi^h\widetilde{\eta}(\cdot,0)\,)^h \right. \\ &\left. + \int_0^T \Big(Z, \tfrac{\partial(\pi^h\widetilde{\eta})}{\partial t}\Big)\,\mathrm{d}t - (\,Z(\cdot,T), \pi^h\widetilde{\eta}(\cdot,T)\,) + (\,Z(\cdot,0), \pi^h\widetilde{\eta}(\cdot,0)\,)\right| \\ \leq C\,\|\mathcal{G}\tfrac{\partial Z}{\partial t}\|_{L^2(0,T;H^1(\Omega))}\,\|\pi^h[\eta-\widetilde{\eta}]\|_{L^2(0,T;H^1(\Omega))} \\ + C\,h\,\|Z\|_{L^\infty(0,T;L^2(\Omega))}\,\|\pi^h\widetilde{\eta}\|_{H^1(0,T;H^1(\Omega))} \\ \leq C\,\|\eta-\widetilde{\eta}\|_{L^2(0,T;H^1(\Omega))} + C\,h\,\|\widetilde{\eta}\|_{H^1(0,T;H^1(\Omega))}\,.\end{aligned} \tag{3.26}$$

Furthermore, it follows from (1.16) and (3.8) that

$$\begin{aligned}\left|\int_0^T \big(\tfrac{\partial Z}{\partial t}, (I-\pi^h)\eta\big)\,\mathrm{d}t\right| &\leq C\,\|\mathcal{G}\tfrac{\partial Z}{\partial t}\|_{L^2(0,T;H^1(\Omega))}\,\|(I-\pi^h)\eta\|_{L^2(0,T;H^1(\Omega))} \\ &\leq C\,\|(I-\pi^h)\eta\|_{L^2(0,T;H^1(\Omega))}.\end{aligned} \tag{3.27}$$

Combining (3.26), the denseness of $H^1(0,T;H^1(\Omega))$ in $L^2(0,T;H^1(\Omega))$, (3.27), (2.19), (1.20), (3.5b), and (3.6a) yields that for $z \equiv u$ and v, respectively,

$$\int_0^T \big(\tfrac{\partial Z}{\partial t}, \pi^h\eta\big)^h \,\mathrm{d}t \to \int_0^T \langle\tfrac{\partial z}{\partial t}, \eta\rangle\,\mathrm{d}t \quad \text{as } h \to 0. \tag{3.28}$$

In view of (2.28), (3.1b), (2.16), (3.7), and (3.8), as $\rho > 0$, we deduce that

$$\begin{aligned}&\left|\int_0^T \Big(\pi^h[(U_\varepsilon^+)^{\frac{1}{2}}\,(U_\varepsilon^-)^{\frac{3}{2}}]\,\Lambda_\varepsilon(V_\varepsilon^+)\,\nabla W_\varepsilon^+, \nabla(I-\pi^h)\eta\Big)\,\mathrm{d}t\right| \\ &\quad\leq \|\pi^h[(U_\varepsilon^+)^{\frac{1}{2}}\,(U_\varepsilon^-)^{\frac{3}{2}}]\,\nabla W_\varepsilon^+\|_{L^2(\Omega_T)}\,\|(I-\pi^h)\eta\|_{L^2(0,T;H^1(\Omega))} \\ &\quad\leq C\,\|U_\varepsilon^+\|^{\frac{1}{2}}_{L^\infty(\Omega_T)}\,\|\,[\pi^h[(U_\varepsilon^-)^3]\,]^{\frac{1}{2}}\,\nabla W_\varepsilon^+\|_{L^2(\Omega_T)}\,\|(I-\pi^h)\eta\|_{L^2(0,T;H^1(\Omega))} \\ &\quad\leq C\,\|(I-\pi^h)\eta\|_{L^2(0,T;H^1(\Omega))},\end{aligned} \tag{3.29a}$$

and similarly

$$\begin{aligned}
&\left|\int_0^T \big(\pi^h[(U_\varepsilon^-)^3]\,\nabla W_\varepsilon^+, \nabla(I-\pi^h)\eta\big)\ \mathrm{d}t\right| + \left|\int_0^T \big(U_\varepsilon^+\,\Lambda_\varepsilon(V_\varepsilon^+)\,\nabla V_\varepsilon^+, \nabla(I-\pi^h)\eta\big)\ \mathrm{d}t\right| \\
&\quad + \left|\int_0^T \big(\pi^h[(U_\varepsilon^-)^2]\,\nabla V_\varepsilon^-, \nabla(I-\pi^h)\eta\big)\ \mathrm{d}t\right| + \left|\int_0^T \big(\nabla V_\varepsilon^+, \nabla(I-\pi^h)\eta\big)\ \mathrm{d}t\right| \\
&\quad + \left|\int_0^T \big(\nabla U_\varepsilon^+, \nabla(I-\pi^h)\eta\big)\ \mathrm{d}t\right| \ \le\ C\,\|(I-\pi^h)\eta\|_{L^2(0,T;H^1(\Omega))}.
\end{aligned} \tag{3.29b}$$

Noting (3.29b), (2.19), (3.5b), and (3.6a), we have for $Z^+ \equiv U_\varepsilon^+$ and V_ε^+, and $z \equiv u$ and v, respectively, that

$$\int_0^T (\nabla Z^+, \nabla(\pi^h\eta)\,)\,\mathrm{d}t \to \int_0^T (\nabla z, \nabla\eta)\,\mathrm{d}t \quad \text{as } h \to 0. \tag{3.30}$$

It also follows from (3.8), (2.28), (1.13), and (3.4a,b) that for all $\widetilde{\eta} \in L^\infty(0,T;W^{1,\infty}(\Omega))$

$$\begin{aligned}
&\left|\int_0^T \big(\,(\pi^h[(U_\varepsilon^-)^2] - u^2)\,\nabla V_\varepsilon^-, \nabla\eta\big)\ \mathrm{d}t\right| \\
&\quad \le \|\pi^h[(U_\varepsilon^-)^2] - u^2\|_{L^\infty(\Omega_T)}\,\|V_\varepsilon^-\|_{L^2(0,T;H^1(\Omega))}\,\|\eta\|_{L^2(0,T;H^1(\Omega))} \\
&\quad \le C\left[\|\pi^h[(U_\varepsilon^-)^2 - u^2]\|_{L^\infty(\Omega_T)} + \|(I-\pi^h)u^2\|_{L^\infty(\Omega_T)}\right]\ \|\eta\|_{L^2(0,T;H^1(\Omega))},
\end{aligned} \tag{3.31a}$$

$$\begin{aligned}
&\left|\int_0^T \big(\,(U_\varepsilon^+\,\Lambda_\varepsilon(V_\varepsilon^+) - u\,\lambda(v)\,)\,\nabla V_\varepsilon^+, \nabla\eta\big)\ \mathrm{d}t\right| \\
&\quad \le \left|\int_0^T \big(\,(U_\varepsilon^+ - u)\,\Lambda_\varepsilon(V_\varepsilon^+)\,\nabla V_\varepsilon^+, \nabla\eta\big)\ \mathrm{d}t\right| \\
&\quad + \left|\int_0^T \Big[\,\big(u\,(\Lambda_\varepsilon(V_\varepsilon^+) - \lambda(v)\,)\,\nabla V_\varepsilon^+, \nabla(\eta - \widetilde{\eta})\,\big) + \big(u\,(\Lambda_\varepsilon(V_\varepsilon^+) - \lambda(v)\,)\,\nabla V_\varepsilon^+, \nabla\widetilde{\eta}\big)\,\Big]\,\mathrm{d}t\right| \\
&\quad \le \|U_\varepsilon^+ - u\|_{L^\infty(\Omega_T)}\,\|V_\varepsilon^+\|_{L^2(0,T;H^1(\Omega))}\,\|\eta\|_{L^2(0,T;H^1(\Omega))} \\
&\qquad + \|u\|_{L^\infty(\Omega_T)}\,\|\Lambda_\varepsilon(V_\varepsilon^+) - \lambda(v)\|_{L^\infty(\Omega_T)}\,\|V_\varepsilon^+\|_{L^2(0,T;H^1(\Omega))}\,\|\eta - \widetilde{\eta}\|_{L^2(0,T;H^1(\Omega))} \\
&\qquad + \|u\|_{L^\infty(\Omega_T)}\,\|\Lambda_\varepsilon(V_\varepsilon^+) - \lambda(v)\|_{L^2(\Omega_T)}\,\|V_\varepsilon^+\|_{L^2(0,T;H^1(\Omega))}\,\|\widetilde{\eta}\|_{L^\infty(0,T;W^{1,\infty}(\Omega))} \\
&\quad \le C\,\Big[\|U_\varepsilon^+ - u\|_{L^\infty(\Omega_T)}\,\|\eta\|_{L^2(0,T;H^1(\Omega))} + \|\eta - \widetilde{\eta}\|_{L^2(0,T;H^1(\Omega))} \\
&\qquad + \|\Lambda_\varepsilon(V_\varepsilon^+) - \lambda(v)\|_{L^2(\Omega_T)}\,\|\widetilde{\eta}\|_{L^\infty(0,T;W^{1,\infty}(\Omega))}\Big].
\end{aligned} \tag{3.31b}$$

Noting that $L^\infty(0,T;W^{1,\infty}(\Omega))$ is dense in $L^2(0,T;H^1(\Omega))$, (3.29b), (2.19), (3.31a,b), (3.5a), (2.13), (3.6b), and (3.6a), we have that

$$\int_0^T \big(\pi^h[(U_\varepsilon^-)^2]\,\nabla V_\varepsilon^-, \nabla(\pi^h\eta)\big)\ \mathrm{d}t \to \int_0^T (u^2\,\nabla v, \nabla\eta)\,\mathrm{d}t \qquad \text{as } h \to 0, \tag{3.32a}$$

$$\int_0^T \big(\,U_\varepsilon^+\,\Lambda_\varepsilon(V_\varepsilon^+)\,\nabla V_\varepsilon^+, \nabla(\pi^h\eta)\big)\ \mathrm{d}t \to \int_0^T (u\,\lambda(v)\,\nabla v, \nabla\eta)\,\mathrm{d}t \quad \text{as } h \to 0. \tag{3.32b}$$

We now show the compactness of $\{W_\varepsilon^+\}_h$ on compact subsets of $\{u > 0\}$. On noting (3.23), (3.8), (2.28), and (2.16), we have for all $h \leq \widehat{h}_0(\alpha)$, similarly to (3.29a), that

$$\begin{aligned}
&\left| \int_{\Omega_T \setminus B_\alpha} \pi^h[(U_\varepsilon^+)^{\frac{1}{2}}\,(U_\varepsilon^-)^{\frac{3}{2}}]\,\Lambda_\varepsilon(V_\varepsilon^+)\,\nabla W_\varepsilon^+ \,.\, \nabla\eta \,\mathrm{d}x\,\mathrm{d}t \right| \\
&\qquad \leq C\,\|U_\varepsilon^+\|_{L^\infty(\Omega_T\setminus B_\alpha)}^{\frac{1}{2}}\,\|\,[\pi^h[(U_\varepsilon^-)^3]\,]^{\frac{1}{2}}\,\nabla W_\varepsilon^+\|_{L^2(\Omega_T)}\,\|\eta\|_{L^2(0,T;H^1(\Omega))} \\
&\qquad \leq C\,\alpha^{\frac{1}{2}}\,\|\eta\|_{L^2(0,T;H^1(\Omega))}
\end{aligned} \tag{3.33a}$$

and similarly

$$\left| \int_{\Omega_T \setminus B_\alpha} \pi^h[(U_\varepsilon^-)^3]\,\nabla W_\varepsilon^+ \,.\, \nabla\eta\,\mathrm{d}x\,\mathrm{d}t \right| \leq C\,\alpha^{\frac{3}{2}}\,\|\eta\|_{L^2(0,T;H^1(\Omega))}. \tag{3.33b}$$

It follows from (3.23), (3.21), and (3.8) that for all $h \leq \widehat{h}_0(\frac{\alpha}{8})$

$$\begin{aligned}
C_1\,\alpha^3 \int_{B_{\frac{\alpha}{4}}} |\nabla W_\varepsilon^+|^2\,\mathrm{d}x\,\mathrm{d}t &\leq C_1\,\alpha^3 \int_{B^h_{\frac{\alpha}{4}}} |\nabla W_\varepsilon^+|^2\,\mathrm{d}x\,\mathrm{d}t \leq \int_{B^h_{\frac{\alpha}{4}}} \pi^h[(U_\varepsilon^-)^3]\,|\nabla W_\varepsilon^+|^2\,\mathrm{d}x\,\mathrm{d}t \\
&\leq C,
\end{aligned} \tag{3.34}$$

where $B_\alpha^h := \{(x,t) \in \overline{\Omega}_T : x \in B_\alpha^h(t)\}$. Similarly to (3.31a,b), it follows from (2.28), (3.34), (1.13), and (3.4a,b) that for all $h \leq \widehat{h}_0(\frac{\alpha}{8})$ and for all $\widetilde{\eta} \in L^\infty(0,T;W^{1,\infty}(\Omega))$

$$\begin{aligned}
&\left| \int_{B_\alpha} (\pi^h[(U_\varepsilon^-)^3] - u^3)\,\nabla W_\varepsilon^+ \,.\, \nabla\eta\,\mathrm{d}x\,\mathrm{d}t \right| \\
&\quad \leq \|\pi^h[(U_\varepsilon^-)^3] - u^3\|_{L^\infty(\Omega_T)}\,\|\nabla W_\varepsilon^+\|_{L^2(B_\alpha)}\,\|\eta\|_{L^2(0,T;H^1(\Omega))} \\
&\quad \leq C\,\alpha^{-\frac{3}{2}}\left[\,\|\pi^h[(U_\varepsilon^-)^3 - u^3]\|_{L^\infty(\Omega_T)} + \|(I-\pi^h)u^3\|_{L^\infty(\Omega_T)}\right]\|\eta\|_{L^2(0,T;H^1(\Omega))},
\end{aligned} \tag{3.35a}$$

$$\begin{aligned}
&\left| \int_{B_\alpha} (\pi^h[(U_\varepsilon^+)^{\frac{1}{2}}\,(U_\varepsilon^-)^{\frac{3}{2}}]\,\Lambda_\varepsilon(V_\varepsilon^+) - u^2\,\lambda(v)\,)\,\nabla W_\varepsilon^+ \,.\, \nabla\eta\,\mathrm{d}x\,\mathrm{d}t \right| \\
&\quad \leq \left| \int_{B_\alpha} (\pi^h[(U_\varepsilon^+)^{\frac{1}{2}}\,(U_\varepsilon^-)^{\frac{3}{2}}] - u^2)\,\Lambda_\varepsilon(V_\varepsilon^+)\,\nabla W_\varepsilon^+ \,.\, \nabla\eta\,\mathrm{d}x\,\mathrm{d}t \right| \\
&\qquad + \left| \int_{B_\alpha} u^2\,(\Lambda_\varepsilon(V_\varepsilon^+) - \lambda(v)\,)\,\nabla W_\varepsilon^+ \,.\, [\,\nabla(\eta-\widetilde{\eta}) + \nabla\widetilde{\eta}\,]\,\mathrm{d}x\,\mathrm{d}t \right| \\
&\quad \leq \|\pi^h[(U_\varepsilon^+)^{\frac{1}{2}}\,(U_\varepsilon^-)^{\frac{3}{2}}] - u^2\|_{L^\infty(\Omega_T)}\,\|\nabla W_\varepsilon^+\|_{L^2(B_\alpha)}\,\|\eta\|_{L^2(0,T;H^1(\Omega))} \\
&\qquad + \|u^2\|_{L^\infty(\Omega_T)}\,\|\Lambda_\varepsilon(V_\varepsilon^+) - \lambda(v)\|_{L^\infty(\Omega_T)}\,\|\nabla W_\varepsilon^+\|_{L^2(B_\alpha)}\,\|\eta-\widetilde{\eta}\|_{L^2(0,T;H^1(\Omega))} \\
&\qquad + \|u^2\|_{L^\infty(\Omega_T)}\,\|\Lambda_\varepsilon(V_\varepsilon^+) - \lambda(v)\|_{L^2(\Omega_T)}\,\|\nabla W_\varepsilon^+\|_{L^2(B_\alpha)}\,\|\widetilde{\eta}\|_{L^\infty(0,T;W^{1,\infty}(\Omega))} \\
&\quad \leq C\,\alpha^{-\frac{3}{2}}\Big[\,\|\pi^h[(U_\varepsilon^+)^{\frac{1}{2}}\,(U_\varepsilon^-)^{\frac{3}{2}} - u^2] - (I-\pi^h)u^2\,\|_{L^\infty(\Omega_T)}\,\|\eta\|_{L^2(0,T;H^1(\Omega))} \\
&\qquad + \|\Lambda_\varepsilon(V_\varepsilon^+) - \lambda(v)\|_{L^2(\Omega_T)}\,\|\widetilde{\eta}\|_{L^\infty(0,T;W^{1,\infty}(\Omega))} + \|\eta-\widetilde{\eta}\|_{L^2(0,T;H^1(\Omega))}\Big].
\end{aligned} \tag{3.35b}$$

From (3.23) we have for all $h \leq \widehat{h}_0(\frac{\alpha}{8})$ and for a.e. $t \in (0,T)$ that $\xi^h(\cdot,t) := U_\varepsilon^+(\cdot,t) \pm \frac{\alpha}{16}\,\zeta^h(\cdot,t)/\|\zeta^h(\cdot,t)\|_{L^\infty(\Omega)} \in K^h$ for any $\zeta^h \in L^2(0,T;S^h)$ with $\mathrm{supp}(\zeta^h) \subset B_{\frac{\alpha}{8}}$.

Choosing $z^h \equiv \xi^h$ in (3.3b) yields, as $\phi^- \equiv 0$, for all $h \leq \widehat{h}_0(\frac{\alpha}{8})$ that

$$\int_0^T \left[\, c\,(\nabla U_\varepsilon^+, \nabla \zeta^h) - (W_\varepsilon^+, \zeta^h)^h \,\right] \mathrm{d}t = 0 \quad \forall\, \zeta^h \in L^2(0,T;S^h) \;\; \text{with}\, \mathrm{supp}(\zeta^h) \subset B_{\frac{\alpha}{8}}. \tag{3.36}$$

Next we derive a bound on W_ε^+ locally on the set $\{u > 0\}$. For any $\alpha \in (0, \alpha_0)$ and any $t \in [0,T]$, we choose a cut-off function $\theta_\alpha(\cdot,t) \in C^\infty(\overline{\Omega})$ such that

$$\begin{aligned} &\theta_\alpha(\cdot,t) \equiv 1 \quad \text{on}\, B_\alpha(t), \qquad 0 \leq \theta_\alpha(\cdot,t) \leq 1 \quad \text{on}\, B_{\frac{\alpha}{2}}(t) \setminus B_\alpha(t), \\ &\theta_\alpha(\cdot,t) \equiv 0 \quad \text{on}\, \overline{\Omega} \setminus B_{\frac{\alpha}{2}}(t) \qquad \text{and} \qquad |\nabla \theta_\alpha(\cdot,t)| \leq C\,\alpha^{-2}. \end{aligned} \tag{3.37}$$

It follows from (3.20) that this last property can be achieved. We have from (3.37) and (3.21) that

$$\mathrm{supp}(\pi^h[\theta_{\frac{\alpha}{2}}^2\, W_\varepsilon^+]\,) \subset B^h_{\frac{\alpha}{4}} \subset B_{\frac{\alpha}{8}} \qquad \forall\, h \leq \widehat{h}_0(\tfrac{\alpha}{8}). \tag{3.38}$$

Next we note, as $d = 1$, that for any $\kappa = (p_j, p_{j+1}) \in \mathcal{T}^h$ and any $z_1, z_2 \in C(\overline{\kappa})$

$$\nabla \pi^h[z_1^2\, z_2] = [\,(z_1\, z_2)(p_j) + (z_1\, z_2)(p_{j+1})\,]\, \nabla \pi^h[z_1] + z_1(p_j)\, z_1(p_{j+1})\, \nabla \pi^h[z_2] \quad \text{on}\ \kappa. \tag{3.39}$$

It follows from (2.1), (3.36), (3.38), (3.39), (3.37), (3.21), and (1.19) that for all $h \leq \widehat{h}_0(\frac{\alpha}{8})$

$$\begin{aligned} \int_{\Omega_T} \pi^h[(\theta_{\frac{\alpha}{2}}\, W_\varepsilon^+)^2]\, \mathrm{d}x\, \mathrm{d}t &= \int_0^T (W_\varepsilon^+, \pi^h[\theta_{\frac{\alpha}{2}}^2\, W_\varepsilon^+]\,)^h\, \mathrm{d}t = \int_0^T c\,(\nabla U_\varepsilon^+, \nabla(\pi^h[\theta_{\frac{\alpha}{2}}^2\, W_\varepsilon^+]\,)\,)\, \mathrm{d}t \\ &\leq C\, \|U_\varepsilon^+\|_{L^2(0,T;H^1(\Omega))} \left[\|\nabla \theta_{\frac{\alpha}{2}}\|_{L^\infty(\Omega_T)} \left[\int_{\Omega_T} \pi^h[(\theta_{\frac{\alpha}{2}}\, W_\varepsilon^+)^2]\, \mathrm{d}x\, \mathrm{d}t \right]^{\frac{1}{2}} + \|\nabla W_\varepsilon^+\|_{L^2(B^h_{\frac{\alpha}{4}})} \right] \\ &\leq C\,(1 + \alpha^{-4})\, \|U_\varepsilon^+\|^2_{L^2(0,T;H^1(\Omega))} + C\, \|\nabla W_\varepsilon^+\|^2_{L^2(B^h_{\frac{\alpha}{4}})}\,. \end{aligned} \tag{3.40}$$

From (3.21), (3.37), and (2.14), we obtain that for all $h \leq \widehat{h}_0(\frac{\alpha}{8})$

$$\int_{\Omega_T} \pi^h[(\theta_{\frac{\alpha}{2}}\, W_\varepsilon^+)^2]\, \mathrm{d}x\, \mathrm{d}t \geq \int_{B^h_\alpha} \pi^h[(W_\varepsilon^+)^2]\, \mathrm{d}x\, \mathrm{d}t \geq \int_{B^h_\alpha} (W_\varepsilon^+)^2\, \mathrm{d}x\, \mathrm{d}t \geq \|W_\varepsilon^+\|^2_{L^2(B_\alpha)}. \tag{3.41}$$

Therefore, combining (3.34), (3.40), (3.41), (3.9), and (3.8) yields that

$$\|W_\varepsilon^+\|_{L^2(0,T;H^1(B_\alpha(t)))} \leq C(\alpha^{-1}) \qquad \forall\, h \leq \widehat{h}_0(\tfrac{\alpha}{8}). \tag{3.42}$$

The bound (3.42) implies the existence of a subsequence and a function $w \in L^2(0,T;H^1(B_\alpha(t)))$ such that

$$W_\varepsilon^+ \to w, \quad \nabla W_\varepsilon^+ \to \nabla w \qquad \text{weakly in}\ L^2(B_\alpha) \quad \text{as}\ h \to 0. \tag{3.43}$$

On noting that $L^\infty(0,T;W^{1,\infty}(\Omega))$ is dense in $L^2(0,T;H^1(\Omega))$, (3.29a,b), (2.19), (3.35a,b), (3.5a), (2.13), (3.6b), and (3.43), we have that as $h \to 0$

$$\int_{B_\alpha} \pi^h[(U_\varepsilon^-)^3]\, \nabla W_\varepsilon^+ . \nabla(\pi^h \eta)\, \mathrm{d}x\, \mathrm{d}t \to \int_{B_\alpha} u^3\, \nabla w\, . \nabla \eta\, \mathrm{d}x\, \mathrm{d}t, \tag{3.44a}$$

$$\int_{B_\alpha} \pi^h[(U_\varepsilon^+)^{\frac{1}{2}}\, (U_\varepsilon^-)^{\frac{3}{2}}]\, \Lambda_\varepsilon(V_\varepsilon^+)\, \nabla W_\varepsilon^+ .\, \nabla(\pi^h \eta)\, \mathrm{d}x\, \mathrm{d}t \to \int_{B_\alpha} u^2 \lambda(v)\, \nabla w\, . \nabla \eta\, \mathrm{d}x\, \mathrm{d}t. \tag{3.44b}$$

Using (2.1), (2.18), and (3.42), we deduce for all $\zeta \in L^2(0,T;H^1(\Omega))$ with $\operatorname{supp}(\zeta) \subset B_\alpha$ and for all $h \leq \widehat{h}_0(\frac{\alpha}{8})$ that

$$\begin{aligned} \left| \int_0^T \left[(W_\varepsilon^+, \pi^h \zeta)^h - (W_\varepsilon^+, \zeta) \right] \mathrm{d}t \right| &= \left| \int_{\Omega_T} (I - \pi^h)(W_\varepsilon^+ \zeta) \, \mathrm{d}x \, \mathrm{d}t \right| \\ &\leq C\, h \int_{\Omega_T} |\nabla (W_\varepsilon^+ \zeta)| \, \mathrm{d}x \, \mathrm{d}t \\ &\leq C\, h \, \|W_\varepsilon^+\|_{L^2(0,T;H^1(B_\alpha(t)))} \, \|\zeta\|_{L^2(0,T;H^1(\Omega))} \\ &\leq C(\alpha^{-1})\, h \, \|\zeta\|_{L^2(0,T;H^1(\Omega))}. \end{aligned} \tag{3.45}$$

It follows from (3.45) and (3.43) that for all $\zeta \in L^2(0,T;H^1(\Omega))$ with $\operatorname{supp}(\zeta) \subset B_\alpha$,

$$\int_0^T (W_\varepsilon^+, \pi^h \zeta)^h \, \mathrm{d}t \to \int_0^T (w, \zeta) \, \mathrm{d}t = \int_{B_\alpha} w \, \zeta \, \mathrm{d}x \, \mathrm{d}t \quad \text{as } h \to 0\,. \tag{3.46}$$

Combining (3.30) for u and (3.46) and noting (3.36) yield that

$$\int_{B_\alpha} [\, c \nabla u \,.\, \nabla \zeta - w\, \zeta \,] \, \mathrm{d}x \, \mathrm{d}t = 0 \quad \forall\, \zeta \in L^2(0,T;H^1(\Omega)) \;\; \text{with } \operatorname{supp}(\zeta) \subset B_\alpha. \tag{3.47}$$

This uniquely defines w in terms of u on the set B_α. Repeating (3.44a,b) for all $\alpha > 0$ and noting (3.33a,b) and (2.19) yield for all $\eta \in L^2(0,T;H^1(\Omega))$ that as $h \to 0$

$$\int_{\Omega_T} \pi^h[(U_\varepsilon^-)^3] \, \nabla W_\varepsilon^+ . \nabla (\pi^h \eta) \, \mathrm{d}x \, \mathrm{d}t \to \int_{B_0} u^3 \, \nabla w \,. \nabla \eta \, \mathrm{d}x \, \mathrm{d}t, \tag{3.48a}$$

$$\int_{\Omega_T} \pi^h[(U_\varepsilon^+)^{\frac{1}{2}} \, (U_\varepsilon^-)^{\frac{3}{2}}] \, \Lambda_\varepsilon(V_\varepsilon^+) \, \nabla W_\varepsilon^+ . \nabla (\pi^h \eta) \, \mathrm{d}x \, \mathrm{d}t \to \int_{B_0} u^2 \, \lambda(v) \, \nabla w \,. \nabla \eta \, \mathrm{d}x \, \mathrm{d}t. \tag{3.48b}$$

Combining (3.3a,c), (3.28), (3.30), (3.32a,b), and (3.48a,b) and repeating (3.47) for all $\alpha > 0$ yield that the functions $\{u,\, w,\, v\}$ satisfy (3.4a,b), (3.24), and (3.25a–c). □

Remark 3.3. If $v^0 \equiv 0$, then (3.25a–c) collapses to

$$\int_0^T \langle \tfrac{\partial u}{\partial t}, \eta \rangle \, \mathrm{d}t - \tfrac{c}{3} \int_{\{u>0\}} u^3 \, \nabla (\Delta u) \,.\, \nabla \eta \, \mathrm{d}x \, \mathrm{d}t = 0 \quad \forall\, \eta \in L^2(0,T;H^1(\Omega)) \tag{3.49}$$

since $v \equiv \lambda(v) \equiv 0$ in Ω_T and $w \equiv -c\, \Delta u$ on $\{u > 0\}$. This is the Bernis–Friedman weak formulation of the degenerate fourth order equation $\frac{\partial u}{\partial t} + \frac{c}{3} \nabla . (u^3 \, \nabla (\Delta u)) = 0$; see [9]. Note that (3.49) incorporates a weak formulation of the boundary condition $\frac{c}{3} \, u^3 \, \frac{\partial \Delta u}{\partial \nu_{\partial\Omega}} = 0$. In addition, (3.25b) implies that $\frac{\partial u}{\partial \nu_{\partial\Omega}}(x,t) = 0$ for $(x,t) \in \partial\Omega \times (0,T)$ whenever $u(x,t) > 0$. Therefore, (3.25a–c) is the natural extension of the Bernis–Friedman weak formulation to the problem (P) in the presence of surfactant ($v^0 \not\equiv 0$).

Remark 3.4. The obstacle formulation in $(\mathrm{P}_\varepsilon^{h,\tau})$ is not crucial in proving well-posedness and convergence of the resulting approximation $\{U_\varepsilon,\, W_\varepsilon,\, V_\varepsilon\}$ to a solution,

$\{u,\, w,\, v\}$, of (P). Replacing $\pi^h[(U_\varepsilon^{n-1})^3]$, $\pi^h[(U_\varepsilon^{n-1})^2]$ by $\pi^h[\,[U_\varepsilon^{n-1}]_+^3]$, $\pi^h[\,[U_\varepsilon^{n-1}]_+^2]$ in (2.11a), the inequality by an equality, K^h by S^h in (2.11b), and $U_\varepsilon^n\,\Lambda_\varepsilon(V_\varepsilon^n)$, $\pi^h[(U_\varepsilon^n)^{\frac{1}{2}}\,(U_\varepsilon^{n-1})^{\frac{3}{2}}]$ by $\pi^h[\,[U_\varepsilon^n]_+]\,\Lambda_\varepsilon(V_\varepsilon^n)$, $\pi^h[\,[U_\varepsilon^n]_+^{\frac{1}{2}}\,[U_\varepsilon^{n-1}]_+^{\frac{3}{2}}]$ in (2.11c), one can easily adapt the proofs of Theorems 2.2, 2.4, and 3.2 and Lemmas 2.3, 2.5, and 3.1. Hence one can pass to a limit $\{u,\, w,\, v\}$ which solves (P) in the sense of (3.25a–c) with u^2 replaced by $[u]_+^2$ in (3.25a) and $u\,\lambda(v)$ replaced by $[u]_+\,\lambda(v)$ in (3.25c). Using $[u]_-$ as a test function in the modified (3.25a), one recovers the nonnegativity of u and hence the weak formulation (3.25a–c). However, as $U_\varepsilon^n(\cdot)$ can now become negative in many disconnected regions where $u(\cdot, t_n) \equiv 0$, this makes the location of the approximate free boundary more difficult.

Remark 3.5. On choosing $U_\varepsilon^0 \equiv \pi^h u^0$ and $V_\varepsilon^0 \equiv \pi^h v^0$, we need the quasi uniformity assumption on the partitioning $\mathcal{T}^h$ only in order to obtain the bound (2.48b) via (2.51), (2.52), and (2.21) and the bound (3.26) via (2.23). However, we can replace this with the far milder assumption that $\{\mathcal{T}^h\}_{h>0}$ is a regular partitioning at the expense of a minimum time step constraint as in [4]. It is easily established from (1.16), (2.22), $\{\mathcal{T}^h\}_{h>0}$ being a regular partitioning, elliptic regularity, assuming that Ω is convex polygonal if $d=2$, and (2.17) that

$$\|(\mathcal{G}-\mathcal{G}^h)z^h\|_1 \leq C\,h\,\|z^h\|_0 \qquad \forall\; z^h \in Z^h\,. \tag{3.50}$$

Then choosing $\chi \equiv \mathcal{G}^h[\frac{U_\varepsilon^n - U_\varepsilon^{n-1}}{\tau_n}]$ and $\chi \equiv \mathcal{G}^h[\frac{V_\varepsilon^n - V_\varepsilon^{n-1}}{\tau_n}]$ in (2.11a) and (2.11c), respectively, we obtain, similarly to (2.51) and (2.52), that

$$\|\mathcal{G}^h\tfrac{\partial U_\varepsilon}{\partial t}\|_{L^2(0,T;H^1(\Omega))} + \|\mathcal{G}^h\tfrac{\partial V_\varepsilon}{\partial t}\|_{L^2(0,T;H^1(\Omega))} \leq C(\|U_\varepsilon\|_{L^\infty(\Omega_T)}) \tag{3.51}$$

on noting (2.53). Combining (3.50) and (3.51) and noting the fifth and sixth bound in (2.48a), it follows for $Z \equiv U_\varepsilon$ or V_ε that

$$\begin{aligned}
\|\mathcal{G}\tfrac{\partial Z}{\partial t}\|_{L^2(0,T;H^1(\Omega))} &\leq \|(\mathcal{G}-\mathcal{G}^h)\tfrac{\partial Z}{\partial t}\|_{L^2(0,T;H^1(\Omega))} + \|\mathcal{G}^h\tfrac{\partial Z}{\partial t}\|_{L^2(0,T;H^1(\Omega))}\\
&\leq C\,h\,\|\tfrac{\partial Z}{\partial t}\|_{L^2(\Omega_T)} + C(\|U_\varepsilon\|_{L^\infty(\Omega_T)})\\
&\leq C(\|U_\varepsilon\|_{L^\infty(\Omega_T)})\,(\tau_{\min}^{-\frac{1}{2}}\,h+1) \leq C(\|U_\varepsilon\|_{L^\infty(\Omega_T)})
\end{aligned}$$

if the mild time step constraint $C\,h^2 \leq \tau_{min} := \min_{n=1\to N}\tau_n$ is satisfied.

3.1. Inclusion of van der Waals forces.

We now extend Lemma 3.1 and Theorem 3.2 to the approximation $(\mathrm{P}_{\delta,\,\varepsilon}^{h,\tau})$.

Lemma 3.6. *Let $d=1, \rho>0$ and $u^0,\, v^0 \in K$, with $u^0(x) \geq \zeta > 0$ for all $x \in \Omega$. Let $\{\mathcal{T}^h, U_\varepsilon^0, V_\varepsilon^0, \tau,\, \varepsilon\}_{h>0}$ be such that assumptions* (i), (ii), *and* (iii) *of Lemma* 3.1 *hold. Then there exist a subsequence of $\{U_\varepsilon, V_\varepsilon\}_h$, where $\{U_\varepsilon,\, W_\varepsilon,\, V_\varepsilon\}$ solve $(\mathrm{P}_{\delta,\,\varepsilon}^{h,\tau})$, and functions $\{u,\, v\}$ satisfying* (3.4a,b) *with $u(x,0) = u^0(x)$ for all $x \in \overline{\Omega}$, $v(\cdot,0) = v^0(\cdot)$ in $(H^1(\Omega))'$, $⨍ u(\cdot,t) = ⨍ u^0 > 0$ for all $t \in [0,T]$, and $⨍ v(\cdot,t) = ⨍ v^0$ for a.e. $t \in [0,T]$, such that as $h \to 0$* (3.5a,b) *and* (3.6a–c) *hold.*

Proof. The proof is exactly the same as that of Lemma 3.1. □

Theorem 3.7. *Let the assumptions of Lemma* 3.6 *hold. Then there exist a subsequence of $\{U_\varepsilon, W_\varepsilon, V_\varepsilon\}_h$, where $\{U_\varepsilon,\, W_\varepsilon,\, V_\varepsilon\}$ solve $(\mathrm{P}_{\delta,\,\varepsilon}^{h,\tau})$, and functions $\{u, w, v\}$ satisfying* (3.4a,b), *$w \in L^2(0,T;H^1(\Omega))$, and $u>0$ on $\overline{\Omega}_T$. In addition, as $h \to 0$,* (3.5a,b), (3.6a–c), *and $W_\varepsilon^+ \to w$ weakly in $L^2(0,T;H^1(\Omega))$ hold. Furthermore, we* *have that* u, v, *and* w *fulfil* $u(\cdot,0) = u^0(\cdot)$, $v(\cdot,0) = v^0(\cdot)$ *and are such that for all* $\eta \in L^2(0,T;H^1(\Omega))$,

$$\int_0^T \langle \tfrac{\partial u}{\partial t}, \eta\rangle \,\mathrm{d}t + \tfrac{1}{3}\int_{\Omega_T} u^3\, \nabla w \,.\, \nabla \eta \,\mathrm{d}x\,\mathrm{d}t + \tfrac{1}{2}\int_{\Omega_T} u^2\, \nabla v \,.\, \nabla\eta \,\mathrm{d}x\,\mathrm{d}t = 0,$$
$$\int_{\Omega_T} [\, c\,\nabla u \,.\, \nabla \eta + \phi(u)\,\eta - w\,\eta\,]\,\mathrm{d}x\,\mathrm{d}t = 0,$$
$$\int_0^T \langle \tfrac{\partial v}{\partial t}, \eta\rangle \,\mathrm{d}t + \int_{\Omega_T} [\,\rho\,\nabla v \,.\, \nabla\eta + u\,\lambda(v)\,\nabla v \,.\, \nabla \eta\,]\,\mathrm{d}x\,\mathrm{d}t + \tfrac{1}{2}\int_{\Omega_T} u^2\,\lambda(v)\,\nabla w \,.\, \nabla\eta \,\mathrm{d}x\,\mathrm{d}t = 0.$$

Proof. Theorem 2.8 and Lemma 2.9 imply that

$$\max_{t\in[0,T]} \int_\Omega \pi^h[\Phi(U_\varepsilon^+)](x,t)\,\mathrm{d}x \leq C. \tag{3.52}$$

From the uniform Hölder continuity of U_ε^+ and (3.52), it follows that there exists $\underline{u} \in \mathbb{R}_{>0}$ independent of h, τ, and ε such that $U_\varepsilon^+(x,t) \geq \underline{u} > 0$ for all $(x,t) \in \overline{\Omega}_T$; see, e.g., [23, Corollary 5.3]. Combining this with (3.5a) yields that u is strictly positive.

The rest of the proof is similar to that of Theorem 3.2 with the following minor modifications. We have that (3.36) holds with the extra term $(\phi^+(U_\varepsilon^+)+\phi^-(U_\varepsilon^-),\zeta^h)^h$ inside the square brackets on the left-hand side. Since u is strictly positive, it is straightforward to show convergence of this extra term to the corresponding term in the weak formulation of the continuous problem. ∎

4. Solution of the nonlinear discrete system. We now discuss algorithms for solving the resulting system of nonlinear equations for $\{U_\varepsilon^n, W_\varepsilon^n, V_\varepsilon^n\}$ at each time level for the approximations $(\mathrm{P}_\varepsilon^{h,\tau})$ and $(\mathrm{P}_{\delta,\varepsilon}^{h,\tau})$. As (2.11a,b) for $(\mathrm{P}_\varepsilon^{h,\tau})$ and (2.11a), (2.12) for $(\mathrm{P}_{\delta,\varepsilon}^{h,\tau})$ are independent of V_ε^n, we first solve these to obtain $\{U_\varepsilon^n, W_\varepsilon^n\}$; then we solve (2.11c) for V_ε^n. First, we consider $(\mathrm{P}_\varepsilon^{h,\tau})$. Adapting the techniques in [4, section 3] we introduce $R_\varepsilon^n \in S^h$ by

$$(R_\varepsilon^n, \chi)^h = c\,(\nabla U_\varepsilon^n, \nabla \chi) + (\phi^-(U_\varepsilon^{n-1}+\varepsilon),\chi)^h - (W_\varepsilon^n,\chi)^h \quad \forall\, \chi\in S^h. \tag{4.1}$$

Hence, for any $\mu \in \mathbb{R}_{>0}$ and on recalling (3.13), $(\mathrm{P}_\varepsilon^{h,\tau})$ is equivalent to the following.

Given $U_\varepsilon^0 \in K^h$, $V_\varepsilon^0 \in S^h$, for $n \geq 1$ find $\{U_\varepsilon^n, W_\varepsilon^n, V_\varepsilon^n\} \in [S^h]^3$ such that (2.11a), (4.1), $R_\varepsilon^n = \pi^h[R_\varepsilon^n - \mu\, U_\varepsilon^n]_+$, and (2.11c) hold. We use this formulation in constructing our iterative method to solve $(\mathrm{P}_\varepsilon^{h,\tau})$.

Given $\{W_\varepsilon^{n,0}, R_\varepsilon^{n,0}\} \in [S^h]^2$, for $k \geq 1$ find $\{U_\varepsilon^{n,k}, W_\varepsilon^{n,k}, R_\varepsilon^{n,k}\} \in [S^h]^3$ such that for all $\chi \in S^h$

$$\begin{aligned}\left(\tfrac{U_\varepsilon^{n,k}-U_\varepsilon^{n-1}}{\tau_n},\chi\right)^h + \tfrac{b^{n-1}}{3}\,(\nabla W_\varepsilon^{n,k},\nabla\chi) &= \tfrac{1}{3}\,((b^{n-1} - \pi^h[(U_\varepsilon^{n-1})^3])\nabla W_\varepsilon^{n,k-1},\nabla\chi)\\ &\quad - \tfrac{1}{2}\,(\pi^h[(U_\varepsilon^{n-1})^2]\,\nabla V_\varepsilon^{n-1},\nabla\chi),\end{aligned} \tag{4.2a}$$

(4.2b)
$$c\,(\nabla U_\varepsilon^{n,k},\nabla\chi) + (\phi^-(U_\varepsilon^{n-1}+\varepsilon),\chi)^h = (W_\varepsilon^{n,k} + R_\varepsilon^{n,k-1},\chi)^h,$$

$$R_\varepsilon^{n,k} = \pi^h[R_\varepsilon^{n,k-1} - \mu\, U_\varepsilon^{n,k}]_+, \tag{4.2c}$$

where $b^{n-1} := |U_\varepsilon^{n-1}|_{0,\infty}^3$.

Then, having obtained $\{U_\varepsilon^n, W_\varepsilon^n\}$, we find V_ε^n as follows. Given $V_\varepsilon^{n,0} \in S^h$, for $k \geq 1$ find $V_\varepsilon^{n,k} \in S^h$ such that

$$\begin{aligned}\Big(\tfrac{V_\varepsilon^{n,k}-V_\varepsilon^{n-1}}{\tau_n},\chi\Big)^h &+ \rho\,(\nabla V_\varepsilon^{n,k},\nabla\chi) + (U_\varepsilon^n\,\Lambda_\varepsilon(V_\varepsilon^{n,k-1})\,\nabla V_\varepsilon^{n,k},\nabla\chi)\\ &= -\tfrac{1}{2}\,(\pi^h[(U_\varepsilon^n)^{\frac{1}{2}}\,(U_\varepsilon^{n-1})^{\frac{3}{2}}]\,\Lambda_\varepsilon(V_\varepsilon^{n,k-1})\,\nabla W_\varepsilon^n,\nabla\chi)\quad\forall\,\chi\in S^h.\end{aligned}\tag{4.3}$$

Equation (4.3) is the natural extension of the iterative procedure proposed in [22] for solving a finite element approximation of the thin film equation. As (4.3) is linear, existence of $V_\varepsilon^{n,k}$ follows from uniqueness; and this is easily established on noting $\rho \geq 0$, (2.5), and $U_\varepsilon^n \in K^h$. Hence the iteration (4.3) is well defined.

The algorithm (4.2a–c) is a simple adaptation of the algorithm in [4, section 3] for problem $(\mathrm{P}_\varepsilon^{h,\tau})$ in the absence of the surfactant and van der Waals forces, i.e., $V_\varepsilon^n \equiv 0$ and $a = 0\,(\phi^- \equiv 0)$.

Defining $A^{n,k-1} \in Z^h$ such that

$$(A^{n,k-1},\chi)^h := \tfrac{1}{3}\,(\pi^h[(U_\varepsilon^{n-1})^3]\,\nabla W_\varepsilon^{n,k-1},\nabla\chi)\quad\forall\,\chi\in S^h\tag{4.4}$$

and $X_\varepsilon^{n-1} \in S^h$ as in (2.30), it follows from (4.2a), (2.22), (4.2b) with $\chi \equiv 1$, (1.17), and (1.14) that

$$\begin{aligned}W_\varepsilon^{n,k} = (I-\fint)W_\varepsilon^{n,k-1} &- \tfrac{3}{b^{n-1}}\,\mathcal{G}^h[\tfrac{U_\varepsilon^{n,k}-U_\varepsilon^{n-1}}{\tau_n} + A^{n,k-1} + X_\varepsilon^{n-1}]\\ &+ \fint\pi^h[\phi^-(U_\varepsilon^{n-1}+\varepsilon)] - \fint R_\varepsilon^{n,k-1}\,.\end{aligned}\tag{4.5}$$

Therefore, (4.2a,b) may be written equivalently as follows: Find $U_\varepsilon^{n,k} \in \overline{S}^h(U_\varepsilon^{n-1}) := \{\,\chi \in S^h : \chi - U_\varepsilon^{n-1} \in Z^h\,\}$ such that

$$\begin{aligned}c\,(\nabla U_\varepsilon^{n,k},\nabla\chi) &+ \tfrac{3}{b^{n-1}}\,(\mathcal{G}^h[\tfrac{U_\varepsilon^{n,k}-U_\varepsilon^{n-1}}{\tau_n}],\chi)^h\\ &= ((I-\fint)(W_\varepsilon^{n,k-1} + R_\varepsilon^{n,k-1} + \overline{X}_\varepsilon^{n,k-1}),\chi)^h\quad\forall\,\chi\in S^h,\end{aligned}\tag{4.6}$$

where $\overline{X}_\varepsilon^{n,k-1} \in S^h$ is such that

$$(\overline{X}_\varepsilon^{n,k-1},\chi)^h := -(\phi^-(U_\varepsilon^{n-1}+\varepsilon) + \tfrac{3}{b^{n-1}}\,\mathcal{G}^h[A^{n,k-1} + X_\varepsilon^{n-1}],\chi)^h\quad\forall\,\chi\in S^h\,.$$

Existence and uniqueness of $U_\varepsilon^{n,k} \in \overline{S}^h(U_\varepsilon^{n-1})$ satisfying (4.6) then follows since, on noting (2.22), this is the Euler–Lagrange equation of the convex minimization problem

$$\min_{\chi\in\overline{S}^h(U_\varepsilon^{n-1})}\Big\{\tfrac{c}{2}\,|\chi|_1^2 + \tfrac{3}{2\,b^{n-1}\,\tau_n}|\nabla\mathcal{G}^h(\chi - U_\varepsilon^{n-1})|_0^2 - (W_\varepsilon^{n,k-1} + R_\varepsilon^{n,k-1} + \overline{X}_\varepsilon^{n,k-1},\chi)^h\Big\}.\tag{4.7}$$

Finally, $W_\varepsilon^{n,k}$ and $R_\varepsilon^{n,k}$ are uniquely defined by (4.5) and (4.2c), respectively. Hence the iterative procedure (4.2a–c) is well defined for any $\mu > 0$.

THEOREM 4.1. *Let the assumptions* (A) *hold. Then there exists a* μ_0 *such that for all* $\mu \in (0,\mu_0)$ *and* $\{W_\varepsilon^{n,0}, R_\varepsilon^{n,0}\} \in [S^h]^2$ *the sequence* $\{U_\varepsilon^{n,k}, W_\varepsilon^{n,k}\}_{k\geq 0}$ *generated by the algorithm* (4.2a–c) *satisfies*

$$U_\varepsilon^{n,k} \to U_\varepsilon^n \quad \textit{and} \quad \int_\Omega \pi^h[(U_\varepsilon^{n-1})^3]\,|\nabla(W_\varepsilon^n - W_\varepsilon^{n,k})|^2\,\mathrm{d}x \to 0 \qquad \textit{as } k\to\infty.\tag{4.8}$$

Proof. This is a simple adaptation of the proof of Theorem 3.1 in [4]. On letting

$$E^k := U_\varepsilon^n - U_\varepsilon^{n,k} \in Z^h, \quad F^k := W_\varepsilon^n - W_\varepsilon^{n,k} \in S^h, \quad \text{and} \quad D^k := R_\varepsilon^n - R_\varepsilon^{n,k} \in S^h,$$

it is an easy exercise to show that

$$\begin{aligned} &|D^k|_h^2 + \tfrac{\mu\,\tau_n}{3}\,|\,[b^{n-1} - \pi^h[(U_\varepsilon^{n-1})^3]\,]^{\frac12}\,\nabla F^k|_0^2 + \tfrac23\,\mu\,\tau_n\,|\,[\pi^h[(U_\varepsilon^{n-1})^3]\,]^{\frac12}\,\nabla F^k|_0^2 \\ &+ (2\,\mu\,c - C\,\mu^2)\,|E^k|_1^2 \;\leq\; |D^{k-1}|_h^2 + \tfrac{\mu\,\tau_n}{3}\,|\,[b^{n-1} - \pi^h[(U_\varepsilon^{n-1})^3]\,]^{\frac12}\,\nabla F^{k-1}|_0^2. \end{aligned} \tag{4.9}$$

Now (4.9) yields that $\{\,|D^k|_h^2 + \frac{\mu\,\tau_n}{3}\,|\,[b^{n-1} - \pi^h[(U_\varepsilon^{n-1})^3]\,]^{\frac12}\,\nabla F^k|_0^2\,\}_{k\geq 0}$ is a decreasing sequence for μ sufficiently small and hence has a limit. Therefore, the desired results (4.8) follow from this and (4.9). □

Remark 4.2. The linear system (4.3) can be solved efficiently using a conjugate gradient algorithm. Although we are unable to show convergence of the iteration (4.3) for V_ε^n, we observed good convergence properties in practice.

4.1. Inclusion of repulsive van der Waals forces. We now consider an algorithm for solving the nonlinear algebraic system at each time level in $(\mathrm{P}_{\delta,\varepsilon}^{h,\tau})$. Our method for $\{U_\varepsilon^n, W_\varepsilon^n\}$ satisfying (2.11a) and (2.12) is based on the general splitting algorithm of [25]; see also [2, 5], where this algorithm has been adapted to solve similar variational inequality problems arising from Cahn–Hilliard systems. V_ε^n satisfying (2.11c) is solved as before using (4.3). We now introduce our algorithm for $\{U_\varepsilon^n, W_\varepsilon^n\}$. Let $\mathcal{B}_n : S^h \to S^h$ be such that for all $q^h \in S^h$, $\chi \in S^h$

$$(\mathcal{B}_n(q^h), \chi)^h := c\,(\nabla q^h, \nabla \chi) + (\phi^-(U_\varepsilon^{n-1}), \chi)^h\,.$$

Hence (2.12) can be rewritten as

$$(\mathcal{B}_n(U_\varepsilon^n) + \phi^+(U_\varepsilon^n), \chi)^h = (W_\varepsilon^n, \chi)^h \qquad \forall\ \chi \in S^h\,. \tag{4.10}$$

Now, for n fixed, multiplying (4.10) with $\mu \in \mathbb{R}_{>0}$, adding $(U_\varepsilon^n, \chi)^h$ to both sides, rearranging on noting (2.11a), and defining $X_\varepsilon^{n-1} \in S^h$ as in (2.30), it follows that $\{U_\varepsilon^n,\, W_\varepsilon^n\} \in [S^h]^2$ solving (2.11a) and (2.12) satisfy for all $\chi \in S^h$

$$\begin{aligned} \left(\tfrac{U_\varepsilon^n - U_\varepsilon^{n-1}}{\tau_n}, \chi\right)^h + \tfrac{b^{n-1}}{3}\,(\nabla W_\varepsilon^n, \nabla \chi) = \tfrac13\,((b^{n-1} - \pi^h[(U_\varepsilon^{n-1})^3])\,\nabla W_\varepsilon^n, \nabla\chi) \\ - (X_\varepsilon^{n-1}, \chi)^h, \end{aligned} \tag{4.11a}$$

$$(U_\varepsilon^n + \mu\,\phi^+(U_\varepsilon^n), \chi)^h = (Y_\varepsilon^n, \chi)^h, \tag{4.11b}$$

where $Y_\varepsilon^n \in S^h$ is such that

$$(Y_\varepsilon^n, \chi)^h := (U_\varepsilon^n, \chi)^h - \mu\,(\mathcal{B}_n(U_\varepsilon^n) - W_\varepsilon^n, \chi)^h \qquad \forall\ \chi \in S^h\,. \tag{4.11c}$$

For later use we introduce also $\overline{Y}_\varepsilon^n \in S^h$ such that

$$(\overline{Y}_\varepsilon^n, \chi)^h := (U_\varepsilon^n, \chi)^h + \mu\,(\mathcal{B}_n(U_\varepsilon^n) - W_\varepsilon^n, \chi)^h \qquad \forall\ \chi \in S^h \tag{4.11d}$$

and note that $\overline{Y}_\varepsilon^n = 2\,U_\varepsilon^n - Y_\varepsilon^n$. We use this as a basis for constructing our iterative procedure to find $\{U_\varepsilon^n,\, W_\varepsilon^n\} \in [S^h]^2$ satisfying (4.11a,b).

Given $\{U_\varepsilon^{n,k-1},\, W_\varepsilon^{n,k-1}\} \in [S^h]^2$ for $k \geq 1$, we define $Y_\varepsilon^{n,k-1} \in S^h$ such that

$$(Y_\varepsilon^{n,k-1}, \chi)^h := (U_\varepsilon^{n,k-1}, \chi)^h - \mu\,(\mathcal{B}_n(U_\varepsilon^{n,k-1}) - W_\varepsilon^{n,k-1}, \chi)^h \qquad \forall\ \chi \in S^h\,. \tag{4.12a}$$

Then we find $U_\varepsilon^{n,k-\frac{1}{2}} \in S^h$ such that

$$(U_\varepsilon^{n,k-\frac{1}{2}} + \mu\,\phi^+(U_\varepsilon^{n,k-\frac{1}{2}}),\chi)^h = (Y_\varepsilon^{n,k-1},\chi)^h \qquad \forall\,\chi \in S^h \tag{4.12b}$$

and find $\{U_\varepsilon^{n,k},\, W_\varepsilon^{n,k}\} \in [S^h]^2$ such that for all $\chi \in S^h$

$$\begin{aligned}\left(\tfrac{U_\varepsilon^{n,k}-U_\varepsilon^{n-1}}{\tau_n},\chi\right)^h &+ \tfrac{b^{n-1}}{3}\,(\nabla W_\varepsilon^{n,k},\nabla\chi)\\ &= \tfrac{1}{3}\,((b^{n-1}-\pi^h[(U_\varepsilon^{n-1})^3])\,\nabla W_\varepsilon^{n,k-1},\nabla\chi) - (X_\varepsilon^{n-1},\chi)^h,\end{aligned} \tag{4.12c}$$

$$(U_\varepsilon^{n,k},\chi)^h + \mu\,(\mathcal{B}_n(U_\varepsilon^{n,k}) - W_\varepsilon^{n,k},\chi)^h = (\overline{Y}_\varepsilon^{n,k},\chi)^h, \tag{4.12d}$$

where $\overline{Y}_\varepsilon^{n,k} := 2\,U_\varepsilon^{n,k-\frac{1}{2}} - Y_\varepsilon^{n,k-1}$. Existence and uniqueness of $U_\varepsilon^{n,k-\frac{1}{2}} > 0$ in (4.12b) follow from the monotonicity of $\varphi : \mathbb{R}_{>0} \to \mathbb{R}$, where $\varphi(s) := s + \mu\,\phi^+(s)$, and the fact that $\lim_{s\to\infty}\varphi(s) = -\lim_{s\searrow 0}\varphi(s) = \infty$.

It remains to show that (4.12c,d) possess a unique solution $\{U_\varepsilon^{n,k},\, W_\varepsilon^{n,k}\} \in [S^h]^2$. This is a simple adaptation of the existence and uniqueness proof for $\{U_\varepsilon^{n,k}, W_\varepsilon^{n,k}, R_\varepsilon^{n,k}\}$ in (4.2a–c). Similarly to (4.5), it follows from (4.12c), (2.22), (4.12d) with $\chi \equiv 1$, (1.17), and (1.14) that

$$\begin{aligned}W_\varepsilon^{n,k} = (I - \fint)W_\varepsilon^{n,k-1} &- \tfrac{3}{b^{n-1}}\,\mathcal{G}^h[\tfrac{U_\varepsilon^{n,k}-U_\varepsilon^{n-1}}{\tau_n} + A^{n,k-1} + X_\varepsilon^{n-1}]\\ &+ \fint\pi^h[\phi^-(U_\varepsilon^{n-1})] + \mu^{-1}\,\fint(U_\varepsilon^{n,k} - \overline{Y}_\varepsilon^{n,k})\,,\end{aligned} \tag{4.13}$$

where $A^{n,k-1} \in Z^h$ is defined as in (4.4). Then similarly to (4.6), (4.12c,d) may be written equivalently as follows: Find $U_\varepsilon^{n,k} \in \overline{S}^h(U_\varepsilon^{n-1}) \equiv S^h(U_\varepsilon^{n-1})$ such that for all $\chi \in S^h$

$$\begin{aligned}(U_\varepsilon^{n,k},(I-\fint)\chi)^h &+ \mu\left[c\,(\nabla U_\varepsilon^{n,k},\nabla\chi) + \tfrac{3}{b^{n-1}}\,(\mathcal{G}^h[\tfrac{U_\varepsilon^{n,k}-U_\varepsilon^{n-1}}{\tau_n}],\chi)^h\right]\\ = (\overline{Y}_\varepsilon^{n,k} &+ \mu\left[W_\varepsilon^{n,k-1} - \phi^-(U_\varepsilon^{n-1}) - \tfrac{3}{b^{n-1}}\,\mathcal{G}^h[A^{n,k-1}+X_\varepsilon^{n-1}]\right],(I-\fint)\chi)^h.\end{aligned} \tag{4.14}$$

Similarly to (4.7), existence and uniqueness of $U_\varepsilon^{n,k} \in \overline{S}^h(U_\varepsilon^{n-1})$ satisfying (4.14) then follow since this is the Euler–Lagrange equation of the convex minimization problem

$$\begin{aligned}\min_{\chi\in\overline{S}^h(U_\varepsilon^{n-1})} &\left\{\tfrac{1}{2}\,|\chi|_h^2 + \mu\left[\tfrac{c}{2}\,|\chi|_1^2 + \tfrac{3}{2\,b^{n-1}\,\tau_n}|\nabla\mathcal{G}^h(\chi - U_\varepsilon^{n-1})|_0^2\right]\right.\\ &\left. -(\overline{Y}_\varepsilon^{n,k} + \mu\left[W_\varepsilon^{n,k-1} - \phi^-(U_\varepsilon^{n-1}) - \tfrac{3}{b^{n-1}}\,\mathcal{G}^h[A^{n,k-1}+X_\varepsilon^{n-1}]\right],\chi)^h\right\}.\end{aligned}$$

Finally, $W_\varepsilon^{n,k}$ is uniquely defined by (4.13). Hence the iterative procedure (4.12a–d) is well defined for any $\mu > 0$.

THEOREM 4.3. *Let the assumptions* (A) *hold. Then for all* $\mu \in \mathbb{R}_{>0}$ *and* $\{U_\varepsilon^{n,0}, W_\varepsilon^{n,0}\} \in [S^h]^2$ *the sequence* $\{U_\varepsilon^{n,k}, W_\varepsilon^{n,k}\}_{k\geq 0}$ *generated by the algorithm* (4.12a–d) *satisfies*

$$U_\varepsilon^{n,k} \to U_\varepsilon^n \quad \textit{and} \quad \int_\Omega \pi^h[(U_\varepsilon^{n-1})^3]\,|\nabla(W_\varepsilon^n - W_\varepsilon^{n,k})|^2\,\mathrm{d}x \to 0 \qquad \textit{as } k \to \infty. \tag{4.15}$$

In addition, it holds that $U_\varepsilon^{n,k-\frac{1}{2}} \to U_\varepsilon^n$ *as* $k \to \infty$.

Proof. A simple adaptation of the proof of Theorem 3.1 in [5] yields that

$$\begin{aligned} & c\,|U_\varepsilon^n - U_\varepsilon^{n,k}|_1^2 + \tfrac{1}{4\mu}\,|Y_\varepsilon^n - Y_\varepsilon^{n,k}|_h^2 + \tfrac{\tau_n}{6}\,|\,[b^{n-1} - \pi^h[(U_\varepsilon^{n-1})^3]\,]^{\frac12}\,\nabla(W_\varepsilon^n - W_\varepsilon^{n,k})|_0^2 \\ & \quad + \tfrac{\tau_n}{3}\,|\,[\pi^h[(U_\varepsilon^{n-1})^3]\,]^{\frac12}\,\nabla(W_\varepsilon^n - W_\varepsilon^{n,k})|_0^2 + (\phi^+(U_\varepsilon^n) - \phi^+(U_\varepsilon^{n,k-\frac12}), U_\varepsilon^n - U_\varepsilon^{n,k-\frac12})^h \\ & \qquad \leq \tfrac{1}{4\mu}\,|Y_\varepsilon^n - Y_\varepsilon^{n,k-1}|_h^2 + \tfrac{\tau_n}{6}\,|\,[b^{n-1} - \pi^h[(U_\varepsilon^{n-1})^3]\,]^{\frac12}\,\nabla(W_\varepsilon^n - W_\varepsilon^{n,k-1})|_0^2\,. \end{aligned} \tag{4.16}$$

Therefore, on noting the monotonicity of ϕ^+, we have that $\{\,\frac{1}{4\mu}\,|Y_\varepsilon^n - Y_\varepsilon^{n,k}|_h^2 + \frac{\tau_n}{6}\,|\,[b^{n-1} - \pi^h[(U_\varepsilon^{n-1})^3]\,]^{\frac12}\,\nabla(W_\varepsilon^n - W_\varepsilon^{n,k})|_0^2\,\}_{k\geq 0}$ is a decreasing sequence for all $\mu > 0$. Since it is bounded below, the sequence has a limit. Hence the desired results follow from this and (4.16). □

Remark 4.4. Note that the algorithm (4.12a–d) can easily be modified to solve the variational inequality that arises at each time step in $(\mathrm{P}_\varepsilon^{h,\tau})$. In particular, let $\overline{\mathcal{B}}_n : S^h \to S^h$ be such that $(\overline{\mathcal{B}}_n(q^h), \chi)^h := c\,(\nabla q^h, \nabla\chi) + (\phi^-(U_\varepsilon^{n-1} + \varepsilon), \chi)^h$ for all $q^h \in S^h$, $\chi \in S^h$, substitute $\overline{\mathcal{B}}_n$ for $\mathcal{B}_n$ in (4.12a–d), and replace (4.12b) with the following: Find $U_\varepsilon^{n,k-\frac12} \in K^h$ such that $(U_\varepsilon^{n,k-\frac12} - Y_\varepsilon^{n,k}, \eta - U_\varepsilon^{n,k-\frac12})^h \geq 0$ for all $\eta \in K^h$. Then this new procedure satisfies the statement of Theorem 4.3 as well; see [5, section 3] for a similar proof. However, we employed algorithm (4.2a–c) to solve $(\mathrm{P}_\varepsilon^{h,\tau})$ since in practice it exhibited superior convergence properties.

Remark 4.5. We see from (4.6) for $(\mathrm{P}_\varepsilon^{h,\tau})$ and (4.14) for $(\mathrm{P}_{\delta,\varepsilon}^{h,\tau})$ that at each iteration for U_ε^n one needs to solve only a fixed linear system with constant coefficients. On a uniform mesh this can be done efficiently using a discrete cosine transform; see [12, section 5], where a similar problem is solved.

5. Numerical results. First, we present numerical experiments in one space dimension in the absence of van der Waals forces, $a = \delta = 0\,(\phi \equiv 0)$. Throughout we chose a uniform partitioning of $\Omega = (-L, L)$, where $L \geq 1$, with mesh points $p_j = -L + (j-1)h$, $j = 1 \to \#J$, where $h = \frac{2L}{\#J-1}$. In addition, we chose uniform time steps $\tau_n = \tau = 1.28 \times 10^{-2}h$ and set the regularization parameter $\varepsilon = 1.28 \times 10^{-3}h$. For the initial profiles $u^0(x)$ and $v^0(x)$, we chose either

$$\text{(i)} \qquad u^0(x) = [\tfrac14 - x^2]_+ \qquad \text{or} \qquad \text{(ii)} \qquad u^0(x) = 1 \tag{5.1a}$$

$$\text{with} \qquad v^0(x) = \tfrac{v_{\max}^0}{2}\,[(1-\gamma) - \tanh(A(|x| - x_0))]_+\,, \tag{5.1b}$$

where $v_{\max}^0 \geq 0$, $\gamma \in [0,1)$, $A > 0$, and $x_0 \in (0, L)$. (i) with $v_{\max}^0 > 0$, (i) with $v_{\max}^0 = 0$, and (ii) with $v_{\max}^0 > 0$ resemble a liquid drop on a plain surface with and without surfactant on top of it and a uniform liquid film with surfactant, respectively. Note that for $\gamma > 0$ the surfactant v^0 has compact support $[-l, l]$, where $l = x_0 + A^{-1}\tanh^{-1}(1-\gamma)$. Throughout we chose $U_\varepsilon^0 \equiv \pi^h u^0$ and $V_\varepsilon^0 \equiv \pi^h v^0$ as the discrete initial data for $(\mathrm{P}_\varepsilon^{h,\tau})$ and $(\mathrm{P}_{\delta,\varepsilon}^{h,\tau})$.

For the iterative algorithms (4.2a–c), (4.12a–d), and (4.3), we set, for $n \geq 1$, $Z^{n,0} \equiv Z^{n-1}$ for $Z = U_\varepsilon, W_\varepsilon, V_\varepsilon$, and R_ε, where

$$R_\varepsilon^0 = 0 \quad \text{and} \quad (W_\varepsilon^0, \chi)^h = c\,(\nabla U_\varepsilon^0, \nabla\chi) + (\phi(U_\varepsilon^0), \chi)^h \qquad \forall\, \chi \in S^h\,,$$

and for each n adopted the stopping criteria

$$|U_\varepsilon^{n,k} - U_\varepsilon^{n,k-1}|_{0,\infty} < tol \quad \text{and} \quad |V_\varepsilon^{n,k} - V_\varepsilon^{n,k-1}|_{0,\infty} < tol, \tag{5.2}$$

TABLE 1
(i) *with* $v^0_{\max} = 0$, *source-type solution errors.*

$\#J$	65	129	257	513	1025	2049
$\max_{n=1\to N} \|\pi^h u(\cdot, t_n) - U^n_\varepsilon(\cdot)\|_{0,\infty} \times 10^4$	74.68	62.90	42.40	8.263	2.538	0.900
$\max_{n=1\to N} \|\pi^h v(\cdot, t_n) - V^n_\varepsilon(\cdot)\|_{0,\infty} \times 10^5$	1.549	1.391	2.041	2.944	3.499	3.989

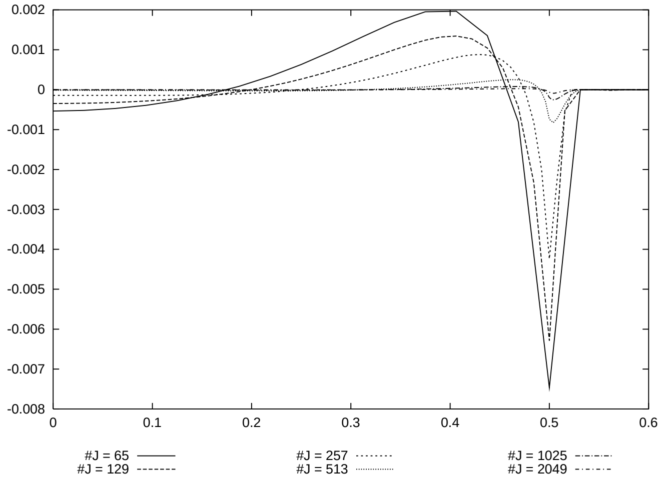


FIG. 1. $\pi^h u^0(x) - U_\varepsilon(x,T)$ *plotted against* x *for* $T = 4$ *with* $v^0 \equiv 0$.

respectively, with $tol = 10^{-8}$. Furthermore, we chose $\mu = \frac{1}{1.13\,h}$ and set $\{U^n_\varepsilon, W^n_\varepsilon, R^n_\varepsilon\} \equiv \{\pi^h[U^{n,k}_\varepsilon]_+, W^{n,k}_\varepsilon, R^{n,k}_\varepsilon\}$ for (4.2a–c), while we used $\mu = 0.625\,h$ and set $\{U^n_\varepsilon, W^n_\varepsilon\} \equiv \{U^{n,k}_\varepsilon, W^{n,k}_\varepsilon\}$ if $U^{n,k}_\varepsilon > 0$ and $\{U^n_\varepsilon, W^n_\varepsilon\} \equiv \{U^{n,k-\frac{1}{2}}_\varepsilon, W^{n,k}_\varepsilon\}$ otherwise for (4.12a–d). For the iteration (4.3) we set $V^n_\varepsilon \equiv V^{n,k}_\varepsilon$.

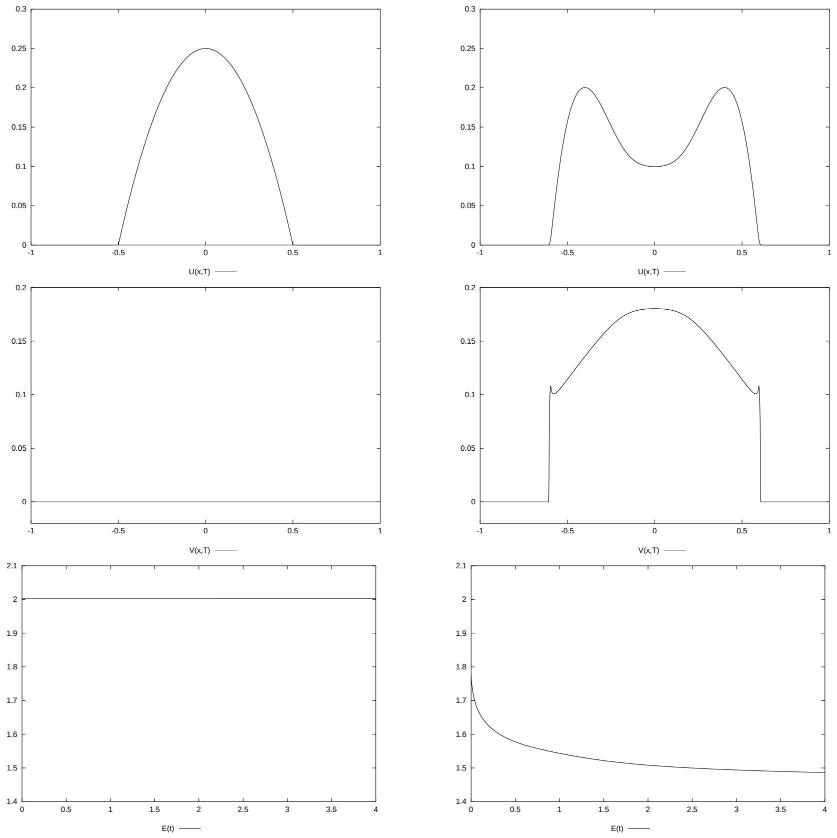
In our first set of experiments we set the parameters $L = 1$, $c = 2 \times 10^{-2}$, $\rho = 0$, and we chose the initial data (i) with $v^0_{\max} = 0$ and a final time $T = 4$. We note from the weak formulation (3.25a–c) that this initial data is a steady state for (P); that is, $u(x,t) = u^0(x)$, $v(x,t) = 0$ for all $(x,t) \in \Omega_T$. The results for various choices of h are displayed in Table 1, where all values are correct to four significant figures.

Remark 5.1. In order to obtain a discretization that leads to a discrete analogue of the energy estimate (1.5) we needed to approximate $\lambda(v)$ in a subtle way; see (2.7). In particular, the resulting scheme does not guarantee that $V_\varepsilon \equiv 0$ if the initial data have this property. However, the results in Table 1 show that the error between V_ε and v is small.

In Figure 1 we plot $\pi^h u^0(x) - U_\varepsilon(x,T)$ for $\#J = 2^k + 1$, $k = 6 \to 11$, on a short interval about the initial free boundary point on the right-hand side, $x = 0.5$. For each $\#J$ there are a few points outside the support of $u^0(x)$ which are much larger than $tol = 10^{-8}$. Outside the region plotted $|\pi^h u^0(x) - U_\varepsilon(x,T)|$ is "zero," i.e., much smaller than tol. This behavior compares with results in [4], where similar errors can be observed. Note that the different choice of c here acts as a time scaling factor.

We see from Figure 2 that for $\#J = 2^{10} + 1$ the initial profile $U^0_\varepsilon \equiv \pi^h u^0$ is graphically preserved in the absence of surfactant. This is underlined by the fact that the energy

$$E(t) := \tfrac{t-t_{n-1}}{\tau_n}\,\mathcal{E}(U^n_\varepsilon, V^n_\varepsilon) + \tfrac{t_n-t}{\tau_n}\,\mathcal{E}(U^{n-1}_\varepsilon, V^{n-1}_\varepsilon), \qquad t \in [t_{n-1}, t_n], \quad n \geq 1,$$



(a) (b)

FIG. 2. *Comparison of* $U_\varepsilon(x,T)$*,* $V_\varepsilon(x,T)$ *at* $T = 4$ *and* $E(t)$ *for* $t \in [0,T]$ *in the* (a) *absence and* (b) *presence of surfactant, respectively.*

remains constant over the whole time period $[0,T]$. When surfactant is included, however, this has a dramatic effect on the shape of the solution $U_\varepsilon(x,T)$. For the plot in Figure 2, we took v^0 as described in (5.1b) with $A = 50$, $x_0 = 0.1$, $\gamma = 10^{-4}$, and $v^0_{\max} = 0.9$ so that the support of $v^0 \subset (-0.2, 0.2) \subset [-0.5, 0.5]$, the support of u^0. Eventually the solutions U_ε and V_ε reach a "numerical steady state"; i.e., we obtain that $\{U^n_\varepsilon, W^n_\varepsilon, V^n_\varepsilon\} \equiv \{U^{n,1}_\varepsilon, W^{n,1}_\varepsilon, V^{n,1}_\varepsilon\}$ for n sufficiently large for the stated stopping criteria on U_ε and V_ε (see (5.2)). For the parameters mentioned above and a stopping tolerance of $tol = 10^{-10}$ this state is reached at $T = 3413$. In Figure 3 we plot $U_\varepsilon(\cdot,t)$ and $V_\varepsilon(\cdot,t)$ for $t = 0$, $t = 50$, $t = 200$, and $t = 3413$, respectively.

Remark 5.2. In the case $\rho = 0$, the only mechanism for surfactant spreading is transport via the fluid velocity. If the support of the initial data of the surfactant is contained in the set of points initially wetted, then one can show that the support of the surfactant at time t is contained in the set $\mathcal{W}(t) := \{x : u(x,\tau) > 0 \text{ for some } \tau \in [0,t]\}$, which is the set of points which have been wetted at some time in the past. For the discrete problem a similar property follows directly from (2.11c), since $V^n_\varepsilon(p_j)$ can only be nonzero if either $V^{n-1}_\varepsilon(p_j) \neq 0$ or $(U^n_\varepsilon, \chi_j) \neq 0$. Therefore, the only modification is that the support of the surfactant can be one mesh point ahead of the discrete analogue of $\mathcal{W}(t)$.

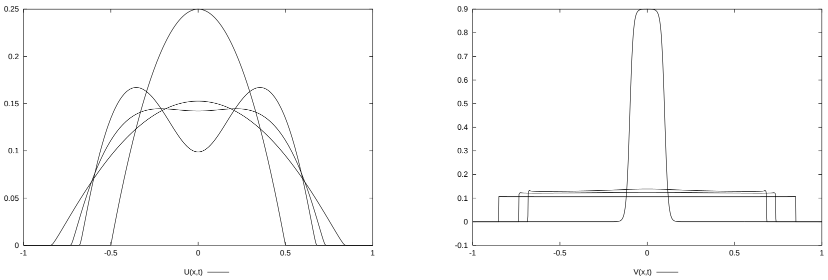


FIG. 3. *$U_\varepsilon(x,t)$ and $V_\varepsilon(x,t)$ for $t = 0, 50, 200, 3413$.*

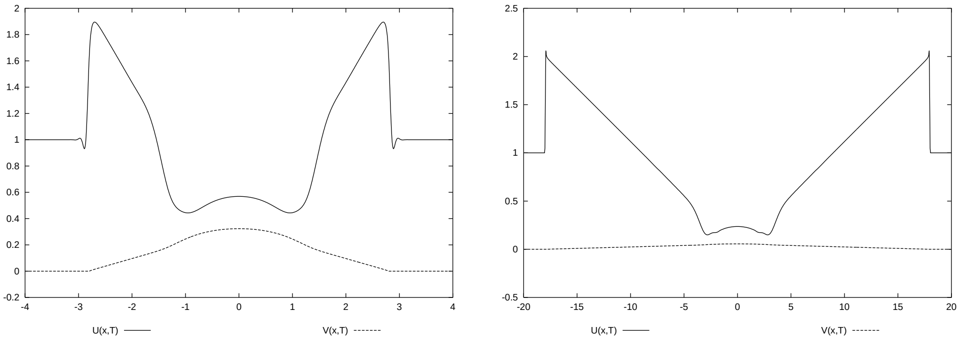


FIG. 4. *$U_\varepsilon(x,T)$ and $V_\varepsilon(x,T)$ plotted against x for $T = 4$ (left) and $T = 1000$ (right).*

Remark 5.3. A parabolic profile for the drop together with a constant surfactant density on the support of the drop is a steady state for the system (1.1a–c) if $\rho = 0$. It is the discrete analogue of such a steady state that we observe for large times in Figure 3.

Remark 5.4. For the thin film equation (in the no-slip case) it is conjectured that the support of the film does not increase. In the case that a surfactant is placed on the film this property does not seem to be true any longer.

In addition, we performed experiments for a uniform liquid layer, i.e., $u^0 \equiv 1$. We chose the parameters similar to the ones reported in [26]. In particular, we took $L = 4$, $T = 4$, $c = 10^{-5}$, $\rho = 2 \times 10^{-4}$, $v^0_{\max} = 1$, $\gamma = 0$, $A = 10$, $x_0 = 0.5$, and $\#J = 2^{10} + 1$. The computed solutions $U_\varepsilon(x,T)$ and $V_\varepsilon(x,T)$ can be seen on the left-hand side of Figure 4.

We note that $U_\varepsilon(x,T)$ and $V_\varepsilon(x,T)$ approach similarity solutions of (P) for the case $\rho = c = a = \delta = 0$; see [24]. This can be seen on the right-hand side of Figure 4, where we plot the two functions for the values $L = 20$, $T = 1000$, $c = 10^{-8}$, $\rho = 0$, $\gamma = 0$, $v^0_{\max} = 1$, $A = 10$, $x_0 = 0.5$, and $\#J = 2^{10} + 1$. We recall that $(\mathrm{P}^{h,\tau}_\varepsilon)$ is only well-posed for $c > 0$. In order to formulate the similarity solutions, we make use of the following transformation of coordinates. Let $\xi := (1+t)^{-\frac{1}{3}}\, x$, $\overline{u}(\xi,t) := U_\varepsilon(x,t)$ and $\overline{v}(\xi,t) := (1+t)^{\frac{1}{3}}\, V_\varepsilon(x,t)$. Then the similarity solutions for $\overline{u}$ and $\overline{v}$ are given by

$$\overline{u}_0(\xi) = \begin{cases} \frac{2\,\xi}{\xi_s}, & 0 \le \xi < \xi_s, \\ 1, & \xi_s \le \xi, \end{cases} \quad \text{and} \quad \overline{v}_0(\xi) = \begin{cases} \frac{\xi_s}{6}\,(\xi_s - \xi), & 0 \le \xi < \xi_s, \\ 0, & \xi_s \le \xi, \end{cases}$$

where $\xi_s := (12 \int_0^L v^0(x)\,\mathrm{d}x)^{\frac{1}{3}}$ is the position of the shock. The corresponding plot is shown on the left-hand side of Figure 5.

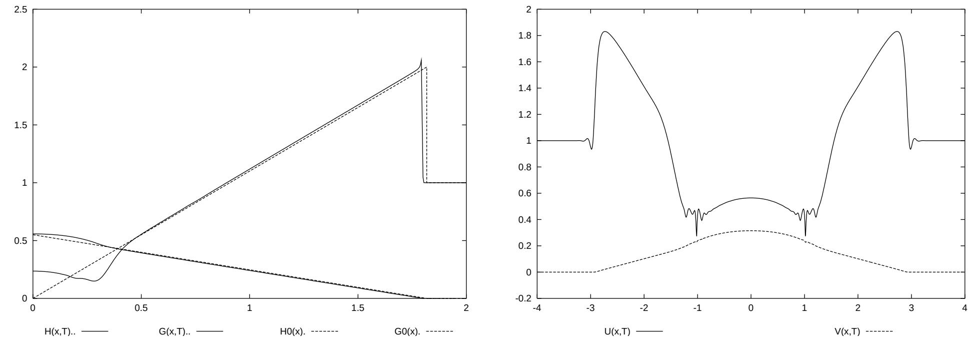


FIG. 5. $\overline{u}(\xi, T)$, $\overline{v}(\xi, T)$ *with the corresponding similarity solutions* $\overline{u}_0(\xi)$, $\overline{v}_0(\xi)$ *plotted against* ξ *for* $T = 1000$ *(left) and* $U_\varepsilon(x, T)$, $V_\varepsilon(x, T)$ *plotted against* x *for* $T = 4.33$ *when* $\phi \not\equiv 0$ *(right).*

5.1. Inclusion of van der Waals forces. In addition, we conducted numerical experiments in which we considered the effect of both attractive and repulsive van der Waals forces being present. Note that this corresponds to $a > 0$ and $\delta > 0$, respectively.

We note that (4.12b) is a decoupled system of $\#J$ equations and that for $\chi \equiv \chi_j$ one has to find $s = U_\varepsilon^{n,k-\frac{1}{2}}(p_j) \in \mathbb{R}_{>0}$ that satisfies

$$\psi(s) := s^2\,(\varphi(s) - Y_\varepsilon^{n,k}(p_j)) = 0, \tag{5.3}$$

where $\varphi(s) := s + \mu\,\phi^+(s)$ as in section 4. To solve $\psi(s) = 0$, we use Newton's method:

$$s^{\ell+1} = s^\ell - [\psi'(s^\ell)]^{-1}\psi(s^\ell), \qquad \ell \geq 0, \tag{5.4}$$

with $|s^{\ell+1} - s^\ell| < tol$ as the stopping criterion and $s^0 = U_\varepsilon^{n,k-\frac{3}{2}}(p_j)$ for $k \geq 2$ and $s^0 = U_\varepsilon^{n,0}(p_j)$ otherwise. Note that we introduced the term s^2 in (5.3) in order to stabilize the Newton iteration. Although other powers of s are possible, this particular choice seemed preferable in practice. In fact, the iteration (5.4) always converged.

On the right-hand side of Figure 5, we plot $U_\varepsilon(x, T)$ and $V_\varepsilon(x, T)$ for $a = 2 \times 10^{-3}$, $\delta = 10^{-5}$, and $\nu = 4$. The other parameters were chosen as follows: $c = 10^{-5}$, $\rho = 2 \times 10^{-3}$, initial data (ii) with $v^0_{\text{max}} = 1$, $\gamma = 0$, $A = 10$, $x_0 = 0.5$, $L = 4$, $T = 4.33$, and $\#J = 2^{10} + 1$. One can clearly see the effect of modeling the van der Waals forces. Once the film thickness reaches a certain threshold, the film tries to rupture in the effected regions. Note that we have plotted the solutions just before such a "rupture" occurs. Although the film height might become extremely thin, it can never actually rupture ($U_\varepsilon = 0$) due to the presence of the repulsive van der Waals forces. We would also like to mention that we repeated the experiment with the parameters mentioned above on a very fine mesh. As we obtained virtually identical results, we are satisfied that the oscillations shown in Figure 5 are not due to mesh effects. In fact, the instabilities are in agreement with linear stability analysis for the thin film equation in the presence of van der Waals forces (see [28, 27]).

5.2. Numerical results for $d = 2$. Finally, we present numerical experiments in two space dimensions with $\Omega = (-L, L) \times (-L, L)$. We took a uniform mesh of squares ς of length $h = \frac{2L}{128}$, each of which was divided into two triangles by its north

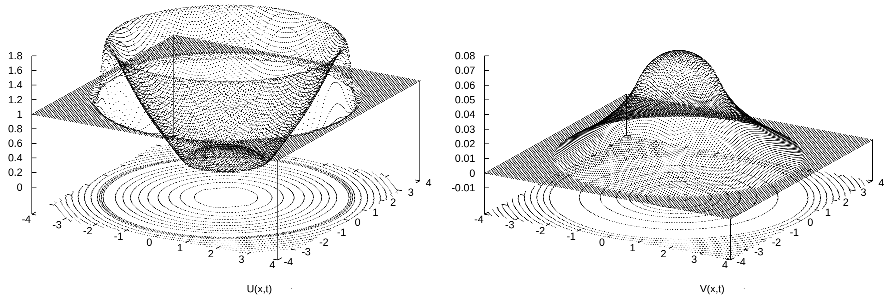


FIG. 6. *$U_\varepsilon(x,T)$ and $V_\varepsilon(x,T)$ plotted against x for $T = 35$.*

east diagonal. We used the modified discrete semi-inner product on $C(\overline{\Omega})$:

$$(\eta_1, \eta_2)^h_* := \int_\Omega \Pi^h(\eta_1(x)\,\eta_2(x))\,\mathrm{d}x\,. \tag{5.5}$$

Here Π^h is the piecewise continuous bilinear interpolant on $\overline{\Omega}$, which on each square ς is bilinear and interpolates at the vertices. Using (5.5) instead of (2.1) enables us to solve (4.6) efficiently using a "discrete cosine transform" approach; see [2]. We note that similarly to (2.1) and (2.14), the semi-inner product (5.5) is equivalent on S^h to the standard L^2 inner product, and in place of (2.17), we have that

$$|(z^h,\chi)^h - (z^h,\chi)^h_*| \leq C h^{1+m}[\ln(\tfrac{1}{h})]^2\,\|z^h\|_m \|\chi\|_1 \qquad \forall\ z^h,\ \chi \in S^h\,, \quad m = 0 \text{ or } 1.$$

Therefore, it is easy to adapt the proofs to show that all the results in this paper remain unchanged with the choice (5.5).

We report on an experiment with the same parameters for (P) as in $d = 1$ for Figure 4. In particular, we set $a = \delta = 0$, $c = 10^{-5}$, $\rho = 2 \times 10^{-4}$, $L = 4$, $T = 35$, $\tau_n = \tau = 10^{-3}$, and $\varepsilon = 10^{-5}$, and for the initial profiles we chose $u^0 \equiv 1$ and (5.1b) for v^0 with $v^0_{\max} = 1$, $\gamma = 0$, $A = 10$, and $x_0 = 0.5$. We set $U^0_\varepsilon \equiv \pi^h u^0$ and $V^0_\varepsilon \equiv \pi^h v^0$. Note that u^0, $v^0 \in W^{1,\infty}(\Omega)$, and hence the results of Lemma 2.5 still hold for the chosen U^0_ε, V^0_ε on noting a standard interpolation result. Note, furthermore, that here we integrate until $T = 35$ as opposed to $T = 4$ in one space dimension. This is due to the slower speed of propagation; e.g., the corresponding similarity solution for $c = \rho = 0$ advances proportionally to $(1+t)^{\frac{1}{4}}$ for $d = 2$ as opposed to $(1+t)^{\frac{1}{3}}$ for $d = 1$ (see [24]). We chose $tol = 10^{-8}$ and $\mu = 400$ for the iterative method (4.2a–c). In Figure 6, we plot $U_\varepsilon(x,T)$ and $V_\varepsilon(x,T)$ for $T = 35$, respectively.

REFERENCES

[1] R. A. ADAMS AND J. FOURNIER, *Cone conditions and properties of Sobolev spaces*, J. Math. Anal. Appl., 61 (1977), pp. 713–734.

[2] J. W. BARRETT AND J. F. BLOWEY, *Finite element approximation of a model for phase separation of a multi-component alloy with non-smooth free energy*, Numer. Math., 77 (1997), pp. 1–34.

[3] J. W. BARRETT AND J. F. BLOWEY, *Finite element approximation of a degenerate Allen–Cahn/Cahn–Hilliard system*, SIAM J. Numer. Anal., 39 (2001), pp. 1598–1624.

[4] J. W. BARRETT, J. F. BLOWEY, AND H. GARCKE, *Finite element approximation of a fourth order nonlinear degenerate parabolic equation*, Numer. Math., 80 (1998), pp. 525–556.

[5] J. W. Barrett, J. F. Blowey, and H. Garcke, *Finite element approximation of the Cahn–Hilliard equation with degenerate mobility*, SIAM J. Numer. Anal., 37 (1999), pp. 286–318.

[6] J. W. Barrett, J. F. Blowey, and H. Garcke, *On fully practical finite element approximations of degenerate Cahn–Hilliard systems*, M2AN Math. Model. Numer. Anal., 35 (2001), pp. 713–748.

[7] J. W. Barrett and R. Nürnberg, *Finite element approximation of a nonlinear degenerate parabolic system describing bacterial pattern formation*, Interfaces Free Bound., 4 (2002), pp. 277–307.

[8] F. Bernis, *Viscous flows, fourth order nonlinear degenerate parabolic equations and singular elliptic problems*, in Free Boundary Problems: Theory and Applications, Pitman Res. Notes Math. Ser. 323, J. I. Diaz, M. A. Herrero, A. Linan, and J. L. Vazquez, eds., Longman, Harlow, UK, 1995, pp. 40–56.

[9] F. Bernis and A. Friedman, *Higher order nonlinear degenerate parabolic equations*, J. Differential Equations, 83 (1990), pp. 179–206.

[10] A. L. Bertozzi, M. P. Brenner, T. F. Dupont, and L. P. Kadanoff, *Singularities and similarities in interface flows*, in Trends and Perspectives in Applied Mathematics, Appl. Math. Sci. 100, Springer-Verlag, New York, 1994, pp. 155–208.

[11] M. Bertsch, R. Dal Passo, H. Garcke, and G. Grün, *The thin viscous flow equation in higher space dimensions*, Adv. Differential Equations, 3 (1998), pp. 417–440.

[12] J. F. Blowey and C. M. Elliott, *The Cahn–Hilliard gradient theory for phase separation with nonsmooth free energy.* II. *Numerical analysis*, European J. Appl. Math., 3 (1992), pp. 147–179.

[13] M. S. Borgas and J. B. Grotberg, *Monolayer flow on a thin film*, J. Fluid Mech., 193 (1988), pp. 151–170.

[14] J. W. Cahn, C. M. Elliott, and A. Novick-Cohen, *The Cahn–Hilliard equation with a concentration dependent mobility: Motion by minus the Laplacian of the mean curvature*, European J. Appl. Math., 7 (1996), pp. 287–301.

[15] J. F. Ciavaldini, *Analyse numérique d'un problème de Stefan à deux phases par une méthode d'elements finis*, SIAM J. Numer. Anal., 12 (1975), pp. 464–487.

[16] C. M. Elliott and H. Garcke, *On the Cahn–Hilliard equation with degenerate mobility*, SIAM J. Math. Anal., 27 (1996), pp. 404–423.

[17] C. M. Elliott and H. Garcke, *Diffusional phase transitions in multicomponent systems with a concentration dependent mobility matrix*, Phys. D, 109 (1997), pp. 242–256.

[18] D. P. Gaver III and J. B. Grotberg, *The dynamics of a localized surfactant on a thin film*, J. Fluid Mech., 213 (1990), pp. 127–148.

[19] H. P. Greenspan, *On the motion of a small viscous droplet that wets a surface*, J. Fluid Mech., 84 (1978), pp. 125–143.

[20] G. Grün, *Degenerate parabolic differential equations of fourth order and a plasticity model with nonlocal hardening*, Z. Anal. Anwendungen, 14 (1995), pp. 541–574.

[21] G. Grün, *On the convergence of entropy consistent schemes for lubrication type equations in multiple space dimensions*, Math. Comp., 72 (2003), pp. 1251–1279.

[22] G. Grün and M. Rumpf, *Nonnegativity preserving numerical schemes for the thin film equation*, Numer. Math., 87 (2000), pp. 113–152.

[23] G. Grün and M. Rumpf, *Simulation of singularities and instabilities in thin film flow*, European J. Appl. Math., 12 (2001), pp. 293–320.

[24] O. E. Jensen and J. B. Grotberg, *Insoluble surfactant spreading on a thin viscous film: Shock evolution and film rupture*, J. Fluid Mech., 240 (1992), pp. 259–288.

[25] P. L. Lions and B. Mercier, *Splitting algorithms for the sum of two nonlinear operators*, SIAM J. Numer. Anal., 16 (1979), pp. 964–979.

[26] O. K. Matar and S. M. Troian, *The development of transient fingering patterns during the spreading of surfactant coated films*, Phys. Fluids, 11 (1999), pp. 3232–3246.

[27] O. K. Matar and S. M. Troian, *Spreading of a surfactant monolayer on a thin liquid film: Onset and evolution of digitated structures*, Chaos, 9 (1999), pp. 141–153.

[28] A. Oron, S. H. Davis, and S. G. Bankoff, *Long-scale evolution of thin liquid films*, Rev. Modern Phys., 69 (1997), pp. 931–980.

[29] M. Renardy and R. C. Rogers, *An Introduction to Partial Differential Equations*, Springer-Verlag, New York, 1992.

[30] A. Schmidt and K. G. Siebert, *Albert–software for scientific computations and applications*, Acta Math. Univ. Comenian (N.S.), 70 (2001), pp. 105–122.

[31] S. Wieland, *Modellierung und mathematische Analyse von Oberflächendiffusion auf dünnen Filmen*, Ph.D. thesis, University of Bonn, Bonn, Germany, 2003 (to appear).

[32] L. Zhornitskaya and A. L. Bertozzi, *Positivity-preserving numerical schemes for lubrication-type equations*, SIAM J. Numer. Anal., 37 (2000), pp. 523–555.

SIAM J. NUMER. ANAL.
Vol. 41, No. 4, pp. 1465–1486

COMPONENT-BY-COMPONENT CONSTRUCTION OF GOOD INTERMEDIATE-RANK LATTICE RULES[*]

F. Y. KUO[†] AND S. JOE[‡]

Abstract. It is known that the generating vector of a rank-1 lattice rule can be constructed component-by-component to achieve strong tractability error bounds in both weighted Korobov spaces and weighted Sobolev spaces. Since the weights for these spaces are nonincreasing, the first few variables are in a sense more important than the rest. We thus propose to copy the points of a rank-1 lattice rule a number of times in the first few dimensions to yield an intermediate-rank lattice rule. We show that the generating vector (and in weighted Sobolev spaces, the shift also) of an intermediate-rank lattice rule can also be constructed component-by-component to achieve strong tractability error bounds. In certain circumstances, these bounds are better than the corresponding bounds for rank-1 lattice rules.

Key words. quasi-Monte Carlo methods, intermediate-rank lattice rules, worst-case error, tractability

AMS subject classifications. 65D30, 65D32, 68Q25

DOI. 10.1137/S0036142902407162

1. Introduction.

The d-dimensional integral

$$I_d(f) = \int_{[0,1]^d} f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$$

may be approximated using rank-1 lattice rules. These are equal-weight rules having quadrature points belonging to the set

$$\left\{ \left\{ \frac{i\boldsymbol{z}}{n} \right\} : 0 \le i \le n-1 \right\}.$$

Here $\boldsymbol{z}$, known as the generating vector, is an integer vector having no factor in common with n, and the braces around a vector indicate that we take the fractional part of each component of the vector. It is shown in [14] that every lattice rule may be written as a multiple sum involving one or more generating vectors; the minimum number of generating vectors required to generate a lattice rule is known as the "rank" of the rule. Besides rank-1 lattice rules involving just one generating vector, there exist lattice rules having rank up to d. More information about lattice rules may be found in [11].

The construction of rank-1 lattice rules for integrands belonging to weighted Korobov and weighted Sobolev spaces has been studied in various papers. These weighted function spaces are tensor product reproducing kernel Hilbert spaces. Recall that a

[*] Received by the editors May 7, 2002; accepted for publication (in revised form) January 21, 2003; published electronically September 9, 2003.

http://www.siam.org/journals/sinum/41-4/40716.html

[†] Department of Mathematics, The University of Waikato, Private Bag 3105, Hamilton, New Zealand. Current address: School of Mathematics, The University of New South Wales, Sydney, NSW 2052, Australia (fkuo@maths.unsw.edu.au).

[‡] Department of Mathematics, The University of Waikato, Private Bag 3105, Hamilton, New Zealand (stephenj@math.waikato.ac.nz).

quasi-Monte Carlo (QMC) rule

$$Q_{n,d}(f) = \frac{1}{n}\sum_{i=0}^{n-1} f(\boldsymbol{x}_i) \tag{1.1}$$

is an equal-weight quadrature rule with the quadrature points chosen in a deterministic way. The "worst-case error" of a QMC rule in some Hilbert space H_d is defined to be

$$e_{n,d}(Q_{n,d}, H_d) := \sup\{|Q_{n,d}(f) - I_d(f)| : \|f\|_{H_d} \le 1, f \in H_d\},$$

and the initial error is

$$e_{0,d}(H_d) := \sup\{|I_d(f)| : \|f\|_{H_d} \le 1, f \in H_d\}.$$

Following the analysis by Sloan and Woźniakowski in [16], the integration problem is said to be "strongly QMC tractable" in the Hilbert space H_d if the minimal number of function evaluations n in a QMC rule (1.1) needed to reduce the initial error $e_{0,d}(H_d)$ by a factor of $\varepsilon > 0$ is bounded by a polynomial in ε^{-1} independently of d.

In [15], a component-by-component algorithm was developed for constructing rank-1 lattice rules in unweighted Korobov spaces. The algorithm was later extended to shifted rank-1 lattice rules (see [12]) in weighted Sobolev spaces, and the rules constructed achieve strong QMC tractability error bounds. Both these constructions assumed that n, the number of quadrature points, was a prime number. The construction was later generalized in [10] to rules with a composite number of points. Construction of rank-1 lattice rules in the randomized setting has been considered in [13]. Recently, it was shown in [9] that the constructions achieve the optimal rate of convergence in the corresponding function spaces.

Lattice rules constructed in this manner are "extensible" in terms of the dimension d; that is, if further dimensions are needed at a later stage, the additional components can be constructed with the existing components kept unchanged. However, if more points are required, then the rules need to be reconstructed from scratch. A recent work [4] showed the existence of good rank-1 lattice rules that are extensible both in terms of the number of points n and the dimension d, but the proof is nonconstructive.

We are interested in "copying" rank-1 lattice rules. Since the weighted function spaces of interest have nonincreasing weights, the first few variables are in a sense more important than the rest. Therefore, it would seem intuitive to copy the points in the first few dimensions. Thus we may copy an n-point d-dimensional rank-1 lattice rule ℓ times in each of the first r dimensions, where $\ell \ge 1$, $\gcd(\ell, n) = 1$, and $0 \le r \le d$. We then obtain the rule with $N = \ell^r n$ points given by

$$Q_{n,d}(f) = \frac{1}{\ell^r n}\sum_{m_r=0}^{\ell-1}\cdots\sum_{m_1=0}^{\ell-1}\sum_{i=0}^{n-1} f\left(\left\{\frac{i\boldsymbol{z}}{n} + \frac{(m_1,\dots,m_r,0,\dots,0)}{\ell}\right\}\right).$$

We call the rule with these points "the (ℓ, r)-copy of a rank-1 lattice rule with generating vector $\boldsymbol{z}$." When $r = 0$ and/or $\ell = 1$, we get just the original n-point rank-1 lattice rule. For $r \ge 1$, the resulting rule is a rank-r lattice rule. These intermediate-rank lattice rules have previously been considered in [7] and [8]. Typically, for reasons of tractability, we will take r to be a fixed number, say, $r = 1$, 2, or 3. For the choice of ℓ it would seem reasonable on practical grounds and theoretical grounds (see Theorem 2.3 and Lemma 2.4) to take ℓ to be 2 in actual calculations. This value of $\ell = 2$ has been used previously in [7] and [8].

Our plan is to construct intermediate-rank lattice rules in both weighted Korobov and weighted Sobolev spaces that achieve strong QMC tractability error bounds. In section 2, we consider intermediate-rank lattice rules in weighted Korobov spaces. We show that the intermediate-rank lattice rule we consider has the same worst-case error as a certain rank-1 lattice rule in a slightly different weighted Korobov space. We then show that there exist intermediate-rank lattice rules with error bounds which are better than the corresponding bounds for rank-1 lattice rules with approximately the same number of points. Moreover, we shall see that the generating vectors constructed component-by-component satisfy strong QMC tractability bounds and achieve the optimal rate of convergence in weighted Korobov spaces. In section 3, we give a brief discussion on the construction of shifted intermediate-rank lattice rules in weighted Sobolev spaces. The final section, section 4, contains numerical results.

Throughout the paper, we will assume that n is a prime number to simplify the analysis. More general results for any positive integer n can be obtained by emulating the more complicated analysis found in [10]. When n is a prime number, $\boldsymbol{z}$ can be chosen from $\mathbb{Z}_n^d$, where $\mathbb{Z}_n := \{1, 2, \ldots, n-1\}$.

2. Intermediate-rank lattice rules in weighted Korobov spaces.

We are interested in the weighted Korobov spaces of periodic functions considered in [10]. These spaces are parameterized by a real parameter $\alpha > 1$ and two sequences of positive weights $\boldsymbol{\beta} = \{\beta_j\}$ and $\boldsymbol{\gamma} = \{\gamma_j\}$ satisfying

$$\frac{\gamma_1}{\beta_1} \geq \frac{\gamma_2}{\beta_2} \geq \cdots .$$

The inner product in these spaces is given by

$$\langle f, g \rangle_d = \sum_{\boldsymbol{h} \in \mathbb{Z}^d} \left(\hat{f}(\boldsymbol{h}) \overline{\hat{g}(\boldsymbol{h})} \prod_{j=1}^{d} r(\alpha, \beta_j, \gamma_j, h_j) \right),$$

where

$$r(\alpha, \beta, \gamma, h) = \begin{cases} \beta^{-1} & \text{if } h = 0, \\ \gamma^{-1} |h|^{\alpha} & \text{otherwise.} \end{cases}$$

Here α is a smoothness parameter characterizing the rate of decay of the Fourier coefficients. Various variations of these spaces have previously been considered in works such as [5], [6], [15], and [16]. The worst-case error in these Korobov spaces for a QMC rule (1.1) is given by

$$e_{n,d}^2(\boldsymbol{x}_0, \ldots, \boldsymbol{x}_{n-1}) = -\prod_{j=1}^{d} \beta_j + \frac{1}{n^2} \sum_{i=0}^{n-1} \sum_{k=0}^{n-1} \prod_{j=1}^{d} \left(\beta_j + \gamma_j \sum_{h=-\infty}^{\infty}{}' \frac{e^{2\pi \mathrm{i} h (x_{i,j} - x_{k,j})}}{|h|^{\alpha}} \right), \tag{2.1}$$

where the $'$ on the sum indicates that we omit the $h = 0$ term. This expression may be written in terms of Bernoulli polynomials if α is chosen to be a positive even number. The initial error is

$$e_{0,d} = \prod_{j=1}^{d} \beta_j^{\frac{1}{2}}.$$

Following the analysis of tractability in [16], it is possible to show that if the weights satisfy

$$\sum_{j=1}^{\infty} \frac{\gamma_j}{\beta_j} < \infty, \tag{2.2}$$

then an upper bound for the square worst-case error of the form

$$\frac{b}{n} \prod_{j=1}^{d} \left(\beta_j + a\gamma_j\right), \tag{2.3}$$

where $a, b > 0$ are bounded independently of d, is enough to ensure strong QMC tractability, with the rate of convergence being $O(n^{-1/2})$. Moreover, the optimal rate of convergence $O(n^{-\alpha/2+\delta})$, for any $\delta > 0$, can be achieved if the weights satisfy a stronger condition,

$$\sum_{j=1}^{\infty} \left(\frac{\gamma_j}{\beta_j}\right)^{\frac{1}{\alpha-2\delta}} < \infty.$$

It is worth mentioning that the condition (2.2) is also necessary for strong QMC tractability (see [6]).

We now consider the (ℓ, r)-copy of a rank-1 lattice rule with generating vector $\boldsymbol{z}$, that is, a rule with points belonging to the set

$$\left\{\left\{\frac{i\boldsymbol{z}}{n} + \frac{(m_1, \dots, m_r, 0, \dots, 0)}{\ell}\right\} : 0 \le i \le n-1,\ 0 \le m_1, \dots, m_r \le \ell - 1\right\},$$

where $\ell \ge 1$, $\gcd(\ell, n) = 1$, and $0 \le r \le d$. An expression for $e_{n,d,\mathrm{copy}(\ell,r)}(\boldsymbol{z})$, the worst-case error for such a rule, is given in the next lemma. Note that though this intermediate-rank lattice rule has $N = \ell^r n$ points, the lemma shows that the worst-case error may be calculated by using a rule having just n points. We will explore this further in the next subsection.

LEMMA 2.1. *We have*

$$e^2_{n,d,\mathrm{copy}(\ell,r)}(\boldsymbol{z}) = -\prod_{j=1}^{d} \beta_j + \frac{1}{n} \sum_{k=0}^{n-1} \left[\prod_{j=1}^{r} \left(\beta_j + \frac{\gamma_j}{\ell^\alpha} {\sum_{h=-\infty}^{\infty}}' \frac{e^{2\pi \mathrm{i} h \ell k z_j/n}}{|h|^\alpha}\right) \times \prod_{j=r+1}^{d} \left(\beta_j + \gamma_j {\sum_{h=-\infty}^{\infty}}' \frac{e^{2\pi \mathrm{i} h k z_j/n}}{|h|^\alpha}\right)\right].$$

Proof. We have from (2.1) that

$$e^2_{n,d,\mathrm{copy}(\ell,r)}(\boldsymbol{z}) = -\prod_{j=1}^{d} \beta_j + \frac{1}{\ell^{2r} n^2} \sum_{q_r=0}^{\ell-1} \cdots \sum_{q_1=0}^{\ell-1} \sum_{m_r=0}^{\ell-1} \cdots \sum_{m_1=0}^{\ell-1} \sum_{i=0}^{n-1} \sum_{k=0}^{n-1} \left[\prod_{j=1}^{r} \left(\beta_j + \gamma_j {\sum_{h=-\infty}^{\infty}}' \frac{e^{2\pi \mathrm{i} h \left(\left\{\frac{iz_j}{n} + \frac{q_j}{\ell}\right\} - \left\{\frac{kz_j}{n} + \frac{m_j}{\ell}\right\}\right)}}{|h|^\alpha}\right) \times \prod_{j=r+1}^{d} \left(\beta_j + \gamma_j {\sum_{h=-\infty}^{\infty}}' \frac{e^{2\pi \mathrm{i} h \left(\left\{\frac{iz_j}{n}\right\} - \left\{\frac{kz_j}{n}\right\}\right)}}{|h|^\alpha}\right)\right].$$

The second term can be written as

$$\frac{1}{n^2}\sum_{i=0}^{n-1}\sum_{k=0}^{n-1}\left[\prod_{j=1}^{r}\left(\frac{1}{\ell^2}\sum_{q=0}^{\ell-1}\sum_{m=0}^{\ell-1}\left(\beta_j+\gamma_j\sum_{h=-\infty}^{\infty}{}' \frac{e^{2\pi \mathrm{i} h((i-k)z_j/n+(q-m)/\ell)}}{|h|^\alpha}\right)\right)\right.$$
$$\left.\times\prod_{j=r+1}^{d}\left(\beta_j+\gamma_j\sum_{h=-\infty}^{\infty}{}' \frac{e^{2\pi \mathrm{i} h(i-k)z_j/n}}{|h|^\alpha}\right)\right]. \tag{2.4}$$

For $0\le q,m\le \ell-1$, the values of $(q-m) \bmod \ell$ are just 0 to $\ell-1$ in some order, with each value occurring ℓ times. Thus we have

$$\frac{1}{\ell^2}\sum_{q=0}^{\ell-1}\sum_{m=0}^{\ell-1}\left(\beta_j+\gamma_j\sum_{h=-\infty}^{\infty}{}' \frac{e^{2\pi \mathrm{i} h((i-k)z_j/n+(q-m)/\ell)}}{|h|^\alpha}\right)$$
$$=\frac{1}{\ell}\sum_{m=0}^{\ell-1}\left(\beta_j+\gamma_j\sum_{h=-\infty}^{\infty}{}' \frac{e^{2\pi \mathrm{i} h((i-k)z_j/n+m/\ell)}}{|h|^\alpha}\right).$$

Now since

$$\sum_{m=0}^{\ell-1} e^{2\pi \mathrm{i} hm/\ell}=\begin{cases}\ell & \text{if } h \text{ is a multiple of } \ell,\\ 0 & \text{otherwise,}\end{cases} \tag{2.5}$$

we have

$$\frac{1}{\ell}\sum_{m=0}^{\ell-1}\left(\beta_j+\gamma_j\sum_{h=-\infty}^{\infty}{}' \frac{e^{2\pi \mathrm{i} h((i-k)z_j/n+m/\ell)}}{|h|^\alpha}\right)$$
$$=\beta_j+\frac{\gamma_j}{\ell}\sum_{h=-\infty}^{\infty}{}'\left(\frac{e^{2\pi \mathrm{i} h(i-k)z_j/n}}{|h|^\alpha}\sum_{m=0}^{\ell-1}e^{2\pi \mathrm{i} hm/\ell}\right)$$
$$=\beta_j+\frac{\gamma_j}{\ell}\sum_{m=-\infty}^{\infty}{}'\left(\frac{e^{2\pi \mathrm{i} m\ell(i-k)z_j/n}}{|m\ell|^\alpha}\cdot\ell\right)$$
$$=\beta_j+\frac{\gamma_j}{\ell^\alpha}\sum_{h=-\infty}^{\infty}{}' \frac{e^{2\pi \mathrm{i} h\ell(i-k)z_j/n}}{|h|^\alpha}.$$

Thus (2.4) can be simplified to

$$\frac{1}{n^2}\sum_{i=0}^{n-1}\sum_{k=0}^{n-1}\left[\prod_{j=1}^{r}\left(\beta_j+\frac{\gamma_j}{\ell^\alpha}\sum_{h=-\infty}^{\infty}{}' \frac{e^{2\pi \mathrm{i} h\ell(i-k)z_j/n}}{|h|^\alpha}\right)\right.$$
$$\left.\times\prod_{j=r+1}^{d}\left(\beta_j+\gamma_j\sum_{h=-\infty}^{\infty}{}' \frac{e^{2\pi \mathrm{i} h(i-k)z_j/n}}{|h|^\alpha}\right)\right],$$

which can be simplified even further to

$$\frac{1}{n}\sum_{k=0}^{n-1}\left[\prod_{j=1}^{r}\left(\beta_j+\frac{\gamma_j}{\ell^\alpha}\sum_{h=-\infty}^{\infty}{}' \frac{e^{2\pi \mathrm{i} h\ell kz_j/n}}{|h|^\alpha}\right)\prod_{j=r+1}^{d}\left(\beta_j+\gamma_j\sum_{h=-\infty}^{\infty}{}' \frac{e^{2\pi \mathrm{i} hkz_j/n}}{|h|^\alpha}\right)\right],$$

since for $0\le i,k\le n-1$, the values of $(i-k) \bmod n$ are just 0 to $n-1$ in some order, with each value occurring n times. This completes the proof. □

2.1. Relationship with rank-1 lattice rules based on worst-case error. Let us define the sequence $\bar{\boldsymbol{\gamma}}$ by

$$\bar{\gamma}_j := \begin{cases} \dfrac{\gamma_j}{\ell^\alpha} & \text{if } 1 \le j \le r, \\ \gamma_j & \text{otherwise} \end{cases}$$

and set $\bar{\boldsymbol{z}}$ to be the d-dimensional vector with components given by

$$\bar{z}_j := \begin{cases} \ell z_j & \text{if } 1 \le j \le r, \\ z_j & \text{otherwise.} \end{cases}$$

Then the expression in Lemma 2.1 may be rewritten in the form

$$e^2_{n,d,\mathrm{copy}(\ell,r)}(\boldsymbol{z}) = -\prod_{j=1}^d \beta_j + \frac{1}{n}\sum_{k=0}^{n-1}\prod_{j=1}^d \left(\beta_j + \bar{\gamma}_j \sum_{h=-\infty}^{\infty}{}' \frac{e^{2\pi \mathrm{i} hk\bar{z}_j/n}}{|h|^\alpha}\right) = e^2_{n,d}(\bar{\boldsymbol{z}}, \bar{\boldsymbol{\gamma}}); \tag{2.6}$$

that is, *the worst-case error of an intermediate-rank lattice rule with generating vector $\boldsymbol{z}$ in the weighted Korobov space with weights $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ is the same as the worst-case error of a rank-*1 *lattice rule with generating vector $\bar{\boldsymbol{z}}$ in the weighted Korobov space with weights $\boldsymbol{\beta}$ and $\bar{\boldsymbol{\gamma}}$.*

Since $\ell \ne 0$ and $\gcd(\ell, n) = 1$, for fixed r there exist a unique $\bar{\boldsymbol{z}}$ for each $\boldsymbol{z}$ and vice versa. Because of this one-to-one correspondence between $\boldsymbol{z}$ and $\bar{\boldsymbol{z}}$, all the known results on rank-1 lattice rules in weighted Korobov spaces can be applied here, with generating vector $\bar{\boldsymbol{z}}$ and weights $\boldsymbol{\beta}$ and $\bar{\boldsymbol{\gamma}}$. Note that the effect of copying in the first r dimensions can be interpreted as a reduction of the first r terms of $\boldsymbol{\gamma}$ by a factor of $1/\ell^\alpha$.

The following theorem is a slight generalization of Lemma 2 in [16]. (There $\boldsymbol{\beta}$ is assumed to be $\mathbf{1}$.)

THEOREM 2.2. *Let n be a prime number, and define $M_{n,d,\mathrm{copy}(\ell,r)}$ to be the mean given by*

$$M_{n,d,\mathrm{copy}(\ell,r)} := \frac{1}{(n-1)^d}\sum_{\boldsymbol{z}\in\mathbb{Z}_n^d} e^2_{n,d,\mathrm{copy}(\ell,r)}(\boldsymbol{z}).$$

Then an expression for $M_{n,d,\mathrm{copy}(\ell,r)}$ is given by

$$\begin{aligned} &-\prod_{j=1}^d \beta_j + \frac{1}{n}\prod_{j=1}^r \left(\beta_j + \frac{2\gamma_j\zeta(\alpha)}{\ell^\alpha}\right)\prod_{j=r+1}^d (\beta_j + 2\gamma_j\zeta(\alpha)) \\ &+ \left(1-\frac{1}{n}\right)\prod_{j=1}^r \left(\beta_j - \frac{2\gamma_j\zeta(\alpha)(1-n^{1-\alpha})}{(n-1)\ell^\alpha}\right)\prod_{j=r+1}^d \left(\beta_j - \frac{2\gamma_j\zeta(\alpha)(1-n^{1-\alpha})}{n-1}\right), \end{aligned}$$

where $\zeta(\alpha)$ is the Riemann zeta function. Moreover, if n satisfies $n \ge 1 + \frac{\gamma_1}{\beta_1}\zeta(\alpha)$, then

$$M_{n,d,\mathrm{copy}(\ell,r)} \le \frac{1}{n}\prod_{j=1}^r \left(\beta_j + \frac{2\gamma_j\zeta(\alpha)}{\ell^\alpha}\right)\prod_{j=r+1}^d (\beta_j + 2\gamma_j\zeta(\alpha)).$$

Clearly there must exist at least one vector $\boldsymbol{z}$ such that

$$e^2_{n,d,\mathrm{copy}(\ell,r)}(\boldsymbol{z}) \le M_{n,d,\mathrm{copy}(\ell,r)} \le \frac{1}{n}\prod_{j=1}^{d}\left(\beta_j + 2\gamma_j\zeta(\alpha)\right).$$

Now let $N = \ell^r n$ denote the total number of quadrature points. It is obvious that this last bound is of the form (2.3) with $a = 2\zeta(\alpha)$, $b = \ell^r$, and $n = N$. Since ℓ and r are fixed, we conclude that there exist intermediate-rank lattice rules that achieve strong QMC tractability error bounds for weighted Korobov spaces.

2.2. Comparison with rank-1 lattice rules based on mean. It follows from Theorem 2.2 with $\ell = 1$ and $n = N$ that for N prime, the mean for rank-1 lattice rules is

$$\widehat{M}_{N,d} = -\prod_{j=1}^{d}\beta_j + \frac{1}{N}\prod_{j=1}^{d}\left(\beta_j + 2\gamma_j\zeta(\alpha)\right) + \left(1-\frac{1}{N}\right)\prod_{j=1}^{d}\left(\beta_j - \frac{2\gamma_j\zeta(\alpha)(1-N^{1-\alpha})}{N-1}\right).$$

Suppose we replace N by $N = \ell^r n$ in this last expression. This is not valid because N is not prime, but calculations using the correct (but more complicated) expression for the mean found in [10] indicate that this yields an underestimate of the true mean.

Now let

$$R_{n,d,\ell,r} := \frac{M_{n,d,\mathrm{copy}(\ell,r)}}{\widehat{M}_{N,d}}.$$

As an indication of whether these intermediate-rank lattice rules are better than rank-1 lattice rules having approximately the same number of points, we would like a result which shows that $R_{n,d,\ell,r} < 1$. A preliminary result of this type is given in the following theorem.

THEOREM 2.3. *Suppose that n is a prime number satisfying $n \ge 1 + \frac{2\gamma_1}{\beta_1}\zeta(\alpha)$. If*

$$\rho_{\ell,r} := \prod_{j=1}^{r}\frac{\ell\beta_j + \frac{2\gamma_j\zeta(\alpha)}{\ell^{\alpha-1}}}{\beta_j + 2\gamma_j\zeta(\alpha)} < 1$$

and

$$\ell^r(n-1)\prod_{j=1}^{r}\left(\beta_j - \frac{2\gamma_j\zeta(\alpha)(1-n^{1-\alpha})}{(n-1)\ell^{\alpha}}\right)\prod_{j=r+1}^{d}\left(\beta_j - \frac{2\gamma_j\zeta(\alpha)(1-n^{1-\alpha})}{n-1}\right)$$

$$< (\ell^r n - 1)\prod_{j=1}^{d}\left(\beta_j - \frac{2\gamma_j\zeta(\alpha)(1-(\ell^r n)^{1-\alpha})}{\ell^r n - 1}\right), \tag{2.7}$$

then

$$R_{n,d,\ell,r} < \rho_{\ell,r}.$$

Proof. By multiplying both $M_{n,d,\mathrm{copy}(\ell,r)}$ and $\widehat{M}_{N,d}$ by $N = \ell^r n$, we can write

$$R_{n,d,\ell,r} = \frac{t_1 + t_2 - c}{b_1 + b_2 - c} \quad \text{and} \quad \rho_{\ell,r} = \frac{t_1}{b_1},$$

where

$$\begin{aligned}
t_1 &= \prod_{j=1}^{r}\left(\ell\beta_j + \frac{2\gamma_j\zeta(\alpha)}{\ell^{\alpha-1}}\right)\prod_{j=r+1}^{d}\left(\beta_j + 2\gamma_j\zeta(\alpha)\right),\\
t_2 &= \ell^r(n-1)\prod_{j=1}^{r}\left(\beta_j - \frac{2\gamma_j\zeta(\alpha)(1-n^{1-\alpha})}{(n-1)\ell^{\alpha}}\right)\prod_{j=r+1}^{d}\left(\beta_j - \frac{2\gamma_j\zeta(\alpha)(1-n^{1-\alpha})}{n-1}\right),\\
b_1 &= \prod_{j=1}^{d}\left(\beta_j + 2\gamma_j\zeta(\alpha)\right),\\
b_2 &= (\ell^r n-1)\prod_{j=1}^{d}\left(\beta_j - \frac{2\gamma_j\zeta(\alpha)(1-(\ell^r n)^{1-\alpha})}{\ell^r n-1}\right),\\
c &= \ell^r n\prod_{j=1}^{d}\beta_j.
\end{aligned}$$

It is not hard to prove that

$$\frac{t_1+t_2-c}{b_1+b_2-c} < \frac{t_1}{b_1}$$

is true if b_1, b_2, t_1, t_2, and c are positive quantities satisfying

$$t_1 < b_1, \quad b_1 + b_2 > c, \quad \text{and} \quad t_2 < b_2 < c. \tag{2.8}$$

Thus the result is proved if we can prove that all these conditions hold.

It may not be obvious that b_2 and t_2 are positive quantities, but one can see that this is the case when $\beta_j - 2\gamma_j\zeta(\alpha)/(n-1) > 0$ for $j = 1, 2\ldots, d$, which is equivalent to the requirement on n given in the statement of the theorem. The requirement that $t_1 < b_1$ comes from the assumption that $\rho_{\ell,r} < 1$, while the requirement that $t_2 < b_2$ comes from the assumption given in (2.7). Also, it is clear that $b_2 < c$.

Let

$$\hat{b}_2 = (\ell^r n - 1)\prod_{j=1}^{d}\left(\beta_j - \frac{2\gamma_j\zeta(\alpha)}{\ell^r n-1}\right).$$

It is clear that $b_2 > \hat{b}_2$. Thus we can prove that $b_1 + b_2 > c$ by proving that $b_1 + \hat{b}_2 - c > 0$. Using the result that

$$\prod_{j=1}^{d}(\beta_j + a_j) = \sum_{\mathfrak{u}\subseteq\mathcal{D}}\left(\prod_{j\notin\mathfrak{u}}\beta_j\prod_{j\in\mathfrak{u}}a_j\right) = \prod_{j=1}^{d}\beta_j + \sum_{\emptyset\neq\mathfrak{u}\subseteq\mathcal{D}}\left(\prod_{j\notin\mathfrak{u}}\beta_j\prod_{j\in\mathfrak{u}}a_j\right),$$

where $\mathcal{D} = \{1, 2, \ldots, d\}$, we have

$$\begin{aligned}
b_1 + \hat{b}_2 - c &= \prod_{j=1}^{d}\left(\beta_j + 2\gamma_j\zeta(\alpha)\right) + (\ell^r n-1)\prod_{j=1}^{d}\left(\beta_j - \frac{2\gamma_j\zeta(\alpha)}{\ell^r n-1}\right) - \ell^r n\prod_{j=1}^{d}\beta_j\\
&= \sum_{\emptyset\neq\mathfrak{u}\subseteq\mathcal{D}}\left(\prod_{j\notin\mathfrak{u}}\beta_j\prod_{j\in\mathfrak{u}}(2\gamma_j\zeta(\alpha))\right) + (\ell^r n-1)\sum_{\emptyset\neq\mathfrak{u}\subseteq\mathcal{D}}\left(\prod_{j\notin\mathfrak{u}}\beta_j\prod_{j\in\mathfrak{u}}\left(-\frac{2\gamma_j\zeta(\alpha)}{\ell^r n-1}\right)\right)\\
&= \sum_{\emptyset\neq\mathfrak{u}\subseteq\mathcal{D}}\left(S(\mathfrak{u})\prod_{j\notin\mathfrak{u}}\beta_j\prod_{j\in\mathfrak{u}}(2\gamma_j\zeta(\alpha))\right),
\end{aligned}$$

where

$$S(\mathfrak{u}) = 1 + (\ell^r n - 1)\left(-\frac{1}{\ell^r n - 1}\right)^{|\mathfrak{u}|}.$$

Clearly $S(\mathfrak{u}) > 0$ if $|\mathfrak{u}|$ is even. For $|\mathfrak{u}| \geq 1$ odd, we have

$$S(\mathfrak{u}) = 1 - (\ell^r n - 1)^{1-|\mathfrak{u}|} \geq 1 - 1 = 0.$$

Thus we conclude that $b_1 + \hat{b}_2 - c > 0$ and hence $b_1 + b_2 > c$. □

In the previous theorem, we made the assumption that $\rho_{\ell,r} < 1$ and that (2.7) was true. Attempts to prove that (2.7) is always true have not been successful. However, all our numerical test calculations with $\ell = 2$, $\alpha = 2$, $\beta_j = 1$, and various choices of γ_j indicate that (2.7) does at least hold for this set of parameters. For other sets of parameters, readers will need to be content with doing their own calculations to see whether it holds or not for their particular situation.

The next result gives some sufficient conditions for $\rho_{2,r}$ to be less than one.

LEMMA 2.4. *Let $\rho_{\ell,r}$ be defined as in Theorem* 2.3*, and set $\ell = 2$. If $\alpha \geq 2$ and*

$$\frac{\gamma_r}{\beta_r} > \frac{1}{(2 - 2^{2-\alpha})\zeta(\alpha)},$$

then $\rho_{2,r} < 1$.

Proof. A product of positive terms is guaranteed to be less than one when each of the terms is less than one. From the definition of $\rho_{\ell,r}$, we see that if $\ell = 2$, then this is the case when

$$\frac{2\beta_j + 2^{2-\alpha}\gamma_j\zeta(\alpha)}{\beta_j + 2\gamma_j\zeta(\alpha)} < 1$$

for all $j = 1, 2, \ldots, r$. When rearranged, this yields

$$\frac{\gamma_j}{\beta_j} > \frac{1}{(2 - 2^{2-\alpha})\zeta(\alpha)}.$$

Since the sequence $\{\frac{\gamma_j}{\beta_j}\}$ is nonincreasing, this completes the proof. □

In the case when $\alpha = 2$, the condition of the lemma becomes $\gamma_r/\beta_r > 1/\zeta(2) = 6/\pi^2 \approx 0.6079$. This suggests that when $\alpha = 2$, it is worthwhile to take r to be at least one when $\gamma_1/\beta_1 > 6/\pi^2$.

In a sense, the quantity $\rho_{\ell,r}$ gives an indication of how much we can gain (or lose) by copying. Later in section 4, we will see that though $\sqrt{\rho_{2,r}}$ is concerned with a ratio of means, the values of $\sqrt{\rho_{2,r}}$ nevertheless provide a measure of the ratios of the worst-case errors between intermediate-rank lattice rules and rank-1 lattice rules with approximately the same number of points.

2.3. Component-by-component construction. We now consider finding the components of the generating vector $\boldsymbol{z}$ one at a time. Keeping in mind the relationship of our intermediate-rank lattice rules with rank-1 lattice rules, we can construct $\bar{\boldsymbol{z}}$ for the rank-1 lattice rule with weights $\boldsymbol{\beta}$ and $\bar{\boldsymbol{\gamma}}$ using the component-by-component Algorithm 8 of [9] from which we can then obtain the corresponding $\boldsymbol{z}$. This yields the same result as constructing $\boldsymbol{z}$ directly using Algorithm 2.5 below.

ALGORITHM 2.5. *Given $1 \leq r \leq d$ and n a prime number:*

1. *Set z_1, the first component of $\boldsymbol{z}$, to* 1.

2. *For $s = 2, 3, \dots, r$, find $z_s \in \mathbb{Z}_n = \{1, 2, \dots, n-1\}$ such that*

$$e^2_{n,s,\mathrm{copy}(\ell,s)}(1, z_2, \dots, z_s) = -\prod_{j=1}^{s} \beta_j + \frac{1}{n}\sum_{k=0}^{n-1}\prod_{j=1}^{s}\left(\beta_j + \frac{\gamma_j}{\ell^\alpha}\sum_{h=-\infty}^{\infty}{}' \frac{e^{2\pi \mathrm{i} h\ell k z_j/n}}{|h|^\alpha}\right)$$

is minimized.

3. *For $s = r+1, r+2, \dots, d$, find $z_s \in \mathbb{Z}_n$ such that*

$$\begin{aligned} &e^2_{n,s,\mathrm{copy}(\ell,r)}(1, z_2, \dots, z_s) \\ &= -\prod_{j=1}^{s}\beta_j + \frac{1}{n}\sum_{k=0}^{n-1}\left[\prod_{j=1}^{r}\left(\beta_j + \frac{\gamma_j}{\ell^\alpha}\sum_{h=-\infty}^{\infty}{}' \frac{e^{2\pi \mathrm{i} h\ell k z_j/n}}{|h|^\alpha}\right)\right. \\ &\qquad\qquad \left.\times \prod_{j=r+1}^{d}\left(\beta_j + \gamma_j \sum_{h=-\infty}^{\infty}{}' \frac{e^{2\pi \mathrm{i} h k z_j/n}}{|h|^\alpha}\right)\right] \end{aligned}$$

is minimized.

Theorem 1 and Corollary 2 in [9] give the theoretical foundation behind such a construction for rank-1 lattice rules. We present the corresponding results here for intermediate-rank lattice rules. Note that Theorem 2.6 below is a slight improvement over the corresponding Theorem 1 of [9]. The proof is thus included in the appendix for completeness. (Such an improvement for rank-1 lattice rules was first obtained in [1] by using a different argument.)

THEOREM 2.6. *Let $\boldsymbol{z} = (1, z_2, \dots, z_d)$ be constructed component-by-component as in Algorithm* 2.5.

(a) *For each $s = 1, 2, \dots, r$, we have*

$$e^2_{n,s,\mathrm{copy}(\ell,s)}(1, z_2, \dots, z_s) \le (n-1)^{-\frac{1}{\lambda}} \prod_{j=1}^{s}\left(\beta_j^\lambda + \frac{2\gamma_j^\lambda \zeta(\alpha\lambda)}{\ell^{\alpha\lambda}}\right)^{\frac{1}{\lambda}}$$

for all λ satisfying $\frac{1}{\alpha} < \lambda \le 1$.

(b) *For each $s = r+1, r+2, \dots, d$, we have*

$$\begin{aligned} &e^2_{n,s,\mathrm{copy}(\ell,r)}(1, z_2, \dots, z_s) \\ &\le (n-1)^{-\frac{1}{\lambda}} \prod_{j=1}^{r}\left(\beta_j^\lambda + \frac{2\gamma_j^\lambda\zeta(\alpha\lambda)}{\ell^{\alpha\lambda}}\right)^{\frac{1}{\lambda}} \prod_{j=r+1}^{s}\left(\beta_j^\lambda + 2\gamma_j^\lambda\zeta(\alpha\lambda)\right)^{\frac{1}{\lambda}} \end{aligned}$$

for all λ satisfying $\frac{1}{\alpha} < \lambda \le 1$.

It can be shown from the bounds above that the intermediate-rank lattice rules constructed using Algorithm 2.5 satisfy strong QMC tractability error bounds and achieve the optimal rate of convergence.

THEOREM 2.7. *For fixed r satisfying $1 \le r \le d$ and n a prime number, let $N = \ell^r n$ denote the total number of quadrature points, and let $\boldsymbol{z}$ be constructed component-by-component as in Algorithm* 2.5. *Then this $\boldsymbol{z}$ satisfies*

$$e_{n,d,\mathrm{copy}(\ell,r)}(\boldsymbol{z}) \le C_d(\delta)\, N^{-\frac{\alpha}{2}+\delta} e_{0,d} \quad \textit{for all} \quad 0 < \delta \le \tfrac{\alpha-1}{2},$$

where

$$C_d(\delta) = (2\ell^r)^{\frac{\alpha}{2}-\delta} \prod_{j=1}^{r} \left[1 + 2\ell^{-\frac{\alpha}{\alpha-2\delta}} \left(\frac{\gamma_j}{\beta_j}\right)^{\frac{1}{\alpha-2\delta}} \zeta\left(\tfrac{\alpha}{\alpha-2\delta}\right)\right]^{\frac{\alpha}{2}-\delta}$$
$$\times \prod_{j=r+1}^{d} \left[1 + 2\left(\frac{\gamma_j}{\beta_j}\right)^{\frac{1}{\alpha-2\delta}} \zeta\left(\tfrac{\alpha}{\alpha-2\delta}\right)\right]^{\frac{\alpha}{2}-\delta}$$

is independent of N. Moreover, if

$$\sum_{j=r+1}^{\infty} \left(\frac{\gamma_j}{\beta_j}\right)^{\frac{1}{\alpha-2\delta}} < \infty,$$

then

$$C_d(\delta) \le C_\infty(\delta) < \infty;$$

that is, $e_{n,d,\mathrm{copy}(\ell,r)}(\boldsymbol{z})$ is $O(N^{-\alpha/2+\delta})$ for $\delta > 0$, independently of d.

Proof. It follows from Theorem 2.6 that the $\boldsymbol{z}$ constructed by Algorithm 2.5 satisfies

$$e_{n,d,\mathrm{copy}(\ell,r)}(\boldsymbol{z})$$
$$\le (n-1)^{-\frac{1}{2\lambda}} \prod_{j=1}^{r} \left(\beta_j^\lambda + \frac{2\gamma_j^\lambda \zeta(\alpha\lambda)}{\ell^{\alpha\lambda}}\right)^{\frac{1}{2\lambda}} \prod_{j=r+1}^{d} \left(\beta_j^\lambda + 2\gamma_j^\lambda \zeta(\alpha\lambda)\right)^{\frac{1}{2\lambda}}$$
$$\le \left(\frac{n\ell^r}{2\ell^r}\right)^{-\frac{1}{2\lambda}} \prod_{j=1}^{r} \left(1 + 2\ell^{-\alpha\lambda} \left(\frac{\gamma_j}{\beta_j}\right)^{\lambda} \zeta(\alpha\lambda)\right)^{\frac{1}{2\lambda}} \prod_{j=r+1}^{d} \left(1 + 2\left(\frac{\gamma_j}{\beta_j}\right)^{\lambda} \zeta(\alpha\lambda)\right)^{\frac{1}{2\lambda}} \prod_{j=1}^{d} \beta_j^{\frac{1}{2}}$$

for all $\frac{1}{\alpha} < \lambda \le 1$. Now with the substitution of

$$-\frac{\alpha}{2} + \delta = -\frac{1}{2\lambda},$$

the condition $\frac{1}{\alpha} < \lambda \le 1$ becomes $0 < \delta \le \frac{\alpha-1}{2}$, and we obtain

$$e_{n,d,\mathrm{copy}(\ell,r)}(\boldsymbol{z}) \le C_d(\delta)\, N^{-\frac{\alpha}{2}+\delta} e_{0,d} \quad \text{for all } 0 < \delta \le \tfrac{\alpha-1}{2},$$

where

$$C_d(\delta) = (2\ell^r)^{\frac{\alpha}{2}-\delta} \prod_{j=1}^{r} \left[1 + 2\ell^{-\frac{\alpha}{\alpha-2\delta}} \left(\frac{\gamma_j}{\beta_j}\right)^{\frac{1}{\alpha-2\delta}} \zeta\left(\tfrac{\alpha}{\alpha-2\delta}\right)\right]^{\frac{\alpha}{2}-\delta}$$
$$\times \prod_{j=r+1}^{d} \left[1 + 2\left(\frac{\gamma_j}{\beta_j}\right)^{\frac{1}{\alpha-2\delta}} \zeta\left(\tfrac{\alpha}{\alpha-2\delta}\right)\right]^{\frac{\alpha}{2}-\delta} \le C_\infty(\delta),$$

and

$$
\begin{aligned}
C_\infty(\delta) &= (2\ell^r)^{\frac{\alpha}{2}-\delta} \prod_{j=1}^{r} \left[1 + 2\ell^{-\frac{\alpha}{\alpha-2\delta}} \left(\frac{\gamma_j}{\beta_j}\right)^{\frac{1}{\alpha-2\delta}} \zeta\left(\tfrac{\alpha}{\alpha-2\delta}\right)\right]^{\frac{\alpha}{2}-\delta} \\
&\quad \times \exp\left(\left(\tfrac{\alpha}{2}-\delta\right) \sum_{j=r+1}^{\infty} \log\left(1 + 2\left(\frac{\gamma_j}{\beta_j}\right)^{\frac{1}{\alpha-2\delta}} \zeta\left(\tfrac{\alpha}{\alpha-2\delta}\right)\right)\right) \\
&\le (2\ell^r)^{\frac{\alpha}{2}-\delta} \prod_{j=1}^{r} \left[1 + 2\ell^{-\frac{\alpha}{\alpha-2\delta}} \left(\frac{\gamma_j}{\beta_j}\right)^{\frac{1}{\alpha-2\delta}} \zeta\left(\tfrac{\alpha}{\alpha-2\delta}\right)\right]^{\frac{\alpha}{2}-\delta} \\
&\quad \times \exp\left((\alpha-2\delta)\, \zeta\left(\tfrac{\alpha}{\alpha-2\delta}\right) \sum_{j=r+1}^{\infty} \left(\frac{\gamma_j}{\beta_j}\right)^{\frac{1}{\alpha-2\delta}}\right),
\end{aligned}
$$

where we have used the fact that $\log(1+x) \le x$ for $x \ge 0$. It is clear from this expression that for $\delta > 0$, $C_\infty(\delta) < \infty$ if

$$
\sum_{j=r+1}^{\infty} \left(\frac{\gamma_j}{\beta_j}\right)^{\frac{1}{\alpha-2\delta}} < \infty.
$$

This completes the proof. □

3. Shifted intermediate-rank lattice rules in weighted Sobolev spaces.

Now we change the function spaces to weighted Sobolev spaces considered in [10]. These spaces are also parameterized by two sequences of positive weights $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ satisfying

$$
\frac{\gamma_1}{\beta_1} \ge \frac{\gamma_2}{\beta_2} \ge \cdots .
$$

The inner product in these spaces is given by

$$
\langle f, g \rangle_d := \sum_{\mathfrak{u} \subseteq \{1,2,\ldots,d\}} \left(\prod_{j \notin \mathfrak{u}} \beta_j^{-1} \prod_{j \in \mathfrak{u}} \gamma_j^{-1} \int_{[0,1]^{|\mathfrak{u}|}} \frac{\partial^{|\mathfrak{u}|}}{\partial \boldsymbol{x}_\mathfrak{u}} f(\boldsymbol{x}_\mathfrak{u}, \mathbf{1}) \frac{\partial^{|\mathfrak{u}|}}{\partial \boldsymbol{x}_\mathfrak{u}} g(\boldsymbol{x}_\mathfrak{u}, \mathbf{1})\, \mathrm{d}\boldsymbol{x}_\mathfrak{u} \right),
$$

where $(\boldsymbol{x}_\mathfrak{u}, \mathbf{1})$ is a d-dimensional vector whose jth component is x_j if $j \in \mathfrak{u}$ and 1 if $j \notin \mathfrak{u}$. Similar spaces have been considered previously (for example, see [12], [13], and [16]). The worst-case error for a QMC rule (1.1) in these spaces is given by

$$
\begin{aligned}
e_{n,d}^2(\boldsymbol{x}_0, \ldots, \boldsymbol{x}_{n-1}) &= \prod_{j=1}^{d} \left(\beta_j + \frac{\gamma_j}{3}\right) - \frac{2}{n} \sum_{i=0}^{n-1} \prod_{j=1}^{d} \left(\beta_j + \frac{\gamma_j}{2}\left(1 - x_{i,j}^2\right)\right) \\
&\quad + \frac{1}{n^2} \sum_{i=0}^{n-1} \sum_{k=0}^{n-1} \prod_{j=1}^{d} \left(\beta_j + \gamma_j \left[1 - \max\left(x_{i,j}, x_{k,j}\right)\right]\right), \qquad (3.1)
\end{aligned}
$$

and the initial error is

$$
e_{0,d} = \prod_{j=1}^{d} \left(\beta_j + \frac{\gamma_j}{3}\right)^{\frac{1}{2}}.
$$

Similar to the weighted Korobov spaces, it can be shown that if the weights satisfy (2.2), then an upper bound for the square worst-case error of the form (2.3) is enough to ensure strong QMC tractability in weighted Sobolev spaces.

We now consider the $\boldsymbol{\Delta}$-shift of the (ℓ, r)-copy of a rank-1 lattice rule with generating vector $\boldsymbol{z}$, that is, a rule with points given by

$$\left\{ \left\{ \frac{i\boldsymbol{z}}{n} + \frac{(m_1, \ldots, m_r, 0, \ldots, 0)}{\ell} + \boldsymbol{\Delta} \right\} : 0 \le i \le n-1,\ 0 \le m_1, \ldots, m_r \le \ell - 1 \right\},$$

where $\ell \ge 1$, $\gcd(\ell, n) = 1$, and $0 \le r \le d$. Let $e_{n,d,\mathrm{copy}(\ell,r)}(\boldsymbol{z}, \boldsymbol{\Delta})$ denote the worst-case error for such a rule. An expression for $e^2_{n,d,\mathrm{copy}(\ell,r)}(\boldsymbol{z}, \boldsymbol{\Delta})$ can be derived from (3.1).

Here we give just the general ideas of the existence and the construction of a good shifted intermediate-rank lattice rule. The full details follow closely the arguments from [12] and [13].

To obtain an upper bound on the square worst-case error, we define the mean of $e^2_{n,d,\mathrm{copy}(\ell,r)}(\boldsymbol{z}, \boldsymbol{\Delta})$ over all values of $\boldsymbol{z} \in \mathbb{Z}_n^d$ and $\boldsymbol{\Delta} \in [0,1]^d$ by

$$M_{n,d,\mathrm{copy}(\ell,r)} := \frac{1}{(n-1)^d} \sum_{\boldsymbol{z} \in \mathbb{Z}_n^d} \left(\int_{[0,1]^d} e^2_{n,d,\mathrm{copy}(\ell,r)}(\boldsymbol{z}, \boldsymbol{\Delta}) \, \mathrm{d}\boldsymbol{\Delta} \right).$$

Using a known relationship between weighted Korobov spaces and weighted Sobolev spaces (see [5]), we see that this mean is exactly the mean given in Theorem 2.2 with α replaced by 2, β_j replaced by $\beta_j + \frac{\gamma_j}{3}$, and γ_j replaced by $\frac{\gamma_j}{2\pi^2}$. An upper bound for $M_{n,d,\mathrm{copy}(\ell,r)}$ follows in the same way from Theorem 2.2:

$$M_{n,d,\mathrm{copy}(\ell,r)} \le \frac{1}{n} \prod_{j=1}^{d} \left(\beta_j + \frac{\gamma_j}{3} + \frac{2\gamma_j}{2\pi^2} \zeta(2) \right) = \frac{1}{n} \prod_{j=1}^{d} \left(\beta_j + \frac{\gamma_j}{2} \right).$$

We thus conclude that there exists at least one pair $(\boldsymbol{z}, \boldsymbol{\Delta})$ such that $e^2_{n,d,\mathrm{copy}(\ell,r)}(\boldsymbol{z}, \boldsymbol{\Delta})$ is bounded by this upper bound on the mean. Since this bound is of the form (2.3), we conclude that shifted intermediate-rank lattice rules achieve strong QMC tractability error bounds in weighted Sobolev spaces.

Let $e_{n,d+1,\ell}(\boldsymbol{x}_0, \ldots, \boldsymbol{x}_{n-1}; z_{d+1}, \Delta_{d+1})$ denote the worst-case error for a QMC rule with the set of points

$$\left\{ \left(\boldsymbol{x}_i, \left\{ \frac{i z_{d+1}}{n} + \frac{m}{\ell} + \Delta_{d+1} \right\} \right) : 0 \le i \le n-1,\ 0 \le m \le \ell - 1 \right\}.$$

These are $(d+1)$-dimensional points obtained by appending $\{\frac{iz_{d+1}}{n} + \frac{m}{\ell} + \Delta_{d+1}\}$ to the existing d components of $\boldsymbol{x}_i$. To construct the pair (z_{d+1}, Δ_{d+1}) component-by-component, we define the following mean:

$$m_{n,d+1,\ell}(\boldsymbol{x}_0, \ldots, \boldsymbol{x}_{n-1}; z_{d+1}) := \int_0^1 e^2_{n,d+1,\ell}(\boldsymbol{x}_0, \ldots, \boldsymbol{x}_{n-1}; z_{d+1}, \Delta_{d+1}) \, \mathrm{d}\Delta_{d+1}.$$

Let us assume that the points $\boldsymbol{x}_0, \ldots, \boldsymbol{x}_{n-1}$ satisfy

$$e^2_{n,d}(\boldsymbol{x}_0, \ldots, \boldsymbol{x}_{n-1}) \le \frac{1}{n} \prod_{j=1}^{d} (\beta_j + \gamma_j).$$

Suppose we choose z_{d+1} from the set $\mathbb{Z}_n$ to minimize $m_{n,d+1,\ell}(\boldsymbol{x}_0, \dots, \boldsymbol{x}_{n-1}; z_{d+1})$ and then choose Δ_{d+1} from the set $\left\{\frac{2m-1}{2n} : 1 \le m \le n-1\right\}$ so that the square worst-case error $e^2_{n,d+1,\ell}(\boldsymbol{x}_0, \dots, \boldsymbol{x}_{n-1}; z_{d+1}, \Delta_{d+1})$ is minimized. Then by using involved algebraic manipulations and the arguments from [12], these choices of z_{d+1} and Δ_{d+1} can be shown to satisfy

$$e^2_{n,d+1,\ell}(\boldsymbol{x}_0, \dots, \boldsymbol{x}_{n-1}; z_{d+1}, \Delta_{d+1}) \le \frac{1}{n} \prod_{j=1}^{d+1} (\beta_j + \gamma_j)\,.$$

Note that the result also holds for $\ell = 1$; that is, there is no "copying" in the $(d+1)$th dimension. For $d = 1$, we can show that there exists (z_1, Δ_1) satisfying

$$e^2_{n,1,\mathrm{copy}(\ell,1)}(z_1, \Delta_1) \le \frac{1}{n} (\beta_1 + \gamma_1)\,.$$

All of the above leads us to the following algorithm for constructing a pair $(\boldsymbol{z}, \boldsymbol{\Delta})$ such that for all $s = 1, \dots, d$,

$$e^2_{n,s,\mathrm{copy}(\ell,\min(s,r))}((z_1, \dots, z_s), (\Delta_1, \dots, \Delta_s)) \le \frac{1}{n} \prod_{j=1}^{s} (\beta_j + \gamma_j)\,.$$

In the following algorithm, the notation

$$m_{n,s,\mathrm{copy}(\ell,r)}((1, z_2, \dots, z_{s-1}), (\Delta_1, \Delta_2, \dots, \Delta_{s-1}); z_s)$$

is used to denote the quantity $m_{n,s,\ell}(\boldsymbol{x}_0, \dots, \boldsymbol{x}_{n-1}; z_s)$ in the situation when $\boldsymbol{x}_0, \dots, \boldsymbol{x}_{n-1}$ are the points from an (ℓ, r)-copy of an $(s-1)$-dimensional rank-1 lattice rule.

```
ALGORITHM 3.1. Given $n$ a prime number and $1 \le r \le d$:
  1. Set $z_1$, the first component of $\boldsymbol{z}$, to 1.
  2. Find $\Delta_1 \in \left\{\frac{1}{2n}, \frac{3}{2n}, \dots, \frac{2n-1}{2n}\right\}$ to minimize $e^2_{n,1,\mathrm{copy}(\ell,1)}(1, \Delta_1)$.
  3. For $s = 2, 3, \dots, r$, do the following:
     (a) Find $z_s \in \{1, 2, \dots, n-1\}$ to minimize
             $m_{n,s,\mathrm{copy}(\ell,s)}((1, z_2, \dots, z_{s-1}), (\Delta_1, \Delta_2, \dots, \Delta_{s-1}); z_s)$.
     (b) Find $\Delta_s \in \left\{\frac{1}{2n}, \frac{3}{2n}, \dots, \frac{2n-1}{2n}\right\}$ to minimize
             $e^2_{n,s,\mathrm{copy}(\ell,s)}((1, z_2, \dots, z_s), (\Delta_1, \Delta_2, \dots, \Delta_s))$.
  4. For $s = r+1, r+2, \dots, d$, do the following:
     (a) Find $z_s \in \{1, 2, \dots, n-1\}$ to minimize
             $m_{n,s,\mathrm{copy}(\ell,r)}((1, z_2, \dots, z_{s-1}), (\Delta_1, \Delta_2, \dots, \Delta_{s-1}); z_s)$.
     (b) Find $\Delta_s \in \left\{\frac{1}{2n}, \frac{3}{2n}, \dots, \frac{2n-1}{2n}\right\}$ to minimize
             $e^2_{n,s,\mathrm{copy}(\ell,r)}((1, z_2, \dots, z_s), (\Delta_1, \Delta_2, \dots, \Delta_s))$.
```

The cost for the construction is $O(n^3 d^2)$ operations, and it is dominated by the construction of the shift. In [13] the idea of using a number of random shifts was introduced. This not only cuts the cost of the construction down to $O(n^2 d^2)$ operations; it also allows error estimation. The reference [3] contains detailed discussions on randomized QMC methods. Following [13], we can construct the generating vector component-by-component by minimizing over the mean

$$F_{n,d,\mathrm{copy}(\ell,r)}(\boldsymbol{z}) := \int_{[0,1]^d} e^2_{n,d,\mathrm{copy}(\ell,r)}(\boldsymbol{z},\boldsymbol{\Delta})\,\mathrm{d}\boldsymbol{\Delta}.$$

ALGORITHM 3.2. *Given n a prime number and $1 \le r \le d$:*

1. *Set z_1, the first component of $\boldsymbol{z}$, to* 1.
2. *For $s = 2, 3, \dots, r$, find $z_s \in \{1, 2, \dots, n-1\}$ to minimize*
$$F_{n,s,\mathrm{copy}(\ell,s)}(1, z_2, \dots, z_s).$$
3. *For $s = r+1, r+2, \dots, d$, find $z_s \in \{1, 2, \dots, n-1\}$ to minimize*
$$F_{n,s,\mathrm{copy}(\ell,r)}(1, z_2, \dots, z_s).$$

Using again the relationship between weighted Korobov spaces and weighted Sobolev spaces, we can obtain the corresponding value of the quantity $\rho_{\ell,r}$ given in Theorem 2.3 for weighted Sobolev spaces by replacing α with 2, β_j with $\beta_j + \frac{\gamma_j}{3}$, and γ_j with $\frac{\gamma_j}{2\pi^2}$. This yields

$$\rho_{\ell,r} = \prod_{j=1}^{r} \frac{\ell\beta_j + \gamma_j\left(\frac{\ell}{3} + \frac{1}{6\ell}\right)}{\beta_j + \frac{\gamma_j}{2}},$$

which is greater than 1 for all $\ell \ge 2$. This means that it is unlikely for the ratio $R_{n,d,\ell,r}$ to be less than 1, and thus copying may not give better results in weighted Sobolev spaces.

4. Numerical results. We will consider weighted Korobov spaces with $\alpha = 2$. In this case, the square worst-case error can be written as

$$e^2_{n,d,\mathrm{copy}(\ell,r)}(\boldsymbol{z}) = -\prod_{j=1}^{d}\beta_j + \frac{1}{n}\sum_{k=0}^{n-1}\left[\prod_{j=1}^{r}\left(\beta_j + \frac{2\pi^2\gamma_j}{\ell^2}B_2\left(\left\{\frac{\ell k z_j}{n}\right\}\right)\right)\right.$$
$$\left.\times\prod_{j=r+1}^{d}\left(\beta_j + 2\pi^2\gamma_j B_2\left(\left\{\frac{k z_j}{n}\right\}\right)\right)\right],$$

where for $x \in [0,1]$, $B_2(x) = x^2 - x + 1/6$ is the Bernoulli polynomial of degree 2. In the implementation of Steps 2 and 3 of Algorithm 2.5, we will consider only values of z_s in $\{1, 2, \dots, (n-1)/2\}$, since

$$B_2\left(\left\{\frac{k z_s}{n}\right\}\right) = B_2\left(\left\{\frac{k(n-z_s)}{n}\right\}\right) \text{ and } B_2\left(\left\{\frac{\ell k z_s}{n}\right\}\right) = B_2\left(\left\{\frac{\ell k(n-z_s)}{n}\right\}\right).$$

For $\alpha = 2$, $\boldsymbol{\beta} = \mathbf{1}$, and two different sequences of $\boldsymbol{\gamma}$,

$$\gamma_j = 0.9^j \quad \text{and} \quad \gamma_j = \frac{1}{j^2},$$

we want to see if intermediate-rank lattice rules are better than rank-1 lattice rules with approximately the same number of points. More precisely, when $\ell = 2$, we

TABLE 4.1
Total number of points close to 4000, $\gamma_j = 0.9^j$.

d	4001	2003×2^1 $= 4006$	1999×2^1 $= 3998$	1009×2^2 $= 4036$	997×2^2 $= 3988$	503×2^3 $= 4024$	499×2^3 $= 3992$
10	2.9726e+00	2.8068e+00	2.8005e+00	2.6666e+00	2.6874e+00	2.5965e+00	2.6068e+00
20	3.8737e+01	3.6466e+01	3.6455e+01	3.4687e+01	3.4922e+01	3.3752e+01	3.3887e+01
30	1.1022e+02	1.0374e+02	1.0372e+02	9.8675e+01	9.9337e+01	9.6009e+01	9.6392e+01
40	1.6309e+02	1.5348e+02	1.5346e+02	1.4598e+02	1.4696e+02	1.4204e+02	1.4260e+02
50	1.8768e+02	1.7661e+02	1.7659e+02	1.6798e+02	1.6911e+02	1.6344e+02	1.6409e+02
60	1.9719e+02	1.8556e+02	1.8554e+02	1.7650e+02	1.7768e+02	1.7172e+02	1.7241e+02
70	2.0063e+02	1.8879e+02	1.8878e+02	1.7957e+02	1.8077e+02	1.7472e+02	1.7542e+02
80	2.0185e+02	1.8994e+02	1.8992e+02	1.8066e+02	1.8187e+02	1.7578e+02	1.7648e+02
90	2.0227e+02	1.9034e+02	1.9032e+02	1.8104e+02	1.8225e+02	1.7615e+02	1.7685e+02
100	2.0242e+02	1.9048e+02	1.9046e+02	1.8118e+02	1.8239e+02	1.7628e+02	1.7698e+02

TABLE 4.2
Total number of points close to 16000, $\gamma_j = 0.9^j$.

d	16007	8009×2^1 $= 16018$	7993×2^1 $= 15986$	4003×2^2 $= 16012$	4001×2^2 $= 16004$	2003×2^3 $= 16024$	1999×2^3 $= 15992$
10	1.4365e+00	1.3566e+00	1.3606e+00	1.2982e+00	1.2973e+00	1.2623e+00	1.2621e+00
20	1.9268e+01	1.8198e+01	1.8231e+01	1.7400e+01	1.7413e+01	1.6905e+01	1.6922e+01
30	5.4841e+01	5.1793e+01	5.1887e+01	4.9516e+01	4.9554e+01	4.8109e+01	4.8157e+01
40	8.1141e+01	7.6628e+01	7.6767e+01	7.3258e+01	7.3314e+01	7.1176e+01	7.1247e+01
50	9.3371e+01	8.8176e+01	8.8337e+01	8.4298e+01	8.4363e+01	8.1902e+01	8.1984e+01
60	9.8103e+01	9.2645e+01	9.2813e+01	8.8570e+01	8.8638e+01	8.6053e+01	8.6138e+01
70	9.9815e+01	9.4261e+01	9.4433e+01	9.0115e+01	9.0184e+01	8.7554e+01	8.7641e+01
80	1.0042e+02	9.4832e+01	9.5004e+01	9.0661e+01	9.0730e+01	8.8084e+01	8.8172e+01
90	1.0063e+02	9.5032e+01	9.5205e+01	9.0852e+01	9.0922e+01	8.8270e+01	8.8358e+01
100	1.0070e+02	9.5102e+01	9.5275e+01	9.0919e+01	9.0988e+01	8.8335e+01	8.8423e+01

TABLE 4.3
Total number of points close to 64000, $\gamma_j = 0.9^j$.

d	64007	32009×2^1 $= 64018$	32003×2^1 $= 64006$	16007×2^2 $= 64028$	16001×2^2 $= 64004$	8009×2^3 $= 64072$	7993×2^3 $= 63944$
10	6.8423e-01	6.4784e-01	6.4773e-01	6.1683e-01	6.1797e-01	5.9937e-01	6.0015e-01
20	9.6190e+00	9.1124e+00	9.0949e+00	8.6927e+00	8.6945e+00	8.4445e+00	8.4523e+00
30	2.7406e+01	2.5958e+01	2.5908e+01	2.4763e+01	2.4768e+01	2.4055e+01	2.4077e+01
40	4.0552e+01	3.8408e+01	3.8336e+01	3.6640e+01	3.6647e+01	3.5592e+01	3.5625e+01
50	4.6664e+01	4.4197e+01	4.4114e+01	4.2162e+01	4.2170e+01	4.0956e+01	4.0995e+01
60	4.9029e+01	4.6436e+01	4.6350e+01	4.4299e+01	4.4307e+01	4.3032e+01	4.3072e+01
70	4.9885e+01	4.7247e+01	4.7159e+01	4.5072e+01	4.5080e+01	4.3783e+01	4.3824e+01
80	5.0187e+01	4.7533e+01	4.7444e+01	4.5345e+01	4.5353e+01	4.4048e+01	4.4089e+01
90	5.0293e+01	4.7633e+01	4.7544e+01	4.5441e+01	4.5449e+01	4.4141e+01	4.4182e+01
100	5.0330e+01	4.7668e+01	4.7579e+01	4.5474e+01	4.5483e+01	4.4173e+01	4.4215e+01

want to know in how many dimensions to copy, that is, which value of $r = 1, 2$, or higher should we choose to get better rules than rank-1 lattice rules. We compare the worst-case errors for rules with approximately 4000, 16000, and 64000 points up to 100 dimensions. (Note that since $\boldsymbol{\beta} = \mathbf{1}$, the initial error $e_{0,d}$ is 1.) The results are presented in Tables 4.1 to 4.6. The second column of each of these tables contains the worst-case error for rank-1 rules, while the other three columns contain the worst-case error for r going from $r = 1$ to $r = 3$. To get a better picture of the results of copying, we divide the worst-case errors of intermediate-rank lattice rules at $d = 100$ by those of rank-1 lattice rules with approximately the same number of points. These ratios are presented in Table 4.7.

We can see from the results that for $\gamma_j = 0.9^j$, copying is good in at least the first three dimensions, but for $\gamma_j = 1/j^2$, it is only good to copy in the first dimension. This seems reasonable as in the first few dimensions the sequence $0.9, 0.81, 0.729, \ldots$

TABLE 4.4
Total number of points close to 4000, $\gamma_j = 1/j^2$.

		2003×2^1	1999×2^1	1009×2^2	997×2^2	503×2^3	499×2^3
d	4001	= 4006	= 3998	= 4036	= 3988	= 4024	= 3992
10	1.9338e-02	1.8362e-02	1.8036e-02	2.0006e-02	2.0218e-02	2.5298e-02	2.5381e-02
20	2.5421e-02	2.3923e-02	2.3776e-02	2.6554e-02	2.6843e-02	3.3678e-02	3.3951e-02
30	2.7770e-02	2.6094e-02	2.6001e-02	2.9126e-02	2.9474e-02	3.7124e-02	3.7290e-02
40	2.9017e-02	2.7262e-02	2.7181e-02	3.0495e-02	3.0864e-02	3.8956e-02	3.9126e-02
50	2.9795e-02	2.7989e-02	2.7913e-02	3.1362e-02	3.1728e-02	4.0095e-02	4.0269e-02
60	3.0326e-02	2.8489e-02	2.8415e-02	3.1954e-02	3.2320e-02	4.0875e-02	4.1051e-02
70	3.0714e-02	2.8853e-02	2.8780e-02	3.2384e-02	3.2751e-02	4.1444e-02	4.1620e-02
80	3.1008e-02	2.9130e-02	2.9058e-02	3.2711e-02	3.3079e-02	4.1875e-02	4.2052e-02
90	3.1240e-02	2.9348e-02	2.9276e-02	3.2970e-02	3.3337e-02	4.2213e-02	4.2390e-02
100	3.1426e-02	2.9523e-02	2.9453e-02	3.3178e-02	3.3545e-02	4.2485e-02	4.2662e-02

TABLE 4.5
Total number of points close to 16000, $\gamma_j = 1/j^2$.

		8009×2^1	7993×2^1	4003×2^2	4001×2^2	2003×2^3	1999×2^3
d	16007	= 16018	= 15986	= 16012	= 16004	= 16024	= 15992
10	7.0679e-03	6.6226e-03	6.7551e-03	7.4726e-03	7.4423e-03	9.2985e-03	9.3671e-03
20	9.7139e-03	9.1387e-03	9.2672e-03	1.0362e-02	1.0344e-02	1.2975e-02	1.3059e-02
30	1.0786e-02	1.0146e-02	1.0266e-02	1.1513e-02	1.1516e-02	1.4491e-02	1.4543e-02
40	1.1364e-02	1.0691e-02	1.0808e-02	1.2139e-02	1.2150e-02	1.5307e-02	1.5349e-02
50	1.1727e-02	1.1033e-02	1.1145e-02	1.2528e-02	1.2543e-02	1.5818e-02	1.5854e-02
60	1.1977e-02	1.1268e-02	1.1378e-02	1.2796e-02	1.2814e-02	1.6168e-02	1.6204e-02
70	1.2159e-02	1.1438e-02	1.1547e-02	1.2993e-02	1.3012e-02	1.6425e-02	1.6461e-02
80	1.2299e-02	1.1567e-02	1.1676e-02	1.3143e-02	1.3163e-02	1.6622e-02	1.6657e-02
90	1.2409e-02	1.1670e-02	1.1778e-02	1.3261e-02	1.3282e-02	1.6778e-02	1.6812e-02
100	1.2498e-02	1.1753e-02	1.1861e-02	1.3357e-02	1.3379e-02	1.6904e-02	1.6938e-02

TABLE 4.6
Total number of points close to 64000, $\gamma_j = 1/j^2$.

		32009×2^1	32003×2^1	16007×2^2	16001×2^2	8009×2^3	7993×2^3
d	64007	= 64018	= 64006	= 64028	= 64004	= 64072	= 63944
10	2.5983e-03	2.4131e-03	2.4454e-03	2.6945e-03	2.6743e-03	3.3548e-03	3.3135e-03
20	3.7412e-03	3.5099e-03	3.5290e-03	3.9372e-03	3.9465e-03	4.9385e-03	4.9097e-03
30	4.2141e-03	3.9582e-03	3.9767e-03	4.4501e-03	4.4667e-03	5.5924e-03	5.5795e-03
40	4.4705e-03	4.2019e-03	4.2200e-03	4.7333e-03	4.7496e-03	5.9532e-03	5.9446e-03
50	4.6325e-03	4.3564e-03	4.3735e-03	4.9111e-03	4.9270e-03	6.1803e-03	6.1750e-03
60	4.7448e-03	4.4624e-03	4.4798e-03	5.0334e-03	5.0484e-03	6.3377e-03	6.3334e-03
70	4.8270e-03	4.5399e-03	4.5573e-03	5.1227e-03	5.1371e-03	6.4533e-03	6.4493e-03
80	4.8901e-03	4.5991e-03	4.6166e-03	5.1912e-03	5.2052e-03	6.5416e-03	6.5378e-03
90	4.9398e-03	4.6460e-03	4.6635e-03	5.2452e-03	5.2593e-03	6.6113e-03	6.6077e-03
100	4.9801e-03	4.6840e-03	4.7014e-03	5.2890e-03	5.3032e-03	6.6678e-03	6.6645e-03

decays more slowly than $1, 1/4, 1/9, \ldots$, and so in the former case, the third variable is still fairly important, while this is not the situation in the latter case.

The phenomenon is also supported by our earlier analysis. Since it may be verified numerically that (2.7) holds, Theorem 2.3 and Lemma 2.4 together suggest that it would be advantageous to copy in the first r dimensions if $\ell = 2$, $\alpha = 2$, and

$$\frac{\gamma_r}{\beta_r} > \frac{6}{\pi^2} \approx 0.6079.$$

For $\gamma_j = 0.9^j$, this is obviously satisfied when $r = 1$, $r = 2$, and $r = 3$. For $\gamma_j = 1/j^2$, this is satisfied only when $r = 1$. Because Lemma 2.4 provides only a sufficient

Table 4.7
Ratios of worst-case errors at $d = 100$.

	Approximate N	$r=1$		$r=2$		$r=3$	
$\gamma_j = 0.9^j$	4000	0.941	0.941	0.895	0.901	0.871	0.874
	16000	0.944	0.946	0.903	0.904	0.877	0.878
	64000	0.947	0.945	0.904	0.904	0.878	0.879
$\gamma_j = 1/j^2$	4000	0.939	0.937	1.056	1.067	1.352	1.358
	16000	0.940	0.949	1.069	1.070	1.353	1.355
	64000	0.941	0.944	1.062	1.065	1.339	1.338

Table 4.8
Values of $\rho_{2,r}$ for $r = 1, 2, 3$.

	$\rho_{2,1}$	$\sqrt{\rho_{2,1}}$	$\rho_{2,2}$	$\sqrt{\rho_{2,2}}$	$\rho_{2,3}$	$\sqrt{\rho_{2,3}}$
$\gamma_j = 0.9^j$	0.879	0.937	0.799	0.894	0.752	0.867
$\gamma_j = 1/j^2$	0.850	0.922	1.124	1.060	1.797	1.340

condition for $\rho_{2,r}$ to be less than one, a direct calculation of $\rho_{2,r}$ was done, and the results (see Table 4.8) show the same conclusion.

If we compare the values of $\sqrt{\rho_{2,r}}$ in Table 4.8 with the ratios in Table 4.7, we see that the values of $\sqrt{\rho_{2,r}}$ are reasonably close to the ratios. So, although $\rho_{2,r}$ is essentially a ratio of means, there is numerical evidence here that it provides a measure of the ratios of the square worst-case errors obtained from intermediate-rank lattice rules and rank-1 lattice rules in the weighted Korobov space setting.

For our choices of weights, it follows from Theorem 2.7 that the rate of convergence is $O(N^{-1+\delta})$ for $\delta > 0$, independently of the dimension d. However, the numerical results presented show a rate of convergence of roughly $O(N^{-1/2})$ for the case $\gamma_j = 0.9^j$ and a somewhat better rate for the case $\gamma_j = 1/j^2$. The observed rates of convergence also appear to be higher for the smaller values of d. This agrees with the numerical results in [9], where the predicted rate of convergence is not observed when moderate values of n are used relative to the dimension. In that situation, the observed rate of convergence depends on the rate of decay of the weights, with faster decaying weights yielding higher convergence rates. To get an observed rate of convergence close to $O(N^{-1})$, we need to have weights that decay much faster, for example, $\gamma_j = 0.1^j$ or $\gamma_j = 1/j^6$. However, if weights such as these were used, the theory would suggest that there would not be much benefit in doing any copying.

Appendix. Let $(1, z_2, \ldots, z_d)$ be constructed using Algorithm 2.5. Here we prove Theorem 2.6; that is, for each $s = 1, 2, \ldots, d$, we have

$$e^2_{n,s,\mathrm{copy}(\ell,\min(s,r))}(1, z_2, \ldots, z_s) \le (n-1)^{-\frac{1}{\lambda}} \prod_{j=1}^{s} \left(\beta_j^\lambda + 2\bar{\gamma}_j^\lambda \zeta(\alpha\lambda)\right)^{\frac{1}{\lambda}}$$

for all λ satisfying $\frac{1}{\alpha} < \lambda \le 1$, where

$$\bar{\gamma}_j := \begin{cases} \dfrac{\gamma_j}{\ell^\alpha} & \text{if } 1 \le j \le r, \\ \gamma_j & \text{otherwise.} \end{cases}$$

The proof makes use of one form of Jensen's inequality (see Theorem 19 of [2]), which states that for $\{a_i\}$ a sequence of positive numbers,

$$\sum a_i \le \left(\sum a_i^\lambda\right)^{\frac{1}{\lambda}} \quad \text{for } 0 < \lambda \le 1.$$

Proof. For $s = 1$, it is not hard to show that for all z_1 we have

$$e^2_{n,1,\text{copy}(\ell,1)}(z_1) = \frac{2\bar{\gamma}_1\zeta(\alpha)}{n^\alpha},$$

and for any λ satisfying $\frac{1}{\alpha} < \lambda \le 1$, we have

$$\frac{2\bar{\gamma}_1\zeta(\alpha)}{n^\alpha} \le n^{-\alpha}\left(\beta_1 + 2\bar{\gamma}_1\zeta(\alpha)\right) \le n^{-\alpha}\left(\beta_1^\lambda + 2^\lambda\bar{\gamma}_1^\lambda[\zeta(\alpha)]^\lambda\right)^{\frac{1}{\lambda}},$$

where the second inequality follows by applying Jensen's inequality to the sum $\beta_1 + 2\bar{\gamma}_1\zeta(\alpha)$. It can be easily verified that $n^{-\alpha} < (n-1)^{-\frac{1}{\lambda}}$, $2^\lambda < 2$, and by Jensen's inequality, $[\zeta(\alpha)]^\lambda \le \zeta(\alpha\lambda)$. Hence the result holds for $s = 1$.

For s satisfying $2 \le s \le d$, suppose that an $(s-1)$-dimensional vector $(1, z_2, \ldots, z_{s-1})$ has already been constructed using Algorithm 2.5 such that it satisfies

$$e^2_{n,s-1,\text{copy}(\ell,\min(s-1,r))}(1, z_2, \ldots, z_{s-1}) \le (n-1)^{-\frac{1}{\lambda}} \prod_{j=1}^{s-1} \left(\beta_j^\lambda + 2\bar{\gamma}_j^\lambda\zeta(\alpha\lambda)\right)^{\frac{1}{\lambda}} \tag{A.1}$$

for all λ satisfying $\frac{1}{\alpha} < \lambda \le 1$. For any $z_s \in \mathbb{Z}_n$, there is a corresponding $\bar{z}_s$ (recall that $\bar{z}_s = \ell z_s$ if $s \le r$ and $\bar{z}_s = z_s$ if $s > r$), and it follows from (2.6) that

$$\begin{aligned} &e^2_{n,s,\text{copy}(\ell,\min(s,r))}(1, z_2, \ldots, z_s) \\ &= \beta_s e^2_{n,s-1,\text{copy}(\ell,\min(s-1,r))}(1, z_2, \ldots, z_{s-1}) + \theta_{n,s}(\alpha, \boldsymbol{\beta}, \bar{\boldsymbol{\gamma}}; 1, z_2, \ldots, z_s), \end{aligned} \tag{A.2}$$

where

$$\begin{aligned} &\theta_{n,s}(\alpha, \boldsymbol{\beta}, \bar{\boldsymbol{\gamma}}; 1, z_2, \ldots, z_s) \\ &= \frac{\bar{\gamma}_s}{n} \sum_{k=0}^{n-1} \left[\prod_{j=1}^{s-1} \left(\beta_j + \bar{\gamma}_j \sideset{}{'}\sum_{h=-\infty}^{\infty} \frac{e^{2\pi \mathrm{i} hk\bar{z}_j/n}}{|h|^\alpha} \right) \sideset{}{'}\sum_{h=-\infty}^{\infty} \frac{e^{2\pi \mathrm{i} hk\bar{z}_s/n}}{|h|^\alpha} \right]. \end{aligned} \tag{A.3}$$

Later we shall prove the following:

(i) For given α, $\boldsymbol{\beta}$, and $\bar{\boldsymbol{\gamma}}$, there exists $z_s = z_s(\alpha, \boldsymbol{\beta}, \bar{\boldsymbol{\gamma}})$ such that

$$\theta_{n,s}(\alpha, \boldsymbol{\beta}, \bar{\boldsymbol{\gamma}}; 1, z_2, \ldots, z_s) \le \frac{2\bar{\gamma}_s\zeta(\alpha)}{n-1} \prod_{j=1}^{s-1} \left(\beta_j + 2\bar{\gamma}_j\zeta(\alpha)\right).$$

(ii) For all $\frac{1}{\alpha} < \lambda \le 1$,

$$\theta_{n,s}(\alpha, \boldsymbol{\beta}, \bar{\boldsymbol{\gamma}}; 1, z_2, \ldots, z_s) \le \left[\theta_{n,s}(\alpha\lambda, \boldsymbol{\beta}^\lambda, \bar{\boldsymbol{\gamma}}^\lambda; 1, z_2, \ldots, z_s)\right]^{\frac{1}{\lambda}},$$

where $\boldsymbol{\beta}^\lambda = \{\beta_j^\lambda\}$ and $\bar{\boldsymbol{\gamma}}^\lambda = \{\bar{\gamma}_j^\lambda\}$.

We see from (i) with α, $\boldsymbol{\beta}$, and $\bar{\boldsymbol{\gamma}}$ replaced by $\alpha\lambda$, $\boldsymbol{\beta}^\lambda$, and $\bar{\boldsymbol{\gamma}}^\lambda$, respectively, that there exists $z_s = z_s(\alpha\lambda, \boldsymbol{\beta}^\lambda, \bar{\boldsymbol{\gamma}}^\lambda)$ such that

$$\theta_{n,s}(\alpha\lambda, \boldsymbol{\beta}^\lambda, \bar{\boldsymbol{\gamma}}^\lambda; 1, z_2, \ldots, z_s) \le \frac{2\bar{\gamma}_s^\lambda\zeta(\alpha\lambda)}{n-1} \prod_{j=1}^{s-1} \left(\beta_j^\lambda + 2\bar{\gamma}_j^\lambda\zeta(\alpha\lambda)\right).$$

For this $z_s = z_s(\alpha\lambda, \boldsymbol{\beta}^\lambda, \bar{\boldsymbol{\gamma}}^\lambda)$, it then follows from (ii) that

$$\theta_{n,s}(\alpha, \boldsymbol{\beta}, \bar{\boldsymbol{\gamma}}; 1, z_2, \dots, z_s) \le \frac{2^{\frac{1}{\lambda}} \bar{\gamma}_s [\zeta(\alpha\lambda)]^{\frac{1}{\lambda}}}{(n-1)^{\frac{1}{\lambda}}} \prod_{j=1}^{s-1} \left(\beta_j^\lambda + 2\bar{\gamma}_j^\lambda \zeta(\alpha\lambda)\right)^{\frac{1}{\lambda}} .$$

Thus it follows from (A.1) and (A.2) that this $z_s = z_s(\alpha\lambda, \boldsymbol{\beta}^\lambda, \bar{\boldsymbol{\gamma}}^\lambda)$ satisfies

$$\begin{aligned} &e^2_{n,s,\mathrm{copy}(\ell,\min(s,r))}(1, z_2, \dots, z_s) \\ &\le \left(\beta_s + 2^{\frac{1}{\lambda}} \bar{\gamma}_s [\zeta(\alpha\lambda)]^{\frac{1}{\lambda}}\right) (n-1)^{-\frac{1}{\lambda}} \prod_{j=1}^{s-1} \left(\beta_j^\lambda + 2\bar{\gamma}_j^\lambda \zeta(\alpha\lambda)\right)^{\frac{1}{\lambda}} \\ &\le \left(\beta_s^\lambda + 2\bar{\gamma}_s^\lambda \zeta(\alpha\lambda)\right)^{\frac{1}{\lambda}} (n-1)^{-\frac{1}{\lambda}} \prod_{j=1}^{s-1} \left(\beta_j^\lambda + 2\bar{\gamma}_j^\lambda \zeta(\alpha\lambda)\right)^{\frac{1}{\lambda}} \\ &= (n-1)^{-\frac{1}{\lambda}} \prod_{j=1}^{s} \left(\beta_j^\lambda + 2\bar{\gamma}_j^\lambda \zeta(\alpha\lambda)\right)^{\frac{1}{\lambda}} , \end{aligned}$$

where the second inequality follows from applying Jensen's inequality to the sum in the first factor. Now since we choose z_s in Algorithm 2.5 to minimize the square worst-case error $e^2_{n,s,\mathrm{copy}(\ell,\min(s,r))}(1, z_2, \dots, z_s)$, this choice of z_s must satisfy the same bound. Hence it follows inductively that the result holds for all $s = 2, 3, \dots, d$.

To complete the proof, we need to prove (i) and (ii).

Proof of (i). Clearly there exists a $z_s = z_s(\alpha, \boldsymbol{\beta}, \bar{\boldsymbol{\gamma}})$ (and hence $\bar{z}_s$) such that

$$\begin{aligned} &\theta_{n,s}(\alpha, \boldsymbol{\beta}, \bar{\boldsymbol{\gamma}}; 1, z_2, \dots, z_s) \\ &\le \frac{1}{n-1} \sum_{z_s=1}^{n-1} \theta_{n,s}(\alpha, \boldsymbol{\beta}, \bar{\boldsymbol{\gamma}}; 1, z_2, \dots, z_s) \\ &= \frac{\bar{\gamma}_s}{n} \sum_{k=0}^{n-1} \left[\prod_{j=1}^{s-1} \left(\beta_j + \bar{\gamma}_j \sum_{h=-\infty}^{\infty}{}' \frac{e^{2\pi i h k \bar{z}_j / n}}{|h|^\alpha} \right) \left(\frac{1}{n-1} \sum_{z_s=1}^{n-1} \sum_{h=-\infty}^{\infty}{}' \frac{e^{2\pi i h k \bar{z}_s / n}}{|h|^\alpha} \right) \right] . \end{aligned} \tag{A.4}$$

Since n is prime and $\gcd(\ell, n) = 1$, it can be shown for $q = 1$ and $q = \ell$ that

$$\frac{1}{n-1} \sum_{z=1}^{n-1} \sum_{h=-\infty}^{\infty}{}' \frac{e^{2\pi i h k q z / n}}{|h|^\alpha} = \begin{cases} 2\zeta(\alpha) & \text{if } k \text{ is a multiple of } n, \\ -\dfrac{2\zeta(\alpha)(1 - n^{1-\alpha})}{n-1} & \text{otherwise.} \end{cases}$$

Upon separating out the $k = 0$ term and using the result above, (A.4) becomes

$$\frac{2\bar{\gamma}_s \zeta(\alpha)}{n} \prod_{j=1}^{s-1} (\beta_j + 2\bar{\gamma}_j \zeta(\alpha)) - \frac{2\bar{\gamma}_s \zeta(\alpha)(1 - n^{1-\alpha})}{n(n-1)} \sum_{k=1}^{n-1} \prod_{j=1}^{s-1} \left(\beta_j + \bar{\gamma}_j \sum_{h=-\infty}^{\infty}{}' \frac{e^{2\pi i h k \bar{z}_j / n}}{|h|^\alpha} \right) .$$

It follows from (2.6) (with the $k = 0$ term separated out) that

$$\begin{aligned} &\frac{1}{n} \sum_{k=1}^{n-1} \prod_{j=1}^{s-1} \left(\beta_j + \bar{\gamma}_j \sum_{h=-\infty}^{\infty}{}' \frac{e^{2\pi i h k \bar{z}_j / n}}{|h|^\alpha} \right) \\ &= e^2_{n,s-1,\mathrm{copy}(\ell,\min(s-1,r))}(1, z_2, \dots, z_{s-1}) + \prod_{j=1}^{s-1} \beta_j - \frac{1}{n} \prod_{j=1}^{s-1} (\beta_j + 2\bar{\gamma}_j \zeta(\alpha)) . \end{aligned}$$

Hence there exists a $z_s = z_s(\alpha, \boldsymbol{\beta}, \bar{\boldsymbol{\gamma}})$ such that

$$\begin{aligned}
&\theta_{n,s}(\alpha, \boldsymbol{\beta}, \bar{\boldsymbol{\gamma}}; 1, z_2, \dots, z_s) \\
&\le \frac{2\bar{\gamma}_s\zeta(\alpha)}{n} \prod_{j=1}^{s-1} (\beta_j + 2\bar{\gamma}_j\zeta(\alpha)) + \frac{2\bar{\gamma}_s\zeta(\alpha)(1-n^{1-\alpha})}{n(n-1)} \prod_{j=1}^{s-1} (\beta_j + 2\bar{\gamma}_j\zeta(\alpha)) \\
&\le \frac{2\bar{\gamma}_s\zeta(\alpha)}{n} \left(1 + \frac{1}{n-1}\right) \prod_{j=1}^{s-1} (\beta_j + 2\bar{\gamma}_j\zeta(\alpha)) \\
&= \frac{2\bar{\gamma}_s\zeta(\alpha)}{(n-1)} \prod_{j=1}^{s-1} (\beta_j + 2\bar{\gamma}_j\zeta(\alpha)) .
\end{aligned}$$

Proof of (ii). Let

$$r(\alpha, \beta, \gamma, h) := \begin{cases} \beta^{-1} & \text{if } h = 0, \\ \gamma^{-1}|h|^\alpha & \text{if } h \neq 0. \end{cases}$$

With this notation we can write $\theta_{n,s}(\alpha, \boldsymbol{\beta}, \bar{\boldsymbol{\gamma}}; 1, z_2, \dots, z_s)$ in (A.3) as

$$\begin{aligned}
&\theta_{n,s}(\alpha, \boldsymbol{\beta}, \bar{\boldsymbol{\gamma}}; 1, z_2, \dots, z_s) \\
&= \frac{\bar{\gamma}_s}{n} \sum_{k=0}^{n-1} \sum_{\substack{\boldsymbol{h}\in\mathbb{Z}^s \\ h_s\neq 0}} \frac{e^{2\pi \mathrm{i} k(h_1,h_2,\dots,h_s)\cdot(\bar{z}_1,\bar{z}_2,\dots,\bar{z}_s)/n}}{|h_s|^\alpha \prod_{j=1}^{s-1} r(\alpha, \beta_j, \bar{\gamma}_j, h_j)} \\
&= \bar{\gamma}_s \sum_{\substack{\boldsymbol{h}\in\mathbb{Z}^s \\ h_s\neq 0 \\ (h_1,h_2,\dots,h_s)\cdot(\bar{z}_1,\bar{z}_2,\dots,\bar{z}_s)\equiv 0 \,(\mathrm{mod}\, n)}} \left(|h_s|^{-\alpha} \prod_{j=1}^{s-1} r(\alpha, \beta_j, \bar{\gamma}_j, h_j)^{-1} \right)
\end{aligned}$$

since

$$\sum_{k=0}^{n-1} e^{2\pi \mathrm{i} k(h_1,h_2,\dots,h_s)\cdot(\bar{z}_1,\bar{z}_2,\dots,\bar{z}_s)/n} = \sum_{k=0}^{n-1} \left(e^{2\pi \mathrm{i}(h_1,h_2,\dots,h_s)\cdot(\bar{z}_1,\bar{z}_2,\dots,\bar{z}_s)/n} \right)^k = 0$$

if $(h_1, h_2, \dots, h_s) \cdot (\bar{z}_1, \bar{z}_2, \dots, \bar{z}_s)$ is not a multiple of n. It now follows from Jensen's inequality that

$$\begin{aligned}
&\theta_{n,s}(\alpha, \boldsymbol{\beta}, \bar{\boldsymbol{\gamma}}; 1, z_2, \dots, z_s) \\
&\le \bar{\gamma}_s \left[\sum_{\substack{\boldsymbol{h}\in\mathbb{Z}^s \\ h_s\neq 0 \\ (h_1,h_2,\dots,h_s)\cdot(\bar{z}_1,\bar{z}_2,\dots,\bar{z}_s)\equiv 0 \,(\mathrm{mod}\, n)}} \left(|h_s|^{-\alpha\lambda} \prod_{j=1}^{s-1} r(\alpha, \beta_j, \bar{\gamma}_j, h_j)^{-\lambda} \right) \right]^{\frac{1}{\lambda}} \\
&= \left[\theta_{n,s}(\alpha\lambda, \boldsymbol{\beta}^\lambda, \bar{\boldsymbol{\gamma}}^\lambda; 1, z_2, \dots, z_s) \right]^{\frac{1}{\lambda}},
\end{aligned}$$

where the last step follows from the property $r(\alpha, \beta, \gamma, h)^\lambda = r(\alpha\lambda, \beta^\lambda, \gamma^\lambda, h)$. This completes the proof. □

REFERENCES

[1] J. Dick, *On the convergence rate of the component-by-component construction of good lattice rules*, J. Complexity, submitted.

[2] G. H. Hardy, J. E. Littlewood, and G. Pólya, *Inequalities*, Cambridge University Press, Cambridge, UK, 1934.

[3] F. J. Hickernell and H. S. Hong, *Quasi-Monte Carlo methods and their randomizations*, in Applied Probability, AMS/IP Stud. Adv. Math. 26, R. Chan, Y.-K. Kwok, D. Yao, and Q Zhang, eds., AMS, Providence, RI, 2002, pp. 59–77.

[4] F. J. Hickernell and H. Niederreiter, *The existence of good extensible rank-*1 *lattices*, J. Complexity, 19 (2003), pp. 286–300.

[5] F. J. Hickernell and H. Woźniakowski, *Integration and approximation in arbitrary dimensions*, Adv. Comput. Math., 12 (2000), pp. 25–58.

[6] F. J. Hickernell and H. Woźniakowski, *Tractability of multivariate integration for periodic functions*, J. Complexity, 17 (2001), pp. 660–682.

[7] S. Joe and S. A. R. Disney, *Intermediate rank lattice rules for multidimensional integration*, SIAM J. Numer. Anal., 30 (1993), pp. 569–582.

[8] S. Joe and I. H. Sloan, *Imbedded lattice rules for multidimensional integration*, SIAM J. Numer. Anal., 29 (1992), pp. 1119–1135.

[9] F. Y. Kuo, *Component-by-component constructions achieve the optimal rate of convergence for multivariate integration in weighted Korobov and Sobolev spaces*, J. Complexity, 19 (2003), pp. 301–320.

[10] F. Y. Kuo and S. Joe, *Component-by-component construction of good QMC rules with a composite number of quadrature points*, J. Complexity, 18 (2002), pp. 943–976.

[11] I. H. Sloan and S. Joe, *Lattice Rules for Multiple Integration*, Clarendon Press, Oxford, UK, 1994.

[12] I. H. Sloan, F. Y. Kuo, and S. Joe, *On the step-by-step construction of quasi-Monte Carlo integration rules that achieve strong tractability error bounds in weighted Sobolev spaces*, Math. Comp., 71 (2002), pp. 1609–1640.

[13] I. H. Sloan, F. Y. Kuo, and S. Joe, *Constructing randomly shifted lattice rules in weighted Sobolev spaces*, SIAM J. Numer. Anal., 40 (2002), pp. 1650–1665.

[14] I. H. Sloan and J. N. Lyness, *The representation of lattice quadrature rules as multiple sums*, Math. Comp., 52 (1989), pp. 81–94.

[15] I. H. Sloan and A. V. Reztsov, *Component-by-component construction of good lattice rules*, Math. Comp., 71 (2002), pp. 263–273.

[16] I. H. Sloan and H. Woźniakowski, *Tractability of multivariate integration for weighted Korobov classes*, J. Complexity, 17 (2001), pp. 697–721.

SIAM J. Numer. Anal.
Vol. 41, No. 4, pp. 1487–1504

NUMERICAL SOLUTION OF A THERMOVISCOELASTIC CONTACT PROBLEM BY A PENALTY METHOD*

M. I. M. COPETTI[†] AND D. A. FRENCH[‡]

Abstract. We consider the numerical approximation of a one-dimensional quasi-static contact problem in linear thermoviscoelasticity. A finite element approximation based on a penalized problem is proposed and analyzed. We furnish an a priori estimate of the difference between the true and numerical solutions. The results of some computations are also presented.

Key words. thermoviscoelasticity, contact problem, penalty method

AMS subject classifications. 65N30, 65N15

DOI. 10.1137/S0036142902403668

1. Introduction.

Consider the quasi-static Signorini contact problem in linear thermoviscoelasticity:

$$\theta_t - \theta_{xx} = -au_{xt}, \qquad 0 < x < 1, \quad t > 0, \tag{1.1}$$

$$\sigma_x = 0, \qquad 0 < x < 1, \quad t > 0, \tag{1.2}$$

with initial conditions

$$\theta(x,0) = \theta_0(x), \quad u(x,0) = u_0(x), \qquad 0 < x < 1, \tag{1.3}$$

and boundary conditions

$$u(0,t) = 0, \quad \theta(0,t) = 0, \quad -\theta_x(1,t) = k(\theta(1,t) - \theta_A), \qquad t > 0, \tag{1.4}$$

$$\sigma(1,t) \le 0, \quad u(1,t) \le g, \quad \sigma(1,t)(u(1,t) - g) = 0, \qquad t > 0, \tag{1.5}$$

with $\sigma = u_x + \zeta u_{xt} - a\theta$. This initial-boundary value problem models the deformations, due to thermal effects, of a homogeneous viscoelastic rod which moves along the x-axis. The temperature of the rod and the displacement from its reference configuration, assumed to be the interval $I = [0,1]$, are denoted by $\theta(x,t)$ and $u(x,t)$, respectively. At its left end the rod is fixed and has zero temperature, while at the right end it is free to expand up to a rigid obstacle, at temperature θ_A, located at distance $g > 0$ from the rest position. When there is no contact with the obstacle, the stress σ vanishes. Here the viscosity is represented by ζu_{xt}, $\zeta > 0$, and describes materials with short memory effect. The constant a is a small positive constant which arises from the physical properties of the rod; we assume $0 < a < 1$ (see [1] for comments on the size of a and [7] for a full nondimensionalization that provides a situation where $a << 1$).

*Received by the editors March 7, 2002; accepted for publication (in revised form) February 18, 2003; published electronically September 9, 2003.
http://www.siam.org/journals/sinum/41-4/40366.html

[†] Laboratório de Análise Numérica e Astrofísica, Departamento de Matemática, Universidade Federal de Santa Maria, 97119-900 Santa Maria, RS, Brazil (mimc@lana.ccne.ufsm.br). This author was a visitor at the Department of Mathematical Sciences, University of Cincinnati, during the course of this work. The research of this author was partially supported by the Brazilian institution CNPq.

[‡] Department of Mathematical Sciences, University of Cincinnati, Cincinnati, OH 45221-0025 (french@math.uc.edu). The research of this author was partially supported by the Taft Foundation at the University of Cincinnati.

The one-dimensional problem with $\zeta = 0$ has received considerable attention from both mathematicians and engineers as a basic model of expansion and contact. The papers [2] and [13] provide examples of early studies of this problem in the engineering literature. Long-time behavior issues are tackled in [12].

In many papers (see [12], for instance), the heat exchange coefficient k is a nonlinear function of the actual distance between the end of the rod and the obstacle when there is no contact and the contact pressure otherwise. In [1], a lengthy proof of solution existence is furnished for the $\zeta = 0$ case with this nonlinear k; this is one of the few papers that addresses this case in a rigorous way. We make the less realistic but more theoretically tractable assumption that $k > 0$ is a constant.

In this paper, we produce error estimates for a finite element approximation based on a penalty formulation. The unilateral constraint $u(1, t) \leq g$ is relaxed by assuming that the obstacle is elastic with rigidity constant $1/\epsilon > 0$. In this situation, the obstacle can be deformed, and the displacement $u(1, t)$ can be greater than g.

We also use the penalty formulation to establish existence and uniqueness results for the continuous problem. Previously, Kuttler and Shillor [11], using tools from convex analysis, proved existence results when the heat transfer coefficient may depend on the gap between the rod and the obstacle.

The contact problem for an elastic rod ($\zeta = 0$) and a rigid obstacle was studied numerically by Copetti [4]. Dirichlet boundary conditions for the temperature at both ends were considered by Copetti and Elliott [6]. In [5], the obstacle was elastic. In these cases, the resulting equations decouple and can be rewritten in terms of the temperature only. This does not appear to be possible for the viscoelastic problem.

We now outline the remainder of this paper. In section 2, we introduce the penalty formulation and show that it has a unique solution. We also identify its regularity properties. In section 3, we prove there exists a unique solution to (1.1)–(1.5) which is a limit of the solution to the penalty problem as the penalty parameter $\epsilon \to 0$. In section 4, we produce an estimate on the size of the difference between the penalty and true solutions. In section 5, we introduce the numerical scheme. Section 6 contains the main result of the paper which is the error estimate for the scheme. Finally, in section 7, we show the results of some numerical simulations using the new method.

We denote the norms of $L^2(I)$ and $H^s(I)$ by $\|\cdot\|$ and $\|\cdot\|_s$, respectively. We will analyze functions in the space $H^1_E(I) = \{\chi \in H^1(I) \mid \chi(0) = 0\}$.

We assume throughout that the initial data is sufficiently regular and compatible. Thus

$$\theta_0 \in H^1_E(I), \quad u_0 \in H^1_E(I), \quad \text{and} \quad u_0(1, \cdot) \leq g. \tag{1.6}$$

Throughout the paper, C denotes positive constants which are allowed to depend on data and norms of u, θ, and σ; these constants are not necessarily the same at each occurrence. Although we do not exploit this fact, note that $\sigma = \sigma(t)$ since $\sigma_x = 0$.

2. Penalized problem—definition and analysis.

In this section, we introduce a penalty formulation for (1.1)–(1.5) which will be used in our existence proof and will be a critical part of the definition of our numerical method.

The following initial-boundary value problem constitutes our penalized problem:

$$\theta^\epsilon_t - \theta^\epsilon_{xx} = -a u^\epsilon_{xt}, \tag{2.1}$$

$$\sigma^\epsilon_x = 0, \tag{2.2}$$

where $\sigma^\epsilon = u^\epsilon_x + \zeta u^\epsilon_{xt} - a\theta^\epsilon$, with initial conditions

$$\theta^\epsilon(x,0) = \theta_0(x), \quad u^\epsilon(x,0) = u_0(x), \tag{2.3}$$

and boundary conditions

$$u^\epsilon(0,t) = 0, \quad \theta^\epsilon(0,t) = 0, \quad -\theta^\epsilon_x(1,t) = k(\theta^\epsilon(1,t) - \theta_A), \tag{2.4}$$

$$\sigma^\epsilon(1,t) = -\frac{1}{\epsilon}[u^\epsilon(1,t) - g]_+. \tag{2.5}$$

The condition (2.5) treats the obstacle as elastic. Note that $\sigma^\epsilon = \sigma^\epsilon(t)$. It is important for us that the functional on the right side is now Lipschitz.

We first prove an existence and uniqueness theorem for the penalized problem, (2.1)–(2.5). This will eventually lead to an existence and uniqueness theorem for (1.1)–(1.5).

THEOREM 2.1. *For any $\epsilon > 0$, there exists a unique $\{\theta^\epsilon,\ u^\epsilon\}$ satisfying* (2.1)–(2.5) *with*

$$\theta^\epsilon \in L^\infty(0,T;H^1_E(I)),\ \theta^\epsilon_t,\ \theta^\epsilon_{xx} \in L^2(0,T;L^2(I)),$$

$$u^\epsilon \in L^\infty(0,T;H^1_E(I)),\ u^\epsilon_t \in L^2(0,T;H^1_E(I)),\ \sigma^\epsilon \in L^2(0,T).$$

Moreover, if $u_0 \in H^2(I)$, then $u^\epsilon_{xx} \in L^\infty(0,T;L^2(I))$.

Remark. Note that the functions u^ϵ, θ^ϵ, and σ^ϵ are bounded in the norms of the above spaces uniformly in ϵ.

Proof. It is helpful to study a modified version of (2.1)–(2.5) in which the dependent variables have zero initial data. Thus consider

$$\tilde{\theta}^\epsilon(x,t) = \theta^\epsilon(x,t) - \theta_0(x) \text{ and } \tilde{u}^\epsilon(x,t) = u^\epsilon(x,t) - u_0(x).$$

Then $\tilde{\theta}^\epsilon,\ \tilde{u}^\epsilon$ satisfy, $\forall\ w,\ v \in H^1_E(I)$,

$$(\tilde{\theta}^\epsilon_t, w) + (\tilde{\theta}^\epsilon_x + \theta_{0x}, w_x) + a(\tilde{u}^\epsilon_{xt}, w) + k(\tilde{\theta}^\epsilon(1,\cdot) + \theta_0(1) - \theta_A)w(1) = 0,$$

$$(\tilde{u}^\epsilon_x + \zeta\tilde{u}^\epsilon_{xt} - a\tilde{\theta}^\epsilon - a\theta_0 + u_{0x}, v_x) + \frac{1}{\epsilon}[\tilde{u}^\epsilon(1,\cdot) + u_0(1) - g]_+ v(1) = 0.$$

Let $\{\phi_i\}_{i=1}^\infty \subset C^\infty(I)$ be an orthogonal basis for $H^1_E(I)$ and orthonormal for $L^2(I)$. Let $V^m = span\{\phi_i\}_{i=1}^m$, and look for

$$\theta^m(x,t) = \sum_{i=1}^m c_i(t)\phi_i(x), \quad u^m(x,t) = \sum_{i=1}^m d_i(t)\phi_i(x),$$

satisfying, $\forall\ w,\ v \in V^m$,

$$(\theta^m_t, w) + (\theta^m_x + \theta_{0x}, w_x) + a(u^m_{xt}, w) + k(\theta^m(1,\cdot) + \theta_0(1) - \theta_A)w(1) = 0,$$

$$(u^m_x + \zeta u^m_{xt} - a\theta^m - a\theta_0 + u_{0x}, v_x) + \frac{1}{\epsilon}[u^m(1,\cdot) + u_0(1) - g]_+ v(1) = 0,$$

with initial conditions $\theta^m(\cdot,0) = u^m(\cdot,0) = 0$. Since the nonlinearity involved is Lipschitz continuous, we know from the theory of ordinary differential equations (see [14]) that there exists a short-time solution (θ^m, u^m). We now proceed to show that this solution exists on a full interval $[0,T]$ for any given T by producing the appropriate a priori estimates.

Taking $w=\theta^m$, $v=u_t^m$ and adding the resulting equations, we find that

$$\frac{1}{2}\frac{d}{dt}\|\theta^m\|^2+\|\theta_x^m\|^2+k(\theta^m(1,\cdot))^2+\frac{1}{2}\frac{d}{dt}\|u_x^m\|^2+\zeta\|u_{xt}^m\|^2+\frac{1}{2\epsilon}\frac{d}{dt}[u^m(1,\cdot)+u_0(1)-g]_+^2$$

$$=k(\theta_A-\theta_0(1))\theta^m(1,\cdot)-(\theta_{0x},\theta_x^m)+(a\theta_0-u_{0x},u_{xt}^m)$$

$$\leq\frac{k}{2}(\theta_A-\theta_0(1))^2+\frac{k}{2}(\theta^m(1,\cdot))^2+\frac{1}{2}\|\theta_{0x}\|^2+\frac{1}{2}\|\theta_x^m\|^2+\frac{1}{2\zeta}\|a\theta_0-u_{0x}\|^2+\frac{\zeta}{2}\|u_{xt}^m\|^2.$$

Thus, integrating the resulting inequality from 0 to T and recalling (1.6), we obtain

$$\|\theta^m(\cdot,T)\|^2+\int_0^T\|\theta_x^m\|^2dt+\int_0^T(\theta^m(1,t))^2dt+\|u_x^m(\cdot,T)\|^2+\int_0^T\|u_{xt}^m\|^2dt$$

$$+\frac{1}{\epsilon}[u^m(1,T)+u_0(1)-g]_+^2\leq C.$$

Choosing $w=\theta_t^m$ yields

$$\|\theta_t^m\|^2+\frac{1}{2}\frac{d}{dt}\|\theta_x^m\|^2+\frac{k}{2}\frac{d}{dt}(\theta^m(1,\cdot)+\theta_0(1)-\theta_A)^2=-a(u_{xt}^m,\theta_t^m)-(\theta_{0x},\theta_{xt}^m)$$

$$\leq\frac{a^2}{2}\|u_{xt}^m\|^2+\frac{1}{2}\|\theta_t^m\|^2-\frac{d}{dt}(\theta_{0x},\theta_x^m),$$

so

$$\int_0^T\|\theta_t^m\|^2dt+\|\theta_x^m(\cdot,T)\|^2+k(\theta^m(1,T)+\theta_0(1)-\theta_A)^2$$

$$\leq a^2\int_0^T\|u_{xt}^m\|^2dt+2\|\theta_{0x}\|\|\theta_x^m(\cdot,T)\|+k(\theta_0(1)-\theta_A)^2.$$

It follows that

$$\int_0^T\|\theta_t^m\|^2dt+\|\theta_x^m(\cdot,T)\|^2+(\theta^m(1,T)+\theta_0(1)-\theta_A)^2\leq C.$$

Therefore,

$$\theta^m,\ u^m \text{ are bounded in } L^\infty(0,T;H_E^1(I)),$$

$$\theta_t^m \text{ is bounded in } L^2(0,T;L^2(I)),$$

$$u_t^m \text{ is bounded in } L^2(0,T;H_E^1(I)),$$

$$\theta^m(1,\cdot),\ [u^m(1,\cdot)+u_0(1)-g]_+ \text{ are bounded in } L^\infty(0,T).$$

We are now in a position to pass to the limit as $m \to \infty$. As a consequence, there exists $\tilde{\theta}^\epsilon,\ \tilde{u}^\epsilon$, and subsequences $\{\theta^m\},\ \{u^m\}$ such that

$$\theta^m \to \tilde{\theta}^\epsilon \text{ in } L^\infty(0,T;H^1_E(I)) \text{ weak-star},$$

$$\theta^m_t \to \tilde{\theta}^\epsilon_t \text{ in } L^2(0,T;L^2(I)) \text{ weakly},$$

$$\theta^m(1,\cdot) \to \tilde{\theta}^\epsilon(1,\cdot) \text{ in } L^\infty(0,T) \text{ weak-star},$$

$$u^m \to \tilde{u}^\epsilon \text{ in } L^\infty(0,T;H^1_E(I)) \text{ weak-star,} \quad \text{and}$$

$$u^m_t \to \tilde{u}^\epsilon_t \text{ in } L^2(0,T;H^1_E(I)) \text{ weakly.}$$

Also, since $u^m(1,\cdot) \to \tilde{u}^\epsilon(1,\cdot)$ in $H^1(0,T)$ weakly and the injection of $H^1(0,T)$ into $L^2(0,T)$ is compact, we have that $u^m(1,\cdot) \to \tilde{u}^\epsilon(1,\cdot)$ in $L^2(0,T)$ strongly.

Because

$$\|[u^m(1,\cdot)+u_0(1)-g]_+ - [\tilde{u}^\epsilon(1,\cdot)+u_0(1)-g]_+\|_{L^2(0,T)} \le \|u^m(1,\cdot)-\tilde{u}^\epsilon(1,\cdot)\|_{L^2(0,T)},$$

we find that

$$[u^m(1,\cdot)+u_0(1)-g]_+ \to [\tilde{u}^\epsilon(1,\cdot)+u_0(1)-g]_+ \text{ in } L^2(0,T) \text{ strongly.}$$

Passing to the limit on m and reversing the change of variables, we find that θ^ϵ, u^ϵ solve, $\forall\ w,\ v \in H^1_E(I)$, the weak form

$$(\theta^\epsilon_t, w) + (\theta^\epsilon_x, w_x) + a(u^\epsilon_{xt}, w) + k(\theta^\epsilon(1,\cdot)-\theta_A)w(1) = 0, \tag{2.6}$$

$$(u^\epsilon_x + \zeta u^\epsilon_{xt} - a\theta^\epsilon, v_x) + \frac{1}{\epsilon}[u^\epsilon(1,\cdot)-g]_+ v(1) = 0. \tag{2.7}$$

Choosing $w,\ v \in C_0^\infty(I)$, it follows that $\theta^\epsilon_{xx} = \theta^\epsilon_t + au^\epsilon_{xt} \in L^2(0,T;L^2(I))$, $(u^\epsilon_x + \zeta u^\epsilon_{xt})_x = a\theta^\epsilon_x \in L^\infty(0,T;L^2(I))$, and (2.1)–(2.2) hold a.e. in Ω_T. It is now straightforward to deduce (2.3)–(2.5) also. In order to obtain H^2 regularity for $u^\epsilon(\cdot,t)$, we observe that, from the defining equations,

$$u^\epsilon_x + \zeta u^\epsilon_{xt} = a\theta^\epsilon + \sigma^\epsilon.$$

If we multiply by the appropriate integrating factor, we can rewrite this as

$$\frac{d}{dt}\left(e^{t/\zeta}u^\epsilon_x(\cdot,t)\right) = \frac{1}{\zeta}(a\theta^\epsilon(\cdot,t)+\sigma^\epsilon(t))e^{t/\zeta}.$$

So, we can integrate and obtain the following representation for u^ϵ_x:

$$u^\epsilon_x(\cdot,t) = e^{-t/\zeta}\left(u^\epsilon_x(\cdot,0) + \frac{1}{\zeta}\int_0^t (a\theta^\epsilon(\cdot,s)+\sigma^\epsilon(s))e^{s/\zeta}ds\right).$$

Differentiating with respect to x, we find that $u^\epsilon_{xx}(\cdot,t) \in L^2(I)$.

It remains to prove uniqueness. We follow the argument suggested in [1]. Let $\{\theta^\epsilon_1, u^\epsilon_1\}$, $\{\theta^\epsilon_2, u^\epsilon_2\}$ be two solutions, and define $\psi = \int_0^t(\theta^\epsilon_1 - \theta^\epsilon_2)ds$, $\eta = u^\epsilon_1 - u^\epsilon_2$. Then we can show from (2.1)–(2.5) that

$$(\psi_t, v) + (\psi_x, v_x) + k\psi(1,\cdot)v(1) + a(\eta_x, v) = 0,$$

$$(\sigma_1(\cdot) - \sigma_2(\cdot), w_x) = -\frac{1}{\epsilon}\left([u_1^\epsilon(1,\cdot) - g]_+ - [u_2^\epsilon(1,\cdot) - g]_+\right) w(1).$$

Now we let $v = \psi_t$, $w = \eta$ and add the resulting equations to obtain the following estimate:

$$\|\psi_t\|^2 + \|\eta_x\|^2 + \frac{1}{2}\frac{d}{dt}(\|\psi_x\|^2 + k\psi(1,\cdot)^2 + \zeta\|\eta_x\|^2) = -\frac{1}{\epsilon}([u_1^\epsilon(1,\cdot) - g]_+$$

$$-[u_2^\epsilon(1,\cdot) - g]_+)(u_1^\epsilon(1,\cdot) - u_2^\epsilon(1,\cdot)) \leq 0,$$

where the last inequality followed from the monotonicity of the $[\cdot]_+$ functional. We can now conclude $\psi_t = \eta = 0$, which finishes the proof. □

3. Existence and uniqueness. In this section, we obtain a solution to (1.1)–(1.5) as the limit of solutions to the penalized problem.

THEOREM 3.1. *There exists a unique* $\{\theta,\ u\}$ *satisfying* (1.1)–(1.5) *with*

$$\theta \in L^\infty(0,T; H_E^1(I)),\ \theta_t,\ \theta_{xx} \in L^2(0,T;L^2(I)),$$

$$u \in L^\infty(0,T;H_E^1(I)),\ u_t \in L^2(0,T;H_E^1(I)),\ \sigma \in L^2(0,T).$$

If in addition $u_0 \in H^2(I)$, *then* $u_{xx} \in L^\infty(0,T;L^2(I))$.

Proof. Setting $w = \theta^\epsilon$ in (2.6) and $v = u_t^\epsilon$ in (2.7), we obtain

$$\|\theta^\epsilon(\cdot,T)\|^2 + \int_0^T \|\theta_x^\epsilon\|^2 dt + \int_0^T (\theta^\epsilon(1,t))^2 dt + \|u_x^\epsilon(\cdot,T)\|^2 + \int_0^T \|u_{xt}^\epsilon\|^2 dt$$

$$+\frac{1}{\epsilon}[u^\epsilon(1,T) - g]_+^2 \leq C.$$

Let $\Delta t > 0$, and define $w(x,t) = (\theta^\epsilon(x,t+\Delta t) - \theta^\epsilon(x,t))/\Delta t$ for $0 \leq t \leq T - \Delta t$ and $w(x,t) = (\theta^\epsilon(x,T) - \theta^\epsilon(x,t))/\Delta t$ for $T - \Delta t \leq t \leq T$. With this choice of w in (2.6), integrating from 0 to T and letting Δt tend to zero, we get

$$\int_0^T \|\theta_t^\epsilon\|^2 dt + \|\theta_x^\epsilon(\cdot,T)\|^2 + (\theta^\epsilon(1,T))^2 \leq C.$$

Thus there exist functions θ and u which are limits of subsequences of $\{\theta^\epsilon\}$ and $\{u^\epsilon\}$ such that, as $\epsilon \to 0$,

$$\theta^\epsilon \to \theta \text{ in } L^\infty(0,T;H_E^1(I)) \text{ weak-star},$$

$$\theta_t^\epsilon \to \theta_t \text{ in } L^2(0,T;L^2(I)) \text{ weakly},$$

$$\theta^\epsilon(1,\cdot) \to \theta(1,\cdot) \text{ in } L^\infty(0,T) \text{ weak-star},$$

$$u^\epsilon \to u \text{ in } L^\infty(0,T;H_E^1(I)) \text{ weak-star}, \quad \text{and}$$

$$u_t^\epsilon \to u_t \text{ in } L^2(0,T;H_E^1(I)) \text{ weakly}.$$

In addition, we have that

$$u^\epsilon \to u \text{ in } L^2(0,T;L^2(I)) \text{ strongly,}$$

$$[u^\epsilon(1,\cdot)-g]_+ \to [u(1,\cdot)-g]_+ \text{ in } L^2(0,T) \text{ strongly, and}$$

$$\theta^\epsilon \to \theta \text{ in } L^2(0,T;L^2(I)) \text{ strongly.}$$

From our first estimate in this proof, we have $[u^\epsilon(1,\cdot)-g]_+^2 \leq C\epsilon$. Thus we conclude that

$$[u^\epsilon(1,\cdot)-g]_+ \to 0 \text{ in } L^2(0,T) \text{ strongly.}$$

As a consequence, $u(1,t)-g \leq 0$ for almost all $t \in (0,T)$. Observing that, for $w \in H_E^1(I)$ which satisfies $w(1) \leq g$, we have

$$-\frac{1}{\epsilon}[u^\epsilon(1,\cdot)-g]_+(w(1)-u^\epsilon(1,\cdot)) = -\frac{1}{\epsilon}[u^\epsilon(1,\cdot)-g]_+(w(1)-g)+\frac{1}{\epsilon}[u^\epsilon(1,\cdot)-g]_+^2 \geq 0,$$

and letting $v = w - u^\epsilon$ in (2.7), we conclude that, $\forall w \in L^2(0,T;H_E^1(I))$ with $w(1,\cdot) \leq g$,

$$\int_0^T (u_x^\epsilon + \zeta u_{xt}^\epsilon - a\theta^\epsilon, w_x - u_x^\epsilon)dt \geq -\int_0^T \frac{1}{\epsilon}[u^\epsilon(1,\cdot)-g]_+(w(1)-u^\epsilon(1,\cdot))dt \geq 0.$$

From this we obtain that

$$\int_0^T (u_x^\epsilon + \zeta u_{xt}^\epsilon, w_x)dt - \int_0^T (a\theta^\epsilon, w_x - u_x^\epsilon)dt \geq \int_0^T \|u_x^\epsilon\|^2 dt$$

$$+\frac{\zeta}{2}\|u_x^\epsilon(\cdot,T)\|^2 - \frac{\zeta}{2}\|u_{0x}\|^2. \tag{3.1}$$

The weak convergence of $u_x^\epsilon(\cdot,t)$ to $u_x(\cdot,t)$ in $L^2(I)$ implies that

$$(u_x^\epsilon(\cdot,t), u_x(\cdot,t)) \to \|u_x(\cdot,t)\|^2,$$

and the Cauchy–Schwarz inequality yields

$$\liminf_{\epsilon\to 0} \|u_x^\epsilon\| \geq \|u_x\|.$$

This allows us to now let $\epsilon \to 0$ on the right side of (3.1) and retain the inequality. We can also let $\epsilon \to 0$ on the first integral on the left side of (3.1) due to the weak convergence properties of u_x^ϵ and u_{xt}^ϵ. We can take the limit on the second integral on the left side due to the strong convergence of θ^ϵ. We can now conclude that θ, u satisfy, $\forall\ w \in H_E^1(I)$ and $\forall\ v \in L^2(0,T;H_E^1(I))$ with $v(1,t) \leq g$,

$$(\theta_t, w) + (\theta_x, w_x) + a(u_{xt}, w) + k(\theta(1,\cdot)-\theta_A)w(1) = 0,$$

$$\int_0^T (u_x + \zeta u_{xt} - a\theta, v_x - u_x)dt \geq 0.$$

Following standard techniques (see [9], [10]), we can now show that θ and u satisfy the claimed regularity properties and (1.1)–(1.5). Uniqueness follows similarly as it did for the penalized problem. □

4. Penalization error. In this section, we provide an estimate of the difference between the true and penalty solutions in terms of the parameter ϵ.

THEOREM 4.1. *There exists a constant C such that for any $T > 0$ the following holds:*

$$\left\| \int_0^T (\theta_x - \theta_x^\epsilon) dt \right\|^2 + \|(u - u^\epsilon)(\cdot, T)\|^2 \leq C\epsilon \|\sigma\|^2_{L^2(0,T)}.$$

Proof. As we did in the uniqueness argument, let $\psi = \int_0^t (\theta - \theta^\epsilon) ds$ and $\eta = u - u^\epsilon$. Then

$$\|\psi_t\|^2 + \|\eta_x\|^2 + \frac{1}{2}\frac{d}{dt}(\|\psi_x\|^2 + k\psi(1,\cdot)^2 + \zeta\|\eta_x\|^2) = (\sigma(\cdot) - \sigma^\epsilon(\cdot))(u(1,\cdot) - u^\epsilon(1,\cdot)).$$

Setting $I = (\sigma(\cdot) - \sigma^\epsilon(\cdot))(u(1,\cdot) - u^\epsilon(1,\cdot))$, we have the following cases:

1. If $u(1,\cdot) - g < 0$ and $u^\epsilon(1,\cdot) - g < 0$, then $\sigma(\cdot) = \sigma^\epsilon(\cdot) = 0$ and $I = 0$.
2. If $u(1,\cdot) - g < 0$ and $u^\epsilon(1,\cdot) - g \geq 0$, then $\sigma(\cdot) = 0$, $\sigma^\epsilon(\cdot) < 0$, and
$$I = -\sigma^\epsilon(\cdot)(u(1,\cdot) - g + g - u^\epsilon(1,\cdot)) \leq 0.$$
3. If $u(1,\cdot) - g = 0$ and $u^\epsilon(1,\cdot) - g < 0$, then $\sigma(\cdot) \leq 0$, $\sigma^\epsilon(\cdot) = 0$, and
$$I = \sigma(\cdot)(g - u^\epsilon(1,\cdot)) \leq 0.$$
4. If $u(1,\cdot) - g = 0$ and $u^\epsilon(1,\cdot) - g \geq 0$, then $\sigma^\epsilon(\cdot) = -(u^\epsilon(1,\cdot) - g)/\epsilon$, so

$$I = \left(\sigma(\cdot) + \frac{1}{\epsilon}(u^\epsilon(1,\cdot) - g)\right)\eta(1,\cdot) = \left(\sigma(\cdot) - \frac{1}{\epsilon}\eta(1,\cdot)\right)\eta(1,\cdot) \leq \frac{\epsilon}{2}\sigma(\cdot)^2 - \frac{1}{2\epsilon}\eta(1,\cdot)^2.$$

Hence, combining these four cases and integrating in t, we have

$$\int_0^T \|\psi_t\|^2 dt + \int_0^T \|\eta_x\|^2 dt + \frac{1}{2}(\|\psi_x(\cdot,T)\|^2 + k\psi(1,T)^2 + \zeta\|\eta_x(\cdot,T)\|^2) + \frac{1}{2\epsilon}\int_0^T \eta(1,t)^2 dt$$

$$\leq \frac{\epsilon}{2}\|\sigma\|^2_{L^2(0,T)}.$$

It follows from the Poincaré inequality that

$$\|\psi_x(\cdot,T)\|^2 + \|\eta(\cdot,T)\|^2 \leq C\epsilon\|\sigma\|^2_{L^2(0,T)},$$

which proves the theorem. □

5. Numerical approximation. We now introduce our finite element numerical method with the backward Euler scheme for the time discretization. We briefly describe the implementation of this method in this section.

Partition the interval $(0,1)$ into subintervals $I_j = (x_{j-1}, x_j)$ of length $h = 1/J$, with $0 = x_0 < x_1 < \cdots < x_J = 1$, and denote by $S_E^h \subset H_E^1(I)$ the space of continuous piecewise linear functions defined on this partition. Our finite element method for

(2.1)–(2.5) on each time step is to find Θ^n, $U^n \in S_E^h$, $n = 1,\dots,N$, such that, $\forall\, W,\ V \in S_E^h$,

$$\frac{1}{\Delta t}(\Theta^n - \Theta^{n-1}, W) + (\Theta^n_x, W_x) + k(\Theta^n(1) - \theta_A)W(1)$$

$$+\frac{a}{\Delta t}(U^n_x - U^{n-1}_x, W) = 0, \tag{5.1}$$

$$(U^n_x - a\Theta^n, V_x) + \frac{\zeta}{\Delta t}(U^n_x - U^{n-1}_x, V_x) + \beta_\epsilon(\Gamma^n)V(1) = 0, \tag{5.2}$$

where $\Theta^0 \in S_E^h$ and $U^0 \in S_E^h$ are given approximations of θ_0 and u_0, respectively, $\beta_\epsilon(\chi) = \frac{1}{\epsilon}[\chi]_+$, $\Gamma^n = U^n(1) - g$, and $\Delta t = T/N$. Assuming that Θ^{n-1} and U^{n-1} are known, we need to iterate to find Θ^n and U^n:

$$\frac{1}{\Delta t}(\Theta^{n,l} - \Theta^{n-1}, W) + (\Theta^{n,l}_x, W_x) + k(\Theta^{n,l}(1) - \theta_A)W(1)$$

$$+\frac{a}{\Delta t}(U^{n,l-1}_x - U^{n-1}_x, W) = 0, \tag{5.3}$$

$$(U^{n,l}_x - a\Theta^{n,l}, V_x) + \frac{\zeta}{\Delta t}(U^{n,l}_x - U^{n-1}_x, V_x) + \beta_\epsilon(\Gamma^{n,l-1})V(1) = 0, \tag{5.4}$$

where $\Theta^{n,0} = \Theta^{n-1}$ and $U^{n,0} = U^{n-1}$. The method defined requires that the systems of algebraic equations

$$(M + \Delta tK + \Delta tkB)\underline{c}^{n,l} = M\underline{c}^{n-1} + \Delta tk\theta_A\underline{e} + aC(\underline{d}^{n-1} - \underline{d}^{n,l-1}),$$

$$(\Delta t + \zeta)K\underline{d}^{n,l} = \zeta K\underline{d}^{n-1} + \Delta taC^T\underline{c}^{n,l} - \frac{\Delta t}{\epsilon}[d_J^{n,l-1} - g]_+\underline{e}$$

be solved at each iteration. We used the representations

$$\Theta^n = \sum_{i=1}^J c_i^n \eta_i, \qquad U^n = \sum_{i=1}^J d_i^n \eta_i,$$

with $\{\eta_i\}_{i=1}^J$ the usual basis for S_E^h, and

$$M_{ij} = (\eta_i, \eta_j),\ K_{ij} = (\eta_{ix}, \eta_{jx}),\ B_{ij} = \eta_i(1)\eta_j(1), C_{ij} = (\eta_i, \eta_{jx}),\ \{\underline{e}\}_i = \eta_i(1).$$

We now show that if Δt is sufficiently small to a reasonable degree, then the iterations are guaranteed to converge for the system (5.1)–(5.2).

THEOREM 5.1. *Suppose that $a < 1$ and $\Delta t < \epsilon\zeta$. Then there exists a unique solution $\{\Theta^n,\ U^n\}$ for problem* (5.1)–(5.2).

Proof. We will show that the iteration converges for the scheme (5.3)–(5.4). Let $e^j = \Theta^{n,j} - \Theta^{n,j-1}$ and $q^j = U^{n,j} - U^{n,j-1}$. Thus e^j and q^j satisfy

$$\|e^j\|^2 + \Delta t\|e^j_x\|^2 + \Delta tk(e^j(1))^2 = -a(q^{j-1}_x, e^j),$$

$$\left(1 + \frac{\zeta}{\Delta t}\right)\|q^j_x\|^2 + (\beta_\epsilon(\Gamma^{n,j-1}) - \beta_\epsilon(\Gamma^{n,j-2}))q^j(1) = a(e^j, q^j_x).$$

Adding the equations, using the Lipschitz continuity of β_ϵ and the Cauchy–Schwarz inequality, yields, for $\delta > 0$ and $\alpha > 0$,

$$\|e^j\|^2 + \Delta t\|e^j_x\|^2 + \Delta tk(e^j(1))^2 + \left(1 + \frac{\zeta}{\Delta t}\right)\|q^j_x\|^2 \leq \frac{\delta}{2}\|q^j_x - q^{j-1}_x\|^2$$

$$+\frac{1}{2\delta}\|e^j\|^2 + \frac{\alpha}{2\epsilon}(q^{j-1}(1))^2 + \frac{1}{2\alpha\epsilon}(q^j(1))^2.$$

Using the fact that $(q^j(1))^2 \leq \|q^j_x\|^2$, we find

$$\left(1 - \frac{1}{2\delta}\right)\|e^j\|^2 + \Delta t\|e^j_x\|^2 + \Delta tk(e^j(1))^2 + \left(1 + \frac{\zeta}{\Delta t} - \delta - \frac{1}{2\alpha\epsilon}\right)\|q^j_x\|^2$$

$$\leq \left(\delta + \frac{\alpha}{2\epsilon}\right)\|q^{j-1}_x\|^2,$$

and taking $\delta = 1/2$ and $\alpha = 1$, we now have

$$\left(\frac{1}{2} + \frac{\zeta}{\Delta t} - \frac{1}{2\epsilon}\right)\|q^j_x\|^2 \leq \left(\frac{1}{2} + \frac{1}{2\epsilon}\right)\|q^{j-1}_x\|^2.$$

Note that our assumption $\Delta t < \epsilon\zeta$ implies $1/\epsilon < \zeta/\Delta t$, so we know the factor on the left side of the inequality above is positive. Moreover,

$$\frac{1}{2} + \frac{\zeta}{\Delta t} - \frac{1}{2\epsilon} > \frac{1}{2} + \frac{1}{\epsilon} - \frac{1}{2\epsilon} = \frac{1}{2} + \frac{1}{2\epsilon},$$

and thus there exists M, with $0 < M < 1$, such that

$$\|q^j_x\|^2 \leq M\|q^{j-1}_x\|^2.$$

One can now show using standard contraction arguments that the sequences $\{\Theta^{n,j}\}$, $\{U^{n,j}\}$ converge to Θ^n, $U^n \in S^h_E$ and that these limits solve the Galerkin approximation (5.1)–(5.2). A similar argument yields the uniqueness of $\{\Theta^n,\ U^n\}$. □

6. Error bound. In this section, we obtain an error bound for the numerical approximation of the contact problem. Set $U^0 = P^h_E u_0$, where $P^h_E : H^1_E(I) \to S^h_E$ is defined by $((\eta - P^h_E\eta)_x, \chi_x) = 0\ \forall\ \chi \in S^h_E$ and satisfies (see [8])

$$\|\eta - P^h_E\eta\| \leq h\|\eta_x\|, \qquad P^h_E\eta(x_i) = \eta(x_i)$$

for $i = 0, 1, \ldots, J$. Also, if η is sufficiently smooth, it follows that

$$\|P^h_E\eta - \eta\| + h\|(P^h_E\eta - \eta)_x\| \leq Ch^2\|\eta\|_2.$$

The error analysis uses a triangle inequality estimate involving $(\theta, u), (\theta^\epsilon, u^\epsilon)$, and (Θ, U). Attempts to obtain an error estimate directly are difficult because of the need to differentiate the monotone graph

$$\beta(v) = \begin{cases} [0,\infty), & v = 0, \\ \emptyset, & v > 0, \\ 0, & v < 0. \end{cases}$$

Approximating this quantity by the penalty formulation allows the discretization analysis to go through more easily since $\beta(v)$ is replaced by the Lipschitz function

$\beta_\epsilon(v) = \epsilon^{-1}[v]_+$. The final estimate, of course, is then subject to the unpleasant ϵ^{-1} factors.

We now state and prove the main theorem of this paper.

THEOREM 6.1. *There exists a constant C such that*

$$\left\| \int_0^{t_n} \theta_x dt - \Delta t \sum_{i=1}^n \Theta_x^i \right\|^2 \leq C \left(\|\Theta^0 - \theta_0\|^2 + \epsilon + h^2 + \frac{h^2}{\Delta t} + h^4 + (\Delta t)^2 + \frac{(\Delta t)^2}{\epsilon^2} \right),$$

$$\|u(\cdot, t_n) - U^n\|^2 \leq C \left(\|\Theta^0 - \theta_0\|^2 + \epsilon + h^2 + \frac{h^2}{\Delta t} + h^4 + (\Delta t)^2 + \frac{(\Delta t)^2}{\epsilon^2} \right),$$

where $t_n = n\Delta t$.

Remark. The $\epsilon, h^2/\Delta t$, and $(\Delta t)^2/\epsilon^2$ terms will surely be the least accurate on the right side of the estimate. (We expect $\|\Theta^0 - \theta_0\| = O(h^2)$.) Balancing these three terms, we find that $\epsilon = h^{4/5}$ and $\Delta t = h^{6/5}$ (which satisfies the condition in Theorem 5.1 for h small enough). Thus the rates of convergence for this theorem are essentially

$$\left\| \int_0^{t_n} \theta_x dt - \Delta t \sum_{i=1}^n \Theta_x^i \right\|^2 \leq C h^{4/5} \quad \text{and} \quad \|u(\cdot, t_n) - U^n\|^2 \leq C h^{4/5}.$$

Proof. First we estimate the error due to the discretization of the penalized problem. Let $\theta^n = \theta^\epsilon(\cdot, t_n)$ and $u^n = u^\epsilon(\cdot, t_n)$. Integrating (2.6) from 0 to t_n and (2.7) from t_{n-1} to t_n, we obtain, $\forall\ w,\ v \in H_E^1(I)$,

$$(\theta^n - \theta_0, w) + (\hat{\theta}_x^n, w_x) + k(\hat{\theta}^n(1) - t_n\theta_A)w(1) + a(u_x^n - u_{0x}, w) = 0,$$

$$(\overline{u}_x^n - a\overline{\theta}^n, v_x) + \frac{\zeta}{\Delta t}(u_x^n - u_x^{n-1}, v_x) + \overline{\beta}_\epsilon^n(\gamma_\epsilon)v(1) = 0,$$

with

$$\hat{\theta}^n = \int_0^{t_n} \theta^\epsilon(\cdot, t)dt, \quad \overline{u}^n = \frac{1}{\Delta t}\int_{t_{n-1}}^{t_n} u^\epsilon(\cdot, t)dt, \quad \overline{\theta}^n = \frac{1}{\Delta t}\int_{t_{n-1}}^{t_n} \theta^\epsilon(\cdot, t)dt,$$

$$\overline{\beta}_\epsilon^n(\gamma_\epsilon) = \frac{1}{\Delta t}\int_{t_{n-1}}^{t_n} \beta_\epsilon(\gamma_\epsilon(t))dt, \quad \text{and} \quad \gamma_\epsilon(t) = u^\epsilon(1, t) - g.$$

Summing (5.1) from 1 to n gives, $\forall\ W \in S_E^h$,

$$(\Theta^n - \Theta^0, W) + \left(\Delta t \sum_{i=1}^n \Theta_x^i, W_x\right) + k\left(\Delta t \sum_{i=1}^n \Theta^i(1) - t_n\theta_A\right)W(1)$$
$$+ a(U_x^n - U_x^0, W) = 0.$$

Thus, $\forall\ W,\ V \in S_E^h$,

$$(\Theta^n - \theta^n, W) + \left(\Delta t \sum_{i=1}^n \Theta_x^i - \hat{\theta}_x^n, W_x\right) + k\left(\Delta t \sum_{i=1}^n \Theta^i(1) - \hat{\theta}^n(1)\right)W(1)$$

$$+ a(U_x^n - u_x^n, W) = (\Theta^0 - \theta_0, W) + a(U_x^0 - u_{0x}, W),$$

and

$$(U_x^n - \overline{u}_x^n, V_x) + \frac{\zeta}{\Delta t}((U_x^n - u_x^n) - (U_x^{n-1} - u_x^{n-1}), V_x) - a(\Theta^n - \overline{\theta}^n, V_x)$$
$$+ (\beta_\epsilon(\Gamma^n) - \overline{\beta}_\epsilon^n(\gamma_\epsilon))V(1) = 0.$$

Defining $\varepsilon^j = \Delta t \sum_{i=1}^j \Theta^i - P_E^h \hat{\theta}^j$, $j = 1, \ldots, n$, $\varepsilon^0 = 0$, $e^n = U^n - P_E^h u^n$, and recalling the properties of P_E^h, we find that, $\forall\ W,\ V \in S_E^h$,

$$\frac{1}{\Delta t}(\varepsilon^n - \varepsilon^{n-1}, W) + (\varepsilon_x^n, W_x) + k\varepsilon^n(1)W(1) = (\Theta^0 - \theta_0, W)$$
$$+ a(U_x^0 - u_{0x}, W) - a(U_x^n - u_x^n, W) + (\theta^n - P_E^h\overline{\theta}^n, W),$$
$$(e_x^n, V_x) + \frac{\zeta}{\Delta t}(e_x^n - e_x^{n-1}, V_x) + (\beta_\epsilon(\Gamma^n) - \beta_\epsilon(\gamma_\epsilon^n))V(1) = \frac{a}{\Delta t}(\varepsilon^n - \varepsilon^{n-1}, V_x)$$
$$+ a(P_E^h\overline{\theta}^n - \overline{\theta}^n, V_x) + (\overline{u}_x^n - u_x^n, V_x) + (\overline{\beta}_\epsilon^n(\gamma_\epsilon) - \beta_\epsilon(\gamma_\epsilon^n))V(1),$$

where $\gamma_\epsilon^n = \gamma_\epsilon(t_n)$. Note that

$$\beta_\epsilon(\Gamma^n) - \beta_\epsilon(\gamma_\epsilon^n) = \beta_\epsilon'(\xi)(U^n(1) - u^n(1)) = \beta_\epsilon'(\xi)(U^n(1) - P_E^h u^n(1)) = \beta_\epsilon'(\xi)e^n(1),$$

where ξ is between Γ^n and γ_ϵ^n. Thus

$$(\beta_\epsilon(\Gamma^n) - \beta_\epsilon(\gamma_\epsilon^n))e^n(1) = \beta_\epsilon'(\xi)(e^n(1))^2 \geq 0,$$
$$a(U_x^n - u_x^n, W) = a(U^n(1) - u^n(1))W(1) - a(U^n - u^n, W_x),$$

and

$$\frac{a}{\Delta t}(\varepsilon^n - \varepsilon^{n-1}, V_x) = \frac{a}{\Delta t}(\varepsilon^n(1) - \varepsilon^{n-1}(1))V(1) - \frac{a}{\Delta t}(\varepsilon_x^n - \varepsilon_x^{n-1}, V).$$

Choosing $W = (\varepsilon^n - \varepsilon^{n-1})/\Delta t$, $V = e^n$ and adding the resulting equations, we obtain

$$\frac{1}{(\Delta t)^2}\|\varepsilon^n - \varepsilon^{n-1}\|^2 + \frac{1}{2\Delta t}(\|\varepsilon_x^n - \varepsilon_x^{n-1}\|^2 + \|\varepsilon_x^n\|^2 - \|\varepsilon_x^{n-1}\|^2)$$
$$+\frac{k}{2\Delta t}\left((\varepsilon^n(1) - \varepsilon^{n-1}(1))^2 + (\varepsilon^n(1))^2 - (\varepsilon^{n-1}(1))^2\right) + \|e_x^n\|^2$$
$$+\frac{\zeta}{2\Delta t}(\|e_x^n - e_x^{n-1}\|^2 + \|e_x^n\|^2 - \|e_x^{n-1}\|^2) \leq \frac{1}{\Delta t}(\Theta^0 - \theta_0, \varepsilon^n - \varepsilon^{n-1})$$
$$-\frac{a}{\Delta t}(U^0 - u_0, \varepsilon_x^n - \varepsilon_x^{n-1}) + \frac{a}{\Delta t}(P_E^h u^n - u^n, \varepsilon_x^n - \varepsilon_x^{n-1})$$
$$+\frac{1}{\Delta t}(\theta^n - P_E^h\overline{\theta}^n, \varepsilon^n - \varepsilon^{n-1}) + a(P_E^h\overline{\theta}^n - \overline{\theta}^n, e_x^n) + (\overline{u}_x^n - u_x^n, e_x^n) + (\overline{\beta}_\epsilon^n(\gamma_\epsilon) - \beta_\epsilon(\gamma_\epsilon^n))e^n(1)$$
$$\leq \|\Theta^0 - \theta_0\|^2 + \frac{2a^2}{\Delta t}\|U^0 - u_0\|^2 + \frac{1}{2(\Delta t)^2}\|\varepsilon^n - \varepsilon^{n-1}\|^2$$
$$+\frac{1}{4\Delta t}\|\varepsilon_x^n - \varepsilon_x^{n-1}\|^2 + \frac{2a^2}{\Delta t}\|P_E^h u^n - u^n\|^2 + 2a^2\|\overline{\theta}^n - P_E^h\overline{\theta}^n\|^2$$
$$+\|\theta^n - P_E^h\overline{\theta}^n\|^2 + 2\|u_x^n - \overline{u}_x^n\| + \frac{1}{4}\|e_x^n\|^2 + (\overline{\beta}_\epsilon^n(\gamma_\epsilon) - \beta_\epsilon(\gamma_\epsilon^n))^2 + \frac{1}{4}(e^n(1))^2.$$

Since $(e^n(1))^2 \le \|e_x^n\|^2$, we have

$$\begin{aligned}&\frac{1}{2(\Delta t)^2}\|\varepsilon^n-\varepsilon^{n-1}\|^2+\frac{1}{4\Delta t}\|\varepsilon_x^n-\varepsilon_x^{n-1}\|^2+\frac{1}{2\Delta t}(\|\varepsilon_x^n\|^2-\|\varepsilon_x^{n-1}\|^2)\\ &\quad+\frac{k}{2\Delta t}((\varepsilon^n(1)-\varepsilon^{n-1}(1))^2+(\varepsilon^n(1))^2-(\varepsilon^{n-1}(1))^2)+\frac{1}{2}\|e_x^n\|^2\\ &\quad+\frac{\zeta}{2\Delta t}(\|e_x^n-e_x^{n-1}\|^2+\|e_x^n\|^2-\|e_x^{n-1}\|^2)\le\|\Theta^0-\theta_0\|^2\\ &\quad+\frac{2a^2}{\Delta t}\|U^0-u_0\|^2+I_1+I_2+I_3+I_4+I_5,\end{aligned}\tag{6.1}$$

where we identify and estimate I_1-I_5 below.

$$I_1=\frac{2a^2}{\Delta t}\|P_E^h u^n-u^n\|^2\le C\frac{h^2}{\Delta t}\|u_x^\epsilon(\cdot,t_n)\|^2,$$

$$I_2=2a^2\|\overline{\theta}^n-P_E^h\overline{\theta}^n\|^2\le C\frac{h^4}{\Delta t}\int_{t_{n-1}}^{t_n}\|\theta^\epsilon\|_2^2dt,$$

$$\begin{aligned}I_3&=\|\theta^n-P_E^h\overline{\theta}^n\|^2=\|\theta^n-\overline{\theta}^n+\overline{\theta}^n-P_E^h\overline{\theta}^n\|^2\\ &=\left\|\frac{1}{\Delta t}\int_{t_{n-1}}^{t_n}\int_t^{t_n}\theta_t^\epsilon(\cdot,s)ds\;dt+\overline{\theta}^n-P_E^h\overline{\theta}^n\right\|^2\\ &\le C\left(\Delta t\int_{t_{n-1}}^{t_n}\|\theta_t^\epsilon\|^2dt+\frac{h^4}{\Delta t}\int_{t_{n-1}}^{t_n}\|\theta^\epsilon\|_2^2dt\right),\end{aligned}$$

$$I_4=2\|u_x^n-\overline{u}_x^n\|^2=2\left\|\frac{1}{\Delta t}\int_{t_{n-1}}^{t_n}(u_x^\epsilon(\cdot,t_n)-u_x^\epsilon(\cdot,t))dt\right\|^2\le C\Delta t\int_{t_{n-1}}^{t_n}\|u_{xt}^\epsilon\|^2dt,$$

$$\begin{aligned}I_5&=(\overline{\beta}_\epsilon^n(\gamma_\epsilon)-\beta_\epsilon(\gamma_\epsilon^n))^2=\left(\frac{1}{\Delta t}\int_{t_{n-1}}^{t_n}(\beta_\epsilon(\gamma_\epsilon)-\beta_\epsilon(\gamma_\epsilon^n))dt\right)^2\\ &\le\left(\frac{1}{\epsilon\Delta t}\int_{t_{n-1}}^{t_n}|u^\epsilon(1,t)-u^\epsilon(1,t_n)|dt\right)^2\le\left(\frac{1}{\epsilon\Delta t}\int_{t_{n-1}}^{t_n}\left|\int_{t_n}^t u_t^\epsilon(1,s)ds\right|dt\right)^2\\ &\le\frac{\Delta t}{\epsilon^2}\int_{t_{n-1}}^{t_n}(u_t^\epsilon(1,\cdot))^2dt\le\frac{\Delta t}{\epsilon^2}\int_{t_{n-1}}^{t_n}\|u_{xt}^\epsilon\|^2dt.\end{aligned}$$

We used the estimate for I_2 on the second term in I_3 and the fact that $|\beta_\epsilon'|\le 1/\epsilon$ on the last series of inequalities. We now multiply (6.1) by Δt, sum from 1 to n, note that $n\Delta t=C$, note that the bounds given by Theorem 2.1 are independent of ϵ, and use the fact that $\varepsilon^0=e^0=0$ to obtain

$$\|\varepsilon_x^n\|^2+\|e_x^n\|^2\le C\left(\|\Theta^0-\theta_0\|^2+\frac{h^2}{\Delta t}+h^4+(\Delta t)^2+\frac{(\Delta t)^2}{\epsilon^2}\right).$$

The Poincaré inequality yields

$$\|U^n - P_E^h u^\epsilon(\cdot, t_n)\|^2 \le C\left(\|\Theta^0 - \theta_0\|^2 + \frac{h^2}{\Delta t} + h^4 + (\Delta t)^2 + \frac{(\Delta t)^2}{\epsilon^2}\right).$$

Using the triangle inequality, we have

$$\begin{aligned}\|u^\epsilon(\cdot, t_n) - U^n\| &\le \|u^\epsilon(\cdot, t_n) - P_E^h u^\epsilon(\cdot, t_n)\| + \|P_E^h u^\epsilon(\cdot, t_n) - U^n\| \\ &\le Ch\|u_x^\epsilon(\cdot, t_n)\| + \|P_E^h u^\epsilon(\cdot, t_n) - U^n\|,\end{aligned}$$

and, as a consequence of the boundedness of u_x^ϵ, it results

$$\|u^\epsilon(\cdot, t_n) - U^n\|^2 \le C\left(\|\Theta^0 - \theta_0\|^2 + h^2 + \frac{h^2}{\Delta t} + h^4 + (\Delta t)^2 + \frac{(\Delta t)^2}{\epsilon^2}\right).$$

Similarly,

$$\left\|\int_0^{t_n} \theta_x^\epsilon dt - \Delta t \sum_{i=1}^n \Theta_x^i\right\| = \|(\hat{\theta}^n - P_E^h\hat{\theta}^n)_x - \varepsilon_x^n\| \le Ch\left(\int_0^{t_n} \|\theta^\epsilon\|_2^2 dt\right)^{1/2} + \|\varepsilon_x^n\|$$

and Theorem 2.1 yield

$$\left\|\int_0^{t_n} \theta_x^\epsilon dt - \Delta t \sum_{i=1}^n \Theta_x^i\right\|^2 \le C\left(\|\Theta^0 - \theta_0\|^2 + h^2 + \frac{h^2}{\Delta t} + h^4 + (\Delta t)^2 + \frac{(\Delta t)^2}{\epsilon^2}\right).$$

Combining these estimates with the penalty errors (Theorem 4.1), we obtain the result. □

Corollary 6.2. *If $u_0 \in H^2(I)$, there exists a constant C such that*

$$\left\|\int_0^{t_n} \theta_x dt - \Delta t \sum_{i=1}^n \Theta_x^i\right\|^2 \le C\left(\|\Theta^0 - \theta_0\|^2 + \epsilon + h^2 + \frac{h^4}{\Delta t} + h^4 + (\Delta t)^2 + \frac{(\Delta t)^2}{\epsilon^2}\right),$$

$$\|u(\cdot, t_n) - U^n\|^2 \le C\left(\|\Theta^0 - \theta_0\|^2 + \epsilon + \frac{h^4}{\Delta t} + h^4 + (\Delta t)^2 + \frac{(\Delta t)^2}{\epsilon^2}\right).$$

Remark. In this case, the rates of convergence are, for $\epsilon = h^{8/5}$ and $\Delta t = h^{12/5}$,

$$\left\|\int_0^{t_n} \theta_x dt - \Delta t \sum_{i=1}^n \Theta_x^i\right\|^2 \le Ch^{8/5} \text{ and } \|u(\cdot, t_n) - U^n\|^2 \le Ch^{8/5}.$$

Proof. If we look at the proof of Theorem 6.1, we see that if $u_0 \in H^2(I)$ and $u^\epsilon(\cdot, t)$ is bounded in $H^2(I)$, independently of ϵ, then the term $h^2/\Delta t$ becomes $h^4/\Delta t$. In the estimate for the displacement, h^2 turns into h^4. □

7. Numerical simulations. In this section, we describe some numerical calculations. In the first experiment, the initial data and the values of a and g, $a = 0.017$, $g = 0.1$, are the same as those in the work of Copetti [4]. Since in [4] the temperature difference is reversed, i.e., the left end of the rod has temperature θ_A and the obstacle is at temperature zero, the numerical scheme was modified accordingly. We

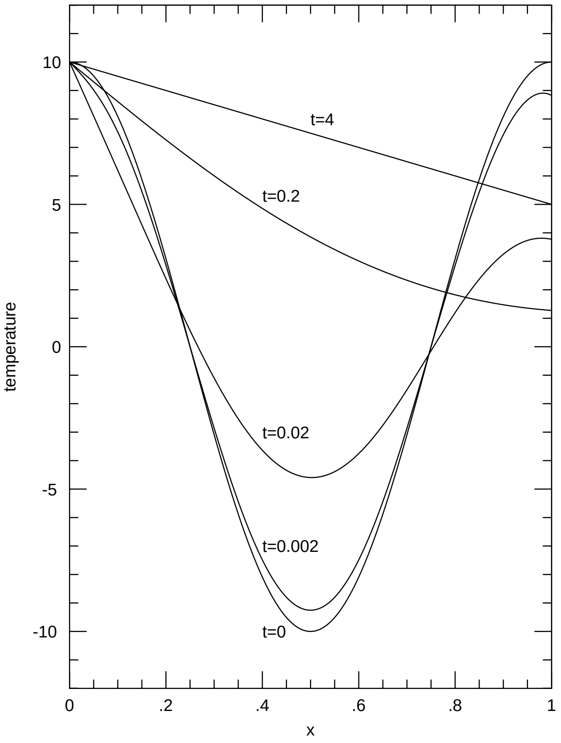


FIG. 7.1. *The evolution in time of the temperature for $\zeta = 10$.*

let $\epsilon = 0.01$, $k = 1$, $h = 1/250$, and $\Delta t = 0.001$, and we took four values for the viscosity coefficient ζ, $\zeta = 0.1$, 0.2, 1, and 10. We observed that the temperature profiles obtained are virtually identical for all values of ζ and are very similar to those presented by Copetti [4], [5], where the rod was only elastic ($\zeta = 0$). This observation was reported also by Copetti [5] for different values of rigidity $1/\epsilon$. Figure 7.1 is a representation of the evolution of temperature in these experiments.

On the other hand, the displacements of the rod are strongly dependent on ζ. When $\zeta = 10$, the deformations are slow in time, and it takes longer for the rod to reach the steady-state configuration (see Figure 7.2). These results support the usual assumption (see Boley and Weiner [3]) that, for small a, the temperature does not depend on the displacement. (Usually the term au_{xt} is then dropped from (1.1).) Note that the final states shown are in agreement with the steady-state solutions as described by Copetti [5].

To examine the error estimates numerically, we performed two experiments with a known solution

$$\theta(x,t) = \exp(t)\cos\left(\frac{3\pi x}{4} + \frac{\pi}{2}\right),$$

$$u(x,t) = \begin{cases} \frac{g \sin x \exp(t)}{\sin 1 \exp(\sqrt{2})}, & 0 < t \leq \sqrt{2}, \\ \frac{g \sin x}{\sin 1}, & \sqrt{2} < t \leq 2, \end{cases}$$

$$\sigma(x,t) = \begin{cases} 0, & 0 < t \leq \sqrt{2}, \\ -t, & \sqrt{2} < t \leq 2, \end{cases}$$

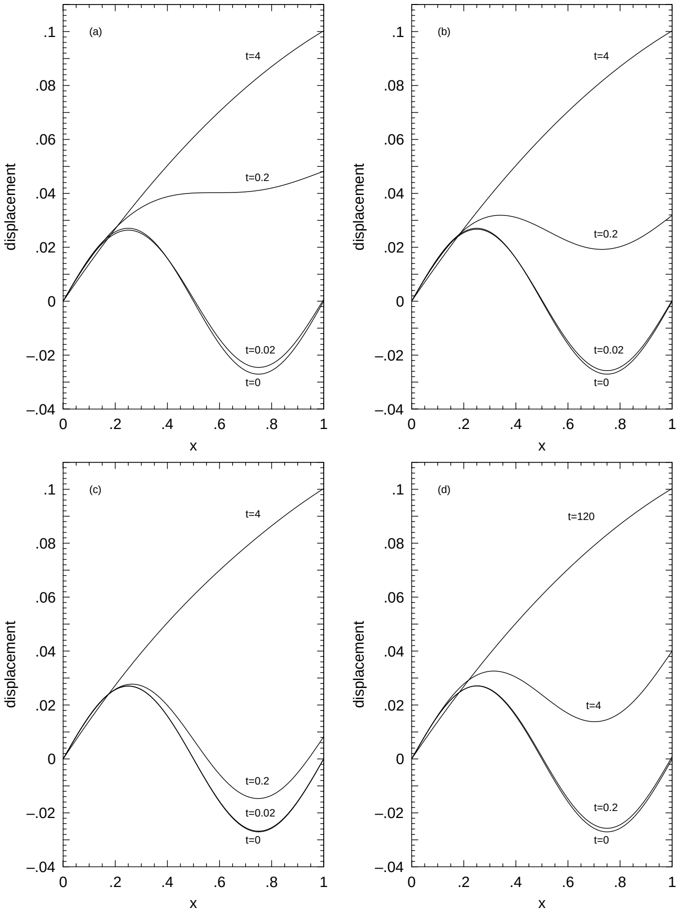


FIG. 7.2. *The evolution in time of the displacement for* (a) $\zeta = 0.1$, (b) $\zeta = 0.2$, (c) $\zeta = 1$, (d) $\zeta = 10$.

satisfying

$$\theta_t - \theta_{xx} = -au_{xt} + f(x,t),$$

$$\sigma_x = 0,$$

where $\sigma = u_x + \zeta u_{xt} - a\theta + g(x,t)$ and θ_0, u_0, f, g, and k are calculated from the exact solution. We let $\theta_A = 0$, $\zeta = 1$, $a = 0.017$, and $g = 0.1$, and Θ^0 was the interpolant of θ_0. In Table 7.1, we report the results of several runs with different values of h, $\epsilon \approx h^{4/5}$, and $\Delta t \approx h^{6/5}$. The computed errors when $\epsilon \approx h^{8/5}$ and $\Delta t \approx h^{12/5}$ are shown in Table 7.2. In both experiments, we observed convergence rates larger than those given by Theorem 6.1 and Corollary 6.2. This is not surprising since we have introduced the penalized problem, and, therefore, our a priori error analysis is probably pessimistic. Moreover, reviewing the proof of Theorem 6.1, we see that the factors $h^2/\Delta t$ and $h^4/\Delta t$ come from the coupling of the equations. If we assume, as suggested in the physical literature, that the temperature does not depend on the displacement, these factors disappear. In addition, for this particular example, the term $\|\Theta^0 - \theta_0\|^2$ converges at the rate h^4. A better-than-predicted rate of convergence

Table 7.1
$\epsilon \approx h^{4/5}$ *and* $\Delta t \approx h^{6/5}$.

h	$\left\Vert\int_0^2 \theta_x dt - \Delta t \sum_{i=1}^N \Theta_x^i\right\Vert^2$	Rate	$\max_i \Vert u(\cdot, t_i) - U^i\Vert^2$	Rate
1/10	9.45×10^{-1}		2.16×10^{-2}	
1/20	1.76×10^{-1}	2.43	8.31×10^{-3}	1.38
1/40	5.21×10^{-2}	1.76	2.89×10^{-3}	1.52
1/80	1.07×10^{-2}	2.29	1.10×10^{-3}	1.39
1/160	3.05×10^{-3}	1.81	3.67×10^{-4}	1.58
1/320	6.53×10^{-4}	2.22	1.29×10^{-4}	1.51

Table 7.2
$\epsilon \approx h^{8/5}$ *and* $\Delta t \approx h^{12/5}$.

h	$\left\Vert\int_0^2 \theta_x dt - \Delta t \sum_{i=1}^N \Theta_x^i\right\Vert^2$	Rate	$\max_i \Vert u(\cdot, t_i) - U^i\Vert^2$	Rate
1/10	6.38×10^{-1}		7.70×10^{-4}	
1/20	1.59×10^{-1}	2	8.93×10^{-5}	3.11
1/40	3.97×10^{-2}	2	9.62×10^{-6}	3.21
1/80	9.93×10^{-3}	2	1.08×10^{-6}	3.16
1/160	2.48×10^{-3}	2	1.20×10^{-7}	3.17
1/320	6.22×10^{-4}	2	1.33×10^{-8}	3.17

was also seen in the simulations run by Copetti and Elliott [6]. Note that, in the last experiment, a small time step was used, and for that situation an explicit-in-time scheme might be competitive. We performed the experiment again using the forward Euler approximation, and we obtained virtually the same results as in Table 7.2.

In our computations, the trapezoidal rule was used to evaluate spatial integrals, and the iterative process (5.3)–(5.4) was stopped when $\Vert\underline{c}^{n,l} - \underline{c}^{n,l-1}\Vert_\infty \leq 1 \times 10^{-7}$ and $\Vert\underline{d}^{n,l} - \underline{d}^{n,l-1}\Vert_\infty \leq 1 \times 10^{-7}$.

Acknowledgment. The authors wish to thank the referees for their valuable comments which improved the manuscript.

REFERENCES

[1] K. T. Andrews, P. Shi, M. Shillor, and S. Wright, *Thermoelastic contact with Barber's heat exchange condition*, Appl. Math. Optim., 28 (1993), pp. 11–48.

[2] J. R. Barber, J. Dundurs, and M. Comninou, *Stability considerations in thermoelastic contact*, ASME J. Appl. Mech., 47 (1980), pp. 871–874.

[3] B. A. Boley and J. H. Weiner, *Theory of Thermal Stresses*, John Wiley, New York, 1960.

[4] M. I. M. Copetti, *Finite element approximation to a contact problem in linear thermoelasticity*, Math. Comp., 68 (1999), pp. 1013–1024.

[5] M. I. M. Copetti, *A one-dimensional thermoelastic problem with unilateral constraint*, Math. Comput. Simulation, 59 (2002), pp. 361–376.

[6] M. I. M. Copetti and C. M. Elliott, *A one-dimensional quasi-static contact problem in linear thermoelasticity*, European J. Appl. Math., 4 (1993), pp. 151–174.

[7] M. I. M. Copetti and D. A. French, *Numerical studies of the stability of steady-state solutions to a contact problem in coupled thermoelasticity*, Appl. Math. Modelling, to appear.

[8] M. Crouzeix and J. Rappaz, *On Numerical Approximation in Bifurcation Theory*, Masson, Paris, 1990.

[9] R. Glowinski, J.-L. Lions, and R. Trémolières, *Numerical Analysis of Variational Inequalities*, North–Holland, Amsterdam, 1981.

[10] S. Jiang and R. Racke, *Evolution Equations in Thermoelasticity*, Chapman and Hall/CRC, Boca Raton, FL, 2000.

[11] K. Kuttler and M. Shillor, *A one-dimensional thermoviscoelastic contact problem*, Adv. Math. Sci. Appl., 4 (1994), pp. 141–159.

[12] J. A. Pelesko, *Nonlinear stability, thermoelastic contact, and the Barber condition*, ASME J. Appl. Mech., 68 (2001), pp. 1–7.

[13] M. G. Srinivasan and D. M. France, *Nonuniqueness in steady-state heat transfer in prestressed duplex tubes-analysis and case history*, ASME J. Appl. Mech., 52 (1985), pp. 257–262.

[14] K. Yosida, *Lectures on Differential and Integral Equations*, Interscience, New York, London, 1960.

SIAM J. Numer. Anal.
Vol. 41, No. 4, pp. 1505–1527

DISCRETE VECTOR POTENTIALS FOR NONSIMPLY CONNECTED THREE-DIMENSIONAL DOMAINS*

FRANCESCA RAPETTI†, FRANÇOIS DUBOIS‡, AND ALAIN BOSSAVIT§

Abstract. In this paper, we focus on the representation of a divergence-free vector field, defined, on a connected nonsimply connected domain $\Omega \subset \mathbb{R}^3$ with a connected boundary Γ, by its curl and its normal component on the boundary. The considered problem is discretized with $H(\mathbf{curl})$- and $H(\mathrm{div})$-conforming finite elements. In order to ensure the uniqueness of the vector potential, we propose a spanning tree methodology to identify the independent edges. The topological features of the domain under consideration are analyzed here by means of the homology groups of first and second order.

Key words. divergence-free vector fields, nonsimply connected domains, edge elements, discrete gauge condition, homology groups

AMS subject classification. 65N30

DOI. 10.1137/S0036142902412646

1. Introduction.

In numerical magnetostatics, an important task is the discretization of the magnetic induction field $\mathbf{u}$, verifying the following equations:

$$\mathbf{curl}\,\mathbf{u} = \omega \quad \text{in } \Omega, \tag{1}$$

$$\mathrm{div}\,\mathbf{u} = 0 \quad \text{in } \Omega, \tag{2}$$

$$\mathbf{u} \cdot \mathbf{n}_\Gamma = g \quad \text{on } \Gamma, \tag{3}$$

where Ω is an open subset of $\mathbb{R}^3$, Γ is its boundary, $\mathbf{n}_\Gamma$ is the outward going normal to Γ, ω is a given current density, and g is a scalar function defined on Γ. A conforming or nonconforming discretization that respects (2) is difficult to obtain with the finite element method [19]. On the other hand, a way to exactly satisfy (2) is to represent $\mathbf{u}$ in terms of a vector potential, i.e., a field $\mathbf{p}$ such that

$$\mathbf{u} = \mathbf{curl}\,\mathbf{p}. \tag{4}$$

The vector $\mathbf{p}$ is not unique but defined up to the gradient of a scalar function. A classical way to ensure the uniqueness of $\mathbf{p}$ is to prescribe a gauge condition such as the Coulomb gauge

$$\mathrm{div}\,\mathbf{p} = 0 \tag{5}$$

and suitable boundary conditions. Moreover, different choices of boundary conditions for the vector field $\mathbf{p}$ are possible, and we refer to [3, 4] for existence and uniqueness results. The vector potential is just a tool for representing the field $\mathbf{u}$ and must be

*Received by the editors August 3, 2002; accepted for publication (in revised form) March 13, 2003; published electronically September 17, 2003.

http://www.siam.org/journals/sinum/41-4/41264.html

†Laboratoire de Mathématiques, UMR 6621 CNRS, Université de Nice et Sophia-Antipolis, Parc Valrose, 06108 Nice cedex 02, France (frapetti@math.unice.fr).

‡Laboratoire des Applications Scientifiques du Calcul Intensif, UPR 9029 CNRS, Université Paris Sud, Bâtiment 506, 91403 Orsay, France (dubois@asci.fr).

§Division Recherche et Développement, Electricité de France, 1 Avenue du Général de Gaulle, 92141 Clamart cedex, France (Alain.Bossavit@der.edfgdf.fr).

easily computable under some constraints on $\mathbf{u}$. In this paper, we choose to fix the current density ω and the magnetic induction flux g across the entire boundary. We remark that, if (2) and (3) are satisfied, then the mean value of the function g across the boundary of Ω is necessarily equal to zero; i.e., $\int_\Gamma g \, d\Gamma = 0$. The chosen problem can be also read in the framework of fluid dynamics: for a given vorticity ω in Ω and mass inflow g across the boundary Γ of Ω, we look for a velocity field $\mathbf{u}$ that satisfies (1) and (3) in the incompressible case, i.e., under the constraint (2).

We restrict ourselves to the previous problem in a nonsimply connected three-dimensional domain with a connected boundary. For the analysis of the mixed formulation of a similar problem in simply connected domains with a nonconnected boundary, see [10].

Concerning the outline of the paper, in section 2, after recalling classical results on vector fields, we split the linear problem in Ω into a homogeneous problem in Ω (i.e., $g = 0$) and a problem on the boundary of Ω (i.e., $g \neq 0$). Then a concrete construction method of a vector potential $\mathbf{p}$ from only the data ω and g is presented. In the discretized problem, the compatibility between these two subproblems requires that the discrete field $\mathbf{p}_m$ associated with a mesh m in Ω has a tangential component on each point of the boundary. The major difficulty is the definition of a "good" discrete space which guarantees the existence of the discrete potential. We start with a short introduction on the homology groups in sections 3 and 4. Then in sections 5 and 6, we adapt the discrete gauge initially proposed by a team of the École Polytechnique [10, 28] to the case of simply connected domains $\Omega \subset \mathbb{R}^3$. Developing the problem presented in [11] in more detail, we generalize in section 7 to the case of proposed nonsimply connected domains. We finally end in section 8 by a short overview on the adopted algorithms and their application in the case where Ω is a torus.

Let us introduce some notation. We consider Ω as a connected bounded domain of $\mathbb{R}^3$, with a connected regular boundary Γ. The scalar product between two vectors $\mathbf{a}$, $\mathbf{b}$ defined in Ω is denoted by $\mathbf{a}\cdot\mathbf{b}$, whereas their vector product is denoted by $\mathbf{a}\times\mathbf{b}$. The tangential component of a vector $\mathbf{v}$ on Γ is $\pi\mathbf{u} = (\mathbf{n}_\Gamma \times \mathbf{u}) \times \mathbf{n}_\Gamma$, and we have a Green formula for regular vector fields $\mathbf{u}$ and $\mathbf{v}$ that reads

$$(6) \qquad \int_\Omega \mathbf{v}\cdot\mathbf{curl}\,\mathbf{u} = \int_\Omega \mathbf{curl}\,\mathbf{v}\cdot\mathbf{u} + \int_\Gamma (\mathbf{n}_\Gamma \times \mathbf{v})\cdot\mathbf{u}.$$

The following operators are defined on Γ as in [9] as follows:

- the surface gradient, $\mathrm{grad}_\Gamma u$, and surface **curl**, $\mathbf{curl}_\Gamma u$, of a scalar function u defined on Γ: $\mathbf{curl}_\Gamma u = \mathrm{grad}_\Gamma u \times \mathbf{n}_\Gamma$;
- the surface curl, $\mathrm{curl}_\Gamma \mathbf{v}$, and surface divergence, $\mathrm{div}_\Gamma \mathbf{v}$, of a tangential vector function $\mathbf{v}$ defined on Γ: $\mathrm{div}_\Gamma \mathbf{v} = \mathrm{curl}_\Gamma(\mathbf{n}_\Gamma \times \mathbf{v})$.

By duality, these operators are also defined for a scalar t or vector $\mathbf{w}$ distributions on Γ as follows:

$$\begin{aligned}
\langle \mathrm{grad}_\Gamma t, \mathbf{v}\rangle_\Gamma &= -\langle t, \mathrm{div}_\Gamma \mathbf{v}\rangle_\Gamma \quad \forall\, \mathbf{v},\\
\langle \mathrm{curl}_\Gamma t, \mathbf{v}\rangle_\Gamma &= \langle t, \mathrm{curl}_\Gamma \mathbf{v}\rangle_\Gamma \quad \forall\, \mathbf{v},\\
\langle \mathrm{curl}_\Gamma \mathbf{w}, u\rangle_\Gamma &= \langle \mathbf{w}, \mathbf{curl}_\Gamma u\rangle_\Gamma \quad \forall\, u,\\
\langle \mathrm{div}_\Gamma \mathbf{w}, u\rangle_\Gamma &= \langle \mathbf{w}, \mathrm{grad}_\Gamma u\rangle_\Gamma \quad \forall\, u.
\end{aligned}$$

Here, the duality product $\langle\cdot,\cdot\rangle_\Gamma$ of two vectors is the scalar product on Γ [9]:

$$\langle \mathbf{w}, \mathbf{v}\rangle_\Gamma = \int_\Gamma \mathbf{w}\cdot\mathbf{v}.$$

The Sobolev spaces $L^2(\Omega)$, $H^1(\Omega)$ are Hilbert spaces with their natural norms $||.||_{0,\Omega}$ and $||.||_{1,\Omega}$, respectively [1]. Following [15], we define

$$H(\mathrm{div},\Omega) = \{\mathbf{u} \in L^2(\Omega)^3 \,|\, \mathrm{div}\,\mathbf{u} \in L^2(\Omega)\},$$
$$H(\mathbf{curl},\Omega) = \{\mathbf{u} \in L^2(\Omega)^3 \,|\, \mathbf{curl}\,\mathbf{u} \in L^2(\Omega)^3\},$$

and associated norms $||.||_{\mathrm{div},\Omega}$ and $||.||_{\mathbf{curl},\Omega}$. We also need to introduce

$$\begin{aligned}
&H(\mathrm{div}_0,\Omega) = \{\mathbf{u} \in H(\mathrm{div},\Omega) \,|\, \mathrm{div}\,\mathbf{u} = 0\},\\
&H_0(\mathrm{div},\Omega) = \{\mathbf{u} \in H(\mathrm{div},\Omega) \,|\, \mathbf{u}\cdot\mathbf{n}_\Gamma = 0\},\\
&H_0(\mathbf{curl},\Omega) = \{\mathbf{u} \in H(\mathbf{curl},\Omega) \,|\, \mathbf{u}\times\mathbf{n}_\Gamma = \mathbf{0}\},\\
&L^2_0(\Omega) = \{u \in L^2(\Omega) \,|\, \textstyle\int_\Omega u = 0\},\\
&\mathcal{C}^{1,1}(\overline{\Omega}) = \{u \in \mathcal{C}^1(\overline{\Omega}) \,|\, \mathbf{grad}\,u \text{ is a vector of Lipschitz functions}\}.
\end{aligned}$$

In a few words, a domain Ω is of class $\mathcal{C}^{1,1}$ if it admits a representation through a $\mathcal{C}^{1,1}(\overline{\Omega})$ map [15]. Note that the boundary of such a domain has a normal vector almost everywhere. In the following, given a space S, the notation $\dim[S]$ denotes the dimension of S. If S is a set, its cardinality, i.e., the number of its elements, is denoted by $\#S$.

2. The continuous problem.

We are interested in the following problem: *given* $g \in L^2_0(\Gamma)$ *and* $\omega \in H(\mathrm{div}_0,\Omega)$, *find* $\mathbf{u} \in H^1(\Omega)^3$ *satisfying*

$$\begin{aligned}
\mathbf{curl}\,\mathbf{u} &= \omega && \text{in } \Omega,\\
\mathrm{div}\,\mathbf{u} &= 0 && \text{in } \Omega,\\
\mathbf{u}\cdot\mathbf{n}_\Gamma &= g && \text{on } \Gamma.
\end{aligned} \tag{7}$$

If Γ is smooth, the continuous problem can be easily analyzed, but the finite elements to discretize it are quite complicated [19, 10]. If Γ is polyhedric, then there are specific difficulties in studying the continuous problem [2, 7], but the finite elements are classical. We recall the main results of regularity for the solution of (7); these results depend on the regularity of the domain Ω. The first result is proven in [2, 15].

PROPOSITION 2.1. *Assume that the bounded domain Ω is of class $\mathcal{C}^{1,1}$ or a convex polyhedron. Then we have the following continuous embedding:*

$$\{\mathbf{v} \in L^2(\Omega)^3 \,|\, \mathbf{curl}\,\mathbf{v} \in L^2(\Omega)^3,\ \mathrm{div}\,\mathbf{v} \in L^2(\Omega),\ \mathbf{v}\cdot\mathbf{n}_\Gamma \in H^{1/2}(\Gamma)\,\} \hookrightarrow H^1(\Omega)^3;$$

as a consequence, problem (7) *has a unique solution, in the sense of distributions, that belongs to* $H^1(\Omega)^3$.

The solution $\mathbf{u}$ of problem (7) is computed as a sum of two functions that are solutions of two simpler problems, i.e., $\mathbf{u} = \mathbf{u}_0 + \hat{\mathbf{u}}$, where $\hat{\mathbf{u}}$ is a divergence-free lifting in $\overline{\Omega}$ of a function $\hat{\mathbf{u}}_\Gamma$ defined on Γ such that

$$\hat{\mathbf{u}}_\Gamma\cdot\mathbf{n}_\Gamma = g, \qquad \mathrm{div}_\Gamma\hat{\mathbf{u}}_\Gamma = 0, \tag{8}$$

and $\mathbf{u}_0$ satisfies

$$\begin{aligned}
\mathbf{curl}\,\mathbf{u}_0 &= \omega - \mathbf{curl}\,\hat{\mathbf{u}} && \text{in } \Omega,\\
\mathrm{div}\,\mathbf{u}_0 &= 0 && \text{in } \Omega,\\
\mathbf{u}_0\cdot\mathbf{n}_\Gamma &= 0 && \text{on } \Gamma\,.
\end{aligned} \tag{9}$$

Note that, thanks to the introduction of a vector potential, (2) is exactly verified, whereas (1) is satisfied in the sense of distributions. Thanks to a trace result proved in [5], problem (7) is well-posed even in nonconvex polyhedra of $\mathbb{R}^3$ (such as a discretized torus).

Proposition 2.2. *Let Γ_i, $i = 1, \dots, L$, be the faces of the boundary Γ of a bounded polyhedron Ω. There exists a real number $s > 1/2$ such that for any function $g \in H^{1/2}(\partial\Gamma_i)$, $i = 1, \dots, L$, problem* (8) *has a unique solution $\hat{\mathbf{u}} \in H^s(\Omega)^3$. In addition, for any $\omega \in H(\mathrm{div}_0, \Omega)$, problem* (9) *has a unique solution $\mathbf{u}_0 \in H^s(\Omega)^3$.*

We end this section by recalling and applying general results on vector fields defined on a regular bounded domain Ω of $\mathbb{R}^3$. We refer to [2, 3] and to their included references for the results. Let us introduce the following spaces:

$$X_T(\Omega) = \{\mathbf{v} \in L^2(\Omega)^3 \,|\, \mathrm{div}\,\mathbf{v} \in L^2(\Omega),\ \mathbf{curl}\,\mathbf{v} \in L^2(\Omega)^3,\ \mathbf{v}\cdot\mathbf{n}_\Gamma \in H^{1/2}(\Gamma)\},$$

$$X_N(\Omega) = \{\mathbf{v} \in L^2(\Omega)^3 \,|\, \mathrm{div}\,\mathbf{v} \in L^2(\Omega),\ \mathbf{curl}\,\mathbf{v} \in L^2(\Omega)^3,\ \mathbf{v}\times\mathbf{n}_\Gamma \in H^{1/2}(\Gamma)^3\},$$

$$H_T(\Omega) = \{\mathbf{v} \in L^2(\Omega)^3 \,|\, \mathrm{div}\,\mathbf{v} = 0,\ \mathbf{curl}\,\mathbf{v} = \mathbf{0},\ \mathbf{v}\cdot\mathbf{n}_\Gamma = 0 \text{ on } \Gamma\},$$

$$H_N(\Omega) = \{\mathbf{v} \in L^2(\Omega)^3 \,|\, \mathrm{div}\,\mathbf{v} = 0,\ \mathbf{curl}\,\mathbf{v} = \mathbf{0},\ \mathbf{v}\times\mathbf{n}_\Gamma = \mathbf{0} \text{ on } \Gamma\},$$

$$P_T : X_T(\Omega) \to H_T(\Omega) \text{ orthogonal projection operator},$$

$$P_N : X_N(\Omega) \to H_N(\Omega) \text{ orthogonal projection operator},$$

$$W^1(\Omega) = \{\mathbf{w} \in H^1(\Omega)^3 \,|\, \mathrm{div}\,\mathbf{w} = 0,\ \mathbf{w}\times\mathbf{n}_\Gamma = \mathbf{0},\ \textstyle\int_\Gamma \mathbf{w}\cdot\mathbf{n}_\Gamma \, d\Gamma = 0\}.$$

Theorem 2.3 (Hodge decomposition). *For a given $\mathbf{u} \in L^2(\Omega)^3$, we have two possible decompositions:*

$$\text{(i)} \qquad \mathbf{u} = \mathbf{grad}\,\phi + \mathbf{curl}\,\mathbf{w} + \theta$$

with $\theta \in H_T(\Omega)$ and a unique $(\phi, \mathbf{w})$ verifying $\phi \in H^1(\Omega) \cap L^2_0(\Omega)$, $\mathbf{w} \in W^1(\Omega)$;

$$\text{(ii)} \qquad \mathbf{u} = \mathbf{grad}\,\psi + \mathbf{curl}\,\mathbf{p} + \eta$$

with $\eta \in H_N(\Omega)$ and a unique $(\psi, \mathbf{p})$ verifying $\psi \in H^1_0(\Omega)$ and

$$\mathbf{p} \in H^1(\Omega)^3,\ \mathrm{div}\,\mathbf{p} = 0,\ \mathbf{p}\cdot\mathbf{n}_\Gamma = 0,\ P_T\mathbf{p} = \mathbf{0}.$$

The decomposition (i) (resp., (ii)) of a field $\mathbf{u}$ is into three orthogonal components of the type $\mathbf{grad}\,\phi$ (resp., $\mathbf{grad}\,\psi$) plus $\mathbf{curl}\,\mathbf{w}$ (resp., $\mathbf{curl}\,\mathbf{p}$) plus a vector lying in $H_T(\Omega)$ (resp., $H_N(\Omega)$).

Theorem 2.4 (Foias, Temam [12]). *Let $\mathbf{u} \in L^2(\Omega)^3$ and $\mathbf{p} \in H(\mathbf{curl}, \Omega)$ such that $\mathbf{u} = \mathbf{curl}\,\mathbf{p}$; then $P_N\mathbf{u} = \mathbf{0}$.*

If a vector $\mathbf{u}$ admits a representation in terms of a vector potential $\mathbf{p}$, i.e., $\mathbf{u} = \mathbf{curl}\,\mathbf{p}$, it clearly satisfies the condition $\mathrm{div}\,\mathbf{u} = 0$. Moreover, for any vector $\mathbf{u}$ of the form (ii), Theorem 2.4 yields $P_N\mathbf{u} = \mathbf{0}$, a condition which precludes flow problems with sources and sinks, as remarked in [10]. Finally, since the scalar $\psi \in H^1_0(\Omega)$, we have $\psi = 0$. The vector potential $\mathbf{p}$ is then the right tool to represent the field $\mathbf{u}$ solution of the considered problem.

In the next section, we recall the main properties of the finite elements we are going to use to discretize $\mathbf{p}$. These finite elements are $H(\mathbf{curl}, \Omega)$-conforming, and by consequence, the field $\mathbf{u}$ will be approximated by $H(\mathrm{div}, \Omega)$-conforming finite elements. Throughout the paper, we treat the three-dimensional case.

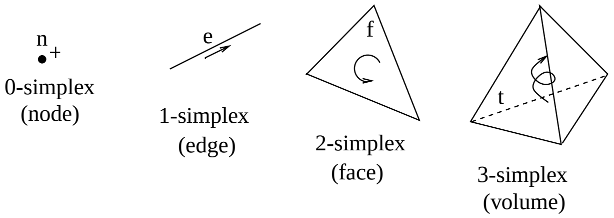


FIG. 1. *Example of oriented p-simplex, $p = 0, \ldots, 3$.*

3. Meshing the domain with cellular complexes. Given a domain $\Omega \subset \mathbb{R}^3$ with boundary Γ, a simplicial mesh m in Ω is a tessellation of $\overline{\Omega}$ by tetrahedra, under the condition that any two of them may intersect along a common face, edge, or node but in no other way. We denote by $\mathcal{N}_m$, $\mathcal{E}_m$, $\mathcal{F}_m$, and $\mathcal{T}_m$ (nodes, edges, faces, and tetrahedra, respectively) the sets of simplices of dimension 0 to 3 thus obtained, and by N_m, E_m, F_m, and T_m their cardinalities. The importance of simplicial meshes lies in the fact that any triangulated domain is homeomorphic to one in which the triangles are flat and the edges straight. Properties on the mesh will hold for the domain, as we are going to present in the following.

First we need to underline some combinatorial properties of the simplicial mesh. Let $\mathcal{M}(r, s)$ denote the set of matrices A whose elements are $A(i, j)$ with $1 \leq i \leq r$ and $1 \leq j \leq s$. In addition to the list of nodes and their positions, the mesh data structure also contains incidence matrices, saying which node belongs to which edge, which edge bounds which face, etc. [6, 14]. There is a notion of orientation for the simplex as in Figure 1 that has to be taken into account to define the incidence matrices. In short, an edge is not only a two-node subset of $\mathcal{N}_m$ but an ordered such set, where the order implies an orientation. Let $e = \{\ell, n\}$ be an edge of the mesh oriented from the node ℓ to n. We can define the incidence numbers $G_{e,n} = 1$, $G_{e,\ell} = -1$, and $G_{e,k} = 0$ for all nodes k other than ℓ and n. These numbers form a rectangular matrix $G \in \mathcal{M}(E_m, N_m)$, which describes how edges connect to nodes. A face $f = \{\ell, n, k\}$ has three vertices which are the nodes ℓ, n, k. Note that $\{n, k, l\}$ and $\{k, l, n\}$ denote the same face f, whereas $\{n, l, k\}$ denotes an oppositely oriented face, which is not supposed to belong to $\mathcal{F}_m$ if f does. An orientation of f induces an orientation of its boundary. So, with respect to the orientation of the face f, the one of the edge $\{l, n\}$ is positive, and the one of $\{k, n\}$ is negative. Then we can define the incidence number $R_{f,e} = 1$ (resp., -1) if the orientation of e matches (resp., does not match) the one on the boundary of f, and $R_{f,e} = 0$ if e is not an edge of f. These numbers form a matrix $R \in \mathcal{M}(F_m, E_m)$. Finally, let us consider the tetrahedron $t = \{k, l, m, n\}$ positively oriented if the three vectors $\{k, l\}$, $\{k, m\}$, and $\{k, n\}$ define a positive frame. ($t' = \{l, m, n, k\}$ has a negative orientation and does not belong to $\mathcal{T}_m$ if t does.) A third matrix $D \in \mathcal{M}(T_m, F_m)$ can be defined by setting $D_{t,f} = \pm 1$ if face f bounds the tetrahedron t, with the sign depending on whether the orientation of f and of the boundary of t match or not, and $D_{t,f} = 0$ in case f does not bound t. For consistency, we attribute an orientation to nodes as well. Implicitly, we have been orienting all nodes the same way (+1) up to now. Note that a sign (-1) to node n changes the sign of all entries of column n in the above G. It can be easily proven that $RG = 0$ and $DR = 0$ [6].

We now define the mixed finite elements we use [6, 23, 24, 26]: they are scalar functions or vector fields associated to all the simplices of the mesh m. We start by denoting φ_n the only continuous, piecewise affine function, which is equal to 1 at n and to 0 at other nodes. We set $W^0_m = \text{span}\,\{\varphi_n \,|\, n \in \mathcal{N}_m\}$. The degree (zero in this case) of the elements of W^0_m refers to the dimension of the simplices they are associated with (i.e., nodes) and not to the degree of φ_n as a polynomial. To the edge e, let us associate the vector field $\mathbf{w}_e$ of the form $\mathbf{a}\times\mathbf{x}+\mathbf{b}$ in each tetrahedron $t \in \mathcal{T}_m$; the two vectors $\mathbf{a}$ and $\mathbf{b}$ are determined by imposing that the circulation of $\mathbf{w}_e$ along $e \in t$ is 1 and 0 along the other edges of t. We denote $W^1_m = \text{span}\,\{\mathbf{w}_e \,|\, e \in \mathcal{E}_m\}$. Similarly, $W^2_m = \text{span}\,\{\mathbf{v}_f \,|\, f \in \mathcal{F}_m\}$ with $\mathbf{v}_f$ the vector of the form $a\,\mathbf{x}+\mathbf{b}$ in each tetrahedron $t \in \mathcal{T}_m$; the scalar a and the vector $\mathbf{b}$ are determined by imposing that the flux of $\mathbf{v}_f$ across the face $f \in t$ is 1 and 0 across the other faces of t. Finally, we introduce $W^3_m = \text{span}\,\{\mu_t \,|\, t \in \mathcal{T}_m\}$, where μ_t is the only scalar whose integral over t is 1 and 0 over the other tetrahedra.[1]

Note that, given two adjacent tetrahedra t and t' sharing a face f, the function φ_n and both the tangential component of $\mathbf{w}_e$ and the normal component of $\mathbf{v}_f$ are continuous across f, whereas the function μ_t is not. Thanks to these continuity properties, $W^0_m \subset H^1(\Omega)$, $W^1_m \subset H(\mathbf{curl},\Omega)$, $W^2_m \subset H(\text{div},\Omega)$, and $W^3_m \subset L^2(\Omega)$. The spaces W^p_m, $p = 0,1,2,3$, have finite dimension given by N_m, E_m, F_m, T_m, respectively, and they play the role of Galerkin approximation spaces for the latter functional spaces.

The properties introduced so far concern the spaces W^p_m taken one by one. There are properties of the structure made of the spaces W^p_m when taken together. We know that the following inclusions hold:

$$\mathbf{grad}\, W^0_m \subset W^1_m, \qquad \mathbf{curl}\, W^1_m \subset W^2_m, \qquad \text{div}\, W^2_m \subset W^3_m.$$

It is natural to ask when the sequence

$$\{0\} \longrightarrow W^0_m \xrightarrow{\mathbf{grad}} W^1_m \xrightarrow{\mathbf{curl}} W^2_m \xrightarrow{\text{div}} W^3_m \longrightarrow \{0\}$$

is exact at levels 1 and 2, i.e., when it happens that

$$\ker(\mathbf{curl}; W^1_m) = \mathbf{grad}\, W^0_m, \qquad \ker(\text{div}; W^2_m) = \mathbf{curl}\, W^1_m,$$

where

$$\ker(\mathbf{curl}; W^1_m) := W^1_m \cap \ker(\mathbf{curl}), \qquad \ker(\text{div}; W^2_m) := W^2_m \cap \ker(\text{div}).$$

At levels 0 and 3, we lose the property of exactitude for the previous sequence because, at level 0, the gradient operator is not injective, and, at level 3, the divergence operator is not surjective. The Poincaré lemma states that, when the domain Ω is contractible [14], the image fills the kernel in both cases. This is not the case with Ω nonsimply connected; for example, we have in fact that $\mathbf{grad}\,(W^0_m)$ is a proper subset of $\ker(\mathbf{curl}; W^1_m)$. This tells us something about the topology of Ω; namely, the

[1] Given the nodes n, l, m, k, the edge $e = \{l, m\}$, the face $f = \{l, m, k\}$, and the tetrahedron $t = \{i, j, k, l\}$, the generators of the spaces W^p_m, $p = 0,1,2,3$, respectively, can also be defined as follows (λ_n is the barycentric coordinate associated to n):

$$\begin{aligned}
\varphi_n &= \lambda_n, \qquad \mathbf{w}_e = \lambda_l\,\mathbf{grad}\,\lambda_m - \lambda_m \mathbf{grad}\,\lambda_l,\\
\mathbf{v}_f &= 2\,(\lambda_l\,\mathbf{grad}\,\lambda_m \times \mathbf{grad}\,\lambda_k + \lambda_m\,\mathbf{grad}\,\lambda_k \times \mathbf{grad}\,\lambda_l + \lambda_k\,\mathbf{grad}\,\lambda_l \times \mathbf{grad}\,\lambda_m),\\
\mu_t &= 6\,(\lambda_i\,\mathbf{grad}\,\lambda_j \times \mathbf{grad}\,\lambda_k \cdot \mathbf{grad}\,\lambda_l + \lambda_j\,\mathbf{grad}\,\lambda_k \times \mathbf{grad}\,\lambda_l \cdot \mathbf{grad}\,\lambda_i\\
&\qquad + \lambda_k\,\mathbf{grad}\,\lambda_l \times \mathbf{grad}\,\lambda_i \cdot \mathbf{grad}\,\lambda_j + \lambda_l\,\mathbf{grad}\,\lambda_i \times \mathbf{grad}\,\lambda_j \cdot \mathbf{grad}\,\lambda_k) = [\text{vol}(t)]^{-1}.
\end{aligned}$$

presence of b_1 "loops," where $b_1 = \dim[\ker(\mathbf{curl}; W_m^1)/\mathbf{grad}\,(W_m^0)]$ is the Betti number of dimension 1 of the domain. Solenoidal fields which are not curls indicate the presence of b_2 "holes," where $b_2 = \dim[\ker(\mathrm{div}; W_m^2)/\mathbf{curl}\,(W_m^1)]$ is the Betti number of dimension 2 of the domain. These are global topological properties of the meshed domain; they do not depend on the mesh that is used to compute them, but they are intrinsic to the considered domain Ω. The sequences are thus an algebraic tool by which the topology of Ω can be explored (and this was Whitney's concern [31]).

The connection between the vector field picture and the cohomological picture in the electromagnetic context has also been considered more recently in [22, 30].

4. Chains, boundary homomorphism, and homology groups. Let m be the simplicial mesh on $\Omega \subset \mathbb{R}^3$. A p-chain c is an assignement to each simplex of dimension p in m of a number α, and we denote by $C_p(m)$ the set of all p-chains. The set $C_p(m)$ has a structure of an abelian group with respect to the addition of p-chains; two p-chains are added by adding the corresponding coefficients.

To give an example, let us consider a path of edges of the mesh m to go from a point n_1 to a point n_2; it is an oriented line. Assigning an integer α_e equal to $+1$ or -1 when the edge e belongs to the path and its orientation is in agreement or in disagreement with that of the path and 0 for all edges e that do not belong to the path, we define a 1-chain. A circuit is a line plus a way to run along it; so, when the line is made of oriented edges, we need to tell the positive direction along each edge, which is precisely what the chain coefficient α_e does. We remark that "chain" is a more general concept than "path," "circuit," etc. In our case, we assume that all coefficients α_i are *relative integers.*

The next concept is the boundary operator $\partial_p : C_p(m) \to C_{p-1}(m)$, $p > 0$. By definition, we have

$$\partial_1(e) = \sum_{n \in \mathcal{N}_m} G_{e,n}\, n, \qquad \partial_2(f) = \sum_{e \in \mathcal{E}_m} R_{f,e}\, e, \qquad \partial_3(t) = \sum_{f \in \mathcal{F}_m} D_{t,f}\, f.$$

Note that ∂_p is represented by a matrix that is G^t, R^t, or D^t depending on the dimension $p > 0$. We remark, in particular, that $\partial_{p+1} \circ \partial_p = 0$, i.e. the boundary of a boundary is the zero chain.

We will say that a p-chain c is closed if $\partial_p c = 0$. Nontrivial closed p-chains are called p-cycles and constitute the subspace $Z_p(m) = \ker(\partial_p; C_p(m))$. A p-chain c is a boundary if there is a $(p+1)$-chain γ such that $c = \partial_{p+1}\gamma$. The p-boundaries constitute the subspace $B_p(m) = \partial_{p+1} C_{p+1}(m)$. Both $Z_p(m)$ and $B_p(m)$ are abelian groups with respect to the addition of p-chains. Boundaries are cycles, but not all cycles are boundaries; we have in fact that $B_p(m) \subset Z_p(m)$.

The quotient space $H_p(m) = [Z_p(m)/B_p(m)]$ is the *homology group* of order p of the mesh m, and the Betti number b_p is equal to $\dim[H_p(m)]$. In particular, we have that $b_0 = \dim[\ker(\mathbf{grad}; W_m^0)]$ is the number of connected components of Ω, and $b_3 = \dim[\mathrm{div}(W_m^3)]$ is the number of connected components of Γ minus one.

Our concern is to determine the cycles that are not boundaries for $p = 1$ and 2, i.e., to computate the generators of $H_1(m)$ and $H_2(m)$. Triangulating a domain reduces the calculation of $H_p(m)$ to a finite procedure (in section 8, we present an algebraic algorithm to define a basis of $H_p(m)$, $p = 1$ and 2); the remarkable thing is that homology groups, in spite of being defined via triangulation, do measure something intrinsic and geometrical (they are topological invariants; i.e., they depend on the domain up to a homeomorphism) that does not depend on the mesh. The homology groups of a surface have a direct link with the possibility of representing curl-free (resp., divergence-free) vectors as gradients (resp., curls). This link is a determinant in the construction of numerical algorithms for solving given problems in terms of scalar or vector potentials, as we are going to see. A key tool in this construction is the Euler–Poincaré characteristic of Ω and its boundary Γ [14].

Remark 4.1. Given a connected domain Ω, the Euler–Poincaré characteristic of Ω is the integer

$$\chi(\Omega) = N_m - E_m + F_m - T_m, \tag{10}$$

where N_m, E_m, F_m, and T_m denote, respectively, the number of nodes, edges, faces, and tetrahedra of the mesh m discretizing $\overline{\Omega}$.

Remark 4.2. Given a connected orientable surface Γ, the Euler–Poincaré characteristic of Γ is the integer

$$\chi(\Gamma) = N_m^\Gamma - E_m^\Gamma + F_m^\Gamma, \tag{11}$$

where N_m^Γ, E_m^Γ, and F_m^Γ denote, respectively, the number of nodes, edges, and triangles of the mesh m^Γ discretizing $\overline{\Gamma}$.

The Euler–Poincaré characteristic is linked to the homology groups' dimension as follows:

$$\chi(\Gamma) = b_0^\Gamma - b_1^\Gamma + b_2^\Gamma, \qquad \chi(\Omega) = b_0 - b_1 + b_2 - b_3,$$

where b_i^Γ, $i = 0, 1, 2$ (resp., b_i, $i = 0, 1, 2, 3$), are the Betti numbers of Γ (resp., Ω). The major point is that these numbers, and consequently the Euler–Poincaré characteristic, are topological invariants. For more details on the subject, see [29].

5. Some discrete spaces and tools. Let us consider a triangulation m of $\overline{\Omega}$, its restriction m^Γ to the boundary Γ of Ω. Let us define the following two functional spaces on Ω and their analogues on the boundary Γ:

$$\begin{aligned} W_{m,0}^2 &= \{\mathbf{v} \in W_m^2 \,|\, \mathbf{v} \cdot \mathbf{n}_\Gamma = 0 \text{ on } \Gamma\}, & W_{m^\Gamma}^2 &= \{\pi\mathbf{v} \,|\, \mathbf{v} \in W_m^2\}, \\ W_{m,0}^1 &= \{\mathbf{v} \in W_m^1 \,|\, \mathbf{v} \times \mathbf{n}_\Gamma = \mathbf{0} \text{ on } \Gamma\}, & W_{m^\Gamma}^1 &= \{\pi\mathbf{v} \,|\, \mathbf{v} \in W_m^1\}. \end{aligned}$$

Note that $W_{m^\Gamma}^1$ is the restriction to Γ of the space W_m^1 in the sense that its vectors are associated to the mesh edges belonging to m^Γ. Similarly, the space $W_{m^\Gamma}^2$, also known as the Raviart–Thomas element space, is the restriction to Γ of W_m^2. Its vectors $\mathbf{v}_e$ are tangential to Γ and, in each triangle $f \in m^\Gamma$, are determined by imposing that the flux of $\mathbf{v}_e$ across the edge $e \in f$ is 1 and 0 across the other edges of f. Vectors of the space $W_{m^\Gamma}^2$ are adapted to represent flux densities that are tangential to Γ.

We look for a discrete approximation $\mathbf{u}_m$ of $\mathbf{u}$ on the mesh m of the form

$$\mathbf{u}_m = \mathbf{curl}\, \mathbf{p}_m \tag{12}$$

with $\mathbf{u}_m$ (resp., $\mathbf{p}_m$) lying in W_m^2 (resp., W_m^1). The uniqueness of the potential $\mathbf{p}_m$ is automatically satisfied if we choose $\mathbf{p}_m \in W_m^1$ and we add a gauge condition. In the following, a linear space for the discrete potential $\mathbf{p}_m$ is proposed; we treat the gauge condition in an entirely algebraic way and obtain the so-called *axial gauge* [16].

To define the discrete space of the vector potential, we need some details on the graph defined in the set of vertices of m by the mesh edges. For a general reference on graph theory, we suggest [14].

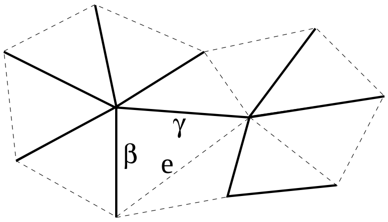


FIG. 2. *In the given mesh m, the thick dark edges compose the spanning tree $\mathcal{T}$, and the dashed ones compose the corresponding cotree $\mathcal{E}_m \setminus \mathcal{T}$. The coedge e closes a circuit together with the tree edges β and γ.*

A set $\mathcal{T}$ of edges of the mesh m such that $C_1(\mathcal{T})$ does not contain any cycle is called a tree. A tree is a spanning tree if there is no strictly larger tree containing it. The set of all left-over edges, i.e., $\mathcal{E}_m \setminus \mathcal{T}$, is called the associated cotree, and its elements are the coedges with respect to $\mathcal{T}$. Coedges thus furnish a basis for 1-cycles in the sense that, given a coedge e, there is a unique way to assign an integer α_ϵ to each edge ϵ of the tree in order to get a closed 1-chain: $\partial(e + \sum_{\epsilon \in \mathcal{T}} \alpha_\epsilon \, \epsilon) = 0$. In short, one says that each coedge $e \in \mathcal{E}_m \setminus \mathcal{T}$ "closes a circuit" C_e in conjunction with edges of the tree. (An example is given in Figure 2, where $C_e = \{e\} \cup \{\beta\} \cup \{\gamma\}$.)

Remark 5.1. For a given mesh m of the domain Ω, the number of edges contained in a spanning tree $\mathcal{T}$ can be expressed in terms of the Betti numbers of the domain by means of the following formula (with easy recursive proof):

$$\#\mathcal{T} = b_1 + (N_m - b_0). \tag{13}$$

For contractible domains, we have $\#\mathcal{T} = N_m - 1$. For noncontractible ones, the spanning tree is enriched with additional edges to take into account that there are 1-cycles that do not bound a surface ($b_1 \neq 0$). (The enriched spanning tree has been called a "belted spanning tree" in [6].)

Now, we explain how to use trees and cotrees to define the proper approximation space to solve the considered problem. In the following sections, $\mathcal{T}$ (resp., $\mathcal{T}^\Gamma$) always represents a spanning tree on Ω (resp., Γ).

6. Approximation of the problem in the simply connected case.

We are interested here in solving problem (7); the domain Ω and its boundary Γ are assumed to be connected and simply connected. We thus assume that Ω is a *sphere* (up to a homeomorphism). The nonsimply connected case is addressed in the next section.

6.1. Lifting the boundary condition for a sphere.

As in [10], given $\mathcal{T}^\Gamma$ and $g \in L^2_0(\Gamma)$, we construct a vector $\hat{\mathbf{u}}^\Gamma_m$ in $W^2_{m^\Gamma}$ such that $\operatorname{div}_\Gamma \hat{\mathbf{u}}^\Gamma_m = 0$, $\hat{\mathbf{u}}^\Gamma_m \cdot \mathbf{n}_\Gamma = g$ face by face on Γ, we show that $\hat{\mathbf{u}}^\Gamma_m$ is unique, and we define $\hat{\mathbf{u}} \in W^2_m$ as the divergence-free lifting of $\hat{\mathbf{u}}^\Gamma_m$ in Ω. Problem (8) is thus well-posed.

PROPOSITION 6.1. *Let us consider a triangulation m of $\overline{\Omega}$, its restriction m^Γ to the boundary Γ of Ω, and a spanning tree $\mathcal{T}^\Gamma$ in m^Γ. Let Γ be a* sphere *and $g \in L^2_0(\Gamma)$. There is a unique divergence-free vector $\hat{\mathbf{u}}^\Gamma_m \in W^2_{m^\Gamma}$ of the form*

$$\hat{\mathbf{u}}^\Gamma_m = \sum_{e \in \mathcal{E}^\Gamma_m \setminus \mathcal{T}^\Gamma} \hat{u}_e \, \mathbf{curl}\, \mathbf{w}_e, \qquad \mathbf{w}_e \in W^1_{m^\Gamma}, \tag{14}$$

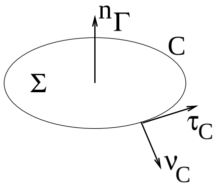


FIG. 3. *Due to the simple connectedness of* Γ*, any circuit* C *in* m^Γ *bounds a surface* $\Sigma \subset \Gamma$.

that satisfies

$$\int_f \hat{\mathbf{u}}_m^\Gamma \cdot \mathbf{n}_\Gamma = \int_f g \quad \forall f \in \mathcal{F}_m^\Gamma. \tag{15}$$

Proof. Let us introduce the two spaces $\ker(\mathrm{div}_\Gamma; W^2_{m^\Gamma})$ and

$$\mathcal{V}(\mathcal{T}^\Gamma) = \mathrm{span}\,\{\mathbf{w}_e \,|\, \mathbf{w}_e \in W^1_{m^\Gamma},\, e \in \mathcal{E}_m^\Gamma \setminus \mathcal{T}^\Gamma\}.$$

The **curl** operator is well defined as $\mathcal{V}(\mathcal{T}^\Gamma) \to \ker(\mathrm{div}_\Gamma; W^2_{m^\Gamma})$.

The **curl** mapping $\mathcal{V}(\mathcal{T}^\Gamma) \to \ker(\mathrm{div}_\Gamma; W^2_{m^\Gamma})$ is injective. Let us consider the vector $\hat{\mathbf{p}}_m^\Gamma \in \mathcal{V}(\mathcal{T}^\Gamma)$ of the form

$$\hat{\mathbf{p}}_m^\Gamma = \sum_{e \in \mathcal{E}_m^\Gamma \setminus \mathcal{T}^\Gamma} \hat{u}_e \, \mathbf{w}_e.$$

Let $\alpha \in \mathcal{E}_m^\Gamma \setminus \mathcal{T}^\Gamma$ be a given coedge, and let $C \subset \{\alpha\} \cup \mathcal{T}^\Gamma$ be the associated cycle; then we have

$$\int_C \hat{\mathbf{p}}_m^\Gamma \cdot \tau_C = \hat{u}_\alpha \int_\alpha \mathbf{w}_\alpha \cdot \tau_C = \hat{u}_\alpha,$$

where τ_C is the tangential vector to C.

On the other hand, since Γ is simply connected, C is the boundary of a surface Σ contained in Γ. On Γ, the normal ν_C and tangential τ_C vectors to C are linked to $\mathbf{n}_\Gamma$ through the relation $\nu_C \times \tau_C = \mathbf{n}_\Gamma$ (see Figure 3). By definition, we have that

$$(\mathbf{curl}\, \hat{\mathbf{p}}_m^\Gamma) \cdot \mathbf{n}_\Gamma = \mathrm{curl}_\Gamma(\pi \hat{\mathbf{p}}_m^\Gamma) = \mathrm{div}_\Gamma(\hat{\mathbf{p}}_m^\Gamma \times \mathbf{n}_\Gamma).$$

The Stokes theorem and the previous tools yield

$$\hat{u}_\alpha = \int_C \hat{\mathbf{p}}_m^\Gamma \cdot \tau_C = \int_C (\hat{\mathbf{p}}_m^\Gamma \times \mathbf{n}_\Gamma) \cdot \nu_C = \int_\Sigma \mathrm{div}_\Gamma(\hat{\mathbf{p}}_m^\Gamma \times \mathbf{n}_\Gamma) = \int_\Sigma (\mathbf{curl}\, \hat{\mathbf{p}}_m^\Gamma) \cdot \mathbf{n}_\Gamma.$$

If $\mathbf{curl}\, \hat{\mathbf{p}}_m^\Gamma = \mathbf{0}$, then $\hat{u}_\alpha = 0$ for all $\alpha \in \mathcal{E}_m^\Gamma \setminus \mathcal{T}^\Gamma$, yielding $\hat{\mathbf{p}}_m^\Gamma = \mathbf{0}$.

Let W be the space composed of vectors in $\ker(\mathrm{div}_\Gamma; W^2_{m^\Gamma})$ verifying (15) with $g \in L_0^2(\Gamma)$. The linear spaces $\mathcal{V}(\mathcal{T}^\Gamma)$ and W have the same dimension. On one hand, we have $\dim[\mathcal{V}(\mathcal{T}^\Gamma)] = E_m^\Gamma - (N_m^\Gamma - 1)$ thanks to Remark 5.1, and on the other hand, $\dim[W] = F_m^\Gamma - 1$. Due to the fact that the Euler–Poincaré characteristic for a spherical surface is 2, Remark 4.2 yields

$$E_m^\Gamma - N_m^\Gamma + 1 = F_m^\Gamma - 1.$$

Given a spanning tree $\mathcal{T}^\Gamma$ and a scalar function $g \in L_0^2(\Gamma)$, there is a unique divergence-free vector of the form (14) and verifying (15). □

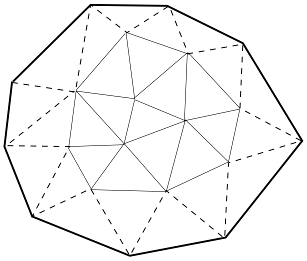


FIG. 4. *In this two-dimensional example, the thick dark edges constitute the set $\mathcal{E}_m^\Gamma$, the dashed edges define $\mathcal{E}_m^\ell$, and the light ones compose $\mathcal{E}_m^{\text{int}}$. Note that here $\mathcal{E}_m^B$ is empty but it is not always the case with more general three-dimensional meshes m.*

6.2. Interior problem for a sphere. At this point we can write

$$\mathbf{u}_m = \hat{\mathbf{u}}_m + \sum_{e \in ?} u_e \,\mathbf{curl}\, \mathbf{w}_e,$$

where $\hat{\mathbf{u}}_m$ is the solution of problem (8) in the sense given in Proposition 6.1 and the symbol "?" in the previous sum is there on purpose to indicate that we do not know yet to which set of internal coedges we have to extend the sum. We remark that

$$\mathcal{E}_m = \mathcal{E}_m^\Gamma \cup \mathcal{E}_m^{\text{int}} \cup \mathcal{E}_m^\ell \cup \mathcal{E}_m^B,$$

where $\mathcal{E}_m^\ell$ is the set of mesh edges having only one extremity on Γ, $\mathcal{E}_m^{\text{int}}$ is the set of mesh edges having both extremities in Ω, and $\mathcal{E}_m^B$ is the set of mesh edges interior to Ω but with both extremities on Γ (see the example in Figure 4).

We denote $\mathcal{T}^{\text{int}}$ a spanning tree contained in $\mathcal{E}_m^{\text{int}}$ and $\mathcal{T}^\ell$ a subset of $\mathcal{E}_m^\ell$ composed of one edge since Γ is connected, linking $\mathcal{T}^{\text{int}}$ to $\mathcal{T}^\Gamma$ and

$$\mathcal{U}_m^0 = (\mathcal{E}_m^{\text{int}} \cup \mathcal{E}_m^\ell \cup \mathcal{E}_m^B) \setminus (\mathcal{T}^{\text{int}} \cup \mathcal{T}^\ell).$$

In the next proposition, we prove that problem (9) is well-posed at the discrete level. In particular, given $\mathcal{T}^{\text{int}} \cup \mathcal{T}^\ell$ and a function $\omega \in \ker(\text{div}; W_m^2)$, we construct a divergence-free vector $\mathbf{u}_m^0 \in W_m^2$ such that $\mathbf{curl}\,\mathbf{u}_m^0 = \omega - \mathbf{curl}\,\hat{\mathbf{u}}_m$ in Ω and $\mathbf{u}_m^0 \cdot \mathbf{n}_\Gamma = 0$ face by face on Γ, and we show that $\mathbf{u}_m^0$ is unique.

PROPOSITION 6.2. *Let us consider a triangulation m of $\overline{\Omega}$ and a spanning tree $\mathcal{T}^{\text{int}} \cup \mathcal{T}^\ell$ in $m \setminus m^\Gamma$. Let us suppose that Ω is a* sphere *and $\omega \in \ker(\text{div}; W_m^2)$. There exists a unique divergence-free vector $\mathbf{u}_m^0 \in W_{m,0}^2$ of the form*

$$\mathbf{u}_m^0 = \sum_{e \in \mathcal{U}_m^0} u_e^0 \,\mathbf{curl}\, \mathbf{w}_e, \qquad \mathbf{w}_e \in W_m^1, \tag{16}$$

that satisfies

$$\mathbf{curl}\,\mathbf{u}_m^0 = \omega - \mathbf{curl}\,\hat{\mathbf{u}}_m \qquad \text{in } \Omega, \tag{17}$$

where $\hat{\mathbf{u}}_m$ is the divergence-free lifting in Ω of $\hat{\mathbf{u}}_m^\Gamma$ defined in Proposition 6.1.

Proof. Let us introduce the two spaces $\ker(\text{div}; W_{m,0}^2)$ and

$$\mathcal{V}(\mathcal{T}^{\text{int}}, \mathcal{T}^\ell) = \text{span}\,\{\mathbf{w}_e \,|\, \mathbf{w}_e \in W_m^1\,,\, e \in \mathcal{U}_m^0\}.$$

The **curl** operator is well defined as $\mathcal{V}(\mathcal{T}^{\text{int}}, \mathcal{T}^\ell) \to \ker(\text{div}; W_{m,0}^2)$.

The **curl** mapping $\mathcal{V}(\mathcal{T}^{\rm int},\mathcal{T}^\ell)\to\ker(\mathrm{div};W^2_{m,0})$ is injective. The proof given for Proposition 6.1 is the same with $\mathbf{p}^0_m\in\mathcal{V}(\mathcal{T}^{\rm int},\mathcal{T}^\ell)$ of the form

$$\mathbf{p}^0_m=\sum_{e\in\mathcal{U}^0_m}\hat{u}_e\,\mathbf{w}_e.$$

The linear spaces $\mathcal{V}(\mathcal{T}^{\rm int},\mathcal{T}^\ell)$ and $\ker(\mathrm{div};W^2_{m,0})$ have the same dimension. Because $\#\mathcal{T}^{\rm int}=(N_m-N^\Gamma_m-1)$, $\#\mathcal{T}^\ell=1$, on one hand we have

$$\dim\left[\mathcal{V}(\mathcal{T}^{\rm int}\,,\,\mathcal{T}^\ell)\right]=E_m-E^\Gamma_m-(N_m-N^\Gamma_m-1+1).$$

On the other hand, because $\dim\left[W^2_m\right]=F_m$ and $\dim\left[W^2_{m,0}\right]=F_m-F^\Gamma_m$, we get

$$\dim\left[\ker(\mathrm{div};W^2_{m,0})\right]=F_m-F^\Gamma_m-(T_m-1).$$

Note that we have T_m-1 independent relations since $\mathrm{div}\,\mathbf{v}=0$ and Ω is simply connected, as it is proved in Lemma 4.2 of [10]. By using the Euler–Poincaré characteristics and Remarks 4.1 and 4.2, we get

$$\begin{aligned}&E_m-E^\Gamma_m-(N_m-N^\Gamma_m)-(F_m-F^\Gamma_m-(T_m-1))\\&\quad=-(N_m-E_m+F_m-T_m)+(N^\Gamma_m-E^\Gamma_m+F^\Gamma_m)-1\\&\quad=-1+2-1=0.\end{aligned}$$

The present proof can also be carried out at an algebraic level. Proposition 6.2 states that, given $\mathcal{T}^{\rm int}\cup\mathcal{T}^\ell$, $\omega\in\ker(\mathrm{div};W^2_m)$, and $\hat{\mathbf{u}}_m$ the divergence-free lifting in Ω of $\hat{\mathbf{u}}^\Gamma_m$ defined in Proposition 6.1, there is a unique divergence-free vector $\mathbf{u}^0_m\in W^2_{m,0}$ of the form (16). Moreover, its coefficients u^0_e, $e\in\mathcal{U}^0_m$, on the chosen basis, are the components of the solution of the linear system

$$\sum_{e\in\mathcal{U}^0_m}u^0_e\int_\Omega\mathbf{curl}\,\mathbf{w}_e\cdot\mathbf{curl}\,\mathbf{w}_\gamma=\int_\Omega\omega\cdot\mathbf{w}_\gamma-\int_\Omega\hat{\mathbf{u}}_m\cdot\mathbf{w}_\gamma\qquad\forall\,\gamma\in\mathcal{U}^0_m.\tag{18}$$

The matrix

$$A=\left(\int_\Omega\mathbf{curl}\,\mathbf{w}_e\cdot\mathbf{curl}\,\mathbf{w}_\gamma\right)_{e,\gamma\in\mathcal{U}^0_m}$$

has full rank; it is in fact the mass matrix for the chosen basis $\{\mathbf{curl}\,\mathbf{w}_e\,|\,e\in\mathcal{U}^0_m\}$ (defined on the coedges) of the space $\ker(\mathrm{div};W^2_{m,0})$. Note that A is a symmetric and positive definite sparse matrix so that the linear system (18) can be solved iteratively by using a conjugate gradient method, as first done by Roux [28]. □

Remark 6.3. Note that $\mathcal{E}^{\rm int}_m$ and $\mathcal{E}^\ell_m$ can be empty. In this case, there is no interior spanning tree $(\mathcal{T}^{\rm int}\cup\mathcal{T}^\ell)$ and $\mathcal{U}^0_m=\mathcal{E}^B_m$. As $N_m=N^\Gamma_m$, Remarks 4.1 and 4.2 yield again

$$(\#\mathcal{U}^0_m=)\quad E_m-E^\Gamma_m=F_m-F^m_m-(T_m-1)\quad(=\dim\left[\ker(\mathrm{div};W^2_{m,0})\right]).$$

A similar remark can be done in the nonsimply connected case.

The function

$$\mathbf{u}_m=\hat{\mathbf{u}}_m+\sum_{e\in\mathcal{U}^0_m}u^0_e\,\mathbf{curl}\,\mathbf{w}_e\tag{19}$$

is then the approximated solution of problem (7). It is natural to ask whether the computed solution depends on the chosen spanning tree. The answer is no, as we state in the next subsection.

6.3. Independence on the spanning tree. We remark that the solution does not depend on the adopted spanning tree if $\omega \in \ker(\mathrm{div}; W_m^2)$. Let us consider two boundary and interior spanning trees as well as two sets of mesh edges:

$$\begin{array}{lll} \mathcal{T}_1^\Gamma, & \mathcal{T}_1 = \mathcal{T}_1^{\mathrm{int}} \cup \mathcal{T}_1^\ell, & \mathcal{U}_{m,1}^0 = (\mathcal{E}_m^{\mathrm{int}} \cup \mathcal{E}_m^\ell \cup \mathcal{E}_m^B) \setminus \mathcal{T}_1, \\ \mathcal{T}_2^\Gamma, & \mathcal{T}_2 = \mathcal{T}_2^{\mathrm{int}} \cup \mathcal{T}_2^\ell, & \mathcal{U}_{m,2}^0 = (\mathcal{E}_m^{\mathrm{int}} \cup \mathcal{E}_m^\ell \cup \mathcal{E}_m^B) \setminus \mathcal{T}_2. \end{array}$$

Let $\mathbf{u}_m^i$ be the solution associated with $\mathcal{E}_m^\Gamma \setminus \mathcal{T}_i^\Gamma$ on the boundary and with $\mathcal{U}_{m,i}^0$ at the interior ($i = 1, 2$). Let us denote $\mathbf{v}_m = \mathbf{u}_m^1 - \mathbf{u}_m^2$. We will prove that $\mathbf{v}_m = \mathbf{0}$.

On the boundary we consider $\mathbf{v}_m = \sum_{e \in \mathcal{E}_m^\Gamma} v_a \,\mathbf{curl}\,\mathbf{w}_e$; since $\mathbf{u}_m^1 \cdot \mathbf{n}_\Gamma = \mathbf{u}_m^2 \cdot \mathbf{n}_\Gamma = g$ on Γ, we have

$$\int_f \mathbf{v}_m \cdot \mathbf{n}_\Gamma = 0 \quad \forall f \in \mathcal{F}_m^\Gamma.$$

Thus $\mathbf{v}_m \cdot \mathbf{n}_\Gamma = 0$ on Γ, since the family $\{\mathbf{curl}\,\mathbf{w}_e \,|\, \mathbf{w}_e \in W_m^1,\ e \in \mathcal{E}_m^\Gamma \setminus \mathcal{T}_i^\Gamma\}$, $i = 1$ or 2, is a basis for the space span$\{\mathbf{curl}\,\mathbf{w}_e \,|\, \mathbf{w}_e \in W_m^1,\ e \in \mathcal{E}_m^\Gamma\}$.

In the interior, because $\omega \in \ker(\mathrm{div}; W_m^2)$, we can write that

$$\omega = \sum_{e \in \mathcal{E}_m} \omega_e \,\mathbf{curl}\,\mathbf{w}_e, \qquad \mathbf{w}_e \in W_m^1.$$

We have that, for all $\mathbf{w}_\gamma \in W_m^1$ with $\gamma \in \mathcal{U}_{m,1}^0$ or $\gamma \in \mathcal{U}_{m,2}^0$,

$$\int_\Omega \omega \cdot \mathbf{w}_\gamma = \sum_{e \in \mathcal{E}_m} \omega_e \int_\Omega \mathbf{w}_e \cdot \mathbf{curl}\,\mathbf{w}_\gamma.$$

The family $\{\mathbf{curl}\,\mathbf{w}_\gamma \,|\, \mathbf{w}_\gamma \in W_m^1,\ \gamma \in \mathcal{U}_{m,i}^0\}$, $i = 1$ or 2, is a basis for $\ker(\mathrm{div}; W_m^2)$, so we have that for all $\gamma \in \mathcal{E}_m^{\mathrm{int}} \cup \mathcal{E}_m^\ell \cup \mathcal{E}_m^B$ and not only for all $\gamma \in \mathcal{U}_{m,i}^0$,

$$\int_\Omega \omega \cdot \mathbf{w}_\gamma = \sum_{e \in \mathcal{E}_m} \omega_e \int_\Omega \mathbf{w}_e \cdot \mathbf{curl}\,\mathbf{w}_\gamma.$$

Remarking that for all $\gamma \in \mathcal{E}_m^{\mathrm{int}} \cup \mathcal{E}_m^\ell \cup \mathcal{E}_m^B$

$$\begin{aligned} \textstyle\int_\Omega \mathbf{u}_m^1 \cdot \mathbf{curl}\,\mathbf{w}_\gamma &= \textstyle\int_\Omega \omega \cdot \mathbf{w}_\gamma, \\ \textstyle\int_\Omega \mathbf{u}_m^2 \cdot \mathbf{curl}\,\mathbf{w}_\gamma &= \textstyle\int_\Omega \omega \cdot \mathbf{w}_\gamma, \end{aligned}$$

we have that, for all $\gamma \in \mathcal{E}_m^{\mathrm{int}} \cup \mathcal{E}_m^\ell \cup \mathcal{E}_m^B$,

$$\int_\Omega \mathbf{v}_m \cdot \mathbf{curl}\,\mathbf{w}_\gamma = 0.$$

That together with $\mathbf{v}_m \in \ker(\mathrm{div}; W_m^2)$ implies $\mathbf{v}_m = \mathbf{0}$.

As we have seen, the final solution does not depend on the particular spanning tree to gauge the potential. In practice, however, the efficiency of the method does via the dependence on the particular tree of the condition number of the "stiffness" matrix A in (18). This dependence is not too dramatic anyway, as underlined by the numerical tests presented in the appendix of [10].

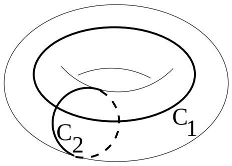


FIG. 5. *Two disjoint loops $\mathcal{C}_1$ and $\mathcal{C}_2$, an example of generators of the first homology group $H_1(m^\Gamma)$ of the torus surface Γ.*

7. Approximation of the problem in the nonsimply connected case.

In the following, we turn our attention to the case where Ω and its boundary are nonsimply connected. As an example of such a situation, we assume that Ω is a *torus*, up to a homeomorphism. For the torus, we have $\chi(\Gamma) = 0$ and $\chi(\Omega) = 0$. Note that Γ is an "empty" torus, while Ω is a "full" torus. Let us look at the differences between the simply connected and nonsimply connected cases.

7.1. Lifting the boundary condition for a torus.

Let $\mathcal{S}^\Gamma$ denote the belted spanning tree on the torus boundary, i.e., $\mathcal{S}^\Gamma = \mathcal{T}^\Gamma \cup \{\Pi_1, \Pi_2\}$, where $\mathcal{T}^\Gamma$ is the usual spanning tree without loops and Π_1, Π_2 are two suitable edges of $\mathcal{E}^\Gamma_m$. In particular, denote $\mathcal{C}_1$ and $\mathcal{C}_2$ two disjoint loops of Γ, as presented in Figure 5, and we have that

$$\mathcal{T}^\Gamma \cup \{\Pi_1\} \text{ contains a loop homologous to } \mathcal{C}_1,$$

$$\mathcal{T}^\Gamma \cup \{\Pi_2\} \text{ contains a loop homologous to } \mathcal{C}_2.$$

Note that, with respect to the simply connected case, the spanning tree on the surface has been enriched according to $\dim[H_1(m^\Gamma)]$ (two edges in the case of the torus surface), as explained in Remark 5.1. Thanks to these added edges, the circuits associated to all remaining coedges do bound a surface contained in Γ, and this is a property that will be exploited during the proof of the following proposition. See [27] for a method to build up a belted spanning tree.

PROPOSITION 7.1. *Let us consider a triangulation m of $\overline{\Omega}$, its restriction m^Γ to the boundary Γ of Ω, and a spanning tree $\mathcal{S}^\Gamma$ in m^Γ. Let Γ be a* torus*, and let $g \in L^2_0(\Gamma)$. There is a unique divergence-free vector $\hat{\mathbf{u}}^\Gamma_m \in W^2_{m^\Gamma}$ of the form*

$$\hat{\mathbf{u}}^\Gamma_m = \sum_{e \in \mathcal{E}^\Gamma_m \setminus \mathcal{S}^\Gamma} \hat{u}_e \, \mathbf{curl}\, \mathbf{w}_e, \qquad \mathbf{w}_e \in W^1_{m^\Gamma}, \tag{20}$$

that satisfies

$$\int_f \hat{\mathbf{u}}^\Gamma_m \cdot \mathbf{n}_\Gamma = \int_f g \quad \forall\, f \in \mathcal{F}^\Gamma_m. \tag{21}$$

Proof. The proof is similar to the proof of Proposition 6.1. We introduce the two spaces $\ker(\mathrm{div}_\Gamma; W^2_{m^\Gamma})$ and

$$\mathcal{V}(\mathcal{S}^\Gamma) = \mathrm{span}\,\{\mathbf{w}_e \,|\, \mathbf{w}_e \in W^1_{m^\Gamma},\, e \in \mathcal{E}^\Gamma_m \setminus \mathcal{S}^\Gamma\}.$$

The curl operator is well defined as $\mathcal{V}(\mathcal{S}^\Gamma) \to \ker(\mathrm{div}_\Gamma; W^2_{m^\Gamma})$.

The curl mapping $\mathcal{V}(\mathcal{S}^\Gamma) \to \ker(\mathrm{div}_\Gamma; W^2_{m^\Gamma})$ is injective. Let us consider the vector $\hat{\mathbf{p}}^\Gamma_m \in \mathcal{V}(\mathcal{S}^\Gamma)$ of the form

$$\hat{\mathbf{p}}^\Gamma_m = \sum_{e \in \mathcal{E}^\Gamma_m \setminus \mathcal{S}^\Gamma} \hat{u}_e \, \mathbf{w}_e.$$

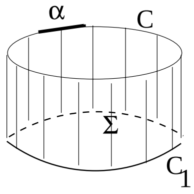


FIG. 6. *When the circuit $\mathcal{C}$ is homologous to one of the two independent loops, say, $\mathcal{C}_1$, it does not bound any surface. In this case, we have to consider the loop $\mathcal{C} \cup \mathcal{C}_1$ that bounds a (lateral, in this case) surface Σ.*

Let $\alpha \in \mathcal{E}^\Gamma_m \setminus \mathcal{S}^\Gamma$ be a given coedge, and let $\mathcal{C} \subset \{\alpha\} \cup \mathcal{S}^\Gamma$ be the associated cycle. It can happen either that $\mathcal{C}$ is the boundary of a surface Σ (in this case we repeat exactly the proof for the simply connected case) or that $\mathcal{C}$ is homologous to one of the two fundamental loops, say, $\mathcal{C}_1$, and so *it does not bound a surface.* To overcome the problem, we have to consider the loop $\mathcal{C} \cup \mathcal{C}_1$; now, this loop *does bound* a surface, and we denote by Σ the corresponding surface (see Figure 6).

We have then

$$\int_{\mathcal{C}\cup\mathcal{C}_1} \hat{\mathbf{p}}^\Gamma_m \cdot \tau_{\mathcal{C}\cup\mathcal{C}_1} = \hat{u}_\alpha \int_\alpha \mathbf{w}_\alpha \cdot \tau_{\mathcal{C}\cup\mathcal{C}_1} = \hat{u}_\alpha,$$

where $\tau_{\mathcal{C}\cup\mathcal{C}_1}$ is the tangential vector to $\mathcal{C} \cup \mathcal{C}_1$. We then conclude that $\hat{u}_\alpha = 0$ by following the same steps of the proof for Proposition 6.1.

Let W again be the space composed of vectors in $\ker(\mathrm{div}_\Gamma; W^2_{m^\Gamma})$ verifying (21) with $g \in L^2_0(\Gamma)$. The linear spaces $\mathcal{V}(\mathcal{S}^\Gamma)$ and W have the same dimension. On one hand, we have $\dim[\mathcal{V}(\mathcal{S}^\Gamma)] = E^\Gamma_m - (N^\Gamma_m - 1 + 2)$ thanks to Remark 5.1, and on the other hand, $\dim[W] = F^\Gamma_m - 1$. Due to the fact that the Euler–Poincaré characteristic for an empty torus is 0, the equality

$$E^\Gamma_m - N^\Gamma_m - 1 = F^\Gamma_m - 1$$

follows from Remark 4.2. $\square$

7.2. Interior problem for a torus. Let $\mathcal{C}_1$ be the loop that does not bound any surface of Ω, and let $\mathcal{C}_2$ be the one that bounds a surface Σ_2 when considered in Ω (see Figure 5). The flux condition

$$\int_{\Sigma_2} \mathbf{u} \cdot \mathbf{n}_\Sigma = \int_{\partial\Sigma_2} \mathbf{p} \cdot \tau_{\mathcal{C}_2} \neq 0$$

yields $\pi\mathbf{p}$ not identically null. In this case, to solve problem (9) we need to "reactivate" one of the two edges Π_1, Π_2 excluded in problem (8) and precisely the one associated with the loop that bounds when we pass from Γ to Ω. In any other case, the degree of freedom associated to Π_2 is zero. In the following, we take into account the more general case where $\pi\mathbf{p} \neq \mathbf{0}$ and we assume that $\Pi^*_2 = \mathrm{supp}(\pi\mathbf{p})$. The degree of freedom associated to this particular edge is equal to the flux of the field $\mathbf{u}$ across the transversal section of the torus. For this feature, from now on, we call Π^*_2 the "flux edge." Denoting by $\mathcal{T}^{\rm int}$ the usual spanning tree without loops, we have $\mathcal{S}^{\rm int} = \mathcal{T}^{\rm int} \cup \{\Pi_1\}$, and the set $\mathcal{U}^0_m$ is now defined as follows:

$$\tilde{\mathcal{U}}^0_m = (\mathcal{E}^{\rm int}_m \cup \mathcal{E}^\ell_m \cup \mathcal{E}^B_m \cup \{\Pi^*_2\}) \setminus (\mathcal{S}^{\rm int} \cup \mathcal{T}^\ell).$$

Proposition 7.2. *Let us consider a triangulation m of $\overline{\Omega}$, together with a spanning tree $\mathcal{S}^{\rm int} \cup \mathcal{T}^{\ell}$ in m. Let us suppose that Ω is a* torus *and $\omega \in \ker(\mathrm{div}; W^2_m)$. There exists a unique divergence-free vector $\mathbf{u}^0_m \in W^2_{m,0}$ of the form*

$$\mathbf{u}^0_m = \sum_{e \in \tilde{\mathcal{U}}^0_m} u^0_e \, \mathbf{curl}\, \mathbf{w}_e \qquad \mathbf{w}_e \in W^1_m, \tag{22}$$

that satisfies

$$\mathbf{curl}\, \mathbf{u}^0_m = \omega - \mathbf{curl}\, \hat{\mathbf{u}}_m \qquad \text{in} \;\; \Omega, \tag{23}$$

where $\hat{\mathbf{u}}_m$ is the divergence-free lifting in Ω of $\hat{\mathbf{u}}^{\Gamma}_m$ defined in Proposition 7.1.

Proof. Let us introduce the two spaces $\ker(\mathrm{div}; W^2_{m,0})$ and

$$\mathcal{V}(\mathcal{S}^{\rm int}, \mathcal{T}^{\ell}) = \mathrm{span}\, \{\mathbf{w}_e \,|\, \mathbf{w}_e \in W^1_m,\, e \in \tilde{\mathcal{U}}^0_m\}.$$

The curl operator is well defined as $\mathcal{V}(\mathcal{S}^{\rm int}, \mathcal{T}^{\ell}) \to \ker(\mathrm{div}; W^2_{m,0})$.

The curl mapping $\mathcal{V}(\mathcal{S}^{\rm int}, \mathcal{T}^{\ell}) \to \ker(\mathrm{div}; W^2_{m,0})$ is injective. The proof given for Proposition 7.1 is the same with $\mathbf{p}^0_m \in \mathcal{V}(\mathcal{S}^{\rm int}, \mathcal{T}^{\ell})$ of the form

$$\mathbf{p}^0_m = \sum_{e \in \tilde{\mathcal{U}}^0_m} \hat{u}_e \, \mathbf{w}_e.$$

The linear spaces $\mathcal{V}(\mathcal{S}^{\rm int}, \mathcal{T}^{\ell})$ and $\ker(\mathrm{div}; W^2_{m,0})$ have the same dimension. In fact, we have

$$\begin{aligned} &\dim\, [\mathcal{V}(\mathcal{S}^{\rm int}, \mathcal{T}^{\ell})] = E_m - E^{\Gamma}_m + 1 - (N_m - N^{\Gamma}_m - 1 + 1 + 1), \\ &\dim\, [\ker(\mathrm{div}; W^2_{m,0})] = F_m - F^{\Gamma}_m - (T_m - 1) - 1. \end{aligned}$$

Note that now, for the presence of "one hole" in Ω, the equation $\mathrm{div}\, \mathbf{u} = 0$ gives only T_m independent conditions. The two are coincident since the Euler–Poincaré characteristic for the "full" torus and its surface is 0.

Once again, the present proof can also be carried out at an algebraic level. Proposition 7.2 states that, given $\mathcal{S}^{\rm int} \cup \mathcal{T}^{\ell}$, $\omega \in \ker(\mathrm{div}; W^2_m)$, and $\hat{\mathbf{u}}_m$ the divergence-free lifting in Ω of $\hat{\mathbf{u}}^{\Gamma}_m$ defined in Proposition 7.1, there is a unique divergence-free vector $\mathbf{u}^0_m \in W^2_{m,0}$ of the form (22). Moreover, its coefficients u^0_e, $e \in \tilde{\mathcal{U}}^0_m$, on the chosen basis, are the components of the solution of the linear system

$$\sum_{e \in \tilde{\mathcal{U}}^0_m} u^0_e \int_{\Omega} \mathbf{curl}\, \mathbf{w}_e \cdot \mathbf{curl}\, \mathbf{w}_{\gamma} = \int_{\Omega} \omega \cdot \mathbf{w}_{\gamma} - \int_{\Omega} \hat{\mathbf{u}}_m \cdot \mathbf{w}_{\gamma} \qquad \forall\, \gamma \in \tilde{\mathcal{U}}^0_m. \tag{24}$$

The matrix

$$\tilde{A} = \left(\int_{\Omega} \mathbf{curl}\, \mathbf{w}_e \cdot \mathbf{curl}\, \mathbf{w}_{\gamma} \right)_{e, \gamma \in \tilde{\mathcal{U}}^0_m}$$

again has full rank, but it is no more sparse due to the presence of the basis function associated to $\{\Pi^*_2\}$. □

The function

$$\mathbf{u}_m = \hat{\mathbf{u}}_m + \sum_{e \in \tilde{\mathcal{U}}^0_m} u^0_e \, \mathbf{curl}\, \mathbf{w}_e \tag{25}$$

is then the approximated solution of problem (7) in the nonsimply connected case.

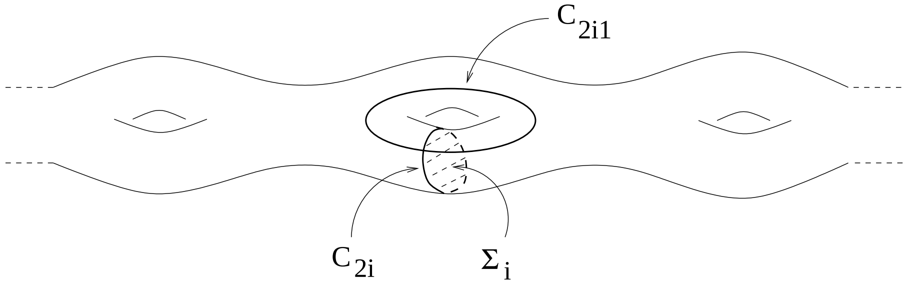


FIG. 7. *Two disjoint loops $\mathcal{C}_{2i-1}$ and $\mathcal{C}_{2i}$, $1 \leq i \leq \kappa$, example of generators of the first homology group of the ith "empty" torus belonging to the surface of the sum of κ tori.*

7.3. Case of the sum of κ tori. The theory that we have presented can be generalized to a domain Ω that is the sum of κ tori, with the integer $\kappa \geq 1$. We have that

$$\chi(\Gamma) = 1 - 2\kappa + 1 = 2(1-\kappa), \qquad \chi(\Omega) = 1 - \kappa.$$

For problem (8), we have to consider

$$\mathcal{S}^{\Gamma} = \left(\mathcal{T}^{\Gamma} \cup \bigcup_{i=1}^{\kappa} \{\Pi_{2i-1}, \Pi_{2i}\} \right),$$

where $\mathcal{T}^{\Gamma}$ is the usual spanning tree without loops and $\{\Pi_{2i-1}, \Pi_{2i}\}$ for $1 \leq i \leq \kappa$ is one pair of suitable edges of $\mathcal{E}_m^{\Gamma}$. In particular, denoting by $\mathcal{C}_{2i-1}$ and $\mathcal{C}_{2i}$ two disjoint loops as presented in Figure 7, we have that

$$\mathcal{T}^{\Gamma} \cup \{\Pi_{2i-1}\} \text{ contains a loop homologous to } \mathcal{C}_{2i-1},$$
$$\mathcal{T}^{\Gamma} \cup \{\Pi_{2i}\} \text{ contains a loop homologous to } \mathcal{C}_{2i}.$$

The spanning tree on the surface has been enriched according to $\dim [H_1(m^{\Gamma})]$ that is now 2κ, as explained in Remark 5.1. The proof of Proposition 7.1 for problem (8) does not change globally; concerning the dimension of the approximation and approximated spaces, we have now

$$\dim [\mathcal{V}(\mathcal{S}^{\Gamma})] = E_m^{\Gamma} - (N_m^{\Gamma} - 1 + 2\kappa), \qquad \dim [W] = F_m^{\Gamma} - 1.$$

The two coincide thanks to the Euler–Poincaré characteristic of Γ.

Similarly, for $1 \leq i \leq \kappa$, let $\mathcal{C}_{2i-1}$ be the loop that does not bound any surface of Ω, and let $\mathcal{C}_{2i}$ be the one that bounds a surface Σ_i when considered in Ω (see Figure 7). For problem (9) we need to "reactivate" one of the two edges Π_{2i-1}, Π_{2i} excluded in problem (8) and precisely the one associated with the loop that bounds a surface when we pass from Γ to Ω. So, because $\Pi_{2i}^{*} =$ supp $(\pi \mathbf{p})$, the flux edges, $1 \leq i \leq \kappa$, with the vector $\mathbf{p}$ as in section 7.2, and denoting by $\mathcal{T}^{\text{int}}$ the usual spanning tree without loops, we have

$$\mathcal{S}^{\text{int}} = \left(\mathcal{T}^{\text{int}} \cup \bigcup_{i=1}^{\kappa} \{\Pi_{2i-1}\} \right),$$

and the set $\mathcal{U}_m^0$ is now defined as follows:

$$\tilde{\mathcal{U}}_m^0 = \left(\mathcal{E}_m^{\mathrm{int}} \cup \mathcal{E}_m^{\ell} \cup \mathcal{E}_m^{B} \cup \bigcup_{i=1}^{\kappa} \{\Pi_{2i}^*\} \right) \setminus (\mathcal{S}^{\mathrm{int}} \cup \mathcal{T}^{\ell}).$$

The proof of Proposition 7.2 for problem (9) is unchanged; the dimension of the approximation and approximated spaces is now

$$\begin{aligned}\dim\,[\mathcal{V}(\mathcal{S}^{\mathrm{int}}, \mathcal{T}^{\ell})] &= E_m - E_m^{\Gamma} + \kappa - (N_m - N_m^{\Gamma} - 1 + \kappa + 1),\\ \dim\,[H_0(\mathrm{div}_0, m)] &= F_m - F_m^{\Gamma} - (T_m - 1) - \kappa,\end{aligned}$$

and the two coincide thanks to the Euler–Poincaré characteristic of Ω and Γ.

Remark 7.3. Another existing strategy to deal with potential problems in nonsimply connected domains relies on the introduction of "cuts" in the domain. The big difficulty with this method is the construction of cuts and understanding where they should be introduced. Kotiuga [21] and coworkers have provided a correct definition of a cut, a constructive algorithm, and an implementation of it (see [18], for example).

The "belted tree" approach proposed in this paper allows us to achieve knowledge of the topological features of the considered domain if this is not given a priori. This knowledge is a preliminary step to the introduction of cuts.

8. Algorithmics and a simple example. From a practical point of view, the determination of the set $\mathcal{U}_m^0$ for simply connected domains is standard, and we refer to [28] for a procedure to construct a particular spanning tree.

This is not the case for the nonsimply connected case. The problem is now to find out the independent loops in order to select explicitly the flux edges previously introduced. We have, in particular, to select the loops that bound a surface when we pass from the boundary to the interior. This question can be summarized by saying that we look for generators of $H_1(m)$ starting from those of $H_1(m^{\Gamma})$.

In section 8.1, we present the algebraic tools, and we explain how to use them in section 8.2. The very first results for a torus are presented in section 8.3. We remark that only the computation of a basis for $H_1(m)$ is useful to our purpose of solving problem (24). In any case, a basis for $H_2(m)$ can be computed with the same tools, and at the end of section 8.2 we give a few indications of how to proceed with it. See [20] for another type of algorithm.

8.1. An integer $\mathcal{QR}$ factorization. In this section, we present a matrix decomposition to compute a set of generators of the homology groups of order $p = 1$ and 2 of $\Omega \subset \mathbb{R}^3$. The same algorithm has been used in [25] to detect mesh defects. The basic idea is to make an integer $\mathcal{QR}$ factorization of the matrices G^t, R^t, and D^t. Given a matrix $A \in \mathcal{M}(r, s)$, we compute a nonsingular unimodular matrix $\mathcal{Q}$ (i.e., $\det(\mathcal{Q}) = \pm 1$) and a permutation matrix $\mathcal{P}$ such that $\mathcal{R} = \mathcal{Q}\, A\, \mathcal{P}$ is upper triangular. As shown later, the two matrices $\mathcal{Q}$ and $\mathcal{P}$ are obtained as products of a certain number of local matrices $\mathcal{Q}_{i,j}$ and $\mathcal{P}_{i,j}$ and exhibit the row and column rank deficiency of A [8]. The key point of the algorithm is the following property [17]: given a matrix $A \in \mathcal{M}(r, s)$ with integer elements, we have

$$Z^r = \ker{(A^t)} \,\oplus\, \mathrm{range}\,(A), \qquad Z^s = \ker{(A)} \,\oplus\, \mathrm{range}\,(A^t). \tag{26}$$

To define $\mathcal{Q}$ and $\mathcal{P}$, we need two elementary operations. First is the transformation π_1 of a vector $v = (\epsilon_i, \epsilon_j)^t$ into the vector $\tilde{v} = (1, 0)^t$. To this purpose, let us introduce the elementary matrices

$$\mathcal{Q}^{\mathrm{el}}_{i,j} = \begin{pmatrix} \epsilon_i & 0 \\ -\epsilon_i & \epsilon_j \end{pmatrix}, \quad (\mathcal{Q}^{\mathrm{el}}_{i,j})^{-1} = \begin{pmatrix} \epsilon_i & 0 \\ \epsilon_j & \epsilon_j \end{pmatrix}$$

and the matrix

$$\mathcal{Q}_{i,j}(\ell, q) = \begin{cases} \delta_{\ell,q}, & \ell \neq i\,, j, \quad q \neq i\,, j, \\ \mathcal{Q}^{\mathrm{el}}_{i,j}(1,1), & \ell = i, \quad q = i, \\ \mathcal{Q}^{\mathrm{el}}_{i,j}(1,2), & \ell = i, \quad q = j, \\ \mathcal{Q}^{\mathrm{el}}_{i,j}(2,1), & \ell = j, \quad q = i, \\ \mathcal{Q}^{\mathrm{el}}_{i,j}(2,2), & \ell = j, \quad q = j. \end{cases}$$

In our case, $\epsilon_i^2 = 1$, and the vector $\tilde{v} = \pi_1(v) = \mathcal{Q}^{\mathrm{el}}_{i,j} v$. Second, we need the permutation π_2 of a vector's components, i.e., the transformation of a vector $v = (\epsilon_i, \epsilon_j)^t$ into the vector $\tilde{v} = (\epsilon_j, \epsilon_i)^t$. To this purpose, we have $\tilde{v} = \pi_2(v) = \mathcal{P}^{\mathrm{el}}_{i,j} v$, where $\mathcal{P}^{\mathrm{el}}_{i,j}$ is a permutation matrix; moreover, we introduce a matrix $\mathcal{P}_{i,j}$ defined similarly to $\mathcal{Q}_{i,j}$ (using $\mathcal{P}^{\mathrm{el}}_{i,j}$ instead of $\mathcal{Q}^{\mathrm{el}}_{i,j}$). We remark that $(\mathcal{P}^{\mathrm{el}}_{i,j})^{-1} = \mathcal{P}^{\mathrm{el}}_{i,j}$, owing to the fact that $\mathcal{P}^{\mathrm{el}}_{i,j}$ is a permutation matrix and that $(\mathcal{Q}_{i,j})^{-1}$ is defined as $\mathcal{Q}_{i,j}$ (using $(\mathcal{Q}^{\mathrm{el}}_{i,j})^{-1}$ instead of $\mathcal{Q}^{\mathrm{el}}_{i,j}$). In the following, I_r denotes the identity matrix of dimension $r > 0$. Now we describe the adopted procedure to build up $\mathcal{Q}$ and $\mathcal{P}$ for a given matrix $A \in \mathcal{M}(r, s)$.

Procedure. We set $\mathcal{Q} = \mathcal{Q}^0 \in \mathcal{M}(r, r)$, $\mathcal{P} = \mathcal{P}^0 \in \mathcal{M}(s, s)$. We loop on the column index j, $1 \leq j \leq s$:

1. We define $\mathcal{V}_j = \{i \,|\, \min\{j, r\} \leq i \leq \min\{s, r\},\, A(i, j) \neq 0\}$, and we put k equal to the cardinality of $\mathcal{V}_j$, i_1 equal to the smallest integer in $\mathcal{V}_j$, and i_2 equal to the smallest integer in $\mathcal{V}_j \setminus \{i_1\}$.
2. In case $k = 0$, let $\mathcal{P}_{j,z}$ be the matrix of the transformation π_2 that permutes the jth column of A with the zth one. The zth column is chosen to be the first column, starting from the last one in A, for which there exists a row index s such that $A(s, z) \neq 0$. If the index z exists, $\mathcal{P} \longleftarrow \mathcal{P}\,\mathcal{P}_{j,z}$, $A \longleftarrow A\,\mathcal{P}_{j,z}$, and we go back to step 1; otherwise we stop the procedure.
3. In case $k \neq 0$ but $A(j, j) = 0$, we apply a partial pivot strategy. Let $\mathcal{Q}_{j,i_1}$ be the matrix of the transformation π_2 that permutes the jth row with the i_1th one; then $\mathcal{Q} \longleftarrow \mathcal{Q}_{j,i_1}\,\mathcal{Q}$, $A \longleftarrow \mathcal{Q}_{j,i_1}\,A$, $i_1 \longleftarrow j$, and we go to step 4.
4. In case $k \geq 2$ and $A(j, j) \neq 0$, let $\mathcal{Q}^{\mathrm{el}}_{i_1,i_2}$ be the matrix of the transformation π_1 applied to the vector $(A(i_1, j), A(i_2, j))^t$, and let $\mathcal{Q}_{i_1,i_2}$ be the associated matrix; then $\mathcal{Q} \longleftarrow \mathcal{Q}_{i_1,i_2}\,\mathcal{Q}$, $A \longleftarrow \mathcal{Q}_{i_1,i_2}\,A$, and we go back to step 1.
5. In case $k = 1$ and $A(j, j) \neq 0$, then $j \longleftarrow j + 1$, and we go back to step 1.

Starting with $\mathcal{Q}^0 = I_r$ and $\mathcal{P}^0 = I_s$, at the end of the procedure, the matrix A has been replaced by $\mathcal{R}$, an upper triangular one. If this new matrix $\mathcal{R}$ does not contain zero rows, then $\dim[\mathrm{range}(\mathcal{R})] = r$. Otherwise, $\dim[\ker(\mathcal{R}^t)] = r - \dim[\mathrm{range}(\mathcal{R})]$. We remark that the procedure converges and its computational cost is similar to that of a $\mathcal{QR}$ decomposition by using Givens transformations.

8.2. Computation of homology group generators. Now, the question is how we can use the previous procedure to compute the generators of $H_p(m)$ for $p = 1$ and 2. To compute a set of generators for $H_1(m)$, we proceed as follows.

(i) We apply the procedure with $A = R^t$, $\mathcal{Q}^0 = I_{E_m}$, and $\mathcal{P}^0 = I_{F_m}$, and we get two invertible matrices $\mathcal{Q}_R$ and $\mathcal{P}_R$ such that $\mathcal{R}_R = \mathcal{Q}_R\,R^t\,\mathcal{P}_R$ is upper

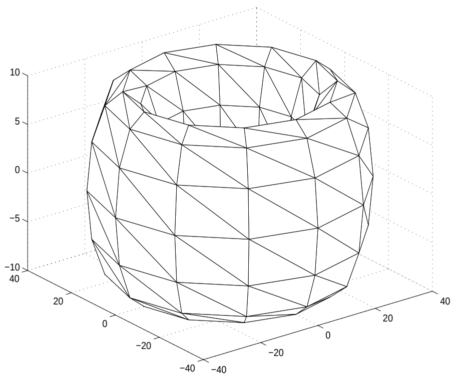


FIG. 8. *An example of surface discretization for the torus.*

triangular. The 1-cycles which bound a surface belong to the image of the matrix R^t that is also the image of $\mathcal{R}_R$.

(ii) We define $\tilde{G}^t = G^t \, \mathcal{Q}_R^{-1}$. In this way we make a change of basis for the 1-chains. Looking at $\tilde{G}^t$, we see immediately from the presence of n_c zero columns that the corresponding columns of $\mathcal{Q}_R^{-1}$ represent vectors that belong to the kernel of G^t. If $\dim[\mathrm{range}\,(R^t)] = \dim[\ker(G^t)]$, then any 1-cycle bounds. In the other case, we apply the procedure with $A = \tilde{G}^t$, $\mathcal{Q}^0 = I_{N_m}$, and $\mathcal{P}^0 = I_{E_m - n_c}$. We then obtain two invertible matrices $\mathcal{Q}_{\tilde{G}}$ and $\mathcal{P}_{\tilde{G}}$ such that $\mathcal{R}_{\tilde{G}} = \mathcal{Q}_{\tilde{G}} \, \tilde{G}^t \, \mathcal{P}_{\tilde{G}}$ is upper triangular.

(iii) The rows in $\mathcal{P}_{\tilde{G}}$, corresponding to zero rows in $\mathcal{R}_{\tilde{G}}$, represent the vectors that complete the kernel of G^t. In fact, we are looking for c such that $G^t \, c = 0$. This is equivalent to $\tilde{G}^t \, v = 0$, where v has zero in the first n_c components and, in the last $E_m - n_c$, any row in $\mathcal{P}_{\tilde{G}}$ corresponding to a new zero row in $\mathcal{R}_{\tilde{G}}$. Then the components of $c = \mathcal{Q}_R^{-1} \, v$ are the coefficients of a 1-chain generator of $H_1(m)$.

To determine the generators of $H_2(m)$, it is sufficient to perform parts (i), (ii), and (iii) with D^t at the place of R^t and R^t at the place of G^t.

8.3. Numerical results on the torus. As an application, we consider the case of a torus. We discretize it by means of a mesh m of 596 tetrahedra and 179 nodes. The discretization of Ω induces a discretization of the surface, denoted m^Γ, composed of 288 triangles and 144 nodes (see Figure 8).

We apply the procedure presented in section 8.2 to the matrices R^t and G^t of the surface Γ. At this point we have a basis for the 1-cycles of the mesh m^Γ that are not boundaries, i.e., a basis for $H_1(m^\Gamma)$ (see Figures 9 and 10). Note that the loops $\mathcal{C}_1$ and $\mathcal{C}_2$ run around the two "holes" of the torus surface.

We want now to make evident the loop that bounds a surface when considered as 1-cycles of the mesh m in Ω. In other terms, in the set of the two computed generators for $H_1(m^\Gamma)$, we look for the one that generates $H_1(m)$.

Let c be an element of $H_1(m^\Gamma)$ and v_c the vector whose components are the coefficients of c for the edges $e \in \mathcal{E}_m^\Gamma$ and zero for $e \in \mathcal{E}_m \setminus \mathcal{E}_m^\Gamma$. We apply the procedure to the matrix R^t associated to the mesh m in Ω, and we transform it into an upper triangular matrix of the form $\mathcal{R}_R = \mathcal{Q}_R \, R^t \, \mathcal{P}_R$. Finally, we consider the vector $w_c = \mathcal{Q}_R^{-1} \, v_c$. If $w_c = 0$, then the 1-chain c is homologous to zero in Ω (see

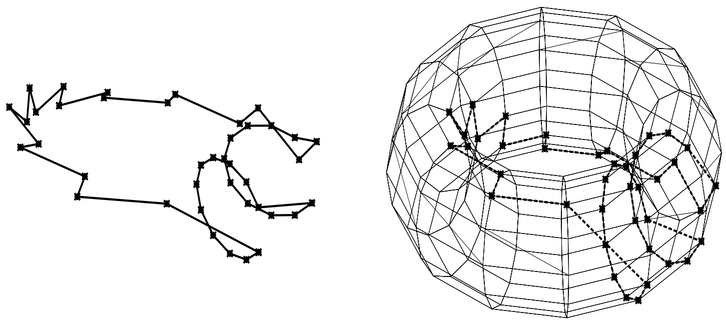


FIG. 9. *Wireframe representation of the loop $\mathcal{C}_2$, one of the two generators of the first homology group of the torus surface.*

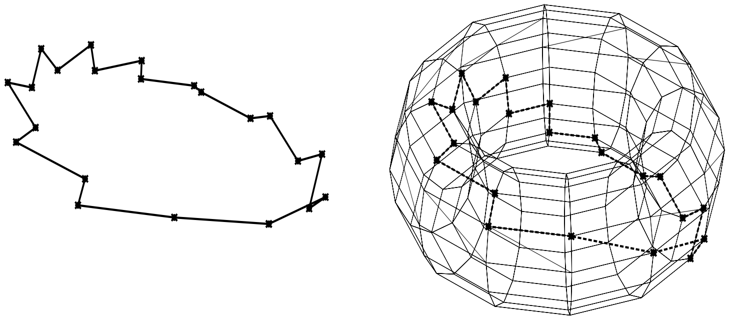


FIG. 10. *Wireframe representation of the loop $\mathcal{C}_1$, one of the two generators of the first homology group of the torus surface.*

Figure 9); if $w_c \neq 0$, then the 1-chain c is also a generator of $H_1(m)$ (see Figure 10). In our case, the loop $\mathcal{C}_1$ is detected to be the element of a basis for $H_1(m)$. Note that the adopted procedure can be optimized in several ways, such as, for example, by looking for those generators with the minimum number of edges or faces, by applying the procedure to suitable portions of the whole meshes, and by using a suitable data format [13].

The two loops $\mathcal{C}_1$ and $\mathcal{C}_2$ are used as follows. For the lifting of the boundary condition, the set $\mathcal{S}^\Gamma$ of "null degrees of freedom" is a belted spanning tree obtained by adding to the standard spanning tree $\mathcal{T}^\Gamma$ two edges, $\Pi_1 \in \mathcal{C}_1$ and $\Pi_2 \in \mathcal{C}_2$, chosen in an arbitrary way, with $\Pi_1 \neq \Pi_2$. These two edges correspond to nonzero components in the vector w_c of the coefficients of the 1-chains c that generate $H_1(m^\Gamma)$.

For the interior problem, the set of active edges $\tilde{\mathcal{U}}_m^0$ is obtained by adding a flux edge Π_2^* to the set $\mathcal{U}_m^0$, the latter selected with the algorithm devoted to a simply connected domain. We conclude by remarking that Π_2^*, the edge that has to be reactivated when passing from the boundary to the interior to solve problem (24), can be chosen to be any edge $e \in \mathcal{E}_m^\Gamma$ for which the corresponding coefficient in the 1-chain c is nonzero.

9. Conclusions. In this paper, we have studied the representation of a solenoidal vector field in terms of a vector potential. The considered problem has been split into two parts—a lifting problem of the boundary condition and an internal problem with homogeneous boundary conditions.

The edge elements are a natural tool to compute vector potentials. On the other hand, the gauge condition, which is necessary to ensure the potential uniqueness, is taken into account in a fully discrete way and expressed in terms of a suitable set of active mesh edges (active in the sense that the associated degree of freedom is a priori different from zero).

According to the authors' knowledge, the problem of the computation of the vector potential is well understood for three-dimensional bounded domains which are connected and simply connected, even with a connected but nonsimply connected boundary. Here, we have presented a method to compute the vector potential for three-dimensional bounded domains which are connected but nonsimply connected, with a connected boundary. The case of three-dimensional bounded domains which are nonsimply connected with a nonconnected boundary has not been considered in the present work.

REFERENCES

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

[2] C. Amrouche, C. Bernardi, M. Dauge, and V. Girault, *Vector potentials in three-dimensional nonsmooth domains*, Math. Methods Appl. Sci., 21 (1998), pp. 823–864.

[3] A. Bendali, J. M. Dominguez, and S. Gallic, *A variational approach for the vector potential formulation of the Stokes and Navier-Stokes problems in three dimensional domains*, J. Math. Anal. Appl., 107 (1985), pp. 537–560.

[4] C. Bernardi, *Méthodes d'éléments finis mixtes pour les équations de Navier-Stockes*, Thèse du 3ème cycle, University of Paris, Paris, 1979.

[5] C. Bernardi, M. Dauge, and Y. Maday, *Compatibilité de traces aux arêtes et coins d'un polyèdre*, C. R. Acad. Sci. Paris Sér. I Math., 331 (2000), pp. 679–684.

[6] A. Bossavit, *Computational Electromagnetism: Variational Formulations, Complementarity, Edge Elements*, Academic Press, New York, 1998.

[7] A. Buffa and P. Ciarlet Jr., *On traces for functional spaces related to Maxwell's equation* II: *Hodge decompositions on the boundary of Lipschitz polyhedra and applications*, Math. Methods Appl. Sci., 24 (2001), pp. 31–48.

[8] T. F. Chan, *Rank revealing QR factorizations*, Linear Algebra Appl., 88/89 (1987), pp. 67–82.

[9] Y. Choquet-Bruhat, *Géometrie Différentielle et Systèmes Extérieurs*, Dunod, Paris, 1968.

[10] F. Dubois, *Discrete vector potential representation of a divergence-free vector field in three-dimensional domains: Numerical analysis of a model problem*, SIAM J. Numer. Anal., 27 (1990), pp. 1103–1141.

[11] F. Dubois and F. Rapetti, *Du tourbillon au champ de vitesse*, in Proceedings of the Workshop at the Conservatoire des Arts et Métiers de Paris on Modèles fluides et représentation en toubillons, Vol. 1, Paris, France, 2000, pp. 127–153; MATAPLI, 65 (2001), pp. 87–88.

[12] C. Foias and R. Temam, *Remarques sur les équations de Navier-Stokes stationnaires et les phénomènes successifs de bifurcation*, Ann. Scuola Norm. Sup. Pisa Cl. Sci., 5 (1978), pp. 29–63.

[13] L. Formaggia, *personal communication*, Ecole Polytechnique Féderale de Lausanne, Lausanne, France, 2002.

[14] P. J. Giblin, *Graphs, Surfaces and Homology. An Introduction to Algebraic Topology*, Chapman and Hall Mathematics Series, Chapman and Hall, London, 1977.

[15] V. Girault and P. A. Raviart, *Finite element methods for Navier-Stokes equations: Theory and Applications*, Springer-Verlag, New York, 1986.

[16] J. Glimm, *personal communication*, Department of Applied Mathematics and Statistics, University of Stony Brook, Stony Brook, NY, 1986.

[17] G. H. Golub and C. F. Van Loan, *Matrix Computation*, John Hopkins University Press, Baltimore, MD, 1983.

[18] P. Gross and P. R. Kotiuga, *Finite element-based algorithms to make cuts for magnetic scalar potentials: Topological constraints and computational complexity*, in Geometric Methods for Computational Electromagnetics, Progress in Electromagnetic Research Ser. 32, F. L. Teixeira, ed., EMW, Cambridge, MA, 2001, pp. 207–245.

[19] F. Hecht, *Construction d'une base de fonctions $\mathbb{P}_1$ non conforme à divergence nulle dans $\mathbb{R}^3$*, RAIRO Anal. Numer., 15 (1980), pp. 315–341.

[20] R. Hiptmair and J. Ostrowski, *Generators of $H_1(\Gamma_h, Z)$ for Triangulated Surfaces: Construction and Classification*, Sonderforschungsbereich 382, Report 160, Universität Tübingen, Tübingen, Germany, 2001.

[21] P. R. Kotiuga, *An algorithm to make cuts for magnetic scalar potentials in tetrahedral meshes based on the finite element method*, IEEE Trans. Magnetics, 25 (1989), pp. 4129–4131.

[22] C. Mattiussi, *An analysis of finite volume, finite element, and finite difference methods using some concepts from algebraic topology*, J. Comput. Phys., 133 (1997), pp. 289–309.

[23] J. C. Nédélec, *Mixed finite elements in $\mathbb{R}^3$*, Numer. Math., 35 (1980), pp. 315–341.

[24] J. C. Nédélec, *A new family of mixed finite elements in $\mathbb{R}^3$*, Numer. Math., 35 (1986), pp. 57–81.

[25] F. Rapetti, F. Dubois, and A. Bossavit, *Integer matrix factorization for mesh defects detection*, C. R. Acad. Sci. Paris Sér. I Math., 334 (2002), pp. 717–720.

[26] S. M. Rao, D. R. Wilton, and A. W. Glisson, *Electromagnetic scattering by surfaces of arbitrary shape*, IEEE Trans. Antennas and Propagation, 30 (1982), pp. 409–418.

[27] Z. Ren and A. Razek, *Boundary edge elements and spanning tree technique in three-dimensional electromagnetic field computation*, Internat. J. Numer. Methods Engrg., 36 (1993), pp. 2877–2893.

[28] F.-X. Roux, lecture, University of Paris, Paris, France, 1984.

[29] J. Stillwell, *Classical Topology and Combinatorial Group Theory*, Grad. Texts in Math. 72, Springer-Verlag, New York, 1993.

[30] F. L. Teixeira and W. C. Chew, *Lattice electromagnetic theory from a topological viewpoint*, J. Math. Phys., 40 (1999), pp. 169–187.

[31] H. Whitney, *Geometric Integration Theory*, Princeton University Press, Princeton, NJ, 1957.

SIAM J. NUMER. ANAL.
Vol. 41, No. 4, pp. 1528–1542

QUASI OPTIMALITY OF THE SUPG METHOD FOR THE ONE-DIMENSIONAL ADVECTION-DIFFUSION PROBLEM[*]

GIANCARLO SANGALLI[†]

Abstract. In this paper, we propose quasi-optimal error estimates, in various norms, for the *streamline-upwind Petrov–Galerkin* (SUPG) method applied to the linear one-dimensional advection-diffusion problem. We follow the classical argument due to Babuška and Brezzi; therefore, the goal of this work is the proof of the inf-sup and of the continuity conditions for the bilinear stabilized variational form, with respect to suitable norms. These norms are suggested by our previous work [G. Sangalli, Technical report 1221, IAN-CNR, Pavia, Italy, 2001; *Numer. Math.*, to appear], in which we analyze the continuous multidimensional advection-diffusion operator. We obtain these results by means of function space interpolation.

Key words. convection-diffusion, finite element method, inf-sup condition

AMS subject classifications. 65G99, 65N30, 76R99

DOI. 10.1137/S0036142902411690

1. Introduction.

In this work, we shall consider the one-dimensional advection-diffusion operator

$$\mathcal{L}_\varepsilon w := -\varepsilon w'' + w' \tag{1.1}$$

and, given a source term f, the related Dirichlet homogeneous boundary value problem

$$\begin{cases} \mathcal{L}_\varepsilon u = f & \text{in } (0,1), \\ u(0) = 0, \\ u(1) = 0. \end{cases} \tag{1.2}$$

We assume the diffusion parameter ε to be positive; when it is small, i.e., in the advection-dominated regime, (1.2) represents one of the simplest examples of singularly perturbed boundary value differential problems. We think of it as a prototype of more general problems, where a skew-symmetric operator (represented by the first-order derivative) is perturbed by a symmetric operator of higher order (the second-order derivative in the example).

The associated variational problem reads

$$\begin{cases} \text{find } u \in H^1_0 \equiv H^1_0(0,1) \text{ such that} \\ a_\varepsilon(u,v) = {}_{H^{-1}}\langle f, v\rangle_{H^1_0} \qquad \forall v \in H^1_0, \end{cases} \tag{1.3}$$

where $a_\varepsilon(w,v) := \varepsilon \int_0^1 w'(x)v'(x)\,dx + \int_0^1 w'(x)v(x)\,dx$ and f is assumed to be in H^{-1}. This problem fits into the Lax–Milgram framework, but its solution, when ε is small, depends on the source term f in a very sensitive way with respect to the usual norms on H^1_0 and H^{-1}; actually,

$$\|\mathcal{L}_\varepsilon^{-1}\|_{L(H^{-1},H^1_0)} := \sup_{w\in H^1_0} \frac{\|w\|_{H^1_0}}{\|\mathcal{L}_\varepsilon w\|_{H^{-1}}} \approx \varepsilon^{-1}.$$

[*] Received by the editors July 20, 2002; accepted for publication (in revised form) March 25, 2003; published electronically September 17, 2003.
http://www.siam.org/journals/sinum/41-4/41169.html

[†] Istituto di Matematica Applicata e Tecnologie Informatiche, Via Ferrata 1, 27100 Pavia, Italy (sangalli@imati.cnr.it).

Nevertheless, problem (1.2) is well posed for any $\varepsilon > 0$, and indeed in [14] and [13] we defined suitable norms $\|\cdot\|_W$ and $\|\cdot\|_V$ such that both the continuity and the inf-sup conditions

$$a_\varepsilon(w, v) \le \kappa \|w\|_W \|v\|_V \quad \forall w \in H_0^1, \forall v \in H_0^1, \tag{1.4}$$

$$\inf_{w \in H_0^1} \sup_{v \in H_0^1} \frac{a_\varepsilon(w, v)}{\|w\|_W \|v\|_V} \ge \gamma > 0, \tag{1.5}$$

hold true with κ and γ independent of ε. The results in this paper are based on the approach proposed in [13]; actually the analysis of [13] is more general since it deals with the multidimensional operator, where the advection term has an anisotropic structure. In section 2, we shall specialize the results of [13] to the simpler one-dimensional problem (1.2).

It is well known that the standard Galerkin numerical method, when applied to (1.2), is unstable (see, e.g., [11]). The most popular methods for (1.2) are actually the *streamline-upwind Petrov–Galerkin* (SUPG) method and its variants—e.g., the *Galerkin least squares* (GaLS) method—introduced by Hughes and coworkers (see [7] and [9]) in the eighties. In this work, we shall consider the one-dimensional version of the SUPG method, whose detailed presentation is postponed to section 2. We recall now only that even though these methods are quite satisfactory for practical situations, their error analysis does not fit into the classical theory due to Babuška and Brezzi (see [1] and [3] and (2.16)–(2.18) in what follows); as a result, it is usually very hard to prove that these methods are *quasi-optimal*, namely, to show that the their numerical solution u_h is close to the exact solution u as the best fit of u in the trial space W_h (up to a multiplicative constant C independent of ε and with respect to a suitable norm $\|\cdot\|$):

$$\|u - u_h\| \le C \inf_{w_h \in W_h} \|u - w_h\|.$$

More recent numerical methods for the advection-diffusion problem are, among others, the *residual-free bubbles* finite element method (*FEM*) (proposed in [6] and analyzed in [4], [5], [8], and [12]) and the method with *negative-order stabilization* (see [2]). Those methods are closely related to the SUPG method—in some cases they lead to the same numerical algorithm—but they actually improve the theoretical understanding of this subject; for our purposes here, we note only that those recent analyses are, roughly speaking, *close* to the ideal Babuška–Brezzi framework, and the methods are proved to be *close* to exhibiting the quasi-optimal behavior.

In this work, we actually prove a family of quasi-optimal error estimates for the SUPG method for solving (1.2). We apply the general theory stated in section 2, by showing that the method verifies the continuity and inf-sup conditions (2.16)–(2.17) with respect to suitable norms, by means of function space interpolation tools. In particular, we show that the method is quasi-optimal (see (3.37)) with respect to a norm whose part independent of ε is of differentiability-order 1/2, which is in accordance with [2], [5], and [13].

We are restricting this analysis to the one-dimensional problem because we are not able, at the present time, to deal with the anisotropic structure of the convection term from the numerical point of view; on the other hand, we will not exploit other special properties of the one-dimensional operator. We refer to section 4 for a discussion on further extensions of our approach.

The outline of the paper is as follows. In section 2, we present the notation and assumptions, recall the results of [13], and specialize and extend them to the one-dimensional case. In section 3, we develop the error analysis for the SUPG method in the present setting. Finally, in section 4, we give some comments about the results proved and outline possible extensions.

2. Preliminaries. We denote by $L^2 \equiv L^2(0,1)$ the usual Lebesgue space endowed with the norm $\|\cdot\|_{L^2}$, by $L^2_0 \equiv L^2_0(0,1)$ its subset containing zero mean-value functions, and by $\Pi_0 : L^2 \to L^2_0$ the L^2-projection onto L^2_0; we also denote by $\overline{w}$ the mean value of a generic function $w \in L^2$ so that $w = \Pi_0 w + \overline{w}$. Moreover, $H^1 \equiv H^1(0,1)$ is the usual Sobolev space endowed with the norm $\|\cdot\|_{H^1}$, and seminorm $|\cdot|_{H^1}$; $H^1_{\#}$ denotes its subspace of functions w such that $w(0) = w(1)$, while H^1_0 denotes the subspace of functions vanishing at 0 and 1, endowed with the norm $|\cdot|_{H^1}$. Finally, $H^{-1} \equiv H^{-1}(0,1) := (H^1_0)^*$ denotes the dual space of H^1_0 endowed with the dual norm $\|\cdot\|_{H^{-1}}$ and the usual pairing $\langle\cdot,\cdot\rangle \equiv {}_{H^{-1}}\langle\cdot,\cdot\rangle_{H^1_0}$; the dual (norm, space) is always denoted by the superscripted star. We shall make use of the interpolation theory of function spaces; more specifically, we shall use the *K-method*, and we refer to [15] for its definition, notation, and properties.

In what follows, C denotes a generic constant whose value, possibly different at various occurrences, does not depend on any other mathematical quantity appearing in the analysis (e.g., ε, θ, p, h, u, w, f, ϕ). We also adopt the notational convention

$$\begin{aligned} \alpha \preceq \beta \quad &\iff \quad \alpha \le C\beta, \\ \alpha \simeq \beta \quad &\iff \quad \alpha \preceq \beta \text{ and } \beta \preceq \alpha. \end{aligned}$$

We now revise the analysis proposed in [13] and specialize it to the one-dimensional case. Following [13], we define

$$\begin{aligned} \|w\|_{A_0} &:= \varepsilon|w|_{H^1} + \|\Pi_0 w\|_{L^2} \qquad && \forall w \in A_0 := H^1_0, \\ \|w\|_{A_1} &:= |w|_{H^1} && \forall w \in A_1 := H^1_0, \end{aligned} \tag{2.1}$$

where we have used $\|\Pi_0 w\|_{L^2}$ instead of the equivalent norm $\|w'\|_{H^{-1}}$. Therefore, one has the equivalence between $\|w\|_{A_0}$ and $\|\mathcal{L}_\varepsilon w\|_{A_1^*}$; i.e.,

$$\varepsilon|w|_{H^1} + \|\Pi_0 w\|_{L^2} \simeq \sup_{v \in H^1_0} \frac{a_\varepsilon(w,v)}{|v|_{H^1}} \quad \forall w \in H^1_0. \tag{2.2}$$

One half of (2.2)—the continuity of $\mathcal{L}_\varepsilon$—is obvious, while the other half actually follows from the coercivity $\varepsilon|w|^2_{H^1} \preceq a_\varepsilon(w,w)$. By means of a duality argument, we obtain from (2.2) the other estimate

$$|w|_{H^1} \simeq \sup_{v \in H^1_0} \frac{a_\varepsilon(w,v)}{\varepsilon|v|_{H^1} + \|\Pi_0 v\|_{L^2}} \quad \forall w \in H^1_0; \tag{2.3}$$

i.e., $\|w\|_{A_1}$ and $\|\mathcal{L}_\varepsilon w\|_{A_0^*}$ are equivalent.

Both (2.2) and (2.3) state that $\mathcal{L}_\varepsilon$ is an isomorphism uniformly with respect to ε; the dependence on ε of the operator has been included in the norms themselves.

We briefly comment on (2.3): it allows control of the H^1_0 norm of the solution u of (1.2) in terms of the source term f, despite the presence of the boundary layer near $x = 1$. It is due to the structure of $\|\cdot\|_{A_0^*}$; when $f \in L^2_0$, then $\|f\|_{A_0^*} \le \|f\|_{L^2}$, and actually there is no boundary layer; otherwise, if $f \in L^2$ has a nonzero mean value, then $\|f\|_{A_0^*}$ behaves asymptotically as $\varepsilon^{-1/2}$ for $\varepsilon \to 0$, and accordingly, the presence of a thin layer on the related solution u makes $|u|_{H^1}$ behave in the same way.

We can infer from (2.2)–(2.3) a family of intermediate estimates: given θ and p with $0 < \theta < 1$, $1 \leq p \leq +\infty$, and denoting by p' the conjugate of p, i.e., $1/p + 1/p' = 1$, by means of the interpolation theory we have (see [13])

$$\varepsilon^{1-\theta}|w|_{H^1} + \|w'\|_{(H^{-1},L_0^2)_{\theta,p}} \simeq \sup_{v\in H_0^1} \frac{a_\varepsilon(w,v)}{\varepsilon^{\theta}|v|_{H^1} + \|v'\|_{(H^{-1},L_0^2)_{1-\theta,p'}}} \quad \forall w \in H_0^1, \tag{2.4}$$

where we have also made use of the equivalence

$$\|w\|_{(A_0,A_1)_{\theta,p}} \simeq \varepsilon^{1-\theta}|w|_{H^1} + \|w'\|_{(H^{-1},L_0^2)_{\theta,p}}, \tag{2.5}$$

that is, the one-dimensional counterpart of [13, Proposition 2]. Condition (2.4) brings our model problem into the framework (1.4)–(1.5).

We also proved the Poincaré-like estimate

$$\|w\|_{L^2} \preceq \|w'\|_{(H^{-1},L_0^2)_{1/2,1}} \qquad \forall w \in H_0^1 \tag{2.6}$$

and, as a consequence,

$$\|w\|_{L^2} \preceq \|w\|_{(A_0,A_1)_{\theta,p}} \, \forall w \in H_0^1 \quad \Leftrightarrow \quad \theta > 1/2 \text{ or } (\theta,p) = (1/2,1), \tag{2.7}$$

$$\|\phi\|_{(A_0,A_1)_{\theta,p}^*} \preceq \|\phi\|_{L^2} \, \forall \phi \in H^{-1} \quad \Leftrightarrow \quad \theta > 1/2 \text{ or } (\theta,p) = (1/2,1). \tag{2.8}$$

Finally, we show the relation between the fractional-order norm appearing in the equivalence (2.4) and a more usual Besov norm. We focus our attention on the case $\theta = 1/2$, which will be of special interest in what follows; the case of a generic θ is similar but more technical (in order to obtain the optimal dependence on θ).

Proposition 2.1. *For $1 \leq p \leq +\infty$, we have*

$$(L^2, H^1_\#)_{1/2,p} = \{w \in L^2 | w' \in (H^{-1}, L_0^2)_{1/2,p}\}$$

and also

$$\|w\|_{(L^2,H^1_\#)_{1/2,p}} \simeq \|w\|_{L^2} + \|w'\|_{(H^{-1},L_0^2)_{1/2,p}} \tag{2.9}$$

for any $w \in (L^2, H^1_\#)_{1/2,p}$.

Proof. Let w be a generic function in L^2 and $p \neq +\infty$; we have, by definition and by the triangle inequality,

$$\begin{aligned}
\|w\|_{(L^2,H^1_\#)_{1/2,p}} &\leq \left[\int_0^{+\infty} \left(t^{-1/2}\|w_0(t)\|_{L^2} + t^{1/2}\|w_1(t)\|_{H^1}\right)^p \frac{dt}{t}\right]^{1/p} \\
&\leq \left[\int_0^{1} \left(t^{-1/2}\|w_0(t)\|_{L^2} + t^{1/2}\|w_1(t)\|_{H^1}\right)^p \frac{dt}{t}\right]^{1/p} \\
&\quad + \left[\int_1^{+\infty} \left(t^{-1/2}\|w_0(t)\|_{L^2} + t^{1/2}\|w_1(t)\|_{H^1}\right)^p \frac{dt}{t}\right]^{1/p} \\
&= \mathrm{I} + \mathrm{II}
\end{aligned}$$

for any $w_0(t)$ and $w_1(t)$ with $w = w_0(t) + w_1(t)$, $w_0(t) \in L^2$, $w_1(t) \in H^1_\#$, and $0 < t < +\infty$; I and II actually depend on $w_0(t)$ and $w_1(t)$. Moreover, we assume, for $0 < t < 1$, that $w_0(t) \in L^2_0$, yielding $\|w_0(t)\|_{L^2} = \|w_0'(t)\|_{H^{-1}}$. Recalling the notation $w_1(t) = \overline{w_1(t)} + \Pi_0 w_1(t)$, where $\overline{w_1(t)} = \overline{w}$, we use the continuity of the mean value in L^2 for $\overline{w}$ and the Bramble–Hilbert lemma for $\Pi_0 w_1(t)$ to get

$$\begin{aligned}\|w_1(t)\|_{H^1} &\le \|\overline{w_1(t)}\|_{H^1} + \|\Pi_0 w_1(t)\|_{H^1}\\ &\preceq \|\overline{w_1(t)}\|_{L^2} + |\Pi_0 w_1(t)|_{H^1}\\ &\preceq \|w\|_{L^2} + \|w_1'(t)\|_{L^2}.\end{aligned}$$

Therefore, we have

$$\begin{aligned}\mathrm{I} &\le \left[\int_0^1 \left(t^{-1/2}\|w_0'(t)\|_{H^{-1}} + t^{1/2}\|w_1'(t)\|_{L^2}\right)^p \frac{dt}{t}\right]^{1/p}\\ &\quad + \left[\int_0^1 \left(t^{1/2}\|w\|_{L^2}\right)^p \frac{dt}{t}\right]^{1/p}\\ &\preceq \left[\int_0^1 \left(t^{-1/2}\|w_0'(t)\|_{H^{-1}} + t^{1/2}\|w_1'(t)\|_{L^2}\right)^p \frac{dt}{t}\right]^{1/p} + \|w\|_{L^2}.\end{aligned} \tag{2.10}$$

Now the key point is that, for any decomposition $w' = \phi_0(t) + \phi_1(t)$, $\phi_0(t) \in H^{-1}$, $\phi_1(t) \in L^2_0$ on $0 < t < 1$, we can define $w_0(t)$ and $w_1(t)$ that satisfy $w_0'(t) = \phi_0(t)$, $w_1'(t) = \phi_1(t)$, and all the conditions given above: simply take, for any $0 < t < 1$, $w_1(t)$ as the primitive of $\phi_1(t)$ with $\overline{w_1(t)} = \overline{w}$ and $w_0(t) = w - w_1(t)$. In particular, $\phi_1(t) \in L^2_0$ yields $w_1(t) \in H^1_\#$. Therefore, we can rewrite (2.10) in terms of $\phi_0(t)$ and $\phi_1(t)$,

$$\mathrm{I} \preceq \left[\int_0^1 \left(t^{-1/2}\|\phi_0(t)\|_{H^{-1}} + t^{1/2}\|\phi_1(t)\|_{L^2}\right)^p \frac{dt}{t}\right]^{1/p} + \|w\|_{L^2}, \tag{2.11}$$

and take the infimum with respect to $\phi_0(t)$ and $\phi_1(t)$, obtaining

$$\mathrm{I} \preceq \|w'\|_{(H^{-1},L^2_0)_{1/2,p}} + \|w\|_{L^2}. \tag{2.12}$$

Otherwise, taking $w_0(t) = w$ and $w_1(t) = 0$ for $1 < t < +\infty$, we have

$$\begin{aligned}\mathrm{II} &\le \left[\int_1^{+\infty} \left(t^{-1/2}\|w\|_{L^2}\right)^p \frac{dt}{t}\right]^{1/p}\\ &\le \frac{2}{p}\|w\|_{L^2}.\end{aligned} \tag{2.13}$$

Therefore, (2.12) and (2.13) yield $\|w\|_{(L^2,H^1_\#)_{1/2,p}} \preceq \|w\|_{L^2} + \|w'\|_{(H^{-1},L^2_0)_{1/2,p}}$ and the inclusion $\{w \in L^2 | w' \in (H^{-1}, L^2_0)_{1/2,p}\} \subset (L^2, H^1_\#)_{1/2,p}$. With obvious modification, one could deal with the case $p = +\infty$.

The remaining part, i.e., the inclusion

$$(L^2, H^1_\#)_{1/2,p} \subset \left\{w \in L^2 | w' \in (H^{-1}, L^2_0)_{1/2,p}\right\}$$

and the related estimate, is given by the interpolation theorem [15, section 1.3.3 (a)] since the derivative operator is both continuous from $H^1_\#$ into L^2_0 and from L^2 into H^{-1}. □

We recall that $(L^2, H^1_\#)_{1/2,p}$ is the space of functions whose periodic extension, say, on $(-1,2)$, belongs to $B^{1/2}_{2,p}(-1,2)$.

It is useful in what follows to notice that (2.6) and (2.9) give a variant of (2.5), as stated in the following corollary.

Corollary 2.2. *We have*

$$\|w\|_{(A_0,A_1)_{1/2,1}} \simeq \varepsilon^{1/2}|w|_{H^1} + \|w\|_{(L^2,H^1_\#)_{1/2,1}} \quad \forall w \in H^1_0. \tag{2.14}$$

Now we turn our attention to the numerical solution of (1.2). We briefly recall the general error estimation theory, due to Babuška and Brezzi (see [10, Proposition 5.5.1]). Let u be the solution of (1.2); the finite element formulation reads

$$\begin{cases} \text{find } u_h \in W_h \text{ such that} \\ a_{\varepsilon,h}(u_h, v_h) = \langle f_h, v_h\rangle \qquad \forall v_h \in V_h, \end{cases} \tag{2.15}$$

where the spaces W_h and V_h are finite-dimensional subsets of H^1_0, while $a_{\varepsilon,h}$ and f_h give a consistent discretization of the continuous problem, namely, $a_{\varepsilon,h}(u, v_h) = \langle f_h, v_h\rangle$ for all $v_h \in V_h$. If there exist constants $\widetilde{\kappa} < +\infty$ and $\widetilde{\gamma} > 0$, independent of ε and h, such that

$$a_{\varepsilon,h}(u - w_h, v_h) \le \widetilde{\kappa}\|u - w_h\|_{W,h}\|v_h\|_{V_h} \quad \forall w_h \in W_h, \forall v_h \in V_h, \tag{2.16}$$

and

$$\inf_{w_h\in W_h}\sup_{v_h\in V_h} \frac{a_{\varepsilon,h}(w_h, v_h)}{\|w_h\|_{W,h}\|v_h\|_{V_h}} \ge \widetilde{\gamma}, \tag{2.17}$$

then the method is *quasi-optimal*:

$$\|u - u_h\|_{W,h} \le (\widetilde{\kappa}\widetilde{\gamma}^{-1} + 1)\inf_{w_h\in W_h}\|u - w_h\|_{W,h}. \tag{2.18}$$

The notation $\|\cdot\|_{W,h}$ refers to a norm on the space W (we postpone the definition of W to (3.20)), which can depend on the discretization. Conditions (2.16)–(2.17) are the analogues of (1.4)–(1.5), and this motivates our interest in (2.4). In order to verify them, we shall look for norms $\|\cdot\|_{W,h}$ and $\|\cdot\|_{V_h}$ that are the discrete counterparts of the ones in (2.4).

We shall consider the very simple case of a uniform subdivision of $(0,1)$ into N open elements T_i of size $h = N^{-1}$,

$$T_i \equiv T_{i,h} := \{x : (i-1)h < x < ih\} \qquad \forall i = 1, 2, \ldots, N, \tag{2.19}$$

and the corresponding space of continuous piecewise linear elements,

$$W_h \equiv V_h := \left\{ \begin{matrix} v \in H^1_0 : v|_{T_i} \text{ is affine} \\ \forall i = 1, \ldots, N \end{matrix} \right\}; \tag{2.20}$$

the SUPG method, proposed by Hughes and coworkers in [7], adds a *weighted residual stabilization* to the continuous variational problem:

$$\begin{aligned} a_{\varepsilon,h}(w, v_h) &:= a_\varepsilon(w, v_h) + \sum_{i=1}^{N} \tau \int_{T_i} (\mathcal{L}_\varepsilon w)(x)\, v_h'(x)\, dx \\ \langle f_h, v_h\rangle &:= \langle f, v_h\rangle + \sum_{i=1}^{N} \tau \int_{T_i} f(x)\, v_h'(x)\, dx. \end{aligned} \tag{2.21}$$

This definition requires both f and w to be regular in the interior of the elements, which is not restrictive for applications. The amount of *streamline*[1] *diffusion* τ is a parameter of the method, and its value is a relevant point. It could depend on ε and h; here we assume that the problem is advection-dominated, namely, $\varepsilon \leq h$, and therefore a usual assumption is that

$$\tau \simeq h. \tag{2.22}$$

We shall return to a more detailed discussion of the optimal value of τ in section 4.

3. Main results. This section is devoted to the error analysis of the SUPG method in the framework (2.16)–(2.18).

First, we analyze the inf-sup condition (2.17).

LEMMA 3.1. *The SUPG method* (2.19)–(2.22) *satisfies the estimates*

$$\varepsilon|w_h|_{H^1} + \|\Pi_0 w_h\|_{L^2} \preceq \sup_{v_h \in W_h} \frac{a_{\varepsilon,h}(w_h, v_h)}{|v_h|_{H^1}} \quad \forall w_h \in W_h, \tag{3.1}$$

$$|w_h|_{H^1} \preceq \sup_{v_h \in W_h} \frac{a_{\varepsilon,h}(w_h, v_h)}{\varepsilon|v_h|_{H^1} + \|\Pi_0 v_h\|_{L^2}} \quad \forall w_h \in W_h. \tag{3.2}$$

Proof. First, recall that any $w_h \in W_h$ is piecewise linear; therefore, the higher-order term in (2.21) vanishes:

$$a_{\varepsilon,h}(w_h, v_h) = (\varepsilon + \tau) \int_0^1 w_h'(x) v_h'(x)\, dx + \int_0^1 w_h'(x) v_h(x)\, dx \quad \forall v_h \in V_h.$$

Thanks to the coercivity of $a_{\varepsilon,h}$,

$$\varepsilon|w_h|_{H^1}^2 \leq a_{\varepsilon,h}(w_h, w_h), \tag{3.3}$$

we have immediately

$$\varepsilon|w_h|_{H^1} \preceq \sup_{v_h \in W_h} \frac{a_{\varepsilon,h}(w_h, v_h)}{|v_h|_{H^1}}. \tag{3.4}$$

We have therefore to prove that

$$\|\Pi_0 w_h\|_{L^2} \preceq \sup_{v_h \in W_h} \frac{a_{\varepsilon,h}(w_h, v_h)}{|v_h|_{H^1}}. \tag{3.5}$$

In order to do this, we shall define $\widetilde{w}_h \in W_h$, depending on w_h, such that

$$a_{\varepsilon,h}(w_h, \widetilde{w}_h) = \|\Pi_0 w_h\|_{L^2}^2, \tag{3.6}$$

$$|\widetilde{w}_h|_{H^1} \preceq \|\Pi_0 w_h\|_{L^2}; \tag{3.7}$$

such a $\widetilde{w}_h$ is the solution of the discrete variational problem

$$a_{\varepsilon,h}(v_h, \widetilde{w}_h) = \int_0^1 (\Pi_0 w_h)(x) v_h(x)\, dx \quad \forall v_h \in W_h. \tag{3.8}$$

[1] In the one-dimensional case, there are no *streamline directions*; we are just following the general terminology.

To verify (3.7), we define $\widetilde{w} \in H_0^1$ as the solution of

$$-(\varepsilon+\tau)\widetilde{w}'' - \widetilde{w}' = \Pi_0 w_h, \tag{3.9}$$

such that $a_{\varepsilon,h}(v_h, \widetilde{w}_h - \widetilde{w}) = 0$ for all $v_h \in W_h$; moreover, (3.9) has the same structure as (1.2), with $\varepsilon + \tau \simeq h$ instead of ε, so that we can make use of the analogue of (2.3):

$$\begin{aligned}
|\widetilde{w}|_{H^1} &\preceq \sup_{v \in H_0^1} \frac{(\varepsilon+\tau)\int_0^1 \widetilde{w}'(x)v'(x)\,dx \;-\int_0^1 \widetilde{w}'(x)v(x)\,dx}{(\varepsilon+\tau)|v|_{H^1} + \|\Pi_0 v\|_{L^2}} \\
&= \sup_{v \in H_0^1} \frac{\int_0^1 \Pi_0 w_h(x)v(x)\,dx}{(\varepsilon+\tau)|v|_{H^1} + \|\Pi_0 v\|_{L^2}} \\
&\leq \sup_{v \in H_0^1} \frac{\int_0^1 \Pi_0 w_h(x)v(x)\,dx}{\|\Pi_0 v\|_{L^2}} \\
&= \sup_{v \in H_0^1} \frac{\int_0^1 \Pi_0 w_h(x)\Pi_0 v(x)\,dx}{\|\Pi_0 v\|_{L^2}} \\
&= \|\Pi_0 w_h\|_{L^2}.
\end{aligned} \tag{3.10}$$

We also need to introduce the nodal interpolant $\widetilde{w}_I \in W_h$ of $\widetilde{w}$, which satisfies

$$|\widetilde{w} - \widetilde{w}_I|_{H^1} + h^{-1}\|\widetilde{w} - \widetilde{w}_I\|_{L^2} \preceq |\widetilde{w}|_{H^1}. \tag{3.11}$$

Therefore, we have, by using the coercivity of $a_{\varepsilon,h}$, (3.8), (3.9), and (3.11),

$$\begin{aligned}
|\widetilde{w} - \widetilde{w}_h|_{H^1}^2 &= (\varepsilon+\tau)^{-1} a_{\varepsilon,h}(\widetilde{w} - \widetilde{w}_h, \widetilde{w} - \widetilde{w}_h) \\
&= (\varepsilon+\tau)^{-1} a_{\varepsilon,h}(\widetilde{w} - \widetilde{w}_I, \widetilde{w} - \widetilde{w}_h) \\
&\leq |\widetilde{w} - \widetilde{w}_h|_{H^1}\left[|\widetilde{w} - \widetilde{w}_I|_{H^1} + (\varepsilon+\tau)^{-1}\|\widetilde{w} - \widetilde{w}_I\|_{L^2}\right] \\
&\preceq |\widetilde{w} - \widetilde{w}_h|_{H^1}|\widetilde{w}|_{H^1},
\end{aligned}$$

which gives $|\widetilde{w} - \widetilde{w}_h|_{H^1} \preceq |\widetilde{w}|_{H^1}$, and then, by using (3.10),

$$\begin{aligned}
|\widetilde{w}_h|_{H^1} &\leq |\widetilde{w}|_{H^1} + |\widetilde{w} - \widetilde{w}_h|_{H^1} \\
&\preceq |\widetilde{w}|_{H^1} \\
&\preceq \|\Pi_0 w_h\|_{L^2},
\end{aligned}$$

which gives (3.7); (3.6) follows immediately from (3.8), and (3.1) is satisfied.

We obtain (3.2) from (3.1) by a duality argument. We now associate to a generic $w_h \in W_h$ the function $\widetilde{w}_h \in W_h$, which satisfies

$$a_{\varepsilon,h}(v_h, \widetilde{w}_h) = \int_0^1 w_h'(x)v_h'(x)\,dx \quad \forall v_h \in W_h. \tag{3.12}$$

The left-hand side of (3.12) is the dual of $a_{\varepsilon,h}(\widetilde{w}_h, v_h)$. We can proceed as before to obtain the analogue of (3.1); in particular,

$$\begin{aligned}
\varepsilon|\widetilde{w}_h|_{H^1} + \|\Pi_0 \widetilde{w}_h\|_{L^2} &\preceq \sup_{v_h \in W_h} \frac{a_{\varepsilon,h}(v_h, \widetilde{w}_h)}{|v_h|_{H^1}} \\
&= \sup_{v_h \in W_h} \frac{\int_0^1 w_h'(x)v_h'(x)\,dx}{|v_h|_{H^1}} \\
&= |w_h|_{H^1};
\end{aligned}$$

then

$$\begin{aligned} |w_h|_{H^1} &= \frac{\int_0^1 w_h'(x) w_h'(x)\,dx}{|w_h|_{H^1}} \\ &= \frac{a_{\varepsilon,h}(w_h, \widetilde{w}_h)}{|w_h|_{H^1}} \\ &\preceq \frac{a_{\varepsilon,h}(w_h, \widetilde{w}_h)}{\varepsilon |\widetilde{w}_h|_{H^1} + \|\Pi_0 \widetilde{w}_h\|_{L^2}}, \end{aligned}$$

which gives (3.2). □

The estimates (3.1)–(3.2) state inf-sup conditions with respect to the same norms as in (2.2)–(2.3). Focusing on (3.1), for example, we see that one could replace $\varepsilon|w_h|_{H^1}$ on the left-hand side with $(\varepsilon+\tau)|w_h|_{H^1}$, as it seems natural from (3.3)–(3.4). Actually this leads to an equivalent estimate at the discrete level because of the inverse inequality $(\varepsilon+\tau)|w_h|_{H^1} \preceq \|\Pi_0 w_h\|_{L^2}$. On the other hand, in order to obtain in what follows a meaningful error estimate (2.18), our aim here is to make use of the "natural" norms (for the continuous problem).

Let us now define the discrete counterpart of (2.1):

$$\begin{aligned} &\|w_h\|_{A_{0,h}} := \varepsilon |w_h|_{H^1} + \|\Pi_0 w_h\|_{L^2} \qquad \forall w_h \in A_{0,h} := W_h, \\ &\|w_h\|_{A_{1,h}} := |w_h|_{H^1} \qquad \forall w_h \in A_{1,h} := W_h. \end{aligned} \tag{3.13}$$

We construct from (3.1)–(3.2) a family of intermediate inf-sup conditions by means of function space interpolation.

PROPOSITION 3.2. *Let $0<\theta<1$ and $1 \le p \le +\infty$. The one-dimensional SUPG method* (2.19)–(2.22) *satisfies the estimates*

$$\|w_h\|_{(A_{0,h},A_{1,h})_{\theta,p}} \preceq \sup_{v_h \in W_h} \frac{a_{\varepsilon,h}(w_h, v_h)}{\|v_h\|_{(A_{0,h},A_{1,h})_{1-\theta,p'}}} \quad \forall w_h \in W_h, \tag{3.14}$$

where $1/p + 1/p' = 1$.

Proof. The bilinear form $a_{\varepsilon,h} : W_h \times W_h \to \mathbb{R}$ induces the linear operator $\mathcal{L}_{\varepsilon,h} : W_h \to W_h^*$ in the usual way,

$$_{W_h^*}\langle \mathcal{L}_{\varepsilon,h} w_h, v_h \rangle_{W_h} := a_{\varepsilon,h}(w_h, v_h) \quad \forall w_h, v_h \in W_h,$$

which turns out to be invertible, thanks to (3.1)–(3.2); in particular,

$$\|\mathcal{L}_{\varepsilon,h}^{-1} \phi_h\|_{A_{0,h}} \preceq \|\phi_h\|_{A_{1,h}^*}, \tag{3.15}$$

$$\|\mathcal{L}_{\varepsilon,h}^{-1} \phi_h\|_{A_{1,h}} \preceq \|\phi_h\|_{A_{0,h}^*} \tag{3.16}$$

for any $\phi_h \in W_h^*$. Therefore, by means of interpolation (see [15, section 1.3.3 (a)]), we obtain

$$\|\mathcal{L}_{\varepsilon,h}^{-1} \phi_h\|_{(A_{0,h},A_{1,h})_{\theta,p}} \preceq \|\phi_h\|_{(A_{1,h}^*,A_{0,h}^*)_{\theta,p}}. \tag{3.17}$$

By means of [15, section 1.11.2], we also have that the norm on $(A_{1,h}^*, A_{0,h}^*)_{\theta,p}$ is actually equivalent to the dual norm on $(A_{1,h}, A_{0,h})_{\theta,p'} \equiv (A_{0,h}, A_{1,h})_{1-\theta,p'}$; notice in particular that, for the case $p=1$ and $p'=+\infty$, the mentioned result follows because $(A_{1,h}, A_{0,h})_{\theta,+\infty} \equiv (A_{1,h}, A_{0,h})^0_{\theta,+\infty}$ in the algebraic sense. Actually, from the algebraic point of view $A_{0,h} \equiv A_{1,h} \equiv V_h$, we are just defining norms that have a different dependence on the parameter ε. Finally,

$$\|w_h\|_{(A_{0,h},A_{1,h})_{\theta,p}} \preceq \|\mathcal{L}_{\varepsilon,h} w_h\|_{(A_{0,h},A_{1,h})^*_{1-\theta,p}}, \tag{3.18}$$

which is just (3.14). □

Now we turn our attention to the continuity of $a_{\varepsilon,h}$ (i.e., estimate (2.16)), which is, contrary to expectation, the most difficult point. For the sake of clarity, we write $a_{\varepsilon,h}(\cdot,\cdot) = a_\varepsilon(\cdot,\cdot) + s(\cdot,\cdot) + c(\cdot,\cdot)$, where $s : H^1_0 \times H^1_0 \to \mathbb{R}$ denotes the stabilizing term

$$s(w,v) := \tau \int_0^1 w'(x)\, v'(x)\, dx, \tag{3.19}$$

while the term $c(\cdot,\cdot)$ makes the numerical formulation consistent. For the definition of $c(\cdot,\cdot)$, we need the trial functions w to be regular inside any element. Therefore, we set

$$W := \left\{ w \in H^1_0 | w''_{|T_i} \in L^2(T_i),\ i = 1, \dots, N \right\}, \tag{3.20}$$

equipped with the graph norm, and we define $c : W \times H^1_0 \to \mathbb{R}$ as

$$c(w,v) = -\tau\varepsilon \sum_{i=1}^{N} \int_{T_i} w''(x)\, v'(x)\, dx. \tag{3.21}$$

First, we consider $a_\varepsilon(\cdot,\cdot)$.

LEMMA 3.3. *Assume* $0 < \theta < 1$, $1 \le p \le +\infty$, *and* $1/p + 1/p' = 1$*; we have*

$$a_\varepsilon(w, v_h) \preceq \|w\|_{(A_0,A_1)_{\theta,p}} \|v_h\|_{(A_{0,h},A_{1,h})_{1-\theta,p'}} \quad \forall w \in H^1_0, \forall v_h \in W_h. \tag{3.22}$$

Proof. Let $\widetilde{\mathcal{L}}_\varepsilon : H^1_0 \to W^*_h$ be the linear operator given by

$$_{W^*_h}\langle \widetilde{\mathcal{L}}_\varepsilon w, v_h \rangle_{W_h} := a_\varepsilon(w, v_h) \quad \forall w \in H^1_0, \forall v_h \in W_h.$$

(Notice that it differs from $\mathcal{L}_\varepsilon$ because we are now considering discrete test functions.) The Cauchy–Schwarz inequality gives, for any $w \in H^1_0$,

$$\|\widetilde{\mathcal{L}}_\varepsilon w\|_{A^*_{1,h}} \preceq \|w\|_{A_0}$$

and

$$\|\widetilde{\mathcal{L}}_\varepsilon w\|_{A^*_{0,h}} \preceq \|w\|_{A_1};$$

therefore, proceeding as in the proof of Lemma 3.2, we obtain

$$\|\widetilde{\mathcal{L}}_\varepsilon w\|_{(A_{1,h},A_{0,h})^*_{\theta,p}} \preceq \|w\|_{(A_0,A_1)_{\theta,p}},$$

which is (3.22). □

We need the following inverse inequalities for the forthcoming analysis.

LEMMA 3.4. *We have*

$$h\|v'_h\|_{L^2} \preceq \|v'_h\|_{H^{-1}} \quad \forall v_h \in W_h. \tag{3.23}$$

Proof. Actually, (3.23) follows from $\|v_h'\|_{L^2} = \|(\Pi_0 v_h)'\|_{L^2}$, $\|v_h'\|_{H^{-1}} = \|\Pi_0 v_h\|_{L^2}$, and from the more usual inverse inequality $h\|(\Pi_0 v_h)'\|_{L^2} \preceq \|\Pi_0 v_h\|_{L^2}$. □

LEMMA 3.5. *We have*

$$h\|v_h'\|_{(L^2,H_0^1)_{1/2,+\infty}} \preceq \|v_h'\|_{(H^{-1},L^2)_{1/2,+\infty}} \quad \forall v_h \in W_h. \tag{3.24}$$

Proof. We shall show that

$$h\|v_h'\|_{(L^2,H_0^1)_{1/2,+\infty}} \preceq h^{1/2}\|v_h'\|_{L^2} \quad \forall v_h \in W_h, \tag{3.25}$$

$$h\|v_h'\|_{(L^2,H_0^1)_{1/2,+\infty}} \preceq h^{-1/2}\|v_h'\|_{H^{-1}} \quad \forall v_h \in W_h, \tag{3.26}$$

which give (3.24) by means of the interpolation theorem [15, section 1.3.3 (a)]. We focus only on (3.25), as (3.26) follows from (3.25) and (3.23).

We have, by definition,

$$\|v_h'\|_{(L^2,H_0^1)_{1/2,+\infty}} \leq \sup_{0<t<+\infty} \left(t^{-1/2}\|\phi_0(t)\|_{L^2} + t^{1/2}\|\phi_1(t)\|_{H^1} \right) \tag{3.27}$$

for any $\phi_0(t)$ and $\phi_1(t)$ with $v_h' = \phi_0(t) + \phi_1(t)$, $\phi_0(t) \in L^2$, $\phi_1(t) \in H_0^1$, and $0 < t < +\infty$. We now choose suitable $\phi_0(t)$ and $\phi_1(t)$. If $t > h$, it suffices to take $\phi_0(t) = v_h'$ and, accordingly, $\phi_1(t) = 0$ to obtain

$$t^{-1/2}\|\phi_0(t)\|_{L^2} + t^{1/2}\|\phi_1(t)\|_{H^1} \leq h^{-1/2}\|v_h'\|_{L^2} \quad \forall v_h \in W_h. \tag{$t > h$ case}$$

Otherwise, when $t \leq h$, set $\delta \equiv \delta(t,x) := \min\{x, 1-x, t/2\}$ and

$$\phi_1(t)(x) := t^{-1} \int_{x-\delta}^{x+\delta} v_h'(\xi)\, d\xi;$$

the effect of this definition is shown in Figure 3.1. As a result, one has

$$t^{-1/2}\|\phi_0(t)\|_{L^2} + t^{1/2}\|\phi_1(t)\|_{H^1} \preceq \left[\sum_{i=0}^{N} \left(v_h'|_{T_{i+1}} - v_h'|_{T_i} \right)^2 \right]^{1/2},$$

where $v_h'|_{T_0} = v_h'|_{T_{N+1}} := 0$ by convention, and

$$\begin{aligned} \sum_{i=0}^{N} \left(v_h'|_{T_{i+1}} - v_h'|_{T_i} \right)^2 &\preceq \sum_{i=0}^{N} \left(v_h'|_{T_i} \right)^2 \\ &= h^{-1}\|v_h'\|_{L^2}^2; \end{aligned}$$

therefore,

$$t^{-1/2}\|\phi_0(t)\|_{L^2} + t^{1/2}\|\phi_1(t)\|_{H^1} \preceq h^{-1/2}\|v_h'\|_{L^2} \quad \forall v_h \in W_h. \tag{$t \leq h$ case}$$

Collecting the ($t > h$ case) and ($t \leq h$ case) with (3.27), we obtain (3.25). □

LEMMA 3.6. *Assume* $1/2 < \theta < 1$ *and* $1 \leq p \leq +\infty$ *or* $\theta = 1/2$ *and* $p = 1$*; we have*

$$s(w, v_h) \preceq \|w\|_{(A_0,A_1)_{\theta,p}} \|v_h\|_{(A_{0,h},A_{1,h})_{1-\theta,p'}} \quad \forall w \in H_0^1, \forall v_h \in W_h, \tag{3.28}$$

where $1/p + 1/p' = 1$.

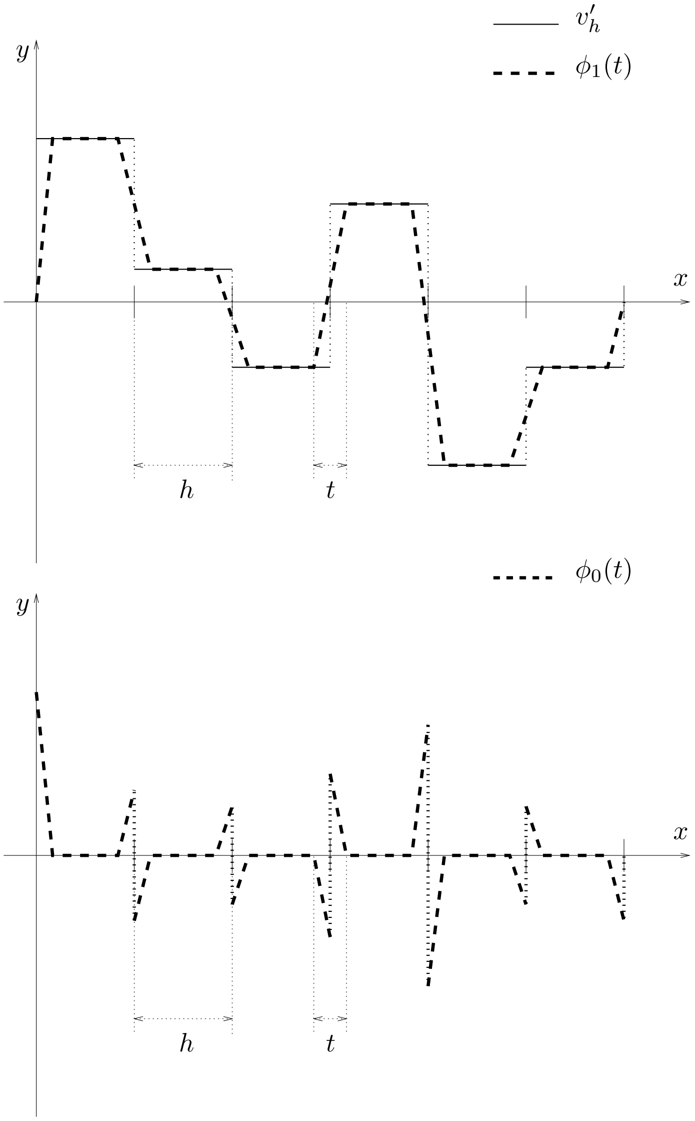


FIG. 3.1. *Construction of $\phi_0(t)$ and $\phi_1(t)$ inside the proof of Lemma* 3.5.

Proof. Let $w \in H_0^1$ and $v_h \in W_h$. Assume for a moment that the estimates

$$s(w, v_h) \preceq \|w'\|_{(H^{-1},L_0^2)_{1/2,1}} \|v_h'\|_{(H^{-1},L_0^2)_{1/2,+\infty}}, \tag{3.29}$$

$$s(w, v_h) \preceq \|w'\|_{L^2} \|v_h'\|_{H^{-1}} \tag{3.30}$$

hold true. Therefore, recalling (2.5) and as $\|v_h\|_{(A_0,A_1)_{1/2,+\infty}} \leq \|v_h\|_{(A_{0,h},A_{1,h})_{1/2,+\infty}}$, we also have

$$s(w, v_h) \preceq \|w\|_{(A_0,A_1)_{1/2,1}} \|v_h\|_{(A_{0,h},A_{1,h})_{1/2,+\infty}},$$
$$s(w, v_h) \preceq \|w\|_{A_1} \|v_h\|_{A_{0,h}},$$

and we can apply a new interpolation, with parameters $0 < \eta < 1$ and $1 \leq q \leq +\infty$, reasoning as in the proof of Lemma 3.3, in order to obtain

$$s(w,v_h) \preceq \|w\|_{((A_0,A_1)_{1/2,1},A_1)_{\eta,q}} \|v_h\|_{((A_{0,h},A_{1,h})_{1/2,+\infty},A_{0,h})_{\eta,q'}},$$

where $1/q + 1/q' = 1$. This gives (3.28), thanks to the reiteration theorem (see [15, section 1.10.2]), with $\theta = 1/2 + \eta/2$ and $p = q$.

First, let us focus on (3.30): it follows from the Cauchy–Schwarz inequality, Lemma 3.4, and (2.22).

Now we focus on (3.29). The Cauchy–Schwarz inequality yields

$$s(w,v_h) \preceq \|w'\|_{(H^{-1},L^2)_{1/2,1}} \|\tau v_h'\|_{(H^{-1},L^2)^*_{1/2,1}}, \tag{3.31}$$

and, of course, we have

$$\|w'\|_{(H^{-1},L^2)_{1/2,1}} \leq \|w'\|_{(H^{-1},L^2_0)_{1/2,1}}; \tag{3.32}$$

on the other hand, thanks to [15, section 1.11.2], (2.22), and Lemma 3.5, we also have

$$\begin{aligned} \|\tau v_h'\|_{(H^{-1},L^2)^*_{1/2,1}} &\preceq \|\tau v_h'\|_{(L^2,H^1_0)_{1/2,+\infty}} \\ &\preceq \|v_h'\|_{(H^{-1},L^2)_{1/2,+\infty}} \\ &\preceq \|v_h'\|_{(H^{-1},L^2_0)_{1/2,+\infty}}. \end{aligned} \tag{3.33}$$

Finally, (3.31)–(3.33) give (3.29). □

It is worth noting that the stabilizing term $s(\cdot,\cdot)$ is continuous, with respect to the norms $\|\cdot\|_{(A_0,A_1)_{\theta,p}}$ and $\|\cdot\|_{(A_{0,h},A_{1,h})_{1-\theta,p'}}$, for any values of θ and p; nevertheless the uniformity with respect to ε requires the restrictions stated in Lemma 3.6. In this sense, these restrictions are optimal: notice that the norms in the left-hand side of (3.24) cannot be replaced by stronger norms since v_h' is discontinuous.

In order to deal with $c(\cdot,\cdot)$, we define an ad hoc seminorm:

$$|||\phi|||_{\theta-1,p} = \sup_{v_h \in W_h} \frac{\sum_{i=1}^{N} \tau \int_{T_i} \phi(x)\, v_h'(x)\, dx}{\|v_h\|_{(A_{0,h},A_{1,h})_{1-\theta,p'}}}. \tag{3.34}$$

The continuity of $c(\cdot,\cdot)$ follows immediately.

LEMMA 3.7. *Assume* $0 < \theta < 1$, $1 \leq p \leq +\infty$, *and* $1/p + 1/p' = 1$*; we have*

$$c(w,v_h) \preceq \varepsilon |||w''|||_{\theta-1,p} \|v_h\|_{(A_{0,h},A_{1,h})_{1-\theta,p'}}. \tag{3.35}$$

The framework is complete, and we can now state our main result.

THEOREM 3.8. *Assume that* $f \in L^2$, $1/2 < \theta < 1$, *and* $1 \leq p \leq +\infty$ *or* $\theta = 1/2$ *and* $p = 1$. *The one-dimensional SUPG method* (2.19)–(2.22) *satisfies the classical continuity and* inf-sup *conditions* (2.16)–(2.17) *with respect to the norms*

$$\begin{aligned} \|w\|_{W,h} &:= \|w\|_{(A_0,A_1)_{\theta,p}} + \varepsilon |||w''|||_{\theta-1,p}, \\ \|v_h\|_{V_h} &:= \|v_h\|_{(A_{0,h},A_{1,h})_{1-\theta,p'}}. \end{aligned}$$

Proof. The continuity (2.16) follows from Lemmas 3.3, 3.6, and 3.7. Moreover, since $\|w_h\|_{(A_0,A_1)_{\theta,p}} \leq \|w_h\|_{(A_{0,h},A_{1,h})_{\theta,p}}$ and $\varepsilon |||w_h''|||_{\theta-1,p} = 0$ for all $v_h \in W_h$, we get

$$\|w\|_{W,h} \leq \|w\|_{(A_{0,h},A_{1,h})_{\theta,p}}, \tag{3.36}$$

whence the inf-sup condition (2.17) follows from Proposition 3.2 and (3.36). □

As a result, we can state the *quasi optimality* (2.18) of the one-dimensional SUPG method with respect to the norm $\|\cdot\|_{W,h}$ (for $1/2<\theta<1$ or $\theta=1/2$ and $p=1$); the most interesting case is for $\theta=1/2$. Thanks to (2.6) and (2.9), we can state the following result.

COROLLARY 3.9. *Assume that $f\in L^2$; let u be the solution of* (1.2) *and u_h be the numerical solution given by the one-dimensional SUPG method* (2.19)–(2.22). *Then*

$$\begin{aligned}&\varepsilon^{1/2}|u-u_h|_{H^1}+\|u-u_h\|_{(L^2,H^1_{\#})_{1/2,1}}+\varepsilon|\!|\!|(u-u_h)''|\!|\!|_{-1/2,1}\\&\quad\preceq\inf_{w_h\in V_h}\left[\varepsilon^{1/2}|u-w_h|_{H^1}+\|u-w_h\|_{(L^2,H^1_{\#})_{1/2,1}}+\varepsilon|\!|\!|(u-w_h)''|\!|\!|_{-1/2,1}\right].\end{aligned}\tag{3.37}$$

Estimate (3.37) is interesting because the norm appearing there is "natural" for the problem, in the sense that $\varepsilon^{1/2}|u|_{H^1}+\|u\|_{(L^2,H^1_{\#})_{1/2,1}}+\varepsilon|\!|\!|u''|\!|\!|_{-1/2,1}$ does not grow as a negative power of ε when $\varepsilon\to0$. Actually, it is not uniformly bounded with respect to ε but behaves as $\log(\varepsilon)$ in the worst case. Indeed, we have

$$\begin{aligned}\sum_{i=1}^N\tau\int_{T_i}\varepsilon u''(x)\,v_h'(x)\,dx&=+\sum_{i=1}^N\tau\int_{T_i}u'(x)\,v_h'(x)\,dx\\&\quad-\sum_{i=1}^N\tau\int_{T_i}f\,v_h'(x)\,dx\\&\preceq s(u,v_h)+\|f\|_{L^2}\|\tau v_h'\|_{L^2}.\end{aligned}\tag{3.38}$$

This, together with (3.28), (2.22), and (3.23), yields $\varepsilon|\!|\!|u''|\!|\!|_{-1/2,1}\preceq\|u\|_{(A_0,A_1)_{1/2,1}}+\|f\|_{L^2}$; then, thanks to (2.4), (2.5), (2.8), (2.14), and [13, equation (22)], we finally get

$$\varepsilon^{1/2}|u|_{H^1}+\|u\|_{(L^2,H^1_{\#})_{1/2,1}}+\varepsilon|\!|\!|u''|\!|\!|_{-1/2,1}\preceq(1+|\log\varepsilon|)\,\|f\|_{L^2}.\tag{3.39}$$

Usually, one can infer the convergence of the numerical method from an estimate like (2.18) (in particular, (3.37)). This is not the case here. We recall that we are interested in uniform convergence with respect to ε in the advection-dominated regime $\varepsilon<h$. In fact, since we are using piecewise linear elements, we easily see that $|\!|\!|(u-u_h)''|\!|\!|_{-1/2,1}=|\!|\!|u''|\!|\!|_{-1/2,1}$. Furthermore, $\varepsilon^{1/2}|u-u_h|_{H^1}$, as well as $\|u-u_h\|_{(L^2,H^1_{\#})_{1/2,1}}$, cannot vanish uniformly with respect to ε when $h\to0$. This is because the boundary layer cannot be captured within the discrete space W_h when $\varepsilon\ll h$. On the other hand, this is not surprising; convergence results are indeed obtained from estimates like (2.18) by assuming *extra* regularity on the solution u. In our case, this is not possible since, for example, $\|u\|_{(A_0,A_1)_{\theta,p}}$ is strongly dependent on ε for $\theta>1/2$.

4. Conclusion and further extensions.

In this paper, we proved the *quasi optimality* of the SUPG method for the one-dimensional advection-diffusion problem on a uniform grid. Actually, it is a very simple case. Most of our analysis is based on our previous work [13] and therefore is suitable for an extension to the multidimensional case; in other parts it depends on some special properties of the one-dimensional problem. We do not know at the moment whether the SUPG method preserves its quasi optimality in the two-dimensional case or whether a modification of the method is required for this purpose. We shall focus on possible extensions of the theory in further works.

We have assumed the amount of *streamline diffusion* (or, better, *artificial diffusion*) τ to be proportional to the *mesh size h*. It is well known that, from a practical point of view, the particular choice of τ is relevant for the accuracy of the method. Our analysis does not give any suggestion for this because, for the sake of simplicity, our final estimate implicitly contains generic constants whose dependence on τ is not investigated. This investigation is indeed a very technical task, but one could perform this kind of analysis by a computational procedure. This has been done in a previous work [14] (see, in particular, section 3 therein), where we perform a fine-tuning of τ based on a similar idea.

REFERENCES

[1] I. Babuška and A. K. Aziz, *Survey lectures on the mathematical foundations of the finite element method*, in The Mathematical Foundations of the Finite Element Method with Applications to Partial Differential Equations, Academic Press, New York, 1972, pp. 1–359. With the collaboration of G. Fix and R. B. Kellogg.

[2] S. Bertoluzza, C. Canuto, and A. Tabacco, *Stable discretizations of convection-diffusion problems via computable negative-order inner products*, SIAM J. Numer. Anal., 38 (2000), pp. 1034–1055.

[3] F. Brezzi, *On the existence, uniqueness and approximation of saddle-point problems arising from Lagrangian multipliers*, Rev. Française Automat. Informat. Recherche Opérationnelle Sér. Rouge, 8 (1974), pp. 129–151.

[4] F. Brezzi, T. J. R. Hughes, L. D. Marini, A. Russo, and E. Süli, *A priori error analysis of residual-free bubbles for advection-diffusion problems*, SIAM J. Numer. Anal., 36 (1999), pp. 1933–1948.

[5] F. Brezzi, D. Marini, and E. Süli, *Residual-free bubbles for advection-diffusion problems: The general error analysis*, Numer. Math., 85 (2000), pp. 31–47.

[6] F. Brezzi and A. Russo, *Choosing bubbles for advection-diffusion problems*, Math. Models Methods Appl. Sci., 4 (1994), pp. 571–587.

[7] A. N. Brooks and T. J. R. Hughes, *Streamline upwind/Petrov-Galerkin formulations for convection dominated flows with particular emphasis on the incompressible Navier-Stokes equations*, Comput. Methods Appl. Mech. Engrg., 32 (1982), pp. 199–259.

[8] L. P. Franca, A. Nesliturk, and M. Stynes, *On the stability of residual-free bubbles for convection-diffusion problems and their approximation by a two-level finite element method*, Comput. Methods Appl. Mech. Engrg., 166 (1998), pp. 35–49.

[9] T. J. R. Hughes, L. P. Franca, and G. M. Hulbert, *A new finite element formulation for computational fluid dynamics.* VIII. *The Galerkin/least-squares method for advective-diffusive equations*, Comput. Methods Appl. Mech. Engrg., 73 (1989), pp. 173–189.

[10] A. Quarteroni and A. Valli, *Numerical Approximation of Partial Differential Equations*, Springer-Verlag, Berlin, 1994.

[11] H.-G. Roos, M. Stynes, and L. Tobiska, *Numerical Methods for Singularly Perturbed Differential Equations Convection-Diffusion and Flow Problems*, Springer-Verlag, Berlin, 1996.

[12] G. Sangalli, *Global and local error analysis for the residual-free bubbles method applied to advection-dominated problems*, SIAM J. Numer. Anal., 38 (2000), pp. 1496–1522.

[13] G. Sangalli, *Analysis of the Advection-Diffusion Operator Using Fractional Order Norms*, Tech. report 1221, IAN-CNR, Pavia, Italy, 2001; Numer. Math., to appear.

[14] G. Sangalli, *Numerical evaluation of FEM with application to the* 1*-D advection-diffusion problem*, Math. Models Methods Appl. Sci., 12 (2002), pp. 205–228.

[15] H. Triebel, *Interpolation Theory, Function Spaces, Differential Operators*, 2nd ed., Johann Ambrosius Barth, Heidelberg, 1995.

SIAM J. NUMER. ANAL.
Vol. 41, No. 4, pp. 1543–1563

IDENTIFICATION OF A TEMPERATURE DEPENDENT HEAT CONDUCTIVITY FROM SINGLE BOUNDARY MEASUREMENTS[*]

PHILIPP KÜGLER[†]

Abstract. Considering the identification of a temperature dependent conductivity in a quasilinear elliptic heat equation from single boundary measurements, we proof uniqueness in dimensions $n \geq 2$. Taking noisy data into account, we apply Tikhonov regularization in order to overcome the instabilities. By using a problem-adapted adjoint, we give convergence rates under substantially weaker and more realistic conditions than required by the general theory. Our theory is supported by numerical tests.

Key words. nonlinear inverse problem, Tikhonov regularization, nonlinear direct problem

AMS subject classifications. 47A52, 35J60

DOI. 10.1137/S0036142902415900

1. Introduction.

The issue of parameter identification is to determine unknown parameters, appearing, e.g., in state equations, from indirect measurements related to the physical state. This inverse problem can be considered as a (mostly) nonlinear operator equation

$$F(q) = z, \tag{1.1}$$

where the forward operator F maps the parameter q onto the output z. As the physical state often cannot be observed exactly, one finds oneself in the situation of given noisy data z^δ instead of z. Now, as parameter identification problems are frequently ill-posed, the estimation of the parameter can be strongly influenced in a negative way by even only small data noise. Hence, for their stable numerical solution, some type of regularization is required. Regularization techniques replace the ill-posed problem by a family of neighboring well-posed problems, leading to a stable approximation of q, called the regularized solution. Probably the most frequently used approach is Tikhonov regularization, where the regularized solutions are sought as the minimizers of

$$q \to \|F(q) - z^\delta\|^2 + \beta \|q\|^2,$$

with some regularization parameter β.

A careful mathematical analysis of the regularization method is needed in order to give useful guidance, under which conditions it will perform well, and confidence in its numerical results. Since for ill-posed problems, convergence of any numerical algorithm can be arbitrarily slow [21], conditions for convergence rates are of special theoretical interest. They are also practically relevant as they tell us for which problems fast convergence of numerical algorithms can be expected. However, according to the general theory [4], such convergence rates can only be obtained under strong

[*] Received by the editors October 2, 2002; accepted for publication (in revised form) February 18, 2003; published electronically October 2, 2003. This research was supported by the Austrian Fonds zur Förderung der wissenschaftlichen Forschung, projects F013-08 and P13478-INF.

http://www.siam.org/journals/sinum/41-4/41590.html

[†] Industrial Mathematics Institute, Johannes Kepler University Linz, Altenbergerstr. 69, 4040 Linz, Austria (kuegler@indmath.uni-linz.ac.at).

source conditions of the type

$$\exists w \;\; q - q^* = F'(q)^* w, \tag{1.2}$$

where q^* is an a priori guess for q, and $F'(q)^*$ is the adjoint of the Fréchet-derivative of F evaluated at q. This general theory has been applied to various inverse problems including parameter identification; see [14], [4] for elliptic problems and integral equations and [19] for a parabolic equation. All these applications are for one-dimensional problems, since only there, the source condition (1.2) has a rather immediate explicit interpretation (usually requiring some additional smoothness and prescribed boundary behavior for $q^\dagger - q^*$). In [16], condition (1.2) was weakened based on ideas from [8] and then fully interpreted for the identification of a nonlinearity $q(u)$ from distributed measurements of u in arbitrary dimensions.

However, before taking data perturbations and convergence rates into account, one has to consider if the given data z at all contain enough information in order to identify the parameter q, i.e., if the mapping F from q onto z is injective. Often, a limit on the amount of available data is given by the setup of the experiment. Frequently (for example, in nondestructive testing), measurements cannot be done within the material Ω but only on (parts of) the boundary $\partial\Omega$, leading to data containing less information.

Considering the inverse conductivity problem, one is interested in finding the conductivity $q(x)$ in

$$\begin{aligned} -\nabla \cdot (q(x)\nabla u) &= 0 \text{ in } \Omega, \\ u &= g \text{ on } \partial\Omega, \end{aligned} \tag{1.3}$$

given the additional boundary data

$$q(x)\frac{\partial u}{\partial n} = h \text{ on } \Gamma, \tag{1.4}$$

with $\Gamma \subset \partial\Omega$. For this case of single boundary measurements, the unique identifiability is widely investigated for parameters

$$q = 1 + \chi(D), \quad \bar{D} \subset \Omega,$$

where χ is the characteristic function of an unknown domain. Several partial results are given (see, e.g., [1], [10], and [13]); nonetheless a general uniqueness result is still missing. Turning to inverse problems for nonlinear elliptic equations

$$-\Delta u + q(u) = 0 \text{ in } \Omega, \tag{1.5}$$

only local uniqueness results for (small) q are available if, in addition to the Dirichlet data g, Neumann data are prescribed on $\partial\Omega$. See [12] and the references given there.

In order to enhance the chances of identifiability, one often resorts to many boundary measurements: For any Dirichlet data g in (1.3), one is given Neumann data h; in other words, the results of all possible boundary measurements are known. Then the information to identify the parameter is contained in the so-called Dirichlet to Neumann map

$$\Lambda : g \to h. \tag{1.6}$$

Based on these multiple boundary measurements, the aim of impedance tomography is the reconstruction of the conductivity $q(x)$ or $q(x,u)$ in (1.3) within the human body or some material. For the linear case, i.e., $q = q(x)$, global uniqueness was proven in [24] for dimensions $n \geq 3$ and in [18] for two dimensions. The uniqueness result for the quasi-linear case $q = q(x, u)$ can be found in [23] for dimensions $n \geq 2$. There also, the anisotropic case, i.e., q is a matrix, is investigated.

We now specify the inverse problem we are looking at in this paper. Our goal is to identify the temperature dependent heat conductivity q in

$$-\nabla \cdot (q(u)\nabla u) = f \text{ in } \Omega \tag{1.7}$$

from only single boundary measurements of the temperature u. Note that not only the inverse but also the forward problem is nonlinear. We show that the parameter is uniquely identifiable on the temperature range as a function of one variable. Besides, further developments of our Tikhonov regularization analysis from [16] allow us to provide a fully interpretable weak source condition for the convergence rate of the regularized solutions.

In section 2, we briefly discuss the nonlinear direct problem (1.7) with mixed boundary conditions for u. For a positive parameter q of H^1-regularity, we guarantee the existence of a unique weak solution $u \in H^1(\Omega)$. Furthermore, we give an estimate of the temperature range governed by u in Ω, which is equivalent to the one-dimensional domain the parameter q lives on. Together with the temperature data on the boundary, the a priori unknown interval, on which the parameter can be recovered, can then be estimated.

In sections 3 and 4, we investigate the inverse problem. First we show that the parameter is uniquely determined by single boundary temperature measurements. From [23], the uniqueness follows only for the case of multiple measurements. Afterward, we give a stability analysis for Tikhonov regularization and prove convergence rates under much weaker assumptions than required in the general theory by taking advantage of a special adjoint approach. This kind of approach was first introduced in [8] for the identification of a space dependent heat conductivity $q(x)$ from distributed temperature measurements.

Section 5 contains a detailed interpretation of the source condition needed for the convergence rate proof both in two and three dimensions. This is different from [8], where a full interpretation could only be given in the one-dimensional case.

Section 6 sketches variants in the setup for the inverse problem. In section 7, we present results of numerical tests which support our theory.

2. The direct problem.

In many applications modelled by the heat equation, for example, in the context of steel production (see [11], [6], [7]), the heat conductivity q does not vary spatially but rather depends on the temperature u itself. Considering the stationary case, the heat distribution is described by the nonlinear elliptic equation

$$-\nabla \cdot (q(u)\nabla u) = f \text{ in } \Omega \tag{2.1}$$

with, e.g., the boundary conditions

$$q(u)\frac{\partial u}{\partial n} = h \text{ on } \Gamma_1 \tag{2.2}$$

and

$$u(x) = g \text{ on } \Gamma_2. \tag{2.3}$$

Here, Ω is an open bounded connected domain in R^n, $n \geq 2$, with boundary $\partial\Omega \in C^2$. We assume $\Gamma_2 \subset \partial\Omega$ to be connected and to have positive measure and set

$\Gamma_1 = \partial\Omega \backslash \Gamma_2$. f is a given heat source density, h is a given temperature flux, and g is a prescribed (boundary) temperature.

We set

$$V = \left\{ v \in H^1(\Omega) \mid v|_{\Gamma_2} = 0 \right\}$$

and assume—already with respect to the inverse problem—that

$$g \text{ is constant, } h \in C(\Gamma_1), \text{ and } f \in C(\Omega).$$

By the trace theorem we then get a $\tilde{g} \in H^1(\Omega)$ such that $\tilde{g}|_{\Gamma_2} = g$. Now, by integration by parts we derive the variational formulation for problems (2.1)–(2.3).

Find $u \in H^1(\Omega)$ such that

$$u - \tilde{g} \quad \in \quad V \tag{2.4}$$

and

$$\int_\Omega q(u) \nabla u \cdot \nabla v \; dx = \int_\Omega f(x) v \; dx + \int_{\Gamma_1} hv \; d\Gamma_1 \quad \text{for all } \; v \in V \tag{2.5}$$

hold. Under the assumption

$$q \in H^1(R) \; \text{ and } \; 0 < \alpha_1 \le q \le \alpha_2, \tag{2.6}$$

there exists a unique solution to the variational problem (2.5) in $H^1(\Omega)$. The proof—based on the continuous embedding $H^1(R) \subseteq C(R)$ and the theory of quasi-monotone operators—can, for example, be found in [22] (see Proposition 5.1 and the subsequent relaxation to quasi-monotone operators). Furthermore, there is an a priori estimate for the solution, i.e., there is a constant $C > 0$ depending only on α_2, h, f, and g, such that

$$\|u_q\|_{H^1(\Omega)} \le C, \tag{2.7}$$

where, in order to emphasize the fact that the solution u depends on the parameter q, we use the notation u_q or $u_q(x)$.

Note that u_q can also be considered as the weak solution of the linear equation

$$\begin{aligned} -\nabla \cdot (\tilde{q}(x) \nabla u) &= f \; \text{ in } \; \Omega, \\ \tilde{q}(x) \frac{\partial u}{\partial n} &= h \; \text{ on } \; \Gamma_1, \\ u(x) &= g \; \text{ on } \; \Gamma_2, \end{aligned}$$

with $\tilde{q}(x) = q(u_q(x))$. Already with respect to the inverse problem, we cite [3] for the weak maximum principle: If we choose f and h such that

$$\int_\Omega f(x) v \; dx + \int_{\Gamma_1} hv \; d\Gamma_1 \le 0$$

holds for all essentially nonnegative $v \in V$, we get

$$\operatorname{ess.} \sup_{x \in \Omega} u_q \le \operatorname{ess.} \sup_{x \in \Gamma_1} \max \{u_q, 0\} \tag{2.8}$$

(or u_q is a positive constant). Hence, the upper bound of the temperature range covered by u_q in Ω is given by the maximum temperature on Γ_1 (or by 0 if the latter is negative).

3. The inverse problem. Given a single boundary observation of the solution of the direct problem, the inverse problem is to recover the physical parameter q on the real interval covered by the temperature using the observation data. In order to overcome the ill-posedness of this identification problem, we choose Tikhonov regularization for its stabilization.

3.1. The interval of identifiability. Identifying the nonlinearity $q(u)$ is theoretically and numerically challenging, since the interval, on which the parameter can be recovered, is a priori not known. Obviously, the parameter, as a function of one variable, cannot be reconstructed on the whole of R but at the most on the interval $[u_{\min}, u_{\max}]$, where $u_{\min}$ and $u_{\max}$ denote the extremal values of the temperature distributed over $\bar{\Omega}$. Outside this interval, no physical information is available, making the identification impossible in advance.

In the case that only boundary temperature measurements are given, the data need not necessarily cover all of $[u_{\min}, u_{\max}]$. If one still wants to recover the parameter on the whole interval, the following experimental setup for indirect measurements volunteers itself. Assume that the heat conductivity q is known up to a temperature value u_0 from possibly direct measurements but is inaccessible at temperatures above. Then we set

$$g = u_0 \ \ (\text{constant}) \tag{3.1}$$

in (2.3). By tuning f and h in (2.1) and (2.2), we drive the temperature on the boundary Γ_1 to values higher than u_0. Finally, we measure the temperature trace along Γ_1, whose maximum value we call u_1.

We know from the maximum principle (2.8) that

$$u_{\max} = u_1$$

holds (for f and h chosen appropriately). Unfortunately, we cannot guarantee $u_{\min} = u_0$ but have only $u_{\min} \leq u_0$. Nevertheless, since we assume to know q up to u_0, we can consider the identification of q on the interval

$$[u_{\min}, u_{\max}]. \tag{3.2}$$

3.2. Output least squares formulation. Denoting by $z(x)$ the measured temperature trace along Γ_1, we want to identify the true thermal conductivity $q^\dagger$ out of a set of admissible parameters, satisfying

$$\gamma u_{q^\dagger} = z, \tag{3.3}$$

where γ denotes the trace operator

$$\begin{aligned} \gamma : H^1(\Omega) &\to L^2(\Gamma_1), \\ u &\to u|_{\Gamma_1}, \end{aligned}$$

and $u_{q^\dagger}$ is the solution of the direct problem (2.1)–(2.3) with $q = q^\dagger$. We always assume the existence of a true parameter $q^\dagger$, i.e., that the exact data z are attainable. Of course, the measured (noisy) data need not be attainable.

We already mentioned in the previous section that $q^\dagger$ can at most be identified on the range of $u_{q^\dagger}$. Nevertheless, during the numerical solution of the inverse problem, temperature values corresponding to other parameters than $q^\dagger$ may occur. Hence the parameters have to be defined on an even larger range than that of $u_{q^\dagger}$. This crucial numerical point is discussed in [15] and [16] for the case of distributed temperature measurements.

For defining the set of admissible parameters, we choose positive constants α_1 and α_2 such that the temperatures u_{α_1} and u_{α_2} (obtained by solving the direct problem) contain (at least) the minimal and maximal values of the data, i.e., the measured temperature trace on Γ_1, respectively. Since α_1 and α_2 are constant and hence regular parameters, regularity results (see, for instance, [17]) yield that u_{α_1} and u_{α_2} are continuous on $\bar{\Omega}$; i.e., there are constants I_1 and I_2 such that

$$\begin{aligned} I_1 \leq u_{\alpha_1} \leq I_2, \\ I_1 \leq u_{\alpha_2} \leq I_2. \end{aligned}$$

Then we can use the finite interval

$$I = [I_1, I_2], \quad I_1, I_2 \in R, \tag{3.4}$$

in order to define the set of admissible parameters as

$$K = \left\{ q \in H^1(R) \mid \alpha_1 \leq q(\tau) \leq \alpha_2 \text{ for } \tau \in I \text{ and } q \text{ is fixed on } R \backslash I \right\}. \tag{3.5}$$

Here, the attribute fixed has to be understood as

$$q_1(\tau) - q_2(\tau) \equiv 0 \ \text{ on } R \backslash I \tag{3.6}$$

for any $q_1,\ q_2 \in K$. The only requirement for the behavior of q on $R \backslash I$ is that $q \in H^1(R)$ is not violated. Then any $q \in K$ is continuous and bounded due to the continuous embedding $H^1(R) \subseteq C_b(R)$. Of course, we can only identify $q^\dagger$ on a subdomain of I where we have information about the system from the data z. Outside this domain, we have no information, so an identification is impossible in advance. In this sense, we should look in (3.4) for an interval I of minimal length. Again, we refer to [16] for a possible numerical approach.

As we shall see below, this construction of K is mainly needed for technical reasons. Things would simplify if one assumes the existence (but not the exact knowledge) of a finite interval I that covers all temperatures u_q for q belonging to

$$\tilde{K} = \left\{ q \in H^1(I) \mid \alpha_1 \leq q(\tau) \leq \alpha_2 \text{ for } \tau \in I \right\} \tag{3.7}$$

with α_1, α_2 appropriately chosen. This assumption may be supported by the finiteness of physical temperatures.

For later use, we introduce the set of the indefinite integrals of the parameters $q \in K$

$$S = \left\{ Q \in H^2(R) \mid \frac{dQ}{d\tau} \in K \text{ and } Q(g) = 0 \right\}, \tag{3.8}$$

where $g \in I$ is the constant from (2.3), (3.1). Because of (3.5), we have a common Lipschitz constant, namely, α_2, for the functions $Q \in S$:

$$|Q(\tau_1) - Q(\tau_2)| \leq \alpha_2 |\tau_1 - \tau_2|, \ \tau_1,\ \tau_2 \in R, \tag{3.9}$$

for all $Q \in S$.

In applications, the exact data $z(x)$ are not known precisely due to measurement errors. Hence the actual data are available in the form

$$z^\delta(x) = z(x) + \text{noise},$$

where one needs some information

$$\|z - z^\delta\|_{L^2(\Gamma_1)} \leq \delta \tag{3.10}$$

about the noise level. Due to the data noise, the ill-posedness of the inverse problem requires some type of regularization in order to determine $q^\dagger$ in (3.3) in a stable way. Choosing Tikhonov regularization, we consider the following output-least-squares problem.

Let the set K of admissible parameters and noisy data z^δ be given as in (3.5) and (3.10). Assume that the exact data z is attainable from a parameter $q^\dagger \in K$. Then, for $\beta > 0$, find a parameter $q_\beta^\delta \in K$ that minimizes

$$J_\beta(q) = \int_{\Gamma_1} |\gamma u_q - z^\delta|^2 d\Gamma_1 + \beta\|q - q^*\|^2_{H^1(R)} \tag{3.11}$$

over K for an appropriate choice of β and $q^* \in K$. The selection of q^* is crucial for the results about the convergence rate in section 4. Available a priori information about the true parameter $q^\dagger$ should be used for the choice of q^*; i.e., q^* should be interpreted as some kind of a priori guess for $q^\dagger$. Because of (3.6), the $H^1(R)$-norm can be replaced by the $H^1(I)$-norm, which in the following is denoted by $\|\cdot\|_I$. (Other penalty terms are possible; see section 6.) Note that q^* also determines the "identified" parameter $q^\dagger$ outside the domain of information in the case of I chosen too large.

Before discussing aspects of stability and convergence of the regularized solutions q_β^δ toward $q^\dagger$, we make sure that the given boundary data are sufficient to identify the parameter uniquely.

3.3. Identifiability. Investigating the identifiability of q, we are interested in the injectivity of the parameter-to-output map

$$q \to \gamma u_q.$$

We show that the temperature trace on Γ_1 determines the parameter uniquely on that range.

THEOREM 3.1. *Let u_{q_1} and $u_{q_2} \in H^1(\Omega)$ be the solutions of the direct problem corresponding to parameters q_1 and $q_2 \in K$. Then $\gamma u_{q_1} = \gamma u_{q_2}$ implies $q_1 = q_2$ on the range of u_{q_1} on Γ_1.*

Proof. For $i = 1, 2$ we define $w_i = Q_i(u_{q_i})$, where $Q_i \in S$ is the antiderivative to q_i. Then w_i satisfies $w_i|_{\Gamma_2} = 0$ and the linear equation

$$\int_\Omega \nabla w_i \cdot \nabla v dx = \int_\Omega f(x) v dx + \int_{\Gamma_1} hv d\Gamma_1 \quad \text{for all } v \in V.$$

Hence the difference $w = w_1 - w_2$ fulfills

$$\int_\Omega \nabla w \cdot \nabla v dx = 0 \quad \text{for all } v \in V,$$

which gives $w = 0$ according to the unique solvability of the homogeneous problem. Hence we have $Q_1(u_{q_1}) = Q_2(u_{q_2})$ in Ω. From the trace theorem we get $\gamma Q_1(u_{q_1}) = \gamma Q_2(u_{q_2})$; the continuity of Q_i yields $Q_1(\gamma u_{q_1}) = Q_2(\gamma u_{q_2})$. From the assumption $\gamma u_{q_1} = \gamma u_{q_2}$, we then can conclude that $Q_1(\tau) = Q_2(\tau)$, and hence $q_1(\tau) = q_2(\tau)$ for τ out of the range of u_{q_1} on Γ_1. ☐

3.4. Existence, stability, and convergence of the regularized solutions. Returning to problem (3.11), we have to make sure that

- a minimizer q_β^δ exists for any data $z^\delta \in L^2(\Gamma_1)$ (existence);
- for a fixed regularization parameter β, the minimizers of (3.11) depend continuously on the data z^δ (stability);
- the regularized solutions q_β^δ converge toward the true parameter $q^\dagger$ as both the noise level δ and the regularization parameter β (chosen by an a priori rule) tend to zero (convergence).

The proof of the desired properties is standard (see [4], [5], [16], or [14]) once the weak closedness of the mapping $q \rightarrow \gamma u_q$ is provided.

PROPOSITION 3.2 (weak closedness). *For $q_n \rightharpoonup q \in K$ in $H^1(R)$ and $\gamma u_{q_n} \rightharpoonup y$ in $L^2(\Gamma_1)$, we have*

$$\gamma u_q = y. \tag{3.12}$$

Proof. From (2.7) we know that the sequence $\{u_{q_n}\}$ is bounded in $H^1(\Omega)$. Therefore, there exists a subsequence $\{u_{q_{n_k}}\}$ such that

$$u_{q_{n_k}} \rightharpoonup u^* \quad \text{in} \quad H^1(\Omega) \tag{3.13}$$

with $u^*|_{\Gamma_2} = g$. As the embedding of $H^1(\Omega)$ into $L^2(\Omega)$ is compact, we also have

$$u_{q_{n_k}} \rightarrow u^* \quad \text{in} \quad L^2(\Omega). \tag{3.14}$$

First, we prove that $u_{q_{n_k}} \rightharpoonup u_q$ in $H^1(\Omega)$, for which we have to show that $u^* = u_q$.

By the help of the triangle inequality, we get

$$\begin{aligned} &\left| \int_\Omega q_{n_k}(u_{n_k}) \nabla u_{n_k} \cdot \nabla v dx - \int_\Omega q(u^*) \nabla u^* \cdot \nabla v dx \right| \\ &\leq \left| \int_\Omega \{q_{n_k}(u_{n_k}) \nabla u_{n_k} - q(u^*) \nabla u_{n_k}\} \cdot \nabla v dx \right| \qquad (3.15) \\ &\quad + \left| \int_\Omega \{q(u^*) \nabla u_{n_k} - q(u^*) \nabla u^*\} \cdot \nabla v dx \right|. \qquad (3.16) \end{aligned}$$

Defining a linear functional l on $H^1(\Omega)$ by

$$l(u) = \int_\Omega q(u^*) \nabla v \cdot \nabla u dx,$$

we obtain from the weak convergence (3.13) that (3.16) vanishes for $k \rightarrow \infty$.

Applying once more the triangle inequality to (3.15) yields

$$\begin{aligned} &\left| \int_\Omega \{q_{n_k}(u_{n_k}) \nabla u_{n_k} - q(u^*) \nabla u_{n_k}\} \cdot \nabla v dx \right| \\ &\leq \left| \int_\Omega \{q_{n_k}(u_{n_k}) - q(u_{n_k})\} \nabla u_{n_k} \cdot \nabla v dx \right| \qquad (3.17) \\ &\quad + \left| \int_\Omega \{q(u_{n_k}) - q(u^*)\} \nabla u_{n_k} \cdot \nabla v dx \right|. \qquad (3.18) \end{aligned}$$

Because of (3.6) and the Cauchy–Schwarz inequality, we get

$$\begin{aligned} &\left| \int_\Omega (q_{n_k}(u_{n_k}) - q(u_{n_k}))\, \nabla u_{n_k} \cdot \nabla v dx \right| \\ &\le \|q_{n_k} - q\|_{C(I)} \|u_{n_k}\|_{H^1(\Omega)} \|v\|_{H^1(\Omega)} \\ &\le \tilde{C} \|q_{n_k} - q\|_{C(I)} \end{aligned} \tag{3.19}$$

for a constant $\tilde{C}$ not depending on n_k because of (2.7). Since the embedding of $H^1(I)$ into $C(I)$ is compact due to the finiteness of I, (3.17) tends to zero for $k \to \infty$.

Because of the boundedness of q, we also can apply the Cauchy–Schwarz inequality to (3.18) and obtain

$$\begin{aligned} &\left| \int_\Omega \{q(u_{n_k}) - q(u^*)\}\, \nabla u_{n_k} \cdot \nabla v dx \right| \\ &\le C \|q(u_{n_k})\nabla v - q(u^*)\nabla v\|_{L^2(\Omega)} \end{aligned} \tag{3.20}$$

by means of (2.7). Furthermore, the continuity of q and (3.14) yield $\lim_{k\to\infty} q(u_{n_k}(x)) = q(u_*(x))$ for almost every $x \in \Omega$. Because of $\partial v / \partial x_i \in L^2(\Omega)$ and the boundedness of q, the dominated convergence theorem (see [9]) shows that (3.20) vanishes for $k \to \infty$.

Summarizing these results, we obtain for $k \to \infty$ that

$$\int_\Omega q(u^*) \nabla u^* \cdot \nabla v dx = \int_\Omega f(x) v dx + \int_{\Gamma_1} h v d\Gamma_1 \quad \text{for all } \ v \in V$$

with $u^* - \tilde{g} \in V$. Hence u^* is the weak solution of (2.1)–(2.3) for the parameters q, f, g, and h. As the weak solution is unique, we conclude that $u^* = u_q$. Finally, as $u_{q_{n_k}} \rightharpoonup u_q$ holds for any subsequence $u_{q_{n_k}}$, we get

$$u_{q_n} \rightharpoonup u_q \ \text{ in } \ H^1(\Omega). \tag{3.21}$$

According to our assumption, we have $\gamma u_{q_n} \rightharpoonup y$ in $L^2(\Gamma_1)$; because of (3.21) and the continuity of the trace operator γ we also know $\gamma u_{q_n} \rightharpoonup \gamma u_q$ in $L^2(\Gamma_1)$. The uniqueness of the weak limit yields $\gamma u_q = y$. □

Hence existence, stability, and convergence of the regularized solutions are guaranteed. The special construction of the set K was only needed in order to derive estimate (3.19). Furthermore, the proof shows that Proposition 3.2 also holds if one considers $\tilde{K}$ as a set of admissible parameters.

4. Convergence rates.

Opposed to the general theory [4], we introduce a weak source condition for the convergence rate, which allows a full interpretation in section 5. Though based on concepts from [16], both the formulation and the proof of the convergence rate theorem are different from [16] since we now have to deal with boundary terms.

THEOREM 4.1. *Assume that there exists a function*

$$w \in V \tag{4.1}$$

such that

$$\left(q^\dagger - q^*, \psi\right)_I = \int_{\Gamma_1} \Psi(u_{q^\dagger}) \frac{\partial w}{\partial n} d\Gamma_1 \text{ for all } \psi \in H^1(I) \tag{4.2}$$

holds, where Ψ *is the antiderivative to* ψ*, fixed by*

$$\Psi(g) = 0. \tag{4.3}$$

Furthermore, assume that $\frac{\partial w}{\partial n} \in L^2(\Gamma_1)$ *with*

$$\Delta w = 0 \;\; in \;\; \Omega. \tag{4.4}$$

Then, with $\beta \sim \delta$*, we have*

$$\|\gamma u_{q_\beta^\delta} - z^\delta\|_{L^2(\Gamma_1)} = O(\delta)$$

and

$$\|q_\beta^\delta - q^\dagger\|_I = O(\sqrt{\delta}),$$

where q_β^δ *is the minimizer of* (3.11).

Proof. For the sake of simplicity, we now omit the explicit notation of γ. Then, as q_β^δ is a minimizer of (3.11), we get $J_\beta(q_\beta^\delta) \le J_\beta(q^\dagger)$. This implies

$$\|u_{q_\beta^\delta} - z^\delta\|_{L^2(\Gamma_1)}^2 + \beta\|q_\beta^\delta - q^*\|_I^2 \le \delta^2 + \beta\|q^\dagger - q^*\|_I^2,$$

from which we obtain

$$\begin{aligned} &\|u_{q_\beta^\delta} - z^\delta\|_{L^2(\Gamma_1)}^2 + \beta\|q^\dagger - q_\beta^\delta\|_I^2 \\ &\le \delta^2 + \beta\|q^\dagger - q^*\|_I^2 + \beta\left\{\|q^\dagger - q_\beta^\delta\|_I^2 - \|q_\beta^\delta - q^*\|_I^2\right\} \\ &= \delta^2 + 2\beta\left(q^\dagger - q^*, q^\dagger - q_\beta^\delta\right)_I . \end{aligned} \tag{4.5}$$

As integration by parts yields

$$\int_{\partial\Omega} \Psi(u_{q^\dagger})\frac{\partial w}{\partial n}dS - \int_\Omega \Psi(u_{q^\dagger})\Delta w dx = \int_\Omega \psi(u_{q^\dagger})\nabla u_{q^\dagger}\nabla w dx,$$

(4.3) and (4.4) give

$$\int_{\Gamma_1} \Psi(u_{q^\dagger})\frac{\partial w}{\partial n}d\Gamma_1 = \int_\Omega \psi(u_{q^\dagger})\nabla u_{q^\dagger}\nabla w dx$$

for the right-hand side of the source condition (4.2). Hence, choosing $\psi = q^\dagger - q_\beta^\delta$ in the source condition (4.2) leads to

$$\left(q^\dagger - q^*, q^\dagger - q_\beta^\delta\right)_I = \int_\Omega \left(q^\dagger(u_{q^\dagger}) - q_\beta^\delta(u_{q^\dagger})\right)\nabla u_{q^\dagger}\nabla w dx. \tag{4.6}$$

Using the direct problem formulation (see (2.5)) for $u_{q_\beta^\delta}$ and $u_{q^\dagger}$, we see by taking the difference that

$$\int_\Omega \left(q_\beta^\delta(u_{q_\beta^\delta})\nabla u_{q_\beta^\delta} - q^\dagger(u_{q^\dagger})\nabla u_{q^\dagger}\right)\cdot\nabla w dx = 0 \tag{4.7}$$

holds. Multiplying (4.6) by β and adding zero in the form of (4.7), it follows that

$$\begin{aligned} \beta\left(q^\dagger - q^*, q^\dagger - q_\beta^\delta\right)_I = \;\; &\beta\int_\Omega \left(q^\dagger(u_{q^\dagger}) - q_\beta^\delta(u_{q^\dagger})\right)\nabla u_{q^\dagger}\nabla w dx \\ &+ \beta\int_\Omega q_\beta^\delta(u_{q_\beta^\delta})\nabla u_{q_\beta^\delta}\cdot\nabla w dx \\ &- \beta\int_\Omega q^\dagger(u_{q^\dagger})\nabla u_{q^\dagger}\cdot\nabla w dx. \end{aligned} \tag{4.8}$$

We simplify the right-hand side of (4.8) to

$$I_1 = \beta \int_\Omega \left(q^\delta_\beta(u_{q^\delta_\beta})\nabla u_{q^\delta_\beta} - q^\delta_\beta(u_{q^\dagger})\nabla u_{q^\dagger}\right) \cdot \nabla w dx.$$

Using the antiderivative $Q^\delta_\beta \in S$ (see (3.8)) of q^δ_β, we obtain

$$I_1 = \beta \int_\Omega \left(\nabla Q^\delta_\beta(u_{q^\delta_\beta}) - \nabla Q^\delta_\beta(u_{q^\dagger})\right) \cdot \nabla w dx.$$

Integration by parts leads to

$$\begin{aligned} I_1 = \beta &\int_{\partial\Omega} \left(Q^\delta_\beta(u_{q^\delta_\beta}) - Q^\delta_\beta(u_{q^\dagger})\right)\frac{\partial w}{\partial n}\, dS \\ -\beta &\int_\Omega \left(Q^\delta_\beta(u_{q^\delta_\beta}) - Q^\delta_\beta(u_{q^\dagger})\right)\Delta w\, dx. \end{aligned}$$

Because of $u_{q^\delta_\beta}|_{\Gamma_2} = u_{q^\dagger}|_{\Gamma_2} = g$ and (4.4), we finally obtain

$$I_1 = \beta \int_{\Gamma_1} \left(Q^\delta_\beta(u_{q^\delta_\beta}) - Q^\delta_\beta(u_{q^\dagger})\right)\frac{\partial w}{\partial n}\, d\Gamma_1.$$

Next, we estimate I_1 by (3.9) and the Cauchy–Schwarz inequality in order to get

$$|I_1| \le \beta\alpha_2 \|u_{q^\dagger} - u_{q^\delta_\beta}\|_{L^2(\Gamma_1)} \left\|\frac{\partial w}{\partial n}\right\|_{L^2(\Gamma_1)}.$$

Applying the triangle inequality and Young's inequality

$$a \cdot b \le \varepsilon a^2 + \frac{b^2}{4\varepsilon}$$

for any $\varepsilon > 0$, we obtain

$$\begin{aligned} |I_1| &\le \beta\alpha_2 \|u_{q^\delta_\beta} - z^\delta\|_{L^2(\Gamma_1)} \left\|\frac{\partial w}{\partial n}\right\|_{L^2(\Gamma_1)} \\ &\quad + \beta\alpha_2 \|z^\delta - u_{q^\dagger}\|_{L^2(\Gamma_1)} \left\|\frac{\partial w}{\partial n}\right\|_{L^2(\Gamma_1)} \\ &\le \varepsilon\alpha_2^2\delta^2 + \frac{\beta^2}{2\varepsilon}\left\|\frac{\partial w}{\partial n}\right\|^2_{L^2(\Gamma_1)} \\ &\quad + \varepsilon\alpha_2^2 \|u_{q^\delta_\beta} - z^\delta\|^2_{L^2(\Gamma_1)}. \end{aligned}$$

Using this estimate for $|I_1|$, we get from (4.5) that

$$\begin{aligned} &\|u_{q^\delta_\beta} - z^\delta\|^2_{L^2(\Gamma_1)} + \beta\|q^\delta_\beta - q^\dagger\|^2_I \\ &\quad \le \delta^2 + 2\alpha_2^2\varepsilon\|u_{q^\delta_\beta} - z^\delta\|^2_{L^2(\Gamma_1)} \\ &\quad + 2\alpha_2^2\varepsilon\delta^2 + \frac{\beta^2}{\varepsilon}\left\|\frac{\partial w}{\partial n}\right\|^2_{L^2(\Gamma_1)}, \end{aligned}$$

which is equivalent to

$$\begin{aligned}\left\{1-2\alpha_2^2\varepsilon\right\}\|u_{q_\beta^\delta}-z^\delta\|^2_{L^2(\Gamma_1)}&\\ +\beta\|q_\beta^\delta-q^\dagger\|_I^2\le\delta^2+2\varepsilon\alpha_2^2\delta^2&\\ +\frac{\beta^2}{\varepsilon}\left\|\frac{\partial w}{\partial n}\right\|^2_{L^2(\Gamma_1)}&.\end{aligned}$$

With $\varepsilon<\frac{1}{2\alpha_2^2}$ we finally obtain that

$$\|u_{q_\beta^\delta}-z^\delta\|_{L^2(\Gamma_1)}=O(\delta) \tag{4.9}$$

(hence, by (3.10), also $\|u_{q_\beta^\delta}-z\|_{L^2(\Gamma_1)}=O(\delta)$) and

$$\|q_\beta^\delta-q^\dagger\|_I=O(\sqrt{\delta}) \tag{4.10}$$

hold. The constants in the O-terms in (4.9) and (4.10) can be derived from the proof. □

Now the theoretical analysis of our identification problem is complete. We have shown stability, convergence, and given conditions on the rate of convergence. In the next section, we give sufficient conditions for the existence of a source function w that satisfies (4.1)–(4.4), which allows the interpretation of the source condition (4.2). Again, we modify the approach of [16] appropriately.

5. Discussion of the source condition. Opposed to the general theory (compare with (1.2)), where Fréchet-differentiability of the forward operator F and Lipschitz continuity of $F'(q)$ already require a more regular parameter, the formulation of (4.2) itself does not impose any more regularity on q^* and on $q^\dagger$ than that they be in $H^1(I)$. Furthermore, in the general theory, the adjoint of $F'(q^\dagger)$ is needed, which makes the source condition usually very difficult to interpret. The new approach uses only the parameter-to-solution map $u_{q^\dagger}$ itself, which has a direct physical meaning, and not its linearization.

Usually, source conditions such as (1.2) mean severe restrictions on the parameter and are readily interpretable only in the one-dimensional case. We next construct a source function w for (4.2), even for the higher dimensional case, under quite natural conditions. The interpretation is based on our work in [16] but now for single boundary measurements.

If we denote the range of the true temperature $u_{q^\dagger}$ on Γ_1 (and hence on Ω (see section 3.1)) by the interval $I^\dagger=[I^\dagger_{\min},I^\dagger_{\max}]$, i.e.,

$$I^\dagger_{\min}=\min_{x\in\Gamma_1}u_{q^\dagger}\text{ and }I^\dagger_{\max}=\max_{x\in\Gamma_1}u_{q^\dagger},$$

we have

$$I^\dagger\subseteq I.$$

If we can assume to know the true parameter $q^\dagger$ outside the range $I^\dagger$ already from q^*, i.e., for $\rho:=q^\dagger-q^*$ we have

$$\rho=0\ \text{ for }\ \tau\in I\backslash I^\dagger, \tag{5.1}$$

the source condition (4.2) turns to

$$(\rho,\psi)_{I^\dagger} = \int_{\Gamma_1} \Psi(u_{q^\dagger}) \frac{\partial w}{\partial n} d\Gamma_1 \quad \text{for all } \psi \in H^1(I). \tag{5.2}$$

This is a very natural assumption as outside of $I^\dagger$ the parameter is of no use for the physical system. Also, from the inverse point of view, we can expect to identify only the parameter on the range of the true temperature as outside of $I^\dagger$ no information is available. Hence on $I \backslash I^\dagger$ $q^\dagger$ is already determined by the choice of q^*.

We now assume that

$$q^\dagger - q^* \in H^4(I) \tag{5.3}$$

and require the trace of the true temperature $u_{q^\dagger}$ to satisfy

$$\gamma u_{q^\dagger} : \Gamma_1 \to I^\dagger \ \text{ is Lipschitz.} \tag{5.4}$$

Then, because of the compact embedding $H^4(I^\dagger) \subset C^3(I^\dagger)$, our assumptions on the a priori knowledge about $\rho = q^\dagger - q^*$ (see (5.1)) result in

$$\rho^{(j)}(I^\dagger_{\text{min}}) = \rho^{(j)}(I^\dagger_{\text{max}}) = 0 \ \text{ for } \ j = 0, 1, 2, 3. \tag{5.5}$$

Because of assumption (5.4), the change of variables formula (see [9], [16]) can be applied with the transformation $t = \gamma u_{q^\dagger}$, whose level-sets are isotherms on Γ_1. This gives

$$\int_{\Gamma_1} s(x) J\gamma u_{q^\dagger}(x) d\Gamma_1 = \int_{I^\dagger} \left[\int_{\gamma u_{q^\dagger}^{-1}\{\tau\}} s dH^{n-2} \right] d\tau \tag{5.6}$$

for any L^{n-1}-summable function $s : R^{n-1} \to R$, where $J\gamma u_{q^\dagger}$ denotes the Jacobian of $\gamma u_{q^\dagger}$ and H^{n-2} is the $(n-2)$-dimensional Hausdorff measure. Next, we have to find a suitable function s in (5.6) for our purpose. First, we define m to be the $(n-2)$-dimensional Hausdorff-measure of the level sets of $\gamma u_{q^\dagger}$, i.e.,

$$m(\tau) = \int_{\gamma u_{q^\dagger}{}^{-1}\{\tau\}} dH^{n-2} \ \text{ for } \ \tau \in I^\dagger,$$

and assume the trace of $u_{q^\dagger}$ on Γ_1 to behave such that

$$(\rho'''(\gamma u_{q^\dagger}) - \rho'(\gamma u_{q^\dagger})) \cdot \frac{1}{m(\gamma u_{q^\dagger})} \cdot J\gamma u_{q^\dagger} \in L^2(\Gamma_1). \tag{5.7}$$

The only term that might cause a violation of (5.7) is $\frac{1}{m(\gamma u_{q^\dagger})}$. In two dimensions $(n = 2)$, the H^{n-2}-measure is the counting measure. Then we have $m(\tau) \neq 0$ on $I^\dagger$ as every $\tau \in I^\dagger$ is at least attained once by $u_{q^\dagger}$ for $x \in \Gamma_1$ (by definition). This is distinct from [16], where even in two dimensions condition (5.7) cannot be guaranteed a priori. In three dimensions (5.7) is certainly fulfilled if $m(\gamma u_{q^\dagger})$ is bounded away from 0, i.e., if all temperatures in $I^\dagger$ are assumed on sets of nonvanishing H^1-measure and if these measures depend in a reasonable way on the temperatures. This is a (weak) regularity condition on the measures of the isotherms, and it is reasonable, since one cannot expect identifiability of $q^\dagger$ for temperatures which are assumed only on a "small" set (of H^1-measure zero). But even such "small isotherms" are not excluded by (5.7): The only temperature values possibly attained by $\gamma u_{q^\dagger}$ at a set of H^1-measure zero are $I^\dagger_{\min}$ or $I^\dagger_{\max}$. This is a consequence of the continuity of the trace of $u_{q^\dagger}$ on Γ_1 and the intermediate value theorem. However, both the first term in the product (see (5.5)) and $J\gamma u_{q^\dagger}$ (necessary condition for an extremum) vanish in the respective critical situation. Now if the product of these two expressions tends faster to zero than $m(\gamma u_{q^\dagger})$, the L^2-boundedness in (5.7) is maintained even then.

Now we again omit the explicit notation of γ and look at the Poisson equation

$$\Delta w = 0 \text{ in } \Omega, \tag{5.8}$$

$$\frac{\partial w}{\partial n} = (\rho'''(u_{q^\dagger}) - \rho'(u_{q^\dagger})) \cdot \frac{1}{m(u_{q^\dagger})} \cdot J u_{q^\dagger} \text{ on } \Gamma_1, \tag{5.9}$$

$$w = 0 \text{ on } \Gamma_2, \tag{5.10}$$

for which the existence of a unique (weak) solution $w \in V$ is guaranteed because of (5.7). Similarly as in [16], one can show that the solution w of (5.8)–(5.10) satisfies the source condition (4.2). Note that the conditions (4.1) and (4.4) of the convergence rate theorem are automatically fulfilled by this approach. The essential assumptions needed for the proof are

- sufficient smoothness of $q^\dagger - q^*$ as required in (5.3);
- sufficient smoothness of the trace of $u_{q^\dagger}$ on Γ_1: (5.4);
- sufficient knowledge about $q^\dagger$ on the boundary of the temperature interval where measurements are available: (5.5);
- condition (5.7) (needed only for $n = 3$), which essentially says that the isotherms of $u_{q^\dagger}$ on Γ_1 depend in a sufficiently regular way on the temperature level, where this regularity is rather weak.

Since under these conditions the source condition can be verified, the convergence rates from Theorem 4.1 are valid.

In any dimension, the heat dependent conductivity is identified as a function of one variable. Hence it is remarkable that the interpretation edges down more in two dimensions, where the boundary temperature represents only a one-dimensional data manifold, than in three, where a two-dimensional data manifold is available.

6. Variants. In section 2, we introduced a nonlinear mixed boundary problem for which we considered in section 4 a constant boundary temperature u_0 on Γ_2 for technical reasons. From the practical (with respect to industry) point of view, the pure Neumann-type problem

$$-\nabla \cdot (q(u)\nabla u) = f(x) \text{ in } \Omega, \tag{6.1}$$

$$q(u)\frac{\partial u}{\partial n} = h \text{ on } \partial\Omega \tag{6.2}$$

could be more realistic, where (6.2) then describes, e.g., the cooling of a steel strand (see [11]). If we choose the space of test functions V as

$$V = \left\{ v \in H^1(\Omega) \mid \int_\Omega v dx = 0 \right\}$$

and require

$$\int_\Omega f dx + \int_{\partial\Omega} h dS = 0,$$

the existence of a unique weak solution u_q in V to (6.1)–(6.2) can be guaranteed by [22]. Then the theory developed in sections 3 and 4 remains valid if we replace Γ_1 by

$\partial\Omega$. This means that the measurements are now done on all of $\partial\Omega$, and the Tikhonov functional is

$$J_\beta(q) = \int_{\partial\Omega} |u_q - z^\delta|^2 dS + \beta\|q - q^*\|_I^2.$$

The source condition

$$\left(q^\dagger - q^*, \psi\right)_I = \int_{\partial\Omega} \Psi(u_{q^\dagger}) \frac{\partial w}{\partial n} dS \quad \text{for all } \psi \in H^1(I), \tag{6.3}$$

which yields the rates

$$\|u_{q_\beta^\delta} - z^\delta\|_{L^2(\partial\Omega)} = O(\delta)$$

and

$$\|q_\beta^\delta - q^\dagger\|_I = O(\sqrt{\delta}),$$

can be interpreted as in section 5. Note that in order to show the existence of a unique (weak) solution to problem (5.8)–(5.10) in V (with $\Gamma_1 = \partial\Omega$, $\Gamma_2 = \emptyset$), we now in addition have to check

$$\int_{\partial\Omega} \frac{\partial w}{\partial n} dS = 0. \tag{6.4}$$

However, this follows from applying the coarea formula (5.6) with

$$s(x) = (\rho'''(u_{q^\dagger}(x)) - \rho'(u_{q^\dagger}(x))) \frac{1}{m(u_{q^\dagger}(x))},$$

which gives

$$\begin{aligned}\int_{\partial\Omega} \frac{\partial w}{\partial n} dS &= \int_{I^\dagger} (\rho'''(\tau) - \rho'(\tau)) \frac{1}{m(\tau)} m(\tau) d\tau \\ &= \int_{I^\dagger} (\rho''' - \rho') d\tau \\ &= 0, \end{aligned} \tag{6.5}$$

where the last equality holds because of (5.5).

Finally, we change the settings of the direct problem to the pure Dirichlet case:

$$-\nabla \cdot (q(u)\nabla u) = f(x) \ \text{ in } \ \Omega, \tag{6.6}$$

$$u = g \ \text{ on } \ \partial\Omega. \tag{6.7}$$

The existence of a unique solution in $H^1(\Omega)$ is given by the theory quoted in section 2. As now the temperature flux $q(u)\frac{\partial u}{\partial n}$ is measured on (all of) the boundary $\partial\Omega$, we need a higher regularity of the solution u_q than in our previous discussions if we still want to measure our data in the L^2-setting, i.e., consider that the measurements are in L^2. If we choose the set of admissible parameters

$$\hat{K} = \left\{ q \in H^2(I) \mid \alpha_1 \le q(\tau) \le \alpha_2 \text{ for } \tau \in I \ \ \wedge \ \ \left\|\frac{dq}{d\tau}\right\|_{L^\infty(I)} \le \infty \right\}$$

and require

$$g \in H^{3/2}(\partial\Omega),$$

then, for $q \in \hat{K}$, the existence of a unique solution u_q even in $H^2(\Omega)$ can be shown (see [2]). Now, the continuous embedding $H^2(R) \subseteq C^1(R)$ and the trace theorem yield

$$q(u_q)\frac{\partial u_q}{\partial n} \in H^{1/2}(\partial\Omega).$$

Hence the Tikhonov functional

$$J_\beta(q) = \int_{\partial\Omega} |\gamma_q u_q - z^\delta|^2 dS + \beta\|q - q^*\|_I^2$$

with

$$\begin{aligned} \gamma_q : H^2(\Omega) &\to L^2(\partial\Omega), \\ u &\to q(u)\frac{\partial u}{\partial n} \end{aligned}$$

is meaningful. Under the source condition

$$\left(q^\dagger - q^*, \psi\right)_I = \int_{\partial\Omega} \Psi(g)\frac{\partial w}{\partial n} dS \quad \text{for all } \psi \in H^1(I) \tag{6.8}$$

(g now is no longer needed to be constant), the rates

$$\|q_\beta^\delta - q^\dagger\|_I = O(\sqrt{\delta})$$

and

$$\|\gamma_{q_\beta^\delta} u_{q_\beta^\delta} - z^\delta\|_{L^2(\partial\Omega)} = O(\delta)$$

can be shown. Note that now (6.8) does not even depend explicitly on the unknown temperature $u_{q^\dagger}$, which is a major difference from the theory of convergence rate developed so far. Once more, the source function $w \in H^1(\Omega)$ can be found as the solution of

$$\begin{aligned} &\Delta w = 0 \text{ in } \Omega, \\ &\frac{\partial w}{\partial n} = (\rho'''(u_{q^\dagger}) - \rho'(u_{q^\dagger})) \cdot \frac{1}{m(u_{q^\dagger})} \cdot J u_{q^\dagger} \text{ on } \partial\Omega, \end{aligned}$$

for the proof of the convergence rate the variational formulation

$$\int_\Omega q(u)\nabla u \cdot \nabla v dx = \int_\Omega h(x) v dx + \int_{\partial\Omega} q(u)\frac{\partial u}{\partial n} v dS$$

with test functions $v \in H^1(\Omega)$ is needed. The advantage of considering pure Dirichlet data for the direct problem is that the condition

$$\gamma u_{q^\dagger} : \partial\Omega \to I^\dagger \text{ is Lipschitz}$$

can now be automatically satisfied by choosing the problem input g regular enough because of

$$\gamma u_{q^\dagger} = g. \tag{6.9}$$

Furthermore, we now can drive the interval $I^\dagger$ on which we want to identify the parameter in a straightforward way by the choice of g.

Finally, we mention that in all our optimization problems, the H^1-regularization term can be replaced by a L^2-term; i.e., we can consider

$$\min J(q) := \left\{L^2\text{-norm of residual}\right\}^2 + \beta \int_I (q-q^*)^2 d\tau. \tag{6.10}$$

Results for stability, convergence, and rate of convergence can be proven in a completely analogous way, where the H^1-scalar product $(q^\dagger - q^*, \psi)_I$ in (4.2) is replaced by the L^2-scalar product $\int_I (q^\dagger - q^*)\psi d\tau$. A solution to the source condition can be constructed as in section 5 under even weaker regularity assumptions on $q^\dagger$.

The existence of minimizers of (6.10) cannot be guaranteed by Theorem 3.2, but this difficulty can be resolved by incorporating a tolerance η into the minimization, i.e., by replacing minimizers of (6.10) by elements $q^\delta_{\beta,\eta}$ such that

$$J(q^\delta_{\beta,\eta}) \le \inf J(q) + \eta.$$

As long as $\eta = O(\delta^2)$, all proofs carry over (see [5]). This can of course also be done for (3.11).

7. Numerical experiments.

In order to test the identification of the heat conductivity by Tikhonov regularization, we carry out numerical simulations using the temperature trace $u|_{\Gamma_1}$ as data. Considering a rectangular domain $\Omega = [0, 0.5] \times [0, 2]$ with boundaries $\Gamma_1 = \{0\} \times [0, 2] \cup [0, 0.5] \times \{2\} \cup \{0.5\} \times [0, 2]$ and $\Gamma_2 = [0, 0.5] \times \{0\}$ and a temperature field

$$u_{q^\dagger}(x, y) = \frac{y}{2}, \tag{7.1}$$

we want to recover the nonlinearity in

$$\begin{aligned} -\nabla \cdot ((2 + \cos(2\pi u))\nabla u) &= \pi \cdot \sin(\pi y) \ \text{ in } \ \Omega, \\ (2 + \cos(2\pi u))\frac{\partial u}{\partial n} &= 0 \ \text{ on } \ \Gamma_1, \\ u &= 0 \ \text{ on } \ \Gamma_2 \end{aligned}$$

from “observations” of

$$u|_{\Gamma_1}. \tag{7.2}$$

While we already know the true data $u_{q^\dagger}|_{\Gamma_1}$ by construction, a sequence of noisy data z^{δ_i} is generated by artificially perturbing $u_{q^\dagger}|_{\Gamma_1}$ with high frequency noise. Then the regularized solutions $q^{\delta_i}_\beta$ are defined as the minimizers of

$$J_\beta(q) = \int_{\Gamma_1} |\gamma u_q - z^{\delta_i}|^2 d\Gamma_1 + \beta \|q - q^*\|^2_I. \tag{7.3}$$

Though the true data $u_{q^\dagger}|_{\Gamma_1}$ (and hence the temperature distribution $u_{q^\dagger}$) cover only the range $I^\dagger = [0, 1]$, we choose a larger interval

$$I = [-0.2, 1.2] \tag{7.4}$$

in (7.3), as both noisy data and computed forward solutions during the minimization procedure may exceed $I^\dagger$. Of course, we can only expect to recover the heat conductivity on $I^\dagger$; that outside it will be determined by the initial guess q^*. With

$$q^* = 3, \tag{7.5}$$

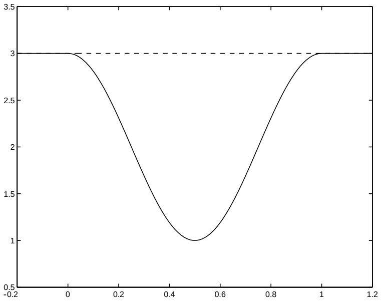


FIG. 1. $q^\dagger$ *and* q^*.

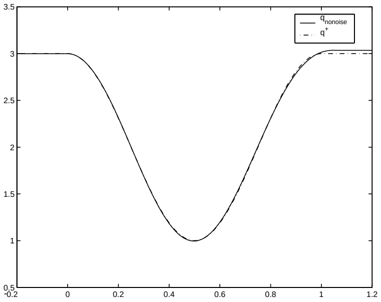


FIG. 2. $q_{nonoise}$ *identified from exact data.*

we then have (see Figure 1)

$$q^\dagger(\tau) = \begin{cases} 2 + \cos(2\pi\tau) & \text{for} \quad \tau \in I^\dagger, \\ 3 & \text{for} \quad \tau \in I \backslash I^\dagger. \end{cases} \tag{7.6}$$

Since we know the exact parameter $q^\dagger$, we can compute the error $\|q^\dagger - q_\beta^{\delta_i}\|_{H^1(I)}$, allowing us to investigate the behavior of Tikhonov regularization with respect to stability and rate of convergence. Note that we at least satisfy condition (5.5) for $j = 1, 2$ by the choice of q^* and in addition conditions (5.4) and (5.7) by the construction of our example.

For the minimization of (7.3) we use a quasi-Newton method, approximating the Hessian matrix of J_β by a BFGS-update formula in each iteration step k. Given a search direction p_k by that rule, the parameter q_k is updated by

$$q_{k+1} = q_k + \alpha_k p_k$$

until a minimum is reached. In order to raise the convergence speed of the optimization procedure, we also use a line search algorithm for the determination of the stepsize α_k. A more detailed discussion can be found in [15].

The first computations were done for $\beta = 0$, i.e., the approach to identify the parameter by simply minimizing the output least squares term

$$\int_{\Gamma_1} |\gamma u_q - z^{(\delta)}|^2 d\Gamma_1.$$

For the case of exact data $z = u_{q^\dagger}|_{\Gamma_1}$, the result $q_{nonoise}$ is shown in Figure 2. As predicted by our theory, the parameter is identifiable from observations of the temper-

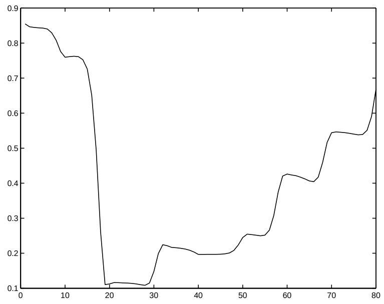


FIG. 3. $\frac{\|q^\dagger - q_k\|_I}{\|q^\dagger\|_I}$ *vs. k.*

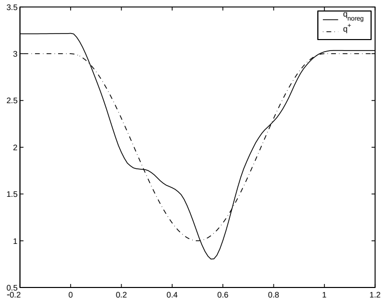


FIG. 4. *Solution* q_{noreg} *after* 80 *steps.*

ature trace on the boundary, but of course only on the interval $I^\dagger$, where the data is available. Outside, the solution is given by the initial guess q^*. Figures 3 and 4 illustrate the ill-posedness of the identification problem. Perturbed data z^{δ_7} with 4.61% noise already have a dramatic impact on the recovery process. On the left-hand side, the relative error $\frac{\|q^\dagger - q_k\|_I}{\|q^\dagger\|_I}$ is plotted vs. the iteration index k in the optimization routine; the right-hand sides records the result q_{noreg} after 80 steps. While the error in the residual $\|\gamma u_{q_k} - z^\delta\|_{L^2(\Gamma_1)}$ (not shown) is monotonically decreasing with k, the error in the parameter starts to increase after some 20 steps, leading to a solution that differs from $q^\dagger$ by more than 60% measured in the $H^1(I)$-norm. Only by introducing the penalty term

$$\beta\|q - q^*\|_I^2$$

(or, alternatively, stopping the iteration at "the right time" (see [4] for an introduction to iterative regularization methods)) these high numerical instabilities can be overcome. Though there are sophisticated methods for choosing the regularization parameter β from the knowledge of the noise level δ and the data z^δ (see [20]) itself, we content ourselves with the a priori choice

$$\beta_i = 4 \cdot 10^{-4} \cdot \delta_i \tag{7.7}$$

for the sequence of perturbed data z^{δ_i}. This relation was found by trial and error, which is sufficient for our purposes. In order to test the rate of convergence behavior of Tikhonov regularization predicted by Theorem 4.1, we need only to meet the requirement $\beta \sim \delta$. Figure 5 shows the error $\|q^\dagger - q_{\beta_i}^{\delta^i}\|_I$ plotted vs. the noise

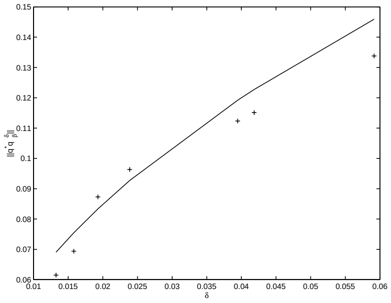


FIG. 5. *Convergence rate* $\|q^\dagger - q^{\delta_i}_{\beta_i}\|_I = O(\sqrt{\delta_i})$.

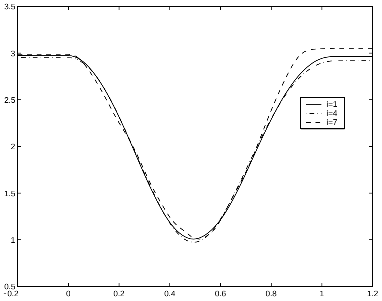


FIG. 6. *Regularized solutions* $q^{\delta_i}_{\beta_i}$.

level $\delta_i = \|\gamma u_{q^\dagger} - z^{\delta_i}\|_{L^2(\Gamma_1)}$. The solid line indicates that the convergence speed $\|q^\dagger - q^{\delta_i}_{\beta_i}\|_I = O(\sqrt{\delta_i})$ from Theorem 4.1 is obeyed, even though not all conditions of section 5 are satisfied by our example. This gives hope that a source function w in Theorem 4.1 can be found under even weaker assumptions than those made in section 5. Finally, Figure 6 shows the regularized solutions $q^{\delta_i}_{\beta_i}$ for $\delta_1 = 0.0132, \delta_4 = 0.0239$, and $\delta_7 = 0.0529$.

Acknowledgment. The author thanks Professor Masahiro Yamamoto (University of Tokyo) and Professor Jun Zou (Chinese University of Hong Kong) for fruitful discussions and the Sohmen Foundation (Linz) for funding the stay in Asia. The numerical part of this work was done while the author was visiting OCIAM (University of Oxford).

REFERENCES

[1] G. ALESSANDRINI AND V. ISAKOV, *Analyticity and uniqueness for the inverse conductivity problem*, Rend. Istit. Mat. Univ. Trieste, 28 (1996), pp. 351–369.

[2] B. BLASCHKE, *Some Newton Type Methods for the Regularization of Nonlinear Ill-Posed Problems*, Dissertation, Johannes Kepler Universität Linz, Linz, Austria, 1996.

[3] D. GILBARG AND N. TRUDINGER, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, New York, 1977.

[4] H. W. ENGL, M. HANKE, AND A. NEUBAUER, *Regularization of Inverse Problems*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1996.

[5] H. W. ENGL, K. KUNISCH, AND A. NEUBAUER, *Convergence rates for Tikhonov regularization of nonlinear ill-posed problems*, Inverse Problems, 5 (1989), pp. 523–540.

[6] H. W. Engl, T. Langthaler, and P. Manselli, *On an inverse problem for a nonlinear heat equation connected with continuous casting of steel*, in Optimal Control of Partial Differential Equations II, Internat. Schriftenreihe Numer. Math. 78, K. H. Hoffmann and W. Krabs, eds., Birkhäuser, Basel, 1987, pp. 67–89.

[7] H. W. Engl and T. Langthaler, *Control of the solidification front by secondary cooling in continuous casting of steel*, in Case Studies in Industrial Mathematics, H. W. Engl, H. Wacker, and W. Zulehner, eds., Teubner, Stuttgart, 1988, pp. 51–77.

[8] H. W. Engl and J. Zou, *A new approach to convergence rate analysis of Tikhonov regularization for parameter identification in heat conduction*, Inverse Problems, 16 (2000), pp. 1907–1923.

[9] L. Evans and R. Gariepy, *Measure Theory and Fine Properties of Functions*, CRC Press, Boca Raton, FL, 1992.

[10] A. Friedman and V. Isakov, *On the uniqueness in the inverse conductivity problem with one measurement*, Indiana Univ. Math. J., 38 (1989), pp. 563–579.

[11] W. Grever, *A nonlinear parabolic initial boundary value problem modelling the continuous casting of steel*, ZAMM Z. Angew. Math. Mech., 78 (1998), pp. 109–119.

[12] V. Isakov, *Inverse Problems for Partial Differential Equations*, Springer-Verlag, New York, 1998.

[13] V. Isakov and J. Powell, *On the inverse conductivity problem with one measurement*, Inverse Problems, 6 (1990), pp. 311–318.

[14] K. Kunisch and W. Ring, *Regularization of nonlinear illposed problems with closed operators*, Numer. Funct. Anal. and Optim., 14 (1993), pp. 389–404.

[15] P. Kügler, *Identification of a Temperature Dependent Heat Conductivity by Tikhonov Regularization*, Diploma Thesis, Johannes Kepler University Linz, Linz, Austria, 2000.

[16] P. Kügler and H. W. Engl, *Identification of a temperature dependent heat conductivity by Tikhonov regularization*, J. Inverse Ill-Posed Probl., 10 (2002), pp. 67–90.

[17] G. Lieberman, *Mixed boundary value problems for elliptic and parabolic differential equations of second order*, J. Math. Anal. Appl., 113 (1986), pp. 422–440.

[18] A. Nachman, *Global uniqueness for a two-dimensional inverse boundary value problem*, Ann. of Math. (2), 143 (1996), pp. 71–96.

[19] O. Scherzer, H. W. Engl, and R. S. Anderssen, *Parameter identification from boundary measurements in a parabolic equation arising from geophysics*, Nonlinear Anal., 20 (1993), pp. 127–156.

[20] O. Scherzer, H. W. Engl, and K. Kunisch, *Optimal a posteriori choice for Tikhonov regularization for solving nonlinear ill-posed problems*, SIAM J. Numer. Anal., 30 (1993), pp. 1796–1838.

[21] E. Schock, *Approximate solution of ill-posed equations: Arbitrarily slow convergence vs. superconvergence*, in Constructive Methods for the Practical Treatment of Integral Equations, G. Hämmerlin and K. Hoffmann, eds., Birkhäuser, Basel, 1985, pp. 234–243.

[22] R. E. Showalter, *Monotone Operators in Banach Space and Nonlinear Partial Differential Equations*, AMS, Providence, RI, 1997.

[23] Z. Sun, *On a quasilinear inverse boundary value problem*, Math. Z., 221 (1996), pp. 293–305.

[24] J. Sylvester and G. Uhlmann, *A global uniqueness theorem for an inverse boundary value problem*, Ann. of Math. (2), 125 (1987), pp. 153–169.

SIAM J. NUMER. ANAL.
Vol. 41, No. 4, pp. 1564–1584

TIME INTEGRATION OF THE DUAL PROBLEM OF ELASTOPLASTICITY BY RUNGE–KUTTA METHODS*

JÖRG BÜTTNER[†] AND BERND SIMEON[‡]

Abstract. Implicit Runge–Kutta methods for the dual problem of elastoplasticity are analyzed and classified. The choice of Runge–Kutta time integration is inspired by the problem structure, which consists of a coupled system of balance equations and unilaterally constrained evolution equations and which can be viewed as an infinite-dimensional differential-algebraic equation. Focussing on the time axis and leaving the space variables continuous, a grid-independent existence and uniqueness result is given for the class of coercive Runge–Kutta methods. Moreover, contractivity preservation and convergence are shown for methods that are also algebraically stable.

Key words. elastoplasticity, implicit Runge–Kutta methods, time integration, PDAE

AMS subject classifications. 65L05, 65L80, 65M12, 74C05

DOI. 10.1137/S0036142902402821

1. Introduction.

We analyze implicit Runge–Kutta methods for an infinite-dimensional system of constrained evolution equations, the so-called dual problem of elastoplasticity. Our approach discretizes only the time axis and leaves the space variables continuous. In this way, we obtain a grid-independent formulation and lay the foundation for an implementation in the fashion of Rothe's method. The main results are first existence and uniqueness of the numerical solution for coercive Runge–Kutta methods, second contractivity preservation for algebraically stable methods, and third convergence for methods that feature both properties as well as a certain stage order.

Elastoplastic models are used for materials where the deformation process shows a time-dependent and irreversible behavior. Applications comprise, e.g., the stretch formation of metal sheets, wear effects in turbine blades, and the behavior of micromechanical devices. The mathematical model consists of a coupled system of balance equations and unilaterally constrained evolution equations, where the first set of equations stands for the balance of momentum and the second for the properties of the material under consideration [16]. We consider here materials that satisfy the principle of maximum plastic dissipation and possess a quadratic internal free energy.

In elastoplasticity, numerical simulation is mostly based on the method of lines [15, 16]. More precisely, one discretizes the balance equations by the finite element method (FEM) and reduces at the same time the evolution equations to the quadrature nodes of the grid. In these nodes, the evolution is integrated in time, with the implicit Euler as standard method and the return-mapping scheme for the handling of the unilateral constraint. Due to this constraint, the so-called yield condition, one deals actually with a differential-algebraic equation (DAE) of index 2 [8, 12].

Numerical experiments [4, 7, 18] indicate that at least order 2 is possible for

*Received by the editors February 18, 2002; accepted for publication (in revised form) February 3, 2003; published electronically October 2, 2003. This work was supported by the German Research Council DFG within SFB 298.

http://www.siam.org/journals/sinum/41-4/40282.html

[†] Institut für Wissenschaftliches Rechnen und Mathematische Modellbildung (IWRMM), University of Karlsruhe, Kaiserstraße 12, 76128 Karlsruhe, Germany (joerg.buettner@imf.fzk.de).

[‡]Technical University of Munich, Boltzmannstraße 3, 85748 Garching, Germany (simeon@mathematik.tu-muenchen.de).

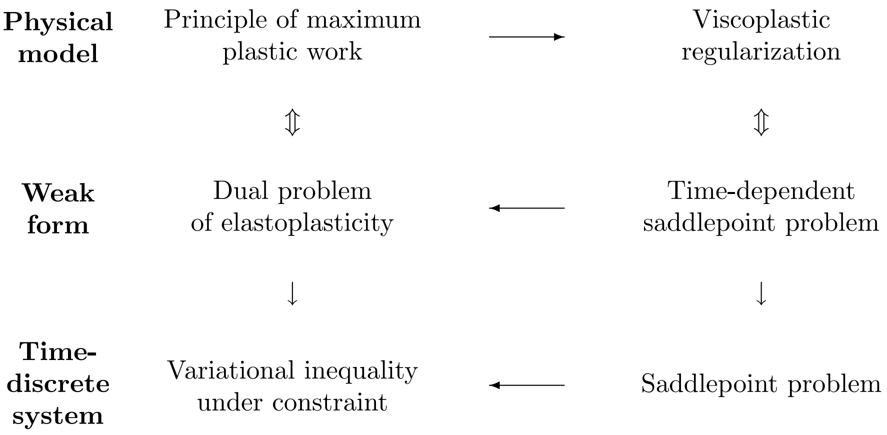

Physical model
Principle of maximum plastic work
Viscoplastic regularization
Weak form
Dual problem of elastoplasticity
Time-dependent saddlepoint problem
Time-discrete system
Variational inequality under constraint
Saddlepoint problem

Fig. 1.1. *Dual problem and viscoplastic regularization.*

the time integration. Clearly, higher order methods can use larger time steps than the implicit Euler method and promise a more efficient simulation, particularly if a variable stepsize algorithm is employed. But which methods should be applied to elastoplasticity?

Before discretization, the equations can be viewed as an infinite-dimensional system of DAEs or as a PDAE (partial differential-algebraic equation). With the numerical analysis of DAEs in mind (see [3, 10, 14]), we focus therefore on the versatile class of implicit Runge–Kutta methods. In combination with the method of lines, this choice turned out to be successful [4, 7], and our objective is now the extension to Rothe's method, i.e., the direct application to the infinite-dimensional problem.

For the special case of the implicit Euler method, existence and uniqueness of the numerical solution have been shown recently [9]. Going one step further, we present results for the large class of implicit Runge–Kutta methods. In particular, we use the notions of coercivity and of algebraic stability to characterize those methods that are applicable in elastoplasticity. Various time integration methods become available in this way, among them Gauss, Radau, Lobatto, and several DIRK (diagonally implicit Runge–Kutta) methods.

Many authors have considered numerical methods in elastoplasticity. We mention here the recent monograph of Han and Reddy [9], which serves as extensive reference and from which we adopted both notation and some fundamental techniques. Wieners [18, 19] concentrated on multigrid methods for large-scale problems but also introduced, in cooperation with Ellsiepen [7], DIRK methods for the time integration. Moreover, Carstensen [5] coupled FEM and BEM for this problem class and studied domain decomposition techniques, while Armero and Pérez-Foguet analyzed closest point projection algorithms [1, 13]. The monograph of Simo and Hughes [16] represents the basic reference of the engineering literature.

Figure 1.1 is a road map of the paper. It illustrates the relation between the original problem formulation and its regularized counterpart, the so-called viscoplastic regularization due to Duvaut and Lions [6]. The latter has no yield condition and is thus much easier to analyze. Using a limit process, the results on the regularized problem both in the time-continuous and in the time-discrete case can be carried over to the original problem.

More specifically, the outline of the paper is as follows: Section 2 summarizes the mathematical model and the necessary framework. On one hand, the principle of maximum plastic dissipation leads to the aforementioned coupled system, which we express, following the approach of [9], in terms of displacement and generalized stress. The corresponding weak form defines the dual problem of elastoplasticity. On the other hand, the viscoplastic regularization yields a time-dependent saddlepoint problem. We also provide in this section some basic results on the existence and uniqueness of solutions.

In section 3, the implicit Runge–Kutta discretization is introduced, and the main results of the paper are stated. It turns out that the coefficient matrix of the Runge–Kutta method plays a crucial role for existence, contractivity, and convergence of the numerical solution. The existence proof makes use of the viscoplastic regularization and starts with the corresponding time-discrete saddlepoint problem. Letting the viscosity parameter tend to zero, we obtain the desired result for the variational inequality that stems from the discretization of the dual problem. The stability and convergence results, however, use techniques from the theory of B-stability for ODEs. For better readability, the proofs are given in separate sections 4 and 5. A conclusion closes the paper.

2. Mathematical model.

Kinematics, dynamics, and material law fully describe the elastoplastic deformation process. More precisely, the kinematic equation determines the geometric properties of the deformation, the dynamic equation expresses the balance of momentum, and the constitutive equations characterize the material under consideration. In this section, we state these three ingredients of the mathematical model and summarize some important properties. We adopt the notation of [9] and refer the interested reader to [9, 11, 16] for more details.

2.1. Kinematic relations and balance law.

Let the undeformed elastoplastic body occupy the region $\Omega \subset \boldsymbol{R}^3$. The deformation maps each material point $x \in \Omega$ at time t to its current position $x + u(x,t)$ with displacement field u. Without loss of generality, homogeneous Dirichlet boundary conditions are assumed,

$$u = 0 \qquad \text{on } \partial\Omega.$$

Here and in the following, we mostly drop the arguments x and t for convenience.

We consider small strains only, and consequently the (total) strain tensor ε is given by

$$\varepsilon = \varepsilon(u) := \frac{1}{2}\left(\nabla u + \nabla u^T\right). \tag{2.1}$$

By a tensor we mean usually a second order tensor in matrix representation, i.e., $\varepsilon = (\varepsilon_{ij})_{3\times 3}$. Like the strain tensor ε, the stress tensor σ also depends on time t, i.e., $\sigma = \sigma(x,t)$.

Material laws or constitutive equations relate stress and strain. We will come to this point in the next subsection. Before that, however, we state the balance of momentum for the quasi-static case,

$$\operatorname{div} \sigma + f = 0 \qquad \text{in } \Omega, \tag{2.2}$$

where $f = f(x,t)$ denotes some given volume force.

2.2. Constitutive equations. The constitutive equations of an elastoplastic material follow from thermodynamical considerations; see [9, 11, 17]. One basic assumption is that the total strain ε of (2.1) splits into an elastic part ε^e and a plastic part ε^p,

$$\varepsilon = \varepsilon^e + \varepsilon^p.$$

Using plastic strain ε^p and a vector of r so-called internal variables $\xi = (\xi_1, \dots, \xi_r)$, the state of a material point at a certain instant of time is modelled by the generalized plastic strain $P := (\varepsilon^p, \xi)$.

Another important quantity is the internal free energy function $\psi = \psi(\varepsilon^e, \xi)$, which comprises the usual stored energy function of elasticity and additional contributions from processes like hardening. Furthermore, the free energy defines the generalized stress $\Sigma := (\sigma, \chi)$ with stress tensor σ and conjugate force χ by

$$\sigma = \frac{\partial}{\partial \varepsilon^e}\psi, \tag{2.3}$$

$$\chi = -\frac{\partial}{\partial \xi}\psi. \tag{2.4}$$

One also says that the generalized stress is conjugate to the generalized strain.

As an example, consider linear hardening, where the quadratic internal free energy reads

$$\psi(\varepsilon^e, \xi) = \underbrace{\frac{1}{2}\varepsilon^e : C : \varepsilon^e}_{=:\psi^e(\varepsilon^e)} + \underbrace{\frac{1}{2}\xi \cdot H \cdot \xi}_{=:\psi^p(\xi)}\ , \tag{2.5}$$

with elasticity tensor C, regular $r \times r$ symmetric positive-definite matrix H of hardening moduli, and hardening variables ξ. Here we abbreviate tensor products by a ":", whereas a "$\cdot$" (or a blank) stands for matrix-vector products.

In case of linear hardening, the derivative (2.3) yields Hooke's law

$$\sigma = C : (\varepsilon - \varepsilon^p), \tag{2.6}$$

and (2.4) becomes

$$\chi = -H\xi.$$

In this paper, we restrict the discussion to materials that possess an internal free energy of the form (2.5).

At this point, all relevant variables of the constitutive equations have been introduced. However, a restriction on the generalized stress Σ has still to be taken into account. More specifically, Σ is required to lie in the set E of admissible generalized stresses. This convex set is characterized by

$$E = \{\Sigma \,|\, \phi(\Sigma) \le 0\} \tag{2.7}$$

with convex function ϕ, the so-called yield function. In the particular example of linear isotropic hardening where we have only one scalar conjugate force χ and a scalar hardening modulus H, the yield function is

$$\phi(\sigma, \chi) = \|\text{dev } \sigma\| - \sqrt{\frac{2}{3}}(\sigma_0 - \chi), \tag{2.8}$$

with given constant $\sigma_0 > 0$ and dev $\sigma = \sigma - 1/3 \text{ tr } \sigma$.

The constitutive equations follow now from the local *principle of maximum plastic work*. The rate of plastic work at a point in the material body is given by

$$W(\dot{P}) = \Sigma : \dot{P} := \sigma : \dot{\varepsilon}^p + \chi \cdot \dot{\xi}\,.$$

Given a rate of change $\dot{P}$ in the generalized plastic strain, the actual stress $\Sigma \in E$ maximizes the plastic work. In other words, every plastic deformation process stores as much energy as possible:

$$W(\dot{P}) = \max_{T \in E} \{T : \dot{P}\}. \tag{2.9}$$

For the current generalized stress Σ the following therefore holds:

$$\Sigma : \dot{P} \geq T : \dot{P} \qquad \forall T \in E. \tag{2.10}$$

Using the method of Lagrange multipliers, it is easily concluded that, for the quadratic free energy (2.5) and the admissible domain (2.7), the Kuhn–Tucker optimality conditions for the maximum principle (2.10) yield the local constitutive equations

$$\begin{aligned}
&\dot{\sigma} = C : \dot{\varepsilon} - C : \frac{\partial}{\partial \sigma}\phi(\sigma,\chi)\gamma,\\
&\dot{\chi} = -H\frac{\partial}{\partial \chi}\phi(\sigma,\chi)\gamma,\\
&0 \geq \phi(\sigma,\chi), \quad \gamma \geq 0,\\
&0 = \phi(\sigma,\chi)\gamma.
\end{aligned} \tag{2.11}$$

With the Lagrange multiplier γ as an additional unknown, the system (2.11) is a differential-algebraic equation with inequality constraint and complementarity condition. The differential equation for σ is the analogue of the flow rule for plastic strain ε^p, and the equation for χ is the hardening rule.

As shown in [4, 12], the index of the DAE (2.11) is 2 if the constraint is active, i.e., if $0 = \phi(\sigma, \chi)$. Then, the Lagrange multiplier γ is positive and the plastic deformation proceeds. Otherwise, if the constraint is not active, the Lagrange multiplier vanishes and we have $\dot{\sigma} = C : \dot{\varepsilon}$ and $\dot{\chi} = 0$, which means that the material is in an elastic phase.

We want to give two remarks here. First, though the DAE (2.11) inspires our choice of time discretization, we prefer to work in a variational setting in the following. There, the Lagrange multipliers actually do not show up anymore. And second, the DAE (2.11) is closely related to a singularly perturbed ODE that is obtained from a regularization technique, the so-called Perzyna formulation. Another regularization due to Duvaut and Lions [6], which we call here viscoplastic regularization, will play a key role in the existence proof in section 4.

2.3. Viscoplastic regularization. In contrast to the elastoplastic case, viscoplastic materials feature a rate-dependent behavior. Their constitutive equations are derived by adding a regularization J_η to the principle (2.9),

$$W_\eta(\dot{P}) = \max_{\Sigma}(\{\Sigma : \dot{P}\} - J_\eta(\Sigma)). \tag{2.12}$$

A common choice for the convex function J_η is the Yosida regularization

$$J_\eta(\Sigma) = \frac{1}{2\eta}\|\Sigma - \Pi\Sigma\|^2, \tag{2.13}$$

where Π denotes the orthogonal projection onto the elastic domain E and η is the regularization parameter [6, 9, 11, 16]. Obviously, for η tending to zero, the penalty term $J_\eta(\Sigma)$ tends to infinity if $\Sigma \notin E$.

For further use below, we mention an important property of the Yosida regularization. Since J_η is convex, its Gâteaux derivative J'_η is monotone and given under the Riesz isomorphism by

$$J'_\eta(\Sigma) = \frac{1}{\eta}(\Sigma - \Pi\Sigma).$$

The monotonicity is expressed as

$$(J'_\eta(\Sigma) - J'_\eta(T), \Sigma - T) \geq 0 \qquad \text{for all generalized stresses } \Sigma, T; \tag{2.14}$$

see [9, Chapter 8.1] for details.

After this outline of the mathematical model, we pass now to the weak form and specify the appropriate function spaces.

2.4. The dual problem of elastoplasticity. We require stress, strain, internal variables, and conjugated forces to be square-integrable on the domain Ω at any time. Consequently, we introduce the space

$$\mathcal{S} := \{\tau = (\tau_{ij})_{3\times 3} \,:\, \tau_{ij} = \tau_{ji},\, \tau_{ij} \in L^2(\Omega)\}$$

for the symmetric tensors, and the space

$$\mathcal{M} := \{\mu = (\mu_j) \,:\, \mu_j \in L^2(\Omega),\, j = 1, \dots, r\}$$

for the conjugated forces. The product space

$$\mathcal{T} := \mathcal{S} \times \mathcal{M},$$

which is like the other spaces endowed with the natural L^2 inner product $(\cdot,\cdot)$, thus contains the square-integrable generalized stresses. Furthermore, we define the convex subset

$$\mathcal{P} := \{T = (\tau, \mu) \in \mathcal{T} \,:\, (\tau, \mu) \in E \quad \text{a.e. in}\, \Omega\}$$

for admissible states T.

In order to derive the variational formulation, we multiply the balance of momentum (2.2) by test functions belonging to

$$V = \{v \in H^1(\Omega)^3 : v = 0 \quad \text{on } \partial\Omega\}.$$

Using the Gauss divergence theorem, we arrive at

$$-\int_\Omega \varepsilon(v) : \sigma \, dx = -\int_\Omega f(t) \cdot v \, dx \qquad \forall v \in V.$$

From the principle of maximum plastic work (2.10), we obtain the variational inequality

$$\begin{aligned}0 &\leq \int_\Omega (\sigma - \tau) : \dot{\varepsilon}^p + (\chi - \mu) \cdot \dot{\xi}\, dx \\ &= -\int_\Omega (\sigma - \tau) : C^{-1}\dot{\sigma} + (\chi - \mu) \cdot H^{-1}\dot{\chi}\, dx + \int_\Omega (\sigma - \tau) : \dot{\varepsilon}\, dx\end{aligned}$$

for all $T = (\tau, \mu) \in \mathcal{P}$. To get a more compact notation, the integrals in the variational inequality are abbreviated by the bilinear forms

$$\begin{aligned} &A : \mathcal{T} \times \mathcal{T} \to \boldsymbol{R}, \qquad A(\Sigma, T) := \int_\Omega \sigma : C^{-1}\tau\, dx + \int_\Omega \chi \cdot H^{-1}\mu\, dx, \\ &b : V \times \mathcal{S} \to \boldsymbol{R}, \qquad b(v, \tau) := -\int_\Omega \varepsilon(v) : \tau\, dx \end{aligned}$$

for $\Sigma = (\sigma, \chi)$ and $T = (\tau, \mu)$. Finally, the integral on the right-hand side of the balance equation is identified with a time-dependent linear operator on the dual space, that is,

$$l(t) : V \to \boldsymbol{R}, \qquad \langle l(t), v \rangle = -\int_\Omega f(t) \cdot v\, dx.$$

Now we can state the dual problem as follows.

DUAL PROBLEM. *Given $l \in H^1(0, t_f; V')$ with $l(0) = 0$, find $(u, \Sigma) = (u, \sigma, \chi) : [0, t_f] \to V \times \mathcal{P}$ with $(u(0), \Sigma(0)) = (0, 0)$ such that for almost all $t \in (0, t_f)$ we have*

$$b(v, \sigma(t)) = \langle l(t), v \rangle \qquad \forall v \in V, \tag{2.15}$$

$$A(\dot{\Sigma}(t), T - \Sigma(t)) + b(\dot{u}(t), \tau - \sigma(t)) \geq 0 \qquad \forall T = (\tau, \mu) \in \mathcal{P}. \tag{2.16}$$

The name "dual," introduced in [9], is related to the use of the conjugate quantities, i.e., the generalized stress, instead of the primal unknowns plastic strain ε^p and internal variables ξ. Note that the dual problem can be viewed as an abstract constrained initial value problem. The generalized stress Σ is the main quantity of interest, whereas the displacement u is not unique since only the velocity $\dot{u}$ appears in the formulation.

2.5. Basic properties. The following assumptions on the shape of the admissible domain $\mathcal{P}$ are a fundamental prerequisite.

SAFE LOAD CONDITION SLC. *There is a constant $c > 0$ such that for any $T_1 = (\tau_1, \mu_1) \in \mathcal{P}$ and any stress tensor $\tau_2 \in \mathcal{S}$ there exists a conjugated force $\mu_2 \in \mathcal{M}$ such that*

$$\|\mu_2\| \leq c\|\tau_2\| \quad \textit{and} \quad (\tau_1, \mu_1) + (\tau_2, \mu_2) \in \mathcal{P}.$$

The safe load condition is fulfilled for various yield conditions, particularly for linear isotropic hardening (2.8). In some sense, it is necessary to complete the dual problem. The same holds true for the next assumption, which refers to the set E from (2.7).

ASSUMPTION A1. *For any $T \in E$ and any $\kappa \in [0, 1)$, we have $\kappa T \in E$ and*

$$\inf_{x \in \Omega} \operatorname{dist}(\kappa T(x), \partial E) > 0. \tag{2.17}$$

Under these assumptions, whose specific forms were first introduced in [9] and are easier to check than other variants, there exists a solution of the dual problem as follows.

THEOREM 2.1. *If the safe load condition SLC and Assumption A*1 *are fulfilled, then the dual problem has a solution* (u, Σ) *and* Σ *is unique.*

We briefly outline the idea of the proof given in [9] since our approach in section 4 will follow the same lines. The key idea is the viscoplastic regularization (2.12), which is an unconstrained optimization problem. Therefore taking the derivative of (2.12) and setting it to zero yields the variational equality

$$A(\dot{\Sigma}(t), T) + \big(J'_\eta(\Sigma(t)), T\big) + b(\dot{u}(t), \tau) = 0 \qquad \forall T = (\tau, \mu) \in \mathcal{T}, \tag{2.18}$$

$$b(v, \sigma(t)) = \langle l(t), v\rangle \qquad \forall v \in V. \tag{2.19}$$

The balance equation (2.15) has been added to show the connection to a saddlepoint problem. Next, for the bilinear form b there exists a constant $\beta_b > 0$ such that

$$\sup_{0\neq\tau\in\mathcal{S}} \frac{|b(v,\tau)|}{\|\tau\|_{\mathcal{S}}} \geq \beta_b \|v\|_V \quad \forall v \in V. \tag{2.20}$$

This is the Babuška–Brezzi condition, but note that b satisfies (2.20) only on $\mathcal{S}$ and not on $\mathcal{T}$. For this reason, we need in addition the SLC. This assumption guarantees a solution σ of (2.19) in the orthogonal complement of the kernel of b such that there exists χ with $(\sigma, \chi) \in \mathcal{P}$. Due to the coercivity of A, which means there exists a constant β_A such that

$$A(T, T) \geq \beta_A \|T\|^2_{\mathcal{T}} \qquad \forall T \in \mathcal{T},$$

and due to the monotonicity of J'_η, that is,

$$(J'_\eta(\Sigma) - J'_\eta(T), \Sigma - T) \geq 0,$$

there exists a solution $(\Sigma, \dot{u})$ of the regularized problem (2.18)–(2.19).

In the second step of the proof, one lets a suitable subsequence of η tend to zero and shows for this sequence that the limit with respect to Σ and $\dot{u}$ exists and solves the dual problem. Uniqueness follows from the coercivity of A.

At the end of this section, we mention another important fact. The solution of the dual problem possesses a contractivity property with respect to the norm

$$\|\cdot\|_A := \sqrt{A(\cdot,\cdot)},$$

the so-called complementary energy norm. The following result holds.

THEOREM 2.2. *Assume that* (u, Σ) *and* $(\widehat{u}, \widehat{\Sigma})$ *are solutions of the dual problem for different initial values. Then*

$$\|\Sigma(t) - \widehat{\Sigma}(t)\|_A \leq \|\Sigma(0) - \widehat{\Sigma}(0)\|_A$$

for all times $t > 0$.

Proofs of this property can be found in [9, 16].

3. Main results. In this section, the implicit Runge–Kutta discretization is introduced and the main results of the paper are stated. It turns out that the coefficient matrix of the Runge–Kutta method plays a crucial role for existence, contractivity, and convergence of the numerical solution. The existence proof makes use of the viscoplastic regularization and starts with the corresponding time-discrete saddlepoint problem. We postpone the details of this proof, however, to section 4.

3.1. Runge–Kutta method. We approximate the solution of the dual problem in time by implicit Runge–Kutta methods. For this purpose, the time derivatives in (2.15) and (2.16) are replaced by the stage derivatives of the Runge–Kutta method. Let s be the number of stages, $(a_{ij}) = \mathcal{A}$ the $s \times s$ coefficient matrix, (b_i) the vector of weights, (c_i) the vector of quadrature nodes, and h the stepsize. Moreover, the numerical approximations at time t_n are denoted by Σ_n and u_n.

One step of the Runge–Kutta method from t_n to $t_{n+1} = t_n + h$ is defined by

$$\Sigma_{n+1} = \Sigma_n + h \sum_{i=1}^{s} b_i \Psi_{n,i}, \qquad u_{n+1} = u_n + h \sum_{i=1}^{s} b_i w_{n,i},$$

where $\Psi_{n,i}$ and $w_{n,i}$ are the stage derivatives at time $t_{n,i} = t_n + c_i h$. The stages $\Sigma_{n,i} = (\sigma_{n,i}, \chi_{n,i}) \in \mathcal{P}$ and the stage derivatives $w_{n,i} \in V$ satisfy the nonlinear system

$$b(v, \sigma_{n,i}) = \langle l_{n,i}, v \rangle \qquad \forall v \in V, \tag{3.1}$$

$$A(\Psi_{n,i}, T - \Sigma_{n,i}) + b(w_{n,i}, \tau - \sigma_{n,i}) \geq 0 \qquad \forall T = (\tau, \nu) \in \mathcal{P} \tag{3.2}$$

for $i = 1, \ldots, s$, with right-hand sides $l_{n,i} = l(t_n + c_i h)$. Stages $\Sigma_{n,i}$ and stage derivatives $\Psi_{n,i}$ are coupled via

$$\Sigma_{n,i} = \Sigma_n + h \sum_{j=1}^{s} a_{ij} \Psi_{n,j}. \tag{3.3}$$

This method definition includes various schemes like implicit Euler, midpoint and generalized midpoint rule (θ-method), DIRK schemes, and further implicit discretizations (the methods of Gauss, Lobatto, or Radau type). Our focus will be on the latter class, but the results are not restricted to it.

3.2. Existence of the numerical solution. Does the nonlinear system (3.1)–(3.3) possesses a solution? The answer depends very much on the coefficient matrix $\mathcal{A}$. If one looks back to the proof of Theorem 2.1, one notices that the coercivity of the bilinear form A was a basic prerequisite. At first look, the coupling (3.3) destroys this property: Consider (3.2) and replace the stage derivatives $\Psi_{n,i}$ by the stages $\Sigma_{n,i}$ using the relation (3.3). Thereby the $\Sigma_{n,i}$ are multiplied in a certain sense by the inverse of the Runge–Kutta matrix before the bilinear form A is evaluated, and we obtain a new bilinear form that depends on the method coefficients. The crucial point is whether this new bilinear form is still coercive.

In the theory of stiff ODEs, similar problems occur [10]. In general, the Runge–Kutta matrix $\mathcal{A}$ is not positive definite, but in many cases it possesses a coercivity property.

DEFINITION 3.1. *We consider the inner product $(u, v)_D := u \cdot Dv$ on $\boldsymbol{R}^s$, where $D = \operatorname{diag}(d_1, \ldots, d_s)$ with positive entries $d_i > 0$. The Runge–Kutta matrix $\mathcal{A}$ is coercive iff there exists a positive diagonal matrix D and a constant α_D such that*

$$(u, \mathcal{A}^{-1} v)_D \geq \alpha_D (u, v)_D$$

for all $u, v \in \boldsymbol{R}^s$. In this case, we set

$$\alpha_0(\mathcal{A}^{-1}) := \sup_{D>0} \alpha_D.$$

TABLE 3.1
Coercivity constants α_D and matrices D for classes of Runge–Kutta methods.

Method	D	Stages	α_D
Gauss	$B(C^{-1} - I)$	s	$\min \frac{1}{2c_i(1-c_i)}$
Radau IA	$B(I - C)$	1 $s > 1$	1 $\frac{1}{2(1-c_2)}$
Radau IIA	BC^{-1}	1 $s > 1$	1 $\frac{1}{2c_{s-1}}$

$(B = \operatorname{diag}(b_1, \dots, b_s), \quad C = \operatorname{diag}(c_1, \dots, c_s))$

From [10, Theorem 14.5], we have the following classification.

THEOREM 3.2. *If the Runge–Kutta method belongs to one of the classes Gauss, Radau* IA*, or Radau* IIA*, then the Runge–Kutta matrix is coercive.*

For these method classes, the matrices D and the constants α_D are listed in Table 3.1.

We remark that DIRK methods with positive coefficients a_{ii} on the diagonal are coercive too [10, Theorem 14.6]. The constant α_D is here given by

$$\alpha_0(A^{-1}) = \min \frac{1}{a_{ii}}.$$

With the notion of coercivity for a Runge–Kutta method at hand, we are able to state the main result of this paper as follows.

THEOREM 3.3. *Let the Runge–Kutta matrix be coercive, and let the SLC and Assumption* A1 *be satisfied. Then the discretization scheme* (3.1)–(3.3) *possesses a solution* $(w_{n,i}, \Sigma_{n,i})$ *for* $i = 1, \dots, s$*, and the stages* Σ_i *of the generalized stresses are unique.*

The proof follows the same lines as the proof of Theorem 2.1. Since we have to consider s stages, however, many technical details need to be addressed. Our starting point is the Runge–Kutta discretization of the viscoplastic regularization (2.18) in combination with the balance equation (2.19). In order to simplify the notation, for the rest of the paper we write Σ_i instead of $\Sigma_{n,i}$ and omit the index n for all other stage variables. Furthermore, to distinguish between the original and the regularized problem, we now use the superscript η for the variables of the regularized problem and discretize it in the following way.

Find $(w_i^\eta, \Sigma_i^\eta) = (w_i^\eta, \sigma_i^\eta, \chi_i^\eta) \in V \times \mathcal{T}$ such that

$$b(v, \sigma_i^\eta) = \langle l_i, v \rangle \qquad \forall v \in V, \tag{3.4}$$

$$A(\Psi_i^\eta, T) + \left(J_\eta'(\Sigma_i^\eta), T\right)_{\mathcal{T}} + b(w_i^\eta, \tau) = 0 \qquad \forall T = (\tau, \nu) \in \mathcal{T} \tag{3.5}$$

for $i = 1, \dots, s$. Stages Σ_i^η and stage derivatives Ψ_i^η are related to each other via

$$\Sigma_i^\eta = \Sigma_n + h \sum_{j=1}^{s} a_{ij} \Psi_j^\eta. \tag{3.6}$$

Note that the solution (u_n, Σ_n) of the last step is taken from the original but not from the regularized problem.

In a certain product space, the discretized regularization represents a saddlepoint problem, and therefore the reasoning of the proof relies very much on the corresponding theory. We postpone the details of the proof to section 4.

3.3. Contractivity. In ODE theory, the differential equation

$$\dot{y} = f(t,y)$$

is called contractive if it satisfies a one-sided Lipschitz condition

$$(f(t,y) - f(t,z), y - z) \leq 0.$$

The elastoplastic contractivity of Theorem 2.2 can be seen as an analogue, and the question is whether time discretizations preserve this property. For generalized midpoint schemes, contractivity of the elastoplastic flow is shown in [15, 16]. Wieners [18] extended the result in the context of DIRK methods and proved that algebraic stability is sufficient to preserve the contractivity.

To be more precise, the notion of algebraic stability is defined as follows (see [10]).

Definition 3.4. *A Runge–Kutta method is called algebraically stable iff*

1. $b_i \geq 0$ *for all* $i = 1, \ldots, s$,
2. $M = (m_{ij}) = (b_i a_{ij} + b_j a_{ji} - b_i b_j)_{i,j=1}^{s}$ *is nonnegative definite.*

The contractivity result of [18] carries immediately over to the more general class considered here; cf. [4, 15, 16].

Theorem 3.5. *Algebraically stable Runge–Kutta methods preserve the contractivity of the elastoplastic flow; that is,*

$$\|\Sigma_n - \widehat{\Sigma}_n\|_A \leq \|\Sigma_0 - \widehat{\Sigma}_0\|_A \tag{3.7}$$

for different initial values Σ_0 *and* $\widehat{\Sigma}_0$.

The notions of algebraic stability and coercivity of a Runge–Kutta method are not equivalent. Remarkably, however, there are several classes, in particular Gauss, Radau IA, and Radau IIA, that feature both properties.

3.4. Convergence. We start by introducing consistency in terms of the stage order q, which is the positive integer such that conditions

$$B(q) \quad : \quad \sum_{i=1}^{s} b_i c_i^{r-1} = \frac{1}{r}, \quad r = 1, \ldots, q,$$

and

$$C(q) \quad : \quad \sum_{j=1}^{s} a_{ij} c_j^{r-1} = \frac{c_i^r}{r}, \quad r = 1, \ldots, q, \; i = 1, \ldots, s,$$

hold true. Thus the stage order defines the accuracy of the quadrature rule that is the basis of the Runge–Kutta method. We assume now that the generalized stresses are elements of the Sobolev space $H^{q+1}(0, t_f; \Omega)$. Then a weak Taylor expansion and conditions $B(q)$ and $C(q)$ imply

$$\Bigg\| \underbrace{\mathbf{\Sigma}(t_{n+1}) - \mathbf{\Sigma}(t_n) - h \sum_{i=1}^{s} b_i \mathbf{\Sigma}(t_n + c_i h)}_{=:Q_n} \Bigg\|_{\mathcal{T}} \leq c h^q \|\mathbf{\Sigma}^{(q+1)}\|_{L^1(t_n, t_{n+1}; \Omega)} =: Q_n^{\mathrm{Err}}$$

TABLE 3.2
The stage order; see Table IV.5.13 *in* [10].

Method	Consistency conditions	Stage order
Gauss	$B(2s), C(s)$	s
Radau IIA	$B(2s-1), C(s)$	s

and

$$\Bigg\| \underbrace{\boldsymbol{\Sigma}(t_n + c_i h) - \boldsymbol{\Sigma}(t_n) - h \sum_{j=1}^{s} a_{ij} \dot{\boldsymbol{\Sigma}}(t_n + c_j h)}_{=:S_n^i} \Bigg\|_{\mathcal{T}} \leq ch^q \|\boldsymbol{\Sigma}^{(q+1)}\|_{L^1(t_n,t_{n+1};\Omega)} =: S_n^{\mathrm{Err}}.$$

Note, that the error estimate on the right-hand side, S_n^{Err}, is independent of the stage number i. Stage orders for Gauss and Radau IIA methods are listed in Table 3.2.

THEOREM 3.6. *Assume that the Runge–Kutta method is coercive and algebraically stable, $B(q)$ and $C(q)$ are fulfilled, and $\Sigma \in H^{(q+1)}(0, t_f; \Omega)$. Then the following holds for the global error on $[0, t_f]$ with $N = [t_f/h]$:*

$$\max_{r=1,\ldots,N} \|\Sigma(t_r) - \Sigma_r\|_{\mathcal{T}} \leq ch^q \|\boldsymbol{\Sigma}^{(q+1)}\|_{L^1(0,t_f;\Omega)}.$$

Hence existence, stability, and consistency of the Runge–Kutta method imply convergence for the dual problem of elastoplasticity. The proof of Theorem 3.6 is postponed to section 5. We remark that the smoothness assumption $\Sigma \in H^{(q+1)}(0, t_f; \Omega)$ is not backed by the regularity provided by the theory; see also the discussion in Chapter 13.1 of [9] on the midpoint rule.

4. Proof of Theorem 3.3. The proof generalizes Theorem 8.12 of [9], which is formulated for the implicit Euler method. We first give an outline and provide some necessary framework.

The existence proof proceeds in three steps. In step one, it is shown that the discrete regularized system (3.4)–(3.6) possesses a solution $(w_i^\eta, \Sigma_i^\eta)$, $i = 1, \ldots, s$, depending on the parameter η. Step two constructs a uniform bound for $(w_i^\eta, \Sigma_i^\eta)$ and selects a weakly convergent subsequence. Finally, step three establishes the limit $\lim_{\eta\to 0}(w_i^\eta, \Sigma_i^\eta) = (w_i, \Sigma_i)$ as a solution of the dual problem. Uniqueness of the stages Σ_i follows from the coercivity of both the bilinear form A and the Runge–Kutta matrix $\mathcal{A}$.

Since the saddlepoint structure plays a crucial role in our first step, we next cite an important result concerning the bilinear form b. Related to b, one defines the operators $B : \mathcal{S} \to V'$ and $B' : V \to \mathcal{S}'$ by

$$\langle B\sigma, v\rangle := \langle B'v, \sigma\rangle := b(v, \sigma) \qquad \text{for } v \in V,\ \sigma \in \mathcal{S}.$$

If the Babuška-Brezzi condition holds (cf. (2.20)), we have the following (see [2, 9]).

PROPOSITION 4.1. *The operator B is an isomorphism from* $(\mathrm{Ker}\ B)^\perp$ *onto V', where*

$$\mathrm{Ker} B = \{\sigma \in \mathcal{S} : b(v, \sigma) = 0 \quad \forall v \in V\},$$

and the operator B' is an isomorphism from (Ker $B)^\circ$ *onto* V, *where*

$$(\mathrm{Ker}B)^\circ = \{f \in \mathcal{S}' : \langle f, \tau\rangle = 0 \quad \forall \tau \in \mathrm{Ker}B\}.$$

The framework is completed by a notational agreement. We abbreviate the array of s stages Σ_i by

$$\vec{\Sigma} := (\Sigma_1, \dots, \Sigma_s)^T.$$

In the same fashion, we extend the initial value Σ_n of the old time step to

$$\vec{\Sigma}_n := (\Sigma_n, \dots, \Sigma_n)^T.$$

Although $\vec{\Sigma}$ and $\vec{\Sigma}_n$ are not vectors of real numbers but of tensor-valued functions, we connect them like standard vectors with the Runge–Kutta matrix $\mathcal{A}$ and skip the introduction of an appropriate Kronecker product. As an example, the relation (3.6) between stages and stage derivatives reads in this notation

$$\vec{\Sigma}^\eta = \vec{\Sigma}_n + h\mathcal{A}\vec{\Psi}^\eta \quad \Leftrightarrow \quad \vec{\Psi}^\eta = \frac{1}{h}\mathcal{A}^{-1}(\vec{\Sigma}^\eta - \vec{\Sigma}_n).$$

4.1. Step one (regularized problem). The bilinear form b satisfies the Babuška–Brezzi condition (2.20). Due to Proposition 4.1 there is for each stage i a unique element $\sigma_i^0 \in (\mathrm{Ker}B)^\perp$ such that

$$b(v, \sigma_i^0) = \langle l_i, v\rangle \qquad \forall v \in V.$$

Taking the constant c as the operator norm of B^{-1}, we have

$$\|\sigma_i^0\|_S \le c\|l_i\|$$

for all stages. Applying the SLC to $2\sigma_i^0$, we find an element in $\mathcal{M}$, denoted by $2\chi_i^0$, such that $(2\sigma_i^0, 2\chi_i^0) \in \mathcal{P}$ and

$$\|\chi_i^0\| \le c\|\sigma_i^0\| \le c\|l_i\|.$$

We set $\Sigma_i^0 = (\sigma_i^0, \chi_i^0)$. Assumption A1 on the structure of the elastic domain ensures that $\Sigma_i^0 \in E$ almost everywhere and (see Figure 4.1)

$$r_i \equiv \inf_{x\in\Omega} \mathrm{dist}(\Sigma_i^0(x), \partial E) > 0. \tag{4.1}$$

Next, we look for a solution of (3.5) in the form $\Sigma_i^\eta = \Sigma_i^0 + \Sigma_i^{\delta\eta}$, where $\Sigma_i^{\delta\eta} = (\sigma_i^{\delta\eta}, \chi_i^{\delta\eta})$ is an element of $\mathrm{Ker}\, B \times \mathcal{M}$. We observe that the balance equation (3.4) still holds if Σ_i^0 is updated by an element of $\mathrm{Ker}\, B \times \mathcal{M}$.

At this point of the proof, the coupling (3.6) becomes important and we pass to the product space $\mathcal{T}^s$. We define a bilinear form on $\mathcal{T}^s$ by

$$\vec{A} : \begin{cases} \mathcal{T}^s \times \mathcal{T}^s & \to \quad \boldsymbol{R}, \\ \left(\vec{\Theta}, \vec{\Upsilon}\right) & \mapsto \quad d_1 A(\Phi_1, \Upsilon_1) + \dots + d_s A(\Phi_s, \Upsilon_s) \quad \text{with } \vec{\Phi} := \mathcal{A}^{-1}\vec{\Theta}, \end{cases}$$

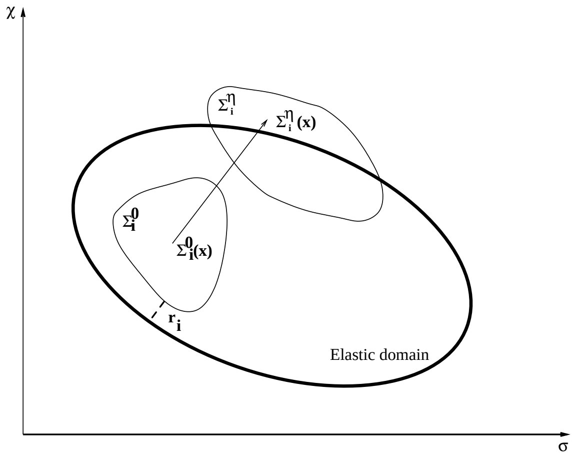


FIG. 4.1. *Situation in step* 1.

where D is the diagonal matrix related to the Runge–Kutta matrix $\mathcal{A}$ according to Definition 3.1. Since the bilinear form A is coercive on $\mathcal{T} \times \mathcal{T}$, its extension $\vec{A}$ defines a coercive bilinear form, that is,

$$\vec{A}(\vec{\Upsilon}, \vec{\Upsilon}) \geq \alpha_0(\mathcal{A}^{-1}) \min\{d_i\} \beta_A (\vec{\Upsilon}, \vec{\Upsilon})_{\mathcal{T}^s}$$

with the scalar product

$$(\vec{\Upsilon}, \vec{\Upsilon})_{\mathcal{T}^s} := (\Upsilon_1, \Upsilon_1)_{\mathcal{T}} + \cdots + (\Upsilon_s, \Upsilon_s)_{\mathcal{T}} \, .$$

In general, $\vec{A}$ is not symmetric. We also need to extend the scalar product of the regularization J'_η to the product space $\mathcal{T}^s$. For this purpose, a weighted sum with the elements of D is introduced, similar to the definition of $\vec{A}$:

$$\vec{J}'_\eta \begin{cases} \mathcal{T}^s \times \mathcal{T}^s & \rightarrow \quad \boldsymbol{R}, \\ \\ \left(\vec{\Theta}, \vec{\Upsilon}\right) & \mapsto \quad d_1 \left(J'_\eta(\Theta_1), \Upsilon_1\right) + \cdots + d_s \left(J'_\eta(\Theta_s), \Upsilon_s\right). \end{cases}$$

Now we can resume the proof and look for an element $\vec{\Sigma}^{\delta\eta} \in (\operatorname{Ker} B \times \mathcal{M})^s$ satisfying

$$\frac{1}{h}\vec{A}(\vec{\Sigma}^{\delta\eta}, \vec{T}) + \vec{J}'_\eta(\vec{\Sigma}^0 + \vec{\Sigma}^{\delta\eta}, \vec{T}) = \frac{1}{h}\vec{A}(\vec{\Sigma}_n - \vec{\Sigma}^0, \vec{T}) \tag{4.2}$$

for all $\vec{T} \in (\operatorname{Ker} B \times \mathcal{M})^s$. This is an operator equation with the nonlinear operator $L : \mathcal{T}^s \rightarrow (\mathcal{T}^s)'$ defined by

$$\langle L\cdot, \vec{T}\rangle := \frac{1}{h}\vec{A}(\cdot, \vec{T}) + \vec{J}'_\eta(\vec{\Sigma}^0 + \cdot, \vec{T}) \qquad \forall \vec{T} \in \mathcal{T}^s.$$

Because $\vec{A}$ is coercive and $\vec{J}'_\eta$ is monotone, the operator L is strongly monotone. [9, Theorem 5.10, p. 107] establishes the existence of a unique solution $\vec{\Sigma}^{\delta\eta}$ of (4.2).

Setting $T_i = T$ and $T_j = 0$ for $j \neq i$, (4.2) yields

$$A(\Psi_i^\eta, T) + \left(J'_\eta(\Sigma_i^\eta), T\right) = 0 \qquad \forall T \in \operatorname{Ker} B \times \mathcal{M}. \tag{4.3}$$

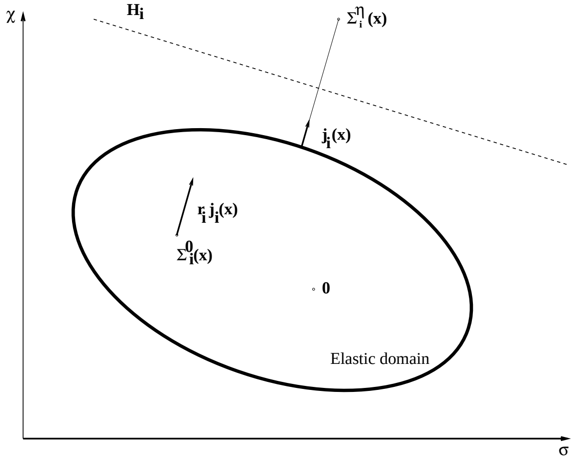


Fig. 4.2. *Situation in step* 2.

The left-hand side of (4.3) defines a continuous linear form on $\mathcal{T}$ and is also part of the left-hand side of (3.5). Though Proposition 4.1 is valid only for stresses $\sigma \in \mathcal{S}$, it is straightforward to extend it to generalized stresses $T \in \mathcal{T}$. Therefore, there exists an element $w_i^\eta \in V$ such that (3.5) is satisfied and consequently the regularized problem (3.4)–(3.6) has a solution $(w_i^\eta, \Sigma_i^\eta)$ for $i = 1, \dots, s$.

4.2. Step two (uniform boundedness). (a) First, we derive a uniform bound for $\vec{\Sigma}^\eta$. Setting $\vec{T} = \vec{\Sigma}^{\delta\eta}$ in (4.2) leads to

$$\frac{1}{h}\vec{A}(\vec{\Sigma}^{\delta\eta}, \vec{\Sigma}^{\delta\eta}) + \vec{J}'_\eta(\vec{\Sigma}^0 + \vec{\Sigma}^{\delta\eta}, \vec{\Sigma}^{\delta\eta}) = \frac{1}{h}\vec{A}(\vec{\Sigma}_n - \vec{\Sigma}^0, \vec{\Sigma}^{\delta\eta}).$$

The convexity of the Yosida regularization J_η implies for $\Sigma_i^0 \in \mathcal{P}$

$$0 = J_\eta(\Sigma_i^0) \geq J_\eta(\Sigma_i^0 + \Sigma_i^{\delta\eta}) + (J'_\eta(\Sigma_i^0 + \Sigma_i^{\delta\eta}), -\Sigma_i^{\delta\eta}).$$

Since $J_\eta(\Sigma_i^0 + \Sigma_i^{\delta\eta}) \geq 0$, we obtain

$$\vec{A}(\vec{\Sigma}^{\delta\eta}, \vec{\Sigma}^{\delta\eta}) \leq \vec{A}(\vec{\Sigma}_n - \vec{\Sigma}^0, \vec{\Sigma}^{\delta\eta}).$$

Next, the coercivity and the continuity of $\vec{A}$ yield the existence of a constant c such that

$$\|\vec{\Sigma}^\eta\| \leq c. \tag{4.4}$$

(b) Second, we show the uniform boundedness of w_i^η. For all $x \in \Omega$ with $\Sigma_i^\eta(x) \in E$, $J'_\eta(\Sigma_i^\eta)(x) = 0$ holds. Therefore we assume $\Sigma_i^\eta(x) \notin E$ and define

$$j_i(x) = \frac{\Sigma_i^\eta(x) - \Pi\Sigma_i^\eta(x)}{|\Sigma_i^\eta(x) - \Pi\Sigma_i^\eta(x)|}.$$

Since E is a closed convex set in the finite-dimensional vector space of generalized forces, $j_i(x)$ is normal to a hyperplane $H_i = \{T \mid j_i(x) : T = \rho_i\}$; see Figure 4.2.

Now we exploit (4.1). Since $\Sigma_i^0(x) + r_i j(x) \in E$ and $0 \in E$, we have

$$j_i(x) : \Sigma_i^\eta(x) \geq \rho_i \quad \text{and} \quad j_i(x) : \left(\Sigma_i^0(x) + r_i j(x)\right) \leq \rho_i.$$

Geometrically this means that $\Sigma_i^\eta(x)$ and $(\Sigma_i^0(x) + r_i j_i(x))$ are separated by the hyperplane H_i; see Figure 4.2. Hence

$$j_i(x) : \left(\Sigma_i^\eta - \Sigma_i^0(x)\right) \geq r_i,$$

that is,

$$|\Sigma_i^\eta(x) - \Pi\Sigma_i^\eta(x)| \leq \frac{1}{r_i}(\Sigma_i^\eta(x) - \Pi\Sigma_i^\eta(x)) : (\Sigma_i^\eta - \Sigma_i^0(x)).$$

Since the latter inequality holds also for all $x \in \Omega$ with $\Sigma_i^\eta(x) \in E$, we obtain

$$\|J_\eta'(\Sigma_i^\eta)\| \leq \frac{1}{r_i}\left(J_\eta'(\Sigma_i^\eta), \Sigma_i^\eta - \Sigma_i^0\right). \tag{4.5}$$

Now we return to (4.3). There we set $T_i = \Sigma_i^\eta - \Sigma_i^0 \in \operatorname{Ker} B \times \mathcal{M}$. Inserting (4.3) into (4.5) yields

$$\|J_\eta'(\Sigma_i^\eta)\|_{\mathcal{T}} \leq -\frac{1}{r_i}A(\Psi_i^\eta, \Sigma_i^\eta - \Sigma_i^0).$$

Since the uniform boundedness of $\vec{\Sigma}^\eta$ implies

$$|A(\Psi_i^\eta, \Sigma_i^\eta - \Sigma_i^0)| \leq \alpha\|\Psi_i^\eta\|_{\mathcal{T}}\|\Sigma_i^\eta - \Sigma_i^0\| \leq \tilde{c},$$

we conclude

$$\|J_\eta'(\Sigma_i^\eta)\|_{\mathcal{T}} \leq \tilde{c}.$$

From the Babuška–Brezzi condition in combination with (3.5) we obtain

$$\begin{aligned}
\beta_b\|w_i^\eta\|_V &\leq \sup_{\tau\in\mathcal{S}}\frac{|b(w_i^\eta,\tau)|}{\|\tau\|_{\mathcal{S}}} \\
&\leq \sup_{\tau\in\mathcal{S}}\frac{|-A(\Psi_i^\eta,(\tau,0)) - J_\eta'\left(\Sigma_i^\eta,(\tau,0)\right)|}{\|\tau\|_{\mathcal{S}}} \\
&\leq \tilde{c}\left(\|\Psi_i^\eta\|_{\mathcal{T}} + \|J_\eta'(\Sigma_i^\eta)\|\right) \leq \tilde{c}.
\end{aligned}$$

Extending this to the product space leads to the uniform bound of $\vec{w}^\eta$.

In parts (a) and (b) of step 2, we established the uniform boundedness of $\vec{\Sigma}^\eta$ and $\vec{w}^\eta$. Thus we can extract a subsequence that converges weakly in the dual space of $\mathcal{T}^s \times V^s$, still denoted by $\vec{\Sigma}^\eta$ and $\vec{w}^\eta$, such that $(\vec{\Sigma}^\eta, \vec{w}^\eta) \rightharpoonup (\vec{\Sigma}, \vec{w})$ as $\eta \to 0$.

4.3. Step three (proving (3.1) and (3.2)). The last part of the existence proof again splits into several parts. In part (a), we show that the limits Σ_i are admissible. In part (b), we prove the equality (3.1), and finally, in part (c), we establish the variational inequality (3.2).

(a) First, we prove that the limits Σ_i are admissible, that is, $\Sigma_i \in \mathcal{P}$. Since

$$\Sigma_i \in \mathcal{P} \Leftrightarrow \|\Sigma_i - \Pi\Sigma_i\| = 0,$$

we investigate the convex functional J_η. The convexity of J_η yields

$$J_\eta(\Sigma_i^\eta) \leq (J_\eta'(\Sigma_i^\eta), \Sigma_i^{\delta\eta}).$$

Equation (4.3) implies $A(\tilde{\Sigma}_i^\eta, \Sigma_i^{\delta\eta}) + J_\eta(\Sigma_i^\eta) \leq 0$, which is equivalent to

$$J_\eta(\Sigma_i^\eta) \leq -A(\Psi_i^\eta, \Sigma_i^{\delta\eta}) = -\frac{1}{h} A(A_i^{-1}(\vec{\Sigma}_n - \vec{\tilde{\Sigma}}_i^\eta), \Sigma_i^\eta - \Sigma_0^i).$$

Since Σ_i^η is uniformly bounded, we have $J_\eta(\Sigma_i^\eta) \leq c$. The Yosida regularization is convex and l.s.c.; hence the function

$$f(\Sigma) := \eta J_\eta = \|\Sigma - \Pi\Sigma\|^2$$

is weakly l.s.c. (see [9, p. 74]). Therefore

$$f(\Sigma_i) \leq \liminf_{\eta \to 0} f(\Sigma_i^\eta) \leq \lim_{\eta \to 0} c\eta = 0.$$

This implies $\|\Sigma_i - \Pi\Sigma_i\| = 0$, that is, $\Sigma_i \in \mathcal{P}$.

(b) Now we show (3.1). Since $b(v, \sigma_i^{\delta\eta}) = 0$ for all $v \in V$, we have

$$b(v, \sigma_i) = b(v, \sigma_i^0) = \langle l_i, v \rangle.$$

(c) Finally we consider the variational inequality (3.2). J_η is a convex function. Hence for all $\bar{T} \in \mathcal{P}$, the convexity implies

$$\begin{aligned} & 0 = J_\eta(\bar{T}) \geq J_\eta(\Sigma_i^\eta) + \left(J'_\eta(\Sigma_i^\eta), \bar{T} - \Sigma_i^\eta\right) \\ \Rightarrow \quad & 0 \geq -J_\eta(\Sigma_i^\eta) \geq \left(J'_\eta(\Sigma_i^\eta), \bar{T} - \Sigma_i^\eta\right). \end{aligned}$$

Setting $\bar{T} = T - \Sigma_i$, we insert the last inequality into (3.5) and arrive at

$$A(\Psi_i^\eta, T - \Sigma_i^\eta) + b(w_i, \tau - \sigma_i^\eta) \geq 0.$$

Now we pass to the product space to obtain

(4.6)
$$\frac{1}{h}\vec{A}(\vec{\Sigma}^\eta, \vec{T} - \vec{\Sigma}^\eta) + d_1 b(w_1^\eta, \tau_1 - \sigma_1^\eta) + \cdots + d_s b(w_s^\eta, \tau_s - \sigma_s^\eta) \geq \frac{1}{h}\vec{A}(\vec{\Sigma}_n, \vec{T} - \vec{\Sigma}^\eta).$$

It is our goal to preserve this inequality while η tends to zero. We proceed as follows:

1. Weak convergence and the relation $\sigma_i^{\delta\eta} \in \operatorname{Ker} B$ lead to

$$b(w_i^\eta, \sigma_i^\eta) = b(w_i^\eta, \sigma^0) \to b(w_i, \sigma^0) = b(w_i^\eta, \sigma_i).$$

2. Weak convergence also implies

$$\begin{aligned} \vec{A}(\vec{\Sigma}^\eta, \vec{T}) &\to \vec{A}(\vec{\Sigma}, \vec{T}), \\ \vec{A}(\vec{\Sigma}_n, \vec{T} - \vec{\Sigma}^\eta) &\to \vec{A}(\vec{\Sigma}_n, \vec{T} - \vec{\Sigma}). \end{aligned}$$

3. The last term of (4.6) needs some more effort. Since $\vec{A}(\cdot,\cdot)$ is not symmetric, we formulate the quadratic identity

(4.7) $$\vec{A}(\vec{\Sigma}^\eta, \vec{\Sigma}) + \vec{A}(\vec{\Sigma}, \vec{\Sigma}^\eta) - \vec{A}(\vec{\Sigma}, \vec{\Sigma}) = \vec{A}(\vec{\Sigma}^\eta, \vec{\Sigma}^\eta) - \vec{A}(\vec{\Sigma}^\eta - \vec{\Sigma}, \vec{\Sigma}^\eta - \vec{\Sigma}).$$

   From the coercivity of $\vec{A}$, passing to the lim inf yields

$$\vec{A}(\vec{\Sigma}, \vec{\Sigma}) \leq \liminf_{\eta \to 0} \vec{A}(\vec{\Sigma}^\eta, \vec{\Sigma}^\eta).$$

Applying the lim inf to (4.6) and inserting the last three limits, we arrive at

$$A(\tilde{\Sigma}_i, T - \Sigma_i) + b(w_i, \tau - \sigma_i) \geq 0.$$

Setting $T_i = T$ and $T_j = 0$ for $j \neq i$ proves the inequality (3.2).

4.4. Uniqueness. Suppose there exist two different solutions (w_i, Σ_i) with $\Sigma_i = (\sigma_i, \chi_i)$ and $(\tilde{w}_i, \tilde{\Sigma}_i)$ with $\tilde{\Sigma}_i = (\tilde{\sigma}_i, \tilde{\chi}_i)$ of the system (3.1)–(3.3).

The linearity of the variational equation (3.1) implies

$$b(v, \sigma_i - \tilde{\sigma}_i) = 0 \qquad \forall v \in V,$$

that is, $\sigma - \tilde{\sigma} \in \operatorname{Ker} B$. Therefore inserting Σ_i into the variational inequality (3.1) and setting $T = \tilde{\Sigma}_i$ yields

$$A(\Psi_i, \tilde{\Sigma}_i - \Sigma_i) \geq 0,$$

where Ψ_i again denotes the stage derivative. The same argument with Σ_i and $\tilde{\Sigma}_i$ interchanged gives

$$A(\tilde{\Psi}_i, \Sigma_i - \tilde{\Sigma}_i) \geq 0.$$

Adding these two equations and passing to the product space $\mathcal{T}^s$ leads to

$$\vec{A}(\vec{\Sigma} - \vec{\tilde{\Sigma}}, \vec{\Sigma} - \vec{\tilde{\Sigma}}) \leq 0.$$

Hence the coercivity of $\vec{A}$ implies $\Sigma_i = \tilde{\Sigma}_i$.

5. Proof of Theorem 3.6.

We use the same notation as in section 4 and extend the proof of Theorem IV.12.4 of [10]. The definitions of the errors Q_n and S_n^i (see subsection 3.4) lead to

$$\begin{aligned}
&\|\Sigma_{n+1} - \Sigma(t_{n+1})\|_{\mathrm{A}}^2 \\
&= \left\| \Sigma_n + h \sum_{j=1}^{s} b_j \Psi_j - \left(\Sigma(t_n) + h \sum_{j=1}^{s} b_j \dot{\Sigma}(t_j) \right) - Q_n \right\|_A^2 \\
&= \|\Sigma_n - \Sigma(t_n)\|_A^2 + 2h \sum_{j=1}^{s} b_j (\Psi_j - \dot{\Sigma}(t_j), \Sigma_j - \Sigma(t_j))_A \\
&\quad - h^2 \sum_{i,j=1}^{s} m_{ij} (\Psi_i - \dot{\Sigma}(t_i), \Psi_j - \dot{\Sigma}(t_j))_A + 2h \sum_{j=1}^{s} b_j (\Psi_j - \dot{\Sigma}(t_j), S_n^j)_A \\
&\quad - 2 \left(\Sigma_n - \Sigma(t_n) + h \sum_{j=1}^{s} b_j (\Psi_j - \dot{\Sigma}(t_j)), Q_n \right)_A + \|Q_n\|_A^2.
\end{aligned}$$

From (3.2) and (2.16) we derive

$$\begin{aligned}
(\Psi_j - \dot{\Sigma}(t_j), \Sigma_j - \Sigma(t_j))_A &\leq b(w_j - w(t_j), \sigma_j - \sigma(t_j)) \\
&= \langle l_j, w_j - w(t_j) \rangle - \langle l(t_j), w_j - w(t_j) \rangle \\
&= 0.
\end{aligned} \tag{5.1}$$

Hence with the coefficient matrix (m_{ij}) being nonnegative definite, we obtain

$$
\begin{aligned}
&\|\Sigma_{n+1} - \Sigma(t_{n+1})\|_A^2 \\
&\quad \le \|\Sigma_n - \Sigma(t_n)\|_A^2 - 2\left(\Sigma_n - \Sigma(t_n) + h\sum_{j=1}^{s} b_j(\Psi_j - \dot\Sigma(t_j)), Q_n\right)_A \\
&\qquad + \|Q_n\|_A^2 + 2h\sum_{j=1}^{s} b_j(\Psi_i - \dot\Sigma(t_i), S_n^j)_A \\
&\quad = \|\Sigma_n - \Sigma(t_n)\|_A^2 - 2(\Sigma_n - \Sigma(t_n), Q_n)_A \\
&\qquad - 2h\sum_{j=1}^{s} b_j(\Psi_j - \dot\Sigma(t_j), Q_n)_A + \|Q_n\|_A^2 + 2h\sum_{j=1}^{s} b_j(\Psi_j - \dot\Sigma(t_j), S_n^j)_A.
\end{aligned} \tag{5.2}
$$

To derive an estimate for the right-hand side of (5.2), we need an upper bound for $h\|\Psi_j - \dot\Sigma(t_j)\|$. Therefore we again pass to the product space $\mathcal{T}^s$ to exploit the coercivity of the Runge–Kutta matrix. In the notation of section 4, we set

$$\vec{\dot\Sigma}(t_j) = (\dot\Sigma(t_1),\dots,\dot\Sigma(t_s))^{\mathrm{T}}, \qquad \vec\Sigma(t_j) = (\Sigma(t_1),\dots,\Sigma(t_s))^{\mathrm{T}},$$

and $\vec\Sigma(t_n) = (\dot\Sigma(t_n),\dots,\dot\Sigma(t_n))^{\mathrm{T}}$. Then for each stage

$$\vec{\dot\Sigma}(t_j) = \frac{1}{h}\mathcal{A}^{-1}(\vec\Sigma(t_j) - \vec\Sigma(t_n)) - \frac{1}{h}\mathcal{A}^{-1}\underbrace{(S_n^1,\dots,S_n^s)^{\mathrm{T}}}_{=:\vec S}.$$

With inequality (5.1), we get

$$\vec A(\vec\Sigma - \vec\Sigma_n - (\vec\Sigma(t_j) - \vec\Sigma(t_n)) + \vec S, \vec\Sigma - \vec\Sigma(t_j)) \le 0$$

and hence

$$\vec A(\vec\Sigma - \vec\Sigma(t_j), \vec\Sigma - \vec\Sigma(t_j)) \le \vec A((\vec\Sigma_n - \vec\Sigma(t_n)) - \vec S, \vec\Sigma - \vec\Sigma(t_j)).$$

The coercivity and continuity of $\vec A$ imply

$$\|\vec\Sigma - \vec\Sigma(t_j)\|_{\mathcal{T}^s}^2 \le c(\|\vec\Sigma_n - \vec\Sigma(t_n)\|_{\mathcal{T}^s} + \|\vec S\|_{\mathcal{T}^s})\|\vec\Sigma - \vec\Sigma(t_j)\|_{\mathcal{T}^s}.$$

Therefore for each stage we have

$$\|\Sigma_j - \Sigma(t_j)\|_{\mathcal{T}} \le c(\|\Sigma_n - \Sigma(t_n)\|_{\mathcal{T}} + S_n^{\mathrm{Err}}). \tag{5.3}$$

The stage error is thus determined by the previous approximation error and the consistency error of the Runge–Kutta method. The identity

$$\vec\Psi - \vec{\dot\Sigma}(t_j) = \frac{1}{h}\mathcal{A}^{-1}(\vec\Sigma - \vec\Sigma(t_j) - \vec\Sigma_n + \vec\Sigma(t_n)) + \frac{1}{h}\mathcal{A}^{-1}\vec S$$

implies for the stage derivatives

$$h\|\Psi_j - \dot\Sigma(t_j)\|_{\mathcal{T}} \le c(\|\Sigma_j - \Sigma(t_j)\|_{\mathcal{T}} + \|\Sigma_n - \Sigma(t_n)\|_{\mathcal{T}} + S_n^{\mathrm{Err}}).$$

Hence with (5.3) we have

$$h\|\Psi_j - \dot\Sigma(t_j)\|_{\mathcal{T}} \le c(\|\Sigma_n - \Sigma(t_n)\|_{\mathcal{T}} + S_n^{\mathrm{Err}}). \tag{5.4}$$

Now we return to (5.2), insert the estimate (5.4), and combine it with the coercivity and continuity of the bilinear form A to obtain

$$\begin{aligned}&\|\Sigma_{n+1}-\Sigma(t_{n+1})\|_{\mathcal{T}}^2\\&\le c\|\Sigma_n-\Sigma(t_n)\|_{\mathcal{T}}(Q_n^{\mathrm{Err}}+S_n^{\mathrm{Err}})+cQ_n^{\mathrm{Err}}S_n^{\mathrm{Err}}+cS_n^{\mathrm{Err}^2}+cQ_n^{\mathrm{Err}^2}.\end{aligned}$$

Summation from 1 to n leads to

$$\begin{aligned}\|\Sigma_{n+1}-\Sigma(t_{n+1})\|_{\mathcal{T}}^2\le c\sum_{r=1}^{n}\|\Sigma_r-\Sigma(t_r)\|_{\mathcal{T}}(Q_r^{\mathrm{Err}}+S_r^{\mathrm{Err}})\\+c\sum_{r=1}^{n}(Q_r^{\mathrm{Err}^2}+S_r^{\mathrm{Err}^2}+S_r^{\mathrm{Err}}Q_r^{\mathrm{Err}}).\end{aligned}$$

With $M:=\max_{r=1,\dots,N}\|\Sigma_{n+1}-\Sigma(t_{n+1})\|_{\mathcal{T}}$ it follows that

$$M^2\le cM\sum_{r=1}^{N-1}(Q_r^{\mathrm{Err}}+S_r^{\mathrm{Err}})+c\sum_{r=1}^{N-1}(Q_r^{\mathrm{Err}^2}+S_r^{\mathrm{Err}^2}+S_r^{\mathrm{Err}}Q_r^{\mathrm{Err}}).$$

Furthermore, due to

$$\|\cdot\|_{L^1(0,t_1;\Omega)}+\dots+\|\cdot\|_{L^1(t_n,t_{n+1};\Omega)}=\|\cdot\|_{L^1(0,t_{n+1};\Omega)}$$

and the Cauchy–Schwarz inequality, we arrive at

$$M^2\le Mch^q\|\Sigma^{(q+1)}\|_{L^1(0,t_f;\Omega)}+ch^{2q}\|\Sigma^{(q+1)}\|_{L^1(0,t_f;\Omega)}.$$

Therefore

$$M\le ch^q\|\Sigma^{(q+1)}\|_{L^1(0,t_f;\Omega)}.$$

6. Conclusions. In this paper, we have extended implicit Runge–Kutta methods to the infinite-dimensional constrained evolution equations of elastoplasticity. Coercivity and algebraic stability, notions that are familiar from the finite-dimensional case, turned out to be sufficient to show existence, uniqueness, contractivity preservation, and convergence. Various time integration methods become available in this way, among them Gauss, Radau, Lobatto, and several DIRK methods.

While the proof for contractivity preservation is very similar to the one for ODEs, the existence proof presented here relies very much on the framework of variational inequalities. Though we have in each material point a DAE of index 2 with a certain Lagrange multiplier (cf. (2.11)), this multiplier showed up neither in the formulation of the dual problem nor in the Runge–Kutta method definition. For implementation, in the fashion of Rothe's method or in the method of lines, however, the multiplier will regain its importance.

REFERENCES

[1] F. Armero and A. Pérez-Foguet, *On the formulation of closest-point projection algorithms in elastoplasticity—Part* I. *The variational structure*, Internat. J. Numer. Methods Engrg., 53 (2002), pp. 297–329.

[2] D. Braess, *Finite Elements: Theory, Fast Solvers and Applications in Solid Mechanics*, Cambridge University Press, Cambridge, UK, 2001.

[3] K. E. Brenan, S. L. Campbell, and L. Petzold, *Numerical Solution of Initial-Value Problems in Differential-Algebraic Equations*, Classics in Appl. Math. 14, SIAM, Philadelphia, 1996.

[4] J. Büttner and B. Simeon, *Runge-Kutta Methods in Elastoplasticity*, Appl. Numer. Math., 41 (2002), pp. 443–458.

[5] C. Carstensen, *Coupling of FEM and BEM for interface problems in viscoplasticity and plasticity with hardening*, SIAM J. Numer. Anal., 33 (1996), pp. 171–207.

[6] G. Duvaut and J. L. Lions, *Inequalities in Mechanics and Physics*, Springer-Verlag, Berlin, New York, 1976.

[7] W. Ehlers, P. Ellsiepen, and M. Amman, *Time- and space-adaptive methods applied to localization phenomena in empty and saturated micropolar and standard porous materials*, Internat. J. Numer. Methods Engrg., 52 (2001), pp. 503–526.

[8] P. Ellsiepen and S. Hartmann, *Remarks on the interpretation of current non-linear finite element analyses as differential-algebraic equations*, Internat. J. Numer. Methods Engrg., 51 (2001), pp. 679–707.

[9] W. Han and B. D. Reddy, *Plasticity, Mathematical Theory and Numerical Analysis*, Springer-Verlag, Berlin, New York, 1999.

[10] E. Hairer and G. Wanner, *Solving Ordinary Differential Equations* II, Springer-Verlag, Berlin, New York, 1996.

[11] J. Lemaitre and J. L. Chaboche, *Mechanics of Solid Materials*, Cambridge University Press, Cambridge, UK, 1990.

[12] P. Papadopoulos and R. Taylor, *On the application of multistep integration methods to infinitesimal elastoplasticity*, Internat. J. Numer. Methods Engrg., 37 (1994), pp. 3169–3184.

[13] A. Pérez-Foguet and F. Armero, *On the formulation of closest-point projection algorithms in elastoplasticity—Part* II. *Globally convergent schemes*, Internat. J. Numer. Methods Engrg., 53 (2002), pp. 331–374.

[14] P. J. Rabier and W. C. Rheinboldt, *Theoretical and numerical analysis of differential-algebraic equations*, in Handbook of Numerical Analysis, Vol. 8, P. G. Ciarlet and J. L. Lions, eds., North–Holland, Amsterdam, 2002.

[15] J. C. Simo, *Numerical analysis and simulation of plasticity*, in Handbook of Numerical Analysis, Vol. 6, P. G. Ciarlet and J. L. Lions, eds., North–Holland, Amsterdam, 1998.

[16] J. C. Simo and T. J. R. Hughes, *Computational Inelasticity*, Springer-Verlag, Berlin, New York, 1998.

[17] R. Temam, *A generalized Norton-Hoff model and the Prandtl-Reuss law of plasticity*, Arch. Ration. Mech. Anal., 95 (1986), pp. 137–183.

[18] Ch. Wieners, *Theorie und Numerik der Prandtl–Reuß Plastizität*, Habilitation thesis, Institut für Computeranwedungen, Stuttgart, 1999.

[19] Ch. Wieners, *Multigrid methods for Prandtl–Reuß Plasticity*, Numer. Linear Algebra Appl., 6 (1999), pp. 457–478.

SIAM J. NUMER. ANAL.
Vol. 41, No. 4, pp. 1585–1594

ELLIPTIC RECONSTRUCTION AND A POSTERIORI ERROR ESTIMATES FOR PARABOLIC PROBLEMS[*]

CHARALAMBOS MAKRIDAKIS[†] AND RICARDO H. NOCHETTO[‡]

Abstract. It is known that the energy technique for a posteriori error analysis of finite element discretizations of parabolic problems yields suboptimal rates in the norm $L^\infty(0,T;L^2(\Omega))$. In this paper, we combine energy techniques with an appropriate pointwise representation of the error based on an elliptic reconstruction operator which restores the optimal order (and regularity for piecewise polynomials of degree higher than one). This technique may be regarded as the "dual a posteriori" counterpart of Wheeler's elliptic projection method in the a priori error analysis.

Key words. a posteriori error estimators, finite elements, semidiscrete parabolic problems, energy technique

AMS subject classification. 65N15

DOI. 10.1137/S0036142902406314

1. Introduction.

A posteriori error estimation and adaptivity are in many cases very successful tools for efficient numerical computations of linear as well as nonlinear PDEs. In particular, a posteriori error control provides a practical, as well as mathematically sound, means of detecting multiscale phenomena and doing reliable computations. Although the a posteriori error analysis of elliptic problems is now mature [2, 3, 6, 7, 18, 23], the time dependent case is still under development. Many papers have appeared for the discontinuous Galerkin method [9, 10, 11, 13, 14, 15, 20, 19] and other schemes [1, 4, 17, 21, 24, 25] mainly for linear parabolic problems.

One of the outstanding issues related to a posteriori estimation of (linear) time dependent problems is the known fact that the energy technique for a posteriori error analysis of finite element discretizations of parabolic problems yields suboptimal rates in the norm $L^\infty(0,T;L^2(\Omega))$. Since the energy method is the most elementary technique for estimating the error in the a priori analysis, the question of whether or not this method can be successfully applied in the a posteriori error analysis is very natural. In addition, we hope that examining this and related issues will enable us to increase our understanding on the important subject of error control for time dependent problems in general.

We will work with the following linear parabolic equation as a model:

$$\begin{aligned} &u_t + Au = f \quad \text{in } \Omega\times[0,T],\\ &u(\cdot,0) = u_0(\cdot) \quad \text{in } \Omega,\\ &u = 0 \quad \text{on } \partial\Omega\times[0,T]. \end{aligned} \tag{1.1}$$

Here A is a linear, symmetric, second order positive definite elliptic operator, and Ω is a bounded domain of $\mathbb{R}^d$ $(d \geq 1)$ with sufficiently smooth boundary for our purposes.

[*] Received by the editors April 24, 2002; accepted for publication (in revised form) October 14, 2002; published electronically October 2, 2003.
http://www.siam.org/journals/sinum/41-4/40631.html

[†] Department of Applied Mathematics, University of Crete, 71409 Heraklion-Crete, Greece and Institute of Applied and Computational Mathematics, Forth, 71110 Heraklion-Crete, Greece (makr@tem.uoc.gr, http://www.tem.uoc.gr/∼makr).

[‡] Department of Mathematics and Institute for Physical Science and Technology, University of Maryland, College Park, MD 20742 (rhn@math.umd.edu, http://www.math.umd.edu/∼rhn). The research of this author was partially supported by NSF grant DMS-9971450.

Let $H := L^2(\Omega)$, $V := H_0^1(\Omega)$, and $V^\star := H^{-1}(\Omega)$ be the dual of V. If $a(\cdot,\cdot)$ is the bilinear form that corresponds to A, our assumptions on A imply that

$$\|v\|_V := a(v,v)^{1/2}$$

defines a norm on V. We denote the norms on H and $V^\star$ by $\|\cdot\|_{V^\star}$ and $\|\cdot\|_H$, respectively, and we indicate with $\langle\cdot,\cdot\rangle$ the duality pairing in either H or $V^* - V$.

We assume that $f \in L^2(0,T;V^\star)$ and $u_0 \in H$ so that (1.1) admits a unique weak solution satisfying

$$\langle u_t(t), v\rangle + a(u(t),v) = \langle f, v\rangle \quad \text{for all } v \in V, \text{ a.e. } t \in [0,T] .$$

In this paper, we consider semidiscrete finite element discretizations of *arbitrary* degree. We combine energy techniques with an appropriate pointwise representation of the error based on a novel *elliptic reconstruction* operator which restores the optimal order in $L^\infty(0,T;L^2(\Omega))$. This technique may be regarded as the dual counterpart of Wheeler's elliptic projection method in the a priori error analysis [27]. In particular, for u_h as the finite element approximation, our estimates exhibit the following properties:

- the estimator is a computable quantity in terms of the approximate solution u_h and the data f, u_0, and Ω, but its actual form and quality depends only on the elliptic estimator at our disposal;
- the order is optimal in $L^\infty(0,T;L^2(\Omega))$ for any polynomial degree ≥ 1, and the regularity is the lowest compatible with (1.1) for polynomial degree > 1;
- the a posteriori estimates mimic completely the corresponding a priori estimates.

Hereafter, the use of "optimal order" and "optimal regularity" is consistent with classical terminology in approximation theory. Consequently, "optimal order" corresponds to the largest exponent r for which the error is $O(h^r)$, where h is the biggest element diameter of the partition. Likewise, "optimal regularity" refers to the lowest regularity which is compatible with (1.1) and an error of $O(h^r)$.

Finite element approximation. For $\mathcal{T}_h$, being a shape-regular partition of Ω, consider the finite element space

$$V_h = \{\chi \in H_0^1(\Omega) : \chi|_K \in \mathbb{P}_k(K) \quad \text{for all } K \in \mathcal{T}_h\},$$

where $\mathbb{P}_k(K)$ is the space of polynomials of degree $\leq k$ over K. The finite element approximation $u_h : [0,T] \to V_h$ of u is defined to satisfy the following linear ODE:

$$\begin{aligned} \langle u_{h,t}, \chi\rangle + a(u_h,\chi) &= \langle f, \chi\rangle \quad \text{for all } \chi \in V_h, \text{ a.e. } t \in [0,T], \\ u_h(\cdot,0) &= u_h^0 \in V_h . \end{aligned} \tag{1.2}$$

A posteriori error estimation. Residual-based a posteriori estimates are usually proved by estimating the linear functional $R \in V^\star$, so-called *residual*,

$$\begin{aligned} -\langle R, v\rangle &= \int_0^T \Big(\langle u_{h,t}, v\rangle + a(u_h,v) - \langle f, v\rangle\Big) dt \\ &= \int_0^T \Big(\langle u_{h,t}, v - I_h v\rangle + a(u_h, v - I_h v) - \langle f, v - I_h v\rangle\Big) dt, \end{aligned} \tag{1.3}$$

in appropriate norms. Here, in the second equality, we have used the definition of the semidiscrete scheme (1.2) and an interpolation operator $I_h : V \to V_h$ stable in V

(e.g., Clement's interpolant). Then, for $e = u - u_h$ as the error to be estimated, we have

$$\frac{1}{2}\|e\,(T)\|_H^2 + \int_0^T a(e,e)dt = \frac{1}{2}\|e\,(0)\|_H^2 + \langle R, e\rangle. \tag{1.4}$$

Due to the presence of $\int_0^T a(u_h, e - I_h e)dt$, which gives rise to the integral of an H^1 elliptic residual, the ensuing a posteriori estimate is of optimal order in $L^2(0,T;H_0^1(\Omega))$, as corresponds to an estimate of $\int_0^T a(e,e)dt$, but is *suboptimal* in $L^\infty(0,T;L^2(\Omega))$. It is well known that an analogous phenomenon occurs in the a priori analysis and that the use of an elliptic projection operator overcomes the difficulty [27]. This is now a standard tool in the finite element analysis.

In this paper, we introduce an *elliptic reconstruction* operator which restores the optimal order in the a posteriori error estimation in $L^\infty(0,T;L^2(\Omega))$. The key properties of the elliptic reconstruction U (cf. Definition 2.1) are (i) $u - U$ satisfies an appropriate pointwise equation (cf. (3.2)) that can be used to derive estimates in terms of $u_{h,t} - U_t$, and (ii) u_h is the finite element solution of an elliptic problem whose exact solution is U, and therefore $u_h - U$ (as well as $u_{h,t} - U_t$) can be estimated in various norms by any given a posteriori elliptic estimator. Note that a similar function U was introduced in [12] for a different purpose.

For clarity of exposition, we present the method in the simplest framework. The ideas of the present paper might be useful for linear problems of nondissipative character as well as for nonlinear dissipative problems. In this direction, they should be explored together with the recent a posteriori results of time discretization of nonlinear problems [17, 19]. The a posteriori analysis of [17, 19] is based on the same principles as those in the present paper, namely, an appropriate pointwise representation of the error and energy arguments.

Although it is possible to derive quasi-optimal order-regularity estimators in $L^\infty(0,T;L^2(\Omega))$ via *parabolic duality* [9, 22], this technique hinges on the parabolic regularizing effect which is not valid for estimates in $L^2(0,T;H_0^1(\Omega))$. For the latter, duality leads invariably to estimators similar to those obtained with the energy approach and which also bound the error in $L^\infty(0,T;L^2(\Omega))$ but with suboptimal order. In contrast, several contributions over the last few years are devoted to estimates that are based on the (forward) energy approach. Picasso [21] derives a posteriori error estimates of residual type that are optimal in $L^2(0,T;H_0^1(\Omega))$ for piecewise linear elements for space discretization and backward Euler for time discretization. Toward overcoming the barrier described above, Babuška, Feistauer, and Šolín [4] derive estimates in $L^2(0,T;L^2(\Omega))$ for (1.2) by a double integration in time; see also [1, 5]. In [24, 25], Verfürth proves a posteriori estimates in $L^r(0,T;L^\rho(\Omega))$, with $1 < r, \rho < \infty$, for fully discrete approximations of quasi-linear parabolic equations.

The paper is organized as follows. We introduce the elliptic reconstruction operator in section 2, and we derive abstract a posteriori error estimates in section 3. In particular, our estimator of Theorem 3.1 depends on an abstract *elliptic estimator function* for elliptic problems; any such estimator can be used. In section 4, we specify the form of the estimates for the classical residual-type elliptic estimators.

2. Elliptic reconstruction.

We now introduce the elliptic reconstruction operator $\mathcal{R} : V_h \to V$. To this end, let $P_h^1 : V \to V_h$ be the elliptic projection operator, i.e.,

$$a(P_h^1 w, \chi) = a(w, \chi) \quad \text{for all } \chi \in V_h, \tag{2.1}$$

and let $P_h^0 : H \to V_h$ be the L^2-projection operator, i.e.,

$$(P_h^0 w, \chi) = \langle w, \chi \rangle \quad \text{for all } \chi \in V_h. \tag{2.2}$$

Let $w \in V$ satisfy the elliptic problem $Aw = g \in V^\star$ or, in weak form,

$$w \in V : \qquad a(w, v) = \langle g, v \rangle \quad \text{for all } v \in V. \tag{2.3}$$

Let $w_h \in V_h$ be the corresponding finite element solution

$$w_h \in V_h : \qquad a(w_h, \chi) = \langle g, \chi \rangle \quad \text{for all } \chi \in V_h; \tag{2.4}$$

hence $w_h = P_h^1 w$. We assume that we have at our disposal a posteriori estimators that control the error $\|w - w_h\|_X$ in the spaces $X = H, V$, or $V^\star$.

ASSUMPTION 2.1. *Let w and w_h be the exact solution and its finite element approximation given in* (2.3) *and* (2.4) *above. We assume that there exists an a posteriori estimator function $\mathcal{E} = \mathcal{E}(w_h, g; X)$, which depends on w_h, g and the space $X = H, V$, or $V^\star$ such that*

$$\|w - w_h\|_X \le \mathcal{E}(w_h, g; X). \tag{2.5}$$

Let $A_h : V_h \to V_h$ be the following discrete version of A:

$$\langle A_h v, \chi \rangle = a(v, \chi) \quad \text{for all } \chi \in V_h. \tag{2.6}$$

Then we have the following definition.

DEFINITION 2.1. *Let u_h be the finite element solution of* (1.2) *and $f_h := P_h^0 f$. We define the elliptic reconstruction $U = \mathcal{R} u_h \in H_0^1(\Omega)$ of u_h to be the solution of the elliptic problem in weak form*

$$a(U(t), v) = \langle g_h(t), v \rangle \quad \textit{for all } v \in H_0^1(\Omega), \ \textit{a.e. } t \in [0, T], \tag{2.7}$$

where

$$g_h := A_h u_h - f_h + f. \tag{2.8}$$

We note that a similar function U was defined at the final time T in [12] in a different context, i.e., in postprocessing the Galerkin method at T with the aim of improving the order of convergence. We observe that U satisfies the strong form

$$AU = A_h u_h - f_h + f \tag{2.9}$$

as well as

$$a(U, \varphi) = a(u_h, \varphi) - \langle f_h - f, \varphi \rangle = a(u_h, \varphi) \quad \text{for all } \varphi \in V_h, \tag{2.10}$$

because $f_h = P_h^0 f$. This relation implies that u_h *is the finite element solution of the elliptic problem whose exact solution is the elliptic reconstruction* U, namely,

$$u_h = P_h^1 U. \tag{2.11}$$

Assume that $f \in H^1(0, T; V^*)$. Since $a(\cdot, \cdot)$ is independent of t, there holds $a(U_t, \varphi) = a(u_{h,t}, \varphi)$ for all $\varphi \in V_h$, or

$$u_{h,t} = P_h^1 U_t. \tag{2.12}$$

In addition,

$$a(U_t, v) = \langle g_{h,t}, v \rangle \quad \text{for all } v \in V. \tag{2.13}$$

3. Abstract a posteriori error analysis. In this section, we establish the improved a posteriori error estimate in H and make several comments about its optimality regarding both order and regularity.

THEOREM 3.1. *Assume that* u *is the solution of* (1.1) *and* u_h *is its finite element approximation* (1.2). *Let* U *be the elliptic reconstruction of* u_h *and* $\mathcal{E}$ *be as defined in Assumption* 2.1. *Then the following a posteriori error bounds hold for* $0 < t \leq T$*:*

$$\|(u-U)(t)\|_H^2 + \int_0^t \|u-U\|_V^2 ds \leq \|u(0)-U(0)\|_H^2 + \int_0^t \mathcal{E}(u_{h,t}, g_{h,t}; V^\star)^2 ds$$

and

$$\begin{aligned}\|(u-u_h)(t)\|_H \leq \|u_0 - u_h^0\|_H + \left(\int_0^t \mathcal{E}(u_{h,t}, g_{h,t}; V^\star)^2 ds\right)^{1/2} \\ + \mathcal{E}(u_h(0), g_h(0); H) + \mathcal{E}(u_h(t), g_h(t); H).\end{aligned}$$

Proof. By virtue of definitions (1.2) and (2.9) of u_h and U, we have

$$u_{h,t} + AU = f,$$

whence U satisfies the following pointwise equation:

$$U_t + AU = f + (U - u_h)_t\,. \tag{3.1}$$

Thus the error equation for $u - U$ reads

$$(u-U)_t + A(u-U) = (u_h - U)_t\,. \tag{3.2}$$

Multiplying by $u - U$ and using standard energy arguments yield

$$\begin{aligned}\|(u-U)(t)\|_H^2 + \int_0^t \|(u-U)(s)\|_V^2 ds \leq \|u(0)-U(0)\|_H^2 \\ + \int_0^t \|(u_{h,t} - U_t)(s)\|_{V^\star}^2 ds\,.\end{aligned} \tag{3.3}$$

Relations (2.12) and (2.13), in conjunction with Assumption 2.1, imply

$$\|u_{ht} - U_t\|_{V^\star} \leq \mathcal{E}(u_{h,t}, g_{h,t}; V^\star)\,,$$

which in turn leads to the first assertion of Theorem 3.1. To show the second one, it suffices to note that (2.11) and Assumption 2.1 yield

$$\|(u_h - U)(t)\|_H \leq \mathcal{E}(u_h(t), g_h(t); H) \quad \text{for all } 0 \leq t \leq T\,, \tag{3.4}$$

which, together with

$$\begin{aligned}\|u(0)-U(0)\|_H &\leq \|u(0)-u_h(0)\|_H + \|P_h^1 U(0) - U(0)\|_H \\ &\leq \|u_0 - u_h^0\|_H + \mathcal{E}(u_h(0), g_h(0); H),\end{aligned}$$

concludes the proof. □

Remark 3.1 (L^2-based estimate). An alternative estimate that follows from the proof of Theorem 3.1 is

$$\begin{aligned}\max_{0\le t\le T}\|u-U\|_H^2 &\le \|u(0)-U(0)\|_H^2+\max_{0\le t\le T}\|u-U\|_H\int_0^T\|u_{h,t}-U_t\|_H dt\\ &\le \max_{0\le t\le T}\|u-U\|_H\left(\|u(0)-U(0)\|_H+\int_0^T\|u_{h,t}-U_t\|_H dt\right).\end{aligned}$$

Therefore, (2.5) and (3.4) imply

$$\max_{0\le t\le T}\|u-U\|_H\le \|u(0)-U(0)\|_H+\int_0^T\mathcal{E}(u_{h,t},g_{h,t};H)dt,$$

along with the corresponding a posteriori error bound

$$\max_{0\le t\le T}\|u-u_h\|_H\le \|u_0-u_h^0\|_H+\mathcal{E}(u_h(0),g_h(0);H)\ +2\int_0^T\mathcal{E}(u_{h,t},g_{h,t};H)dt.$$

Remark 3.2 (a priori vs. a posteriori bounds). Note that the elliptic reconstruction is an “a posteriori dual” to Wheeler's elliptic projection [22, 27]. Furthermore, the two results in Theorem 3.1 are indeed an *a posteriori dual* to the classical a priori estimate for semidiscrete linear parabolic problems [22, 27]

$$\begin{aligned}&\|(u_h-P_h^1u)(t)\|_H^2+\int_0^t\|u_h-P_h^1u\|_V^2ds\\ &\qquad\le \|u_h(0)-P_h^1u(0)\|_H^2+\int_0^t\|u_t-P_h^1u_t\|_{V^\star}^2ds\end{aligned}\tag{3.5}$$

and

$$\begin{aligned}&\|(u-u_h)(t)\|_H\le \|(u-P_h^1u)(t)\|_H\\ &\qquad+\left(\|u_h(0)-P_h^1u(0)\|_H^2+\int_0^t\|u_t-P_h^1u_t\|_{V^\star}^2dt\right)^{1/2}.\end{aligned}\tag{3.6}$$

Remark 3.3 (optimal regularity). The a priori bound in (3.5) (and therefore in (3.6)) is of optimal order. The regularity required is optimal only for polynomial degree $k\ge 2$. Indeed, by exploiting standard results on superconvergence in negative norms of elliptic finite element problems, we see that the following bound for the error of the elliptic projection holds [22, 26]:

$$\|v-P_h^1v\|_{V^\star}\le Ch^{(k+1)}\|v\|_k.\tag{3.7}$$

This estimate follows from the definition of the dual norm $\|w\|_{V^\star}=\sup_{\|z\|_V=1}\langle w,z\rangle$ and a standard duality argument. Using (3.7), we obtain

$$\int_0^T\|u_t-P_h^1u_t\|_{V^\star}^2dt\le C\int_0^T h^{2(k+1)}\|u_t\|_k^2dt\le Ch^{2(k+1)}\int_0^T\|u\|_{k+2}^2dt\,;$$

here $\|\cdot\|_s$ denotes the Sobolev norm of $H^s(\Omega)$, and for simplicity take $A=-\Delta$ and $f=0$. For an (optimal) convergence rate of order $O(h^{k+1})$ in $L^\infty(0,T;L^2(\Omega))$, the minimal regularity required by our finite element space is $u \in L^\infty(0,T;H^{k+1}(\Omega))$. However, it is a simple matter to check that for (1.1) both

$$\int_0^T \|u\|_{k+2}^2 dt \quad \text{and} \quad \max_{0\le t\le T} \|u\|_{k+1}^2$$

are bounded by the same constant depending on data. Thus the classical a priori estimate (3.6) is of optimal order and regularity for $k \geq 2$. The negative norm $\|\cdot\|_{V^*}$ appears in a complete similar fashion in the a posteriori error analysis of Theorem 3.1, and thus for polynomial degree $k \geq 2$ this indicates that the estimator is of *optimal order-regularity.*

4. Application: Residual-type error estimators. In this section, we derive the specific form of the estimates of section 3 in case we choose the classical residual-type estimators for (2.5) [6, 23]. Of course, any other choice, such as solving local problems [2, 7, 18, 23] or averaging techniques [3], is possible according to Theorem 3.1. For simplicity, we assume that $A = -\Delta$ and that Ω is sufficiently smooth in order for (4.2) below to be valid. However, Theorem 3.1 is general enough to allow for geometric singularities and corresponding elliptic estimators. We refer to [16] for *weighted* a posteriori estimators, which account for corner singularities in both H and $V^\star$ in an optimal fashion, as well as to [8], where an error estimator is derived for an elliptic problem with curved boundaries.

We first calculate $\mathcal{E}(u_{h,t}, g_{h,t}; V^\star)$ or, equivalently, estimate

$$\|\rho\|_{V^\star} = \sup_{\|\phi\|_V \le 1} \langle \rho, \phi\rangle, \quad \rho = (U - u_h)_t.$$

We accomplish this via standard duality arguments. Given $\phi \in V$, let $\psi \in V$ be defined by

$$a(\psi, v) = \langle \nabla\psi, \nabla v\rangle = \langle v, \phi\rangle \quad \text{for all } v \in V, \tag{4.1}$$

and suppose there exists a constant $C_\Omega > 0$, depending on the domain Ω, such that

$$\|\psi\|_{H^3(\Omega)} \le C_\Omega \|\phi\|_{H^1(\Omega)}. \tag{4.2}$$

If $\mathcal{T}_h = \{K\}$ is a shape-regular partition of Ω into finite elements K, then $\mathcal{S}_h = \{S\}$ denotes the set of internal interelement sides and $\mathcal{N}_h(E)$ stands for the union of all elements of $\mathcal{T}_h$ intersecting the *closed* set E ($= K$ or S). Then, assuming for the time being that the polynomial degree is $k \geq 2$ and recalling (2.12), we can write

$$\begin{aligned}\langle \rho, \phi\rangle &= a(\psi, \rho) = a(\psi - I_h\psi, \rho) \\ &\le \sum_{K\in\mathcal{T}_h} |(\psi - I_h\psi, \Delta\rho)_K| + \sum_{S\in\mathcal{S}_h} \int_S |\psi - I_h\psi||[\partial_n\rho]|\, ds \\ &\le C_I \sum_{K\in\mathcal{T}_h} h_K^3 |\psi|_{3,\mathcal{N}_h(K)} \|\Delta\rho\|_{L^2(K)} \\ &\quad + C_I \sum_{S\in\mathcal{S}_h} h_S^{5/2} |\psi|_{3,\mathcal{N}_h(S)} \|[\partial_n\rho]\|_{L^2(S)},\end{aligned} \tag{4.3}$$

where $C_I > 0$ is an interpolation constant associated with the local interpolation operator I_h. If we further set

$$\eta_{-1}(u_{h,t})^2 = \sum_{K\in\mathcal{T}_h} h_K^6 \|\Delta\rho\|_{L^2(K)}^2 + \sum_{S\in\mathcal{S}_h} h_S^5 \|[\partial_n u_{h,t}]\|_{L^2(S)}^2$$

and make use of (4.2), then we end up with the a posteriori error estimate

$$\mathcal{E}(u_{h,t}, g_{h,t}; V^\star) = \|\rho\|_{V^\star} \le C_I C_\Omega \eta_{-1}(u_{h,t}),$$

where C_I now contains an additional factor to account for the h-independent overlap of sets $\mathcal{N}_h(E)$ in (4.3).

The form of $\eta_{-1}(u_{h,t})$ can be further simplified upon using the definition of the elliptic reconstruction and the semidiscrete scheme:

$$\Delta\rho = \Delta U_t - \Delta u_{h,t} = -A_h u_{h,t} + f_{h,t} - f_t - \Delta u_{h,t}\,.$$

Since $u_{h,tt} + A_h u_{h,t} = f_{h,t}$, we have

$$\Delta\rho = -f_{h,t} + u_{h,tt} + f_{h,t} - f_t - \Delta u_{h,t} = (u_{h,t} - \Delta u_h - f)_t\,.$$

If we denote the element residuals as

$$r|_K := u_{h,t} - \Delta u_h - f \quad \text{for all } K \in \mathcal{T}_h, \qquad j|_S := [\partial_n u_h] \quad \text{for all } S \in \mathcal{S}_h,$$

we finally get

$$\eta_{-1}(u_{h,t})^2 = \sum_{K\in\mathcal{T}_h} h_K^6 \|r_t\|^2_{L^2(K)} + \sum_{S\in\mathcal{S}_h} h_S^5 \|j_t\|^2_{L^2(S)}, \tag{4.4}$$

and

$$\mathcal{E}(u_{h,t}, g_{h,t}; V^\star) \le C_I C_\Omega \eta_{-1}(u_{h,t}) \quad \text{if } k \ge 2.$$

Using similar arguments, we can derive

$$\mathcal{E}(u_h, g_h; H) \le C_I C_\Omega \eta_0(u_h) \quad \text{if } k \ge 2,$$

where

$$\eta_0(u_h)^2 = \sum_{K\in\mathcal{T}_h} h_K^4 \|r\|^2_{L^2(K)} + \sum_{S\in\mathcal{S}_h} h_S^3 \|j\|^2_{L^2(S)}. \tag{4.5}$$

Note that the constants C_I, C_Ω may have different values now. Finally, in the case $k = 1$, the use of negative norm does not give better results because of the lack of superconvergence in $V^\star$. Hence

$$\mathcal{E}(u_{h,t}, g_{h,t}; V^\star) \le \mathcal{E}(u_{h,t}, g_{h,t}; H) \le C_I C_\Omega \eta_0(u_{h,t})\,. \tag{4.6}$$

In summary, in view of Theorem 3.1, we have derived the following *explicit* error estimate.

THEOREM 4.1 (a posteriori estimators of residual type). *Assume that the domain Ω is sufficiently smooth, and let $t \in (0,T]$. If $k = 1$, then the following a posteriori estimate holds:*

$$\|(u - u_h)(t)\|_H \le \|u^0 - u_h^0\|_H + C_I C_\Omega \left\{ \eta_0(u_h(0)) + \eta_0(u_h(t)) + \left(\int_0^t \eta_0(u_{h,t}(s))^2 ds \right)^{1/2} \right\}.$$

In addition, for $k \geq 2$, we have

$$\|(u-u_h)(t)\|_H \leq \|u^0 - u_h^0\|_H + C_I C_\Omega \left\{ \eta_0(u_h(0)) + \eta_0(u_h(t)) + \left(\int_0^t \eta_{-1}(u_{h,t}(s))^2 ds \right)^{1/2} \right\},$$

where the estimators η_0 and η_{-1} are given by (4.5) *and* (4.4)*, respectively.*

Remark 4.1. The reasoning of Remark 3.3 applies and indicates that the estimator in Theorem 4.1 is of optimal order for polynomial degree $k \geq 1$ and of optimal regularity for $k \geq 2$. We do not actually show an a priori convergence rate for the a posteriori estimators $\eta_0(u_h)$ and $\eta_1(u_h)$; this is of interest but lies outside the scope of this paper.

REFERENCES

[1] S. Adjerid, J. E. Flaherty, and I. Babuška, *A posteriori error estimation for the finite element method-of-lines solution of parabolic problems*, Math. Models Methods Appl. Sci., 9 (1999), pp. 261–286.

[2] M. Ainsworth and J. Oden, *A unified approach to a posteriori error estimation using element residual methods*, Numer. Math., 65 (1993), pp. 23–50.

[3] M. Ainsworth, J. Z. Zhu, A. W. Craig, and O. C. Zienkiewicz, *Analysis of the Zienkiewicz-Zhu a posteriori error estimator in the finite element method*, Internat. J. Numer. Methods Engrg., 28 (1989), pp. 2161–2174.

[4] I. Babuška, M. Feistauer, and P. Šolín, *On one approach to a posteriori error estimates for evolution problems solved by the method-of-lines*, Numer. Math., 89 (2001), pp. 225–256.

[5] I. Babuška and S. Ohnimus, *A posteriori error estimation for the semidiscrete finite element method of parabolic differential equations*, Comput. Methods Appl. Mech. Engrg., 190 (2001), pp. 4691–4712.

[6] I. Babuška and W. C. Rheinboldt, *Error estimates for adaptive finite element computations*, SIAM J. Numer. Anal., 15 (1978), pp. 736–754.

[7] R. E. Bank and A. Weiser, *Some a posteriori error estimators for elliptic partial differential equations*, Math. Comp., 44 (1985), pp. 283–301.

[8] W. Dörfler and M. Rumpf, *An adaptive strategy for elliptic problems including a posteriori error controlled boundary approximation*, Math. Comp., 67 (1998), pp. 1361–1382.

[9] K. Eriksson and C. Johnson, *Adaptive finite element methods for parabolic problems.* I. *A linear model problem*, SIAM J. Numer. Anal., 28 (1991), pp. 43–77.

[10] K. Eriksson and C. Johnson, *Adaptive finite element methods for parabolic problems.* IV. *Nonlinear problems*, SIAM J. Numer. Anal., 32 (1995), pp. 1729–1749.

[11] K. Eriksson, C. Johnson, and S. Larsson, *Adaptive finite element methods for parabolic problems.* VI. *Analytic semigroups*, SIAM J. Numer. Anal., 35 (1998), pp. 1315–1325.

[12] B. García-Archilla and E. S. Titi, *Postprocessing the Galerkin method: The finite-element case*, SIAM J. Numer. Anal., 37 (2000), pp. 470–499.

[13] C. Johnson, *Error estimates and adaptive time-step control for a class of one-step methods for stiff ordinary differential equations*, SIAM J. Numer. Anal., 25 (1988), pp. 908–926.

[14] C. Johnson, *Discontinuous Galerkin finite element methods for second order hyperbolic problems*, Comput. Methods Appl. Mech. Engrg., 107 (1993), pp. 117–129.

[15] C. Johnson, Y. Y. Nie, and V. Thomée, *An a posteriori error estimate and adaptive timestep control for a backward Euler discretization of a parabolic problem*, SIAM J. Numer. Anal., 27 (1990), pp. 277–291.

[16] X. Liao and R. H. Nochetto, *Local a posteriori error estimates and adaptive control of pollution effects*, Numer. Methods Partial Differential Equations, 19 (2003), pp. 421–442.

[17] Ch. Makridakis and R. H. Nochetto, *A posteriori error analysis of a class of dissipative methods for nonlinear evolution problems*, submitted (2002).

[18] P. Morin, R. H. Nochetto, and K. G. Siebert, *Local problems on stars: A posteriori error estimators, convergence, and performance*, Math. Comp., 72 (2003), pp. 1067–1097.

[19] R. H. Nochetto, G. Savaré, and C. Verdi, *A posteriori error estimates for variable time-step discretizations of nonlinear evolution equations*, Comm. Pure Appl. Math., 53 (2000), pp. 525–589.

[20] R. H. Nochetto, A. Schmidt, and C. Verdi, *A posteriori error estimation and adaptivity for degenerate parabolic problems*, Math. Comp., 69 (2000), pp. 1–24.

[21] M. Picasso, *Adaptive finite elements for a linear parabolic problem*, Comput. Methods Appl. Mech. Engrg., 167 (1998), pp. 223–237.

[22] V. Thomée, *Galerkin Finite Element Methods for Parabolic Problems*, Springer-Verlag, Berlin, 1997.

[23] R. Verfürth, *A Posteriori Error Estimation and Adaptive Mesh Refinement Techniques*, Teubner, Stuttgart, 1995.

[24] R. Verfürth, *A posteriori error estimates for nonlinear problems. $L^r(0,T;L^\rho(\omega))$-error estimates for finite element discretizations of parabolic equations*, Math. Comp., 67 (1998), pp. 1335–1360.

[25] R. Verfürth, *A posteriori error estimates for nonlinear problems: $L^r(0,T;W^{1,\rho}(\omega))$-error estimates for finite element discretizations of parabolic equations*, Numer. Methods Partial Differential Equations, 14 (1998), pp. 487–518.

[26] L. B. Wahlbin, *Superconvergence in Galerkin Finite Element Methods*, Springer-Verlag, Berlin, 1995.

[27] M. F. Wheeler, *A priori L_2 error estimates for Galerkin approximations to parabolic partial differential equations*, SIAM J. Numer. Anal., 10 (1973), pp. 723–759.

SIAM J. Numer. Anal.
Vol. 41, No. 5, pp. 1595–1619

A NEW DUAL-PETROV–GALERKIN METHOD FOR THIRD AND HIGHER ODD-ORDER DIFFERENTIAL EQUATIONS: APPLICATION TO THE KDV EQUATION*

JIE SHEN†

Abstract. A new dual-Petrov–Galerkin method is proposed, analyzed, and implemented for third and higher odd-order equations using a spectral discretization. The key idea is to use trial functions satisfying the underlying boundary conditions of the differential equations and test functions satisfying the "dual" boundary conditions. The method leads to linear systems which are sparse for problems with constant coefficients and well conditioned for problems with variable coefficients. Our theoretical analysis and numerical results indicate that the proposed method is extremely accurate and efficient and most suitable for the study of complex dynamics of higher odd-order equations.

Key words. dual-Petrov–Galerkin, KDV equation, Legendre polynomials, spectral approximation

AMS subject classifications. 65M70, 35Q53, 65M12

DOI. 10.1137/S0036142902410271

1. Introduction.

Over the last thirty years, spectral methods have been playing an increasingly important role in scientific and engineering computations. Most work on spectral methods is concerned with elliptic and parabolic-type equations; there has also been active research on spectral methods for hyperbolic problems (see, for instance, [11, 7, 14] and the references therein). However, there is only a limited body of literature on spectral methods for dispersive, namely, third and higher odd-order, equations. In particular, relatively few studies are devoted to third and higher odd-order equations in finite intervals. This is partly due to the fact that direct collocation methods for higher odd-order boundary problems lead to very much higher condition numbers—more precisely, of order N^{2k}, where N is the number of modes and k is the order of the equation—and often exhibit unstable modes if the collocation points are not properly chosen (see, for instance, [17, 21]).

In a sequence of papers [22, 23, 25, 26], the author constructed efficient spectral-Galerkin algorithms for elliptic equations in various situations. In this paper, we extend the main idea for constructing efficient spectral-Galerkin algorithms—using *compact combinations* of orthogonal polynomials, which satisfy *essentially* all the underlying homogeneous boundary conditions, as basis functions—to third and higher odd-order equations. Since the main differential operators in these equations are not symmetric, it is quite natural to employ a Petrov–Galerkin method.

The key idea of the new spectral dual-Petrov–Galerkin method is the innovative choice of the test and trial functional spaces. More precisely, we choose the trial functions to satisfy the underlying boundary conditions of the differential equations, and we choose the test functions to satisfy the "dual" boundary conditions.

Recently, Ma and Sun [19, 20] studied an interesting Legendre–Petrov–Galerkin method for third-order equations. The main difference between this paper and [19, 20]

*Received by the editors June 26, 2002; accepted for publication (in revised form) February 18, 2003; published electronically October 2, 2003. This research was partially supported by NFS grant DMS-0074283.

http://www.siam.org/journals/sinum/41-5/41027.html

†Department of Mathematics, Purdue University, West Lafayette, IN 47907 (shen@math.purdue.edu).

lies in the choice of the test and trial function spaces and their basis functions. The critical feature of test and trial spaces used here is that they allow us to integrate by parts freely without introducing any additional boundary terms. With this property and our choice of using "compact combinations" (minimal interactions) of Legendre polynomials as basis functions for the test and trial spaces, we obtain linear systems which are *compactly sparse* for problems with constant coefficients and *well conditioned* (i.e., the condition number is independent of the number of unknowns) for problems with variable coefficients. This is rather remarkable considering the fact that the problems at hand are nonsymmetric and involve high-order derivatives.

Together with the so-called Chebyshev–Legendre approach [10, 24], i.e., using the Legendre formulation and Chebyshev–Gauss–Lobatto points, our method has a quasi-optimal computational complexity and is well conditioned to permit the use of very large numbers of modes without suffering from large round-off errors, which are necessary for simulations of very complex dynamics of challenging scientific and engineering problems.

The new spectral dual-Petrov–Galerkin method not only leads to quasi-optimal numerical algorithms; it is also equivalent to a *natural* weighted variational formulation for third and higher odd-order equations. By showing that the basis functions for the trial (and test) spaces form a sequence of orthogonal polynomials in a weighted Sobolev space, we are able to establish optimal error estimates in appropriate weighted Sobolev spaces.

The paper is organized as follows. In sections 2 and 3, we study the third-order and fifth-order equations, respectively. As an example of application, we consider the Korteweg–de Vries (KDV) equation on a finite interval in section 4. In section 5, we discuss miscellaneous issues/extensions of the spectral dual-Petrov–Galerkin methods. In section 6, we present various numerical results exhibiting the accuracy and efficiency of our numerical algorithms. We end the paper with a few concluding remarks.

We now introduce some notation. Let $\omega(x)$ be a positive weight function on $I=(-1,1)$. One usually requires that $\omega \in L^1(I)$. However, in this paper, we shall be interested mainly in the case $\omega \notin L^1(I)$. We shall use the weighted Sobolev spaces $H^m_\omega(\Omega)$ $(m=0,\pm1,\dots)$ whose norms are denoted by $\|\cdot\|_{m,\omega}$. In particular, the norm and inner product of $L^2_\omega(\Omega)=L^2_\omega(\Omega)$ are denoted by $\|\cdot\|_\omega$ and $(\cdot,\cdot)_\omega$, respectively. To account for homogeneous boundary conditions, we define

$$H^m_{0,\omega}(\Omega)=\{v\in H^m_\omega(\Omega): v(\pm1)=v'(\pm1)=\cdots=v^{(m-1)}(\pm1)=0\},\quad m=1,2,\dots.$$

The subscript ω will be omitted from the notation in the case where $\omega\equiv1$.

We denote by c a generic constant that is independent of any parameters and functions. In most cases, we shall simply use the expression $A\lesssim B$ to mean that there exists a generic constant c such that $A\le cB$.

Let L_k be the kth degree Legendre polynomial. We now recall some basic properties of Legendre polynomials (cf. [27]) which will be used in this paper.

$$(1.1)\qquad \int_{-1}^{1} L_k(x)L_j(x)dx=\frac{2}{2k+1}\delta_{kj};$$

$$(1.2)\qquad L_n(x)=\frac{1}{2n+1}(L'_{n+1}(x)-L'_{n-1}(x)),\qquad n\ge1;$$

$$L_n'(x) = \sum_{\substack{k=0\\ k+n \text{ odd}}}^{n-1} (2k+1)L_k(x); \tag{1.3a}$$

$$L_n''(x) = \sum_{\substack{k=0\\ k+n \text{ even}}}^{n-2} \left(k+\frac{1}{2}\right)(n(n+1)-k(k+1))\,L_k(x); \tag{1.3b}$$

$$L_n(\pm 1) = (\pm 1)^n, \tag{1.4a}$$

$$L_n'(\pm 1) = \frac{1}{2}(\pm 1)^{n-1}n(n+1), \tag{1.4b}$$

$$L_n''(\pm 1) = (\pm 1)^n (n-1)n(n+1)(n+2)/8. \tag{1.4c}$$

2. Third-order equations.

2.1. Dual-Petrov–Galerkin method. Consider the model third-order equation

$$\begin{aligned} &\alpha u - \beta u_x - \gamma u_{xx} + u_{xxx} = f, \quad x \in I = (-1,1),\\ &u(\pm 1) = u_x(1) = 0, \end{aligned} \tag{2.1}$$

where α, β, γ are given constants. Without loss of generality, we consider only homogeneous boundary conditions, for nonhomogeneous boundary conditions $u(-1) = c_1$, $u(1) = c_2$, and $u_x(1) = c_3$ can be handled easily by considering $v = u - \hat{u}$, where $\hat{u}$ is the unique quadratic polynomial satisfying the nonhomogeneous boundary conditions.

Denoting by P_N the space of polynomials of degree $\le N$, we set

$$V_N = \{u \in P_N:\ u(\pm 1) = u_x(1) = 0\}, \quad V_N^* = \{u \in P_N:\ u(\pm 1) = u_x(-1) = 0\}. \tag{2.2}$$

For any constants a and b, let $\omega^{a,b}(x) = (1-x)^a(1+x)^b$. We also define

$$V = \{u:\ u \in H_0^1(I),\ u_x \in L^2_{\omega^{-2,0}}(I)\}, \quad V^* = \{u:\ u \in H_0^1(I),\ u_x \in L^2_{\omega^{0,-2}}(I)\}. \tag{2.3}$$

It is clear that $V_N \subset V$ and $V_N^* \subset V^*$.

We consider the following Legendre dual-Petrov–Galerkin approximation for (2.1): Find $u_N \in V_N$ such that

$$\alpha(u_N, v_N) - \beta(\partial_x u_N, v_N) + \gamma(\partial_x u_N, \partial_x v_N) + (\partial_x u_N, \partial_x^2 v_N) = (f, v_N) \ \ \forall v_N \in V_N^*, \tag{2.4}$$

where $(u,v) = \int_I uv dx$, $\partial_x u$, and $\partial_x^2 u$ denote $\frac{du}{dx}$ and $\frac{d^2u}{dx^2}$, respectively.

Notice that for any $u_N \in V_N$ we have $\omega^{-1,1}u_N \in V_N^*$. Thus the above dual-Petrov–Galerkin formulation is equivalent to the following weighted spectral-Galerkin approximation: Find $u_N \in V_N$ such that

$$\begin{aligned} \alpha(u_N, v_N)_{\omega^{-1,1}} - \beta(\partial_x u_N, v_N)_{\omega^{-1,1}} + \gamma(\partial_x u_N, \omega^{1,-1}\partial_x(v_N\omega^{-1,1}))_{\omega^{-1,1}} \\ + (\partial_x u_N, \omega^{1,-1}\partial_x^2(v_N\omega^{-1,1}))_{\omega^{-1,1}} = (f, v_N)_{\omega^{-1,1}} \ \ \forall v_N \in V_N, \end{aligned} \tag{2.5}$$

where $(u,v)_{\omega^{-1,1}} = \int_I uv\omega^{-1,1}dx$.

We shall see that the dual-Petrov–Galerkin formulation (2.4) is most suitable for implementation, while the weighted Galerkin formulation (2.5) is more convenient for error analysis.

2.2. Basis functions and projection operators. As suggested in [22, 24], one should choose compact combinations of orthogonal polynomials as basis functions to minimize the bandwidth and the condition number of the coefficient matrix corresponding to (2.5). Let $\{p_k\}$ be a sequence of orthogonal polynomials. As a general rule, for one-dimensional differential equations with m boundary conditions, one should look for basis functions in the form

$$\phi_k(x) = p_k(x) + \sum_{j=1}^{m} a_j^{(k)} p_{k+j}(x), \tag{2.6}$$

where $a_j^{(k)}$ $(j = 1, \dots, m)$ are chosen so that $\phi_k(x)$ satisfy the m homogeneous boundary conditions.

Using (1.4), one verifies readily that

$$\begin{aligned} \phi_k(x) &= L_k(x) - \frac{2k+3}{2k+5}L_{k+1}(x) - L_{k+2}(x) + \frac{2k+3}{2k+5}L_{k+3}(x) \in V_{k+3}, \\ \psi_k(x) &= L_k(x) + \frac{2k+3}{2k+5}L_{k+1}(x) - L_{k+2}(x) - \frac{2k+3}{2k+5}L_{k+3}(x) \in V^*_{k+3}. \end{aligned} \tag{2.7}$$

Therefore, for $N \ge 3$, we have

$$\begin{aligned} V_N &= \operatorname{span}\{\phi_0, \phi_1, \dots, \phi_{N-3}\}; \\ V^*_N &= \operatorname{span}\{\psi_0, \psi_1, \dots, \psi_{N-3}\}. \end{aligned} \tag{2.8}$$

Next, we discuss the properties of $\{\phi_k\}$ and $\{\psi_k\}$ and related projection operators in $L^2_{\omega^{-2,-1}}$ and $L^2_{\omega^{-1,-2}}$. Since the procedures for $L^2_{\omega^{-2,-1}}$ and $L^2_{\omega^{-1,-2}}$ are completely parallel, we shall describe only the results for $L^2_{\omega^{-2,-1}}$. One can obtain the corresponding results for $L^2_{\omega^{-1,-2}}$ by making a change of variable $x \to -x$.

LEMMA 2.1. *Let $\{\phi_k\}$ be defined as in* (2.7). *Then*

$$\int_I \phi_k \phi_j \omega^{-2,-1} dx = 0, \; k \ne j, \tag{2.9}$$

and $\{\phi_k\}$ form a complete orthogonal basis in $L^2_{\omega^{-1,1}}$.

Furthermore, ϕ_k satisfies the following Sturm–Liouville equation:

$$A\phi_k := -(1-x)^2(1+x)\partial_x \left\{(1-x)^{-1}\partial_x \phi_k(x)\right\} = (k+1)(k+3)\phi_k(x). \tag{2.10}$$

Proof. By construction, $p_k(x) := \phi_k(x)\omega^{-2,-1}$ is a polynomial of degree $\le k$. Thanks to the orthogonality of the Legendre polynomials,

$$\int_I \phi_k \phi_j \omega^{-2,-1} dx = \int_I p_k \phi_j dx = 0 \;\; \forall k < j.$$

Hence $\{\phi_k\}$ is a sequence of orthogonal polynomials in $L^2_{\omega^{-2,-1}}$. One can verify that $\phi_k(x)$ is proportional to $(1-x)^2(1+x)J_k^{2,1}(x)$. Thus $\{\phi_k\}$ forms a complete orthogonal basis in $L^2_{\omega^{-2,-1}}$ since $\{J_k^{2,1}\}$ forms a complete orthogonal basis in $L^2_{\omega^{2,1}}$.

It is clear that $A\phi_k(x)$ is a polynomial of degree $\le k+3$ and $\partial_x\{(1-x)^{-1}\partial_x\phi_k(x)\}$ is a polynomial of degree $\le k$. Hence

$$\begin{aligned} \int_I A\phi_k(x)\phi_j(x)\omega^{-2,-1}dx &= -\int_I \partial_x\{(1-x)^{-1}\partial_x\phi_k(x)\}\phi_j dx \\ &= -\int_I \partial_x\{(1-x)^{-1}\partial_x\phi_j(x)\}\phi_k dx = 0 \;\; \forall j < k. \end{aligned}$$

Therefore, $A\phi_k$ must be proportional to ϕ_k; i.e., $A\phi_k = \lambda_k \phi_k$. By comparing the coefficients of x^{k+3}, we find that $\lambda_k = (k+1)(k+3)$. $\square$

Now, let π_N be the $L^2_{\omega^{-2,-1}}$-orthogonal projector $L^2_{\omega^{-2,-1}} \to V_N$ defined by

$$(u - \pi_N u, v_N)_{\omega^{-2,-1}} = 0 \ \ \forall v_N \in V_N. \tag{2.11}$$

We also define

$$B^m_{\omega^{-2,-1}}(I) = \{u \in L^2_{\omega^{-2,-1}}(I) : \partial^l_x u \in L^2_{\omega^{l-2,l-1}}(I), \ 1 \le l \le m\}. \tag{2.12}$$

Then, we have the following error estimates.

Theorem 2.1.

$$\|\partial^l_x (u - \pi_N u)\|_{\omega^{l-2,l-1}} \lesssim N^{l-m} \|\partial^m_x u\|_{\omega^{m-2,m-1}} \ \ \forall u \in B^m_{\omega^{-2,-1}}, \ 0 \le l \le m.$$

Proof. We recall that for $a, b > -1$, the Jacobi polynomials satisfy the following relations:

$$\partial^l_x J^{a,b}_k(x) = \kappa^{a,b}_{k,l} J^{a+l,b+l}_{k-l}(x), \ a, b > -1, \ k \ge l, \tag{2.13}$$

where

$$\kappa^{a,b}_{k,l} = \frac{\Gamma(k+l+a+b+1)}{2^l \Gamma(k+a+b+1)};$$

$$\int_I J^{a,b}_k(x) J^{a,b}_j(x) \omega^{a,b} dx = \gamma^{a,b}_k \delta_{kj}, \tag{2.14}$$

where

$$\gamma^{a,b}_k = \frac{2^{a+b+1}\Gamma(k+a+1)\Gamma(k+b+1)}{(2k+a+b+1)\Gamma(k+1)\Gamma(k+a+b+1)}.$$

We shall extend the definition of the Jacobi polynomials to $(a, b) = (-2, -1)$ such that the relations (2.13)–(2.14) are still valid. To this end, we define

$$\begin{aligned} J^{-1,0}_k(x) &= -\frac{1}{2}(1-x) J^{1,0}_{k-1}(x), \ k \ge 1, \\ J^{-2,-1}_k(x) &= \frac{1}{2}(k-2) \int_{-1}^{x} J^{-1,0}_{k-1}(t) dt, \ k \ge 3. \end{aligned} \tag{2.15}$$

One derives immediately that $\{J^{-1,0}_k\}$ are mutually orthogonal in $L^2_{\omega^{-1,0}}$. Note that $\{L_{k-1} - L_k\}$ are also mutually orthogonal in $L^2_{\omega^{-1,0}}$. Hence $J^{-1,0}_k$ is proportional to $\{L_{k-1} - L_k\}$. One can also derive from the properties of Legendre polynomials that $J^{-2,-1}_k(\pm 1) = \partial_x J^{-2,-1}_k(1) = 0$. We then derive from (1.2) and (2.15) that $J^{-2,-1}_k$ must be proportional to ϕ_{k-3}. Hence $\{J^{-2,-1}_k\}$ are mutually orthogonal in $L^2_{\omega^{-2,-1}}$. Moreover, one can verify that $J^{-1,0}_k$ and $J^{-2,-1}_k$ satisfy the relations (2.13)–(2.14).

For any $u \in L^2_{\omega^{-2,-1}}$, we write

$$u(x) = \sum_{k=3}^{\infty} \tilde{u}_k J^{-2,-1}_k(x) \ \text{ with } \ \tilde{u}_k = (u, J^{-2,-1}_k)_{\omega^{-2,-1}} / \|J^{-2,-1}_k\|^2_{\omega^{-2,-1}}.$$

Hence $u - \pi_{_N} u = \sum_{k=N+1}^{\infty} \tilde{u}_k J_k^{-2,-1}$. Let us define

$$C_{N,l,m} = \max_{k>N} \frac{(\kappa_{k,l}^{-2,-1})^2 \gamma_{k-l}^{l-2,l-1}}{(\kappa_{k,m}^{-2,-1})^2 \gamma_{k-m}^{m-2,m-1}}. \tag{2.16}$$

Then, by using (2.13)–(2.14), we find

$$\begin{aligned} \|\partial_x^l (u - \pi_{_N} u)\|^2_{\omega^{l-2,l-1}} &= \sum_{k=N+1}^{\infty} \tilde{u}_k^2 (\kappa_{k,l}^{-2,-1})^2 \|J_{k-l}^{l-2,l-1}\|^2_{\omega^{l-2,l-1}} \\ &\le C_{N,l,m} \sum_{k=N+1}^{\infty} \tilde{u}_k^2 (\kappa_{k,m}^{-2,-1})^2 \|J_{k-m}^{m-2,m-1}\|^2_{\omega^{m-2,m-1}} \\ &\le C_{N,l,m} \|\partial_x^m u\|^2_{\omega^{m-2,m-1}}. \end{aligned} \tag{2.17}$$

The desired results follow from the above inequality and the fact that

$$C_{N,l,m} \lesssim N^{2(l-m)}. \qquad \square$$

2.3. Error estimates. Let us first prove the following generalized Poincaré inequalities.

LEMMA 2.2.

$$\begin{aligned} \int_I \frac{u^2}{(1-x)^4} dx &\le \frac{4}{9} \int_I \frac{u_x^2}{(1-x)^2} dx \quad \forall u \in V_N, \\ \int_I \frac{u^2}{(1-x)^3} dx &\le \int_I \frac{u_x^2}{1-x} dx \quad \forall u \in V_N. \end{aligned} \tag{2.18}$$

Proof. Let $u \in V_N$ and $h \le 2$. Then, for any constant q, we have

$$\begin{aligned} 0 &\le \int_I \left(\frac{u}{1-x} + q u_x \right)^2 \frac{1}{(1-x)^h} dx \\ &= \int_I \left(\frac{u^2}{(1-x)^{2+h}} + q \frac{(u^2)_x}{(1-x)^{1+h}} + q^2 \frac{u_x^2}{(1-x)^h} \right) dx \\ &= (1 - (1+h)q) \int_I \frac{u^2}{(1-x)^{2+h}} dx + q^2 \int_I \frac{u_x^2}{(1-x)^h} dx. \end{aligned}$$

We obtain the first inequality by taking $h = 2$ and $q = \frac{2}{3}$ and the second inequality with $h = 1$ and $q = 1$. $\square$

Remark 2.1. We note that with a change of variable $x \to -x$ in the above lemma, we have corresponding inequalities for $u \in V_N^*$.

LEMMA 2.3.

$$\frac{1}{3} \|u_x\|^2_{\omega^{-2,0}} \le (u_x, (u\omega^{-1,1})_{xx}) \le 3 \|u_x\|^2_{\omega^{-2,0}} \ \forall u \in V_N. \tag{2.19}$$

Proof. For any $u \in V_N$, we have $u\omega^{-1,1} \in V_N^*$. Thanks to the homogeneous boundary conditions built into the spaces V_N and V_N^*, all the boundary terms from the integration by parts of the third-order term would vanish. Therefore, using the identity $\partial_x^k \omega^{-1,1}(x) = \frac{2\,k!}{(1-x)^{k+1}}$ and Lemma 2.2, we find

$$\begin{aligned}(u_x, (u\omega^{-1,1})_{xx}) &= (u_x, u_{xx}\omega^{-1,1} + 2u_x\omega_x^{-1,1} + u\omega_{xx}^{-1,1})\\ &= \frac{1}{2}\int_I \left((u_x^2)_x\omega^{-1,1} + (u^2)_x\omega_{xx}^{-1,1} + 4u_x^2\omega_x^{-1,1}\right) dx\\ &= \int_I \left(\frac{3}{2}u_x^2\omega_x^{-1,1} - \frac{1}{2}u^2\omega_{xxx}^{-1,1}\right) dx\\ &= 3\int_I \frac{u_x^2}{(1-x)^2}dx - 6\int_I \frac{u^2}{(1-x)^4}dx \geq \frac{1}{3}\int_I \frac{u_x^2}{(1-x)^2}dx.\end{aligned}$$

The desired results follow immediately from the above. □

Before we proceed with the error estimates, we make the following simple but important observation.

Lemma 2.4. *Let π_N be defined in* (2.11). *Then*

$$(\partial_x(u - \pi_N u), \partial_x^2 v_N) = 0 \ \ \forall u \in V, \ v_N \in V_N^*.$$

Proof. The result is a direct consequence of (2.11), the identity

$$(\partial_x(u - \pi_N u), \partial_x^2 v_N) = -(u - \pi_N u, \omega^{2,1}\partial_x^3 v_N)_{\omega^{-2,-1}},$$

and the fact that $\omega^{2,1}\partial_x^3 v_N \in V_N$. □

Let us denote $\hat{e}_N = \pi_N u - u_N$ and $e_N = u - u_N = (u - \pi_N u) + \hat{e}_N$.

Theorem 2.2. *For any α, $\beta \geq 0$ and $-\frac{1}{3} < \gamma < \frac{1}{6}$, there exists a unique solution for the system* (2.4). *Furthermore, for $u \in B^m_{\omega^{-2,-1}}$, we have*

$$\alpha\|e_N\|_{\omega^{-1,1}} + N^{-1}\|(e_N)_x\|_{\omega^{-1,0}} \lesssim (1 + |\gamma|N)N^{-m}\|\partial_x^m u\|_{\omega^{m-2,m-1}}, \ m \geq 1.$$

Proof. We derive from (2.1), (2.5), and Lemma 2.4 that

$$\begin{aligned}\alpha(e_N, v_N)_{\omega^{-1,1}} - \beta(\partial_x e_N, v_N)_{\omega^{-1,1}} + \gamma(\partial_x e_N, \omega^{1,-1}\partial_x(v_N\omega^{-1,1}))_{\omega^{-1,1}}\\ + (\partial_x \hat{e}_N, \omega^{1,-1}\partial_x^2(v_N\omega^{-1,1}))_{\omega^{-1,1}} = 0 \ \ \forall v_N \in V_N.\end{aligned} \tag{2.20}$$

Taking $v_N = \hat{e}_N$ in the above and using Lemma 2.3 and the identities

$$\begin{aligned}&-(v_x, v)_{\omega^{-1,1}} = -\frac{1}{2}\int_I (v^2)_x\omega^{-1,1}dx = \|v\|^2_{\omega^{-2,0}} \ \ \forall v \in V_N,\\ &(v_x, (v\omega^{-1,1})_x) = (v_x, v_x\omega^{-1,1} + 2v\omega^{-2,0}) = \|v_x\|^2_{\omega^{-1,1}} - 2\|v\|^2_{\omega^{-3,0}} \ \ \forall v \in V_N,\end{aligned} \tag{2.21}$$

we obtain

$$\begin{aligned}\alpha\|\hat{e}_N\|^2_{\omega^{-1,1}} + \beta\|\hat{e}_N\|^2_{\omega^{-2,0}} + \gamma\|(\hat{e}_N)_x\|^2_{\omega^{-1,1}} - 2\gamma\|\hat{e}_N\|^2_{\omega^{-3,0}} + \frac{1}{3}\|(\hat{e}_N)_x\|^2_{\omega^{-2,0}}\\ \leq -\alpha(u - \pi_N u, \hat{e}_N)_{\omega^{-1,1}} + \beta(\partial_x(u - \pi_N u), \hat{e}_N)_{\omega^{-1,1}}\\ - \gamma(\partial_x(u - \pi_N u), \partial_x(\hat{e}_N\omega^{-1,1})).\end{aligned}$$

The right-hand side can be bounded by using Lemma 2.2, the Cauchy–Schwarz inequality, and the fact that $\omega^{-1,2} \leq 2\omega^{-1,1} \leq 2\omega^{-2,0}$:

$$\begin{aligned}(u - \pi_N u, \hat{e}_N)_{\omega^{-1,1}} &\leq \|\hat{e}_N\|_{\omega^{-1,1}}\|u - \pi_N u\|_{\omega^{-1,1}} \leq 2\|\hat{e}_N\|_{\omega^{-1,1}}\|u - \pi_N u\|_{\omega^{-2,-1}},\\ ((u - \pi_N u)_x, \hat{e}_N)_{\omega^{-1,1}} &= (u - \pi_N u, \partial_x\hat{e}_N\omega^{-1,1} + 2\hat{e}_N\omega^{-2,0})\\ &\lesssim \|u - \pi_N u\|_{\omega^{-2,-1}}\|\partial_x\hat{e}_N\|_{\omega^{-2,0}},\\ ((u - \pi_N u)_x, (\hat{e}_N\omega^{-1,1})_x) &= ((u - \pi_N u)_x, (\hat{e}_N)_x\omega^{-1,1} + 2\hat{e}_N\omega^{-2,0})\\ &\leq \|(u - \pi_N u)_x\|_{\omega^{-1,0}}\|(\hat{e}_N)_x\|_{\omega^{-2,0}}.\end{aligned}$$

For $0 \leq \gamma < \frac{1}{6}$, we choose δ sufficiently small such that $\frac{1}{3} - 2\gamma - \delta > 0$. Combining the above inequalities, using the inequality

$$ab \leq \epsilon a^2 + \frac{1}{4\epsilon} b^2 \quad \forall \epsilon > 0, \tag{2.22}$$

and dropping some unnecessary terms, we get

$$\begin{aligned}\frac{\alpha}{2}\|\hat{e}_{\scriptscriptstyle N}\|^2_{\omega^{-1,1}} + \left(\frac{1}{3} - 2\gamma - \delta\right)&\|(\hat{e}_{\scriptscriptstyle N})_x\|^2_{\omega^{-2,0}} \\ &\lesssim \|u - \pi_{\scriptscriptstyle N} u\|^2_{\omega^{-2,-1}} + \gamma\|(u - \pi_{\scriptscriptstyle N} u)_x\|^2_{\omega^{-1,0}} \\ &\lesssim (1 + \gamma N^2) N^{-2m} \|\partial_x^m u\|_{\omega^{m-2,m-1}}.\end{aligned}$$

The last inequality follows from Theorem 2.1.

For $-\frac{1}{3} < \gamma < 0$, we choose δ sufficiently small such that $\frac{1}{3} + \gamma - \delta > 0$, and we derive similarly

$$\frac{\alpha}{2}\|\hat{e}_{\scriptscriptstyle N}\|^2_{\omega^{-1,1}} + \left(\frac{1}{3} + \gamma - \delta\right)\|(\hat{e}_{\scriptscriptstyle N})_x\|^2_{\omega^{-2,0}} \lesssim (1 + |\gamma| N^2) N^{-2m} \|\partial_x^m u\|_{\omega^{m-2,m-1}}.$$

The desired results follow from the triangular inequality, Theorem 2.1, and the fact that $\|u\|_{\omega^{-1,0}} \leq 2\|u\|_{\omega^{-1,0}}$. □

Remark 2.2. Note that the error estimate in the above theorem is optimal for $\gamma = 0$ but suboptimal for $\gamma \neq 0$.

2.4. Linear system and its coefficient matrices. Hence, by setting

$$\begin{aligned}&u_{\scriptscriptstyle N} = \sum_{k=0}^{N-3} \tilde{u}_k \phi_k, \quad \bar{u} = (\tilde{u}_0, \tilde{u}_1, \dots, \tilde{u}_{N-3})^t, \\ &\tilde{f}_k = (f, \psi_k), \quad \bar{f} = (\tilde{f}_0, \tilde{f}_1, \dots, \tilde{f}_{N-3})^t, \\ &m_{ij} = (\phi_j, \psi_i), \quad p_{ij} = -(\phi_j', \psi_i), \quad q_{ij} = (\phi_j', \psi_i'), \quad s_{ij} = (\phi_j', \psi_i''),\end{aligned} \tag{2.23}$$

the linear system (2.4) becomes

$$(\alpha M + \beta P + \gamma Q + S)\bar{u} = \bar{f}, \tag{2.24}$$

where M, P, Q, and S are $(N-2) \times (N-2)$ matrices with entries m_{ij}, p_{ij}, q_{ij}, and s_{ij}, respectively.

Thanks to the orthogonality of the Legendre polynomials, we have $m_{ij} = 0$ for $|i - j| > 3$. Therefore, M is a seven-diagonal matrix. We note that the homogeneous "dual" boundary conditions satisfied by ϕ_j and ψ_i allow us to integrate by parts freely without introducing additional boundary terms; namely, we have

$$s_{ij} = (\phi_j', \psi_i'') = (\phi_j''', \psi_i) = -(\phi_j, \psi_i''').$$

Thanks to the compact form of ϕ_j and ψ_i, we have $s_{ij} = 0$ for $i \neq j$. So S is a diagonal matrix. Similarly, we see that P is a pentadiagonal matrix and Q is a tridiagonal matrix. It is an easy matter to derive that

$$s_{ii} = 2(2i + 3)^2. \tag{2.25}$$

Nonzero elements of M, P, Q can be easily determined from the properties of Legendre polynomials. Hence the linear system (2.24), under the condition of Theorem 2.2, can be easily formed and inverted.

3. Fifth-order equations.

In this section, we shall consider an example of fifth-order equations. We shall follow essentially the same procedures as in the previous section and will omit some repetitive details.

3.1. Dual-Petrov–Galerkin method.

Consider the model fifth-order equation:

$$\begin{aligned} &\alpha u + \beta u_{xxx} - u_{xxxxx} = f, \quad x \in I = (-1,1), \\ &u(\pm 1) = u_x(\pm 1) = u_{xx}(1) = 0, \end{aligned} \tag{3.1}$$

where α and β are given constants. For the sake of simplicity and with the fifth-order KDV equation in mind, we included only zeroth- and third-order linear terms in the equation. Other linear terms as well as nonhomogeneous boundary conditions can be treated as in the previous section.

Similarly to the third-order equation, we define

$$\begin{aligned} W_N &= \{u \in P_N : u(\pm 1) = u_x(\pm 1) = u_{xx}(1) = 0\}, \\ W_N^* &= \{u \in P_N : u(\pm 1) = u_x(\pm 1) = u_{xx}(-1) = 0\}. \end{aligned} \tag{3.2}$$

We also define

$$\begin{aligned} W &= \{u : \ u \in H_0^2(I), \ u_{xx} \in L^2_{\omega^{-2,0}}\}, \\ W^* &= \{u : \ u \in H_0^2(I), \ u_{xx} \in L^2_{\omega^{0,-2}}\}. \end{aligned} \tag{3.3}$$

It is clear that $W_N \subset W$ and $W_N^* \subset W^*$.

We consider the following Legendre dual-Petrov–Galerkin approximation for (3.1): Find $u_N \in W_N$ such that

$$\alpha(u_N, v_N) - \beta(\partial_x^2 u_N, \partial_x v_N) + (\partial_x^2 u_N, \partial_x^3 v_N) = (f, v_N) \quad \forall v_N \in W_N^*. \tag{3.4}$$

Once again, the above dual-Petrov–Galerkin formulation is equivalent to the following weighted spectral-Galerkin approximation: Find $u_N \in W_N$ such that

$$\begin{aligned} &\alpha(u_N, v_N)_{\omega^{-1,1}} - \beta(\partial_x^2 u_N, \omega^{1,-1}\partial_x(v_N \omega^{-1,1}))_{\omega^{-1,1}} \\ &\qquad + (\partial_x^2 u_N, \omega^{1,-1}\partial_x^3(v_N \omega^{-1,1}))_{\omega^{-1,1}} = (f, v_N)_{\omega^{-1,1}} \quad \forall v_N \in W_N. \end{aligned} \tag{3.5}$$

Note in particular that $W \subset V$ and $W^* \subset V^*$. Hence the results proved in the previous section are still valid here.

3.2. Basis functions and projection operators.

The construction of suitable basis functions for W_N and W_N^* follows the general principle (2.6); i.e., we look for

$$\Phi_k = L_k + a_1^{(k)} L_{k+1} + a_2^{(k)} L_{k+2} + a_3^{(k)} L_{k+3} + a_4^{(k)} L_{k+4} + a_5^{(k)} L_{k+5} \tag{3.6}$$

such that $\Phi_k \in W$. Using (1.4) and after some simplifications, we find that $\{a_j^{(k)}\}$ satisfy the following relations:

$$\begin{aligned} a_2^{(k)} + a_4^{(k)} &= -1, \\ (k+2)(k+3)a_2^{(k)} + (k+4)(k+5)a_4^{(k)} &= -k(k+1), \end{aligned} \tag{3.7}$$

and

$$\begin{aligned} a_1^{(k)} + a_3^{(k)} + a_5^{(k)} &= 0, \\ (k+1)(k+2)a_1^{(k)} + (k+3)(k+4)a_3^{(k)} + (k+5)(k+6)a_5^{(k)} &= 0, \\ (k+1)^2(k+2)^2 a_1^{(k)} + (k+3)^2(k+4)^2 a_3^{(k)} + (k+5)^2(k+6)^2 a_5^{(k)} &= g_k, \end{aligned} \tag{3.8}$$

where $g_k = -k^2(k+1)^2 - (k+2)^2(k+3)^2 a_2^{(k)} - (k+4)^2(k+5)^2 a_4^{(k)}$.

One derives immediately from (3.7) that

$$a_2^{(k)} = -\frac{2(2k+5)}{2k+7},\ a_4^{(k)} = \frac{2k+3}{2k+7}. \tag{3.9}$$

We can then determine $a_1^{(k)},\ a_3^{(k)},\ a_5^{(k)}$ by solving the 3×3 system (3.8).

It is easy to verify that

$$\Psi_k = L_k - a_1^{(k)} L_{k+1} + a_2^{(k)} L_{k+2} - a_3^{(k)} L_{k+3} + a_4^{(k)} L_{k+4} - a_5^{(k)} L_{k+5} \in W^*. \tag{3.10}$$

Hence, for $N \geq 5$, we have

$$\begin{aligned} W_N &= \operatorname{span}\{\Phi_0, \Phi_1, \dots, \Phi_{N-5}\};\\ W_N^* &= \operatorname{span}\{\Psi_0, \Psi_1, \dots, \Psi_{N-5}\}. \end{aligned} \tag{3.11}$$

We define

$$L^2_{\omega^{-3,-2}}(I) = L^2_{\omega^{-3,-2}}(I) \cap W,\ L^2_{\omega^{-2,-3}}(I) = L^2_{\omega^{-2,-3}}(I) \cap W^* \tag{3.12}$$

equipped with the norm of $\|\cdot\|_{\omega^{-3,-2}}$ and $\|\cdot\|_{\omega^{-2,-3}}$, respectively. We shall only summarize the results for $L^2_{\omega^{-3,-2}}$ below. The corresponding results for $L^2_{\omega^{-2,-3}}$ are obtained by using the transform $x \to -x$. The proofs are essentially the same as in the previous section.

Lemma 3.1. *Let $\{\Phi_k\}$ be defined as in* (3.6). *Then*

$$\int_I \Phi_k \Phi_j \omega^{-3,-2} dx = 0,\ k \neq j, \tag{3.13}$$

and $\{\Phi_k\}$ forms a complete orthogonal basis in $L^2_{\omega^{-3,-2}}$.

Furthermore, Φ_k satisfies the following Sturm–Liouville equation:

(3.14)
$$B\Phi_k := -(1-x)^3(1+x)^2\partial_x\left\{(1-x)^{-2}(1+x)^{-1}\partial_x\Phi_k(x)\right\} = (k+1)(k+5)\Phi_k(x).$$

Now, let Π_N be the $L^2_{\omega^{-3,-2}}$-orthogonal projector $L^2_{\omega^{-3,-2}} \to W_N$ defined by

$$(u - \Pi_N u, v_N)_{\omega^{-3,-2}} = 0\ \ \forall v_N \in W_N. \tag{3.15}$$

We define

$$H^m_{\omega^{-3,-2}}(I) = \{u \in L^2_{\omega^{-3,-2}}(I) : \partial_x^l u \in L^2_{\omega^{l-3,l-2}}(I),\ 1 \leq l \leq m\}. \tag{3.16}$$

Theorem 3.1.

$$\|\partial_x^l(u - \Pi_N u)\|_{\omega^{l-3,l-2}} \lesssim N^{l-m}\|\partial_x^m u\|_{\omega^{m-3,m-2}}\ \ \forall u \in H^m_{\omega^{-3,-2}},\ 0 \leq l \leq m.$$

Proof. Let us define

$$J_k^{-3,-2}(x) = \frac{1}{2}(k-4)\int_{-1}^x J_{k-1}^{-2,-1}(t)dt,\ k \geq 5. \tag{3.17}$$

One can show that $J_k^{-3,-2}(\pm 1) = \partial_x J_k^{-3,-2}(\pm 1) = \partial_x^2 J_k^{-3,-2}(1) = 0$. Since $J_k^{-2,-1}$ is proportional to ϕ_{k-3}, we then derive from (1.2) and (3.17) that $J_k^{-3,-2}$ must be proportional to Φ_{k-5} so $\{J_k^{-3,-2}\}$ are mutually orthogonal in $L^2_{\omega^{-3,-2}}$. Moreover, one can verify that $J_k^{-3,-2}$ satisfies the relations (2.13)–(2.14). Thus the desired results follow from the same arguments as those in Theorem 2.1. ∎

3.3. Error estimates. We first prove the following generalized Poincaré inequalities.

LEMMA 3.2.

$$\int_I \frac{u^2}{(1-x)^6}dx \leq \frac{4}{25}\int_I \frac{u_x^2}{(1-x)^4}dx \leq \frac{16}{225}\int_I \frac{u_{xx}^2}{(1-x)^2}dx \;\; \forall u \in W_N, \tag{3.18}$$

and

$$\int_I \frac{u_x^2}{(1-x)^4}dx \leq \frac{1}{7}\int_I \frac{u_{xx}^2}{(1-x)^2}dx + 2\int_I \frac{u^2}{(1-x)^6}dx \;\; \forall u \in W_N. \tag{3.19}$$

Proof. The proof is similar to that of Lemma 2.2. Letting $u \in W_N$, for any constant q,

$$\begin{aligned} 0 \leq \int_I \left(\frac{u}{1-x} + qu_x\right)^2 \frac{1}{(1-x)^4}dx &= \int_I \left(\frac{u^2}{(1-x)^6} + q\frac{(u^2)_x}{(1-x)^5} + q^2\frac{u_x^2}{(1-x)^4}\right)dx \\ &= (1-5q)\int_I \frac{u^2}{(1-x)^6}dx + q^2\int_I \frac{u_x^2}{(1-x)^4}dx. \end{aligned}$$

We obtain the first part of (3.18) by taking $q = \frac{2}{5}$. The second part is a direct consequence of Lemma 2.2 since $u_x \in V_N$.

For (3.19), we consider the following relation with any constants q and r:

$$\begin{aligned} 0 &\leq \int_I \left(\frac{u}{(1-x)^2} + q\frac{u_x}{1-x} + ru_{xx}\right)^2 \frac{1}{(1-x)^2}dx \\ &= (1-5q+20r)\int_I \frac{u^2}{(1-x)^6}dx + (q^2-2r-4qr)\int_I \frac{u_x^2}{(1-x)^4}dx + r^2\int_I \frac{u_{xx}^2}{(1-x)^2}dx. \end{aligned}$$

We obtain (3.19) by taking $q = \frac{3}{2}$ and $r = \frac{1}{2}$. □

LEMMA 3.3.

$$\frac{5}{7}\int_I \frac{u_{xx}^2}{(1-x)^2}dx \leq (\partial_x^2 u, \partial_x^3(u\omega^{-1,1})) \leq \frac{203}{15}\int_I \frac{u_{xx}^2}{(1-x)^2}dx \;\; \forall u \in W_N. \tag{3.20}$$

Proof. For any $u \in W_N$, we set $u = \Phi(1-x)$ with $\Phi(\pm 1) = \Phi_x(\pm 1) = 0$. Then, by integrating by parts and using the fact that all boundary terms are zero, we find

$$\begin{aligned} (\partial_x^2 u, \partial_x^3(u\omega^{-1,1})) &= -(\partial_x^3 u, \partial_x^2(u\omega^{-1,1})) = -(\Phi_{xxx}(1-x) - 3\Phi_{xx}, \Phi_{xx}(1+x) + 2\Phi_x) \\ &= \int_I \left\{-\frac{1}{2}(\Phi_{xx}^2)_x(1-x^2) + 3\Phi_{xx}^2(1+x) + 3(\Phi_x^2)_x + 2\Phi_{xx}(\Phi_x(1-x))_x\right\}dx \\ &= 5\int_I \Phi_{xx}^2 dx = 5\int_I \left\{\partial_x^2\left(\frac{u}{1-x}\right)\right\}^2 dx. \end{aligned}$$

Expanding $\partial_x^2(\frac{u}{1-x})$ and integrating by parts, we get

$$\int_I \left\{\partial_x^2\left(\frac{u}{1-x}\right)\right\}^2 dx = \int_I \frac{u_{xx}^2}{(1-x)^2}dx - 6\int_I \frac{u_x^2}{(1-x)^4}dx + 24\int_I \frac{u^2}{(1-x)^6}dx. \tag{3.21}$$

We conclude by applying (3.18) and (3.19) to the above. □

Let Π_N be defined in (3.15). By definition, we have

$$(\partial_x^2(u-\Pi_N u), \partial_x^3 v_N) = (u-\Pi_N u, \omega^{3,2}\partial_x^5 v_N)_{\omega^{-3,-2}} = 0 \quad \forall u \in W,\ v_N \in W_N^*. \tag{3.22}$$

Letting u and u_N be, respectively, the solution of (3.1) and (3.4), we denote $\hat{e}_N = \Pi_N u - u_N$ and $e_N = u - u_N = (u - \Pi_N u) + \hat{e}_N$.

THEOREM 3.2. *For any α, $\beta \geq 0$, there exists a unique solution for the system* (3.4). *Furthermore, for $u \in H^m_{\omega^{-2,-1}}$, we have*

$$\begin{aligned}\alpha\|e_N\|_{\omega^{-1,1}} + \beta N^{-1}\|(e_N)_x\|_{\omega^{-2,0}} + N^{-2}\|(e_N)_{xx}\|_{\omega^{-1,0}}&\\ \lesssim (1+\beta N)N^{-m}\|\partial_x^m u\|_{\omega^{m-3,m-2}},\ m \geq 2.&\end{aligned}$$

Proof. We derive from (3.1), (3.5), and (3.22) that

$$\alpha(e_N, v_N)_{\omega^{-1,1}} - \beta(\partial_x^2 e_N, \partial_x(v_N\omega^{-1,1})) + (\partial_x^2\hat{e}_N, \partial_x^3(v_N\omega^{-1,1})) = 0 \quad \forall v_N \in W_N.$$

Taking $v_N = \hat{e}_N$ in the above and using Lemmas 2.3 and 3.3, we obtain

$$\begin{aligned}\alpha\|\hat{e}_N\|^2_{\omega^{-1,1}} + \frac{\beta}{9}\|(\hat{e}_N)_x\|^2_{\omega^{-2,0}} + \frac{5}{7}\|(\hat{e}_N)_{xx}\|^2_{\omega^{-2,0}}&\\ \leq -\alpha(u-\Pi_N u, \hat{e}_N)_{\omega^{-1,1}} + \beta(\partial_x^2(u-\Pi_N u), \partial_x(\hat{e}_N\omega^{-1,1}))&\\ = -\alpha(u-\Pi_N u, \hat{e}_N)_{\omega^{-1,1}}&\\ - \beta(\partial_x(u-\Pi_N u), (\partial_x^2\hat{e}_N\omega^{-1,1} + 4\partial_x\hat{e}_N\omega^{-2,0} + 4\hat{e}_N\omega^{-3,0})).&\end{aligned}$$

Using the Cauchy–Schwarz inequality, we bound the right-hand side as follows:

$$(u-\Pi_N u, \hat{e}_N)_{\omega^{-1,1}} \leq \|\hat{e}_N\|_{\omega^{-1,1}}\|u-\Pi_N u\|_{\omega^{-1,1}} \lesssim \|\hat{e}_N\|_{\omega^{-1,1}}\|u-\Pi_N u\|_{\omega^{-3,-2}};$$

and thanks to Lemma 2.2,

$$\begin{aligned}((u-\Pi_N u)_x, (\partial_x^2\hat{e}_N\omega^{-1,1} + 4\partial_x\hat{e}_N\omega^{-2,0} + 4\hat{e}_N\omega^{-3,0}))&\\ \lesssim \|(u-\Pi_N u)_x\|_{\omega^{-2,-1}}(\|\partial_x^2\hat{e}_N\|_{\omega^{0,3}} + \|\partial_x\hat{e}_N\|_{\omega^{-2,1}} + \|\hat{e}_N\|_{\omega^{-4,1}})&\\ \lesssim \|(u-\Pi_N u)_x\|_{\omega^{-2,-1}}(\|\partial_x^2\hat{e}_N\|_{\omega^{-2,0}} + \|\partial_x\hat{e}_N\|_{\omega^{-2,0}}).&\end{aligned}$$

Using (2.22) and combining the above inequalities, we arrive at

$$\begin{aligned}\frac{\alpha}{2}\|\hat{e}_N\|^2_{\omega^{-1,1}} + \frac{\beta}{18}\|\partial_x\hat{e}_N\|^2_{\omega^{-2,0}} + \frac{5}{14}\|\partial_x^2\hat{e}_N\|^2_{\omega^{-2,0}}&\\ \lesssim \|u-\Pi_N u\|^2_{\omega^{-3,-2}} + \beta\|(u-\Pi_N u)_x\|^2_{\omega^{-2,-1}}&\\ \lesssim (1+\beta N^2)N^{-2m}\|\partial_x^m u\|_{\omega^{m-3,m-2}}.&\end{aligned}$$

The last inequality follows from Theorem 3.1.

We can then conclude from the above inequality and the triangular inequality. □

Remark 3.1. Similarly as in the previous section, the error estimate in the above theorem is optimal for $\beta = 0$ but suboptimal for $\beta \neq 0$.

3.4. Linear system and its coefficient matrices. Similarly to the third-order equation, we set

$$\begin{aligned}&u_N = \sum_{k=0}^{N-5}\tilde{u}_k\Phi_k,\quad \bar{u} = (\tilde{u}_0, \tilde{u}_1, \dots, \tilde{u}_{N-5})^t,\\ &\tilde{f}_k = (f, \Psi_k),\quad \bar{u} = (\tilde{f}_0, \tilde{f}_1, \dots, \tilde{f}_{N-5})^t,\\ &m_{ij} = (\Phi_j, \Psi_i),\quad M = (m_{ij})_{i,j=0,1,\dots,N-5},\\ &p_{ij} = -(\Phi_j'', \Psi_i'),\quad P = (p_{ij})_{i,j=0,1,\dots,N-5},\\ &s_{ij} = (\Phi_j'', \Psi_i'''),\quad S = (s_{ij})_{i,j=0,1,\dots,N-5}\end{aligned} \tag{3.23}$$

so that the linear system (3.4) becomes

$$(\alpha M + \beta P + S)\bar{u} = \bar{f}. \tag{3.24}$$

Obviously, M is an eleven nonzero diagonal matrix, and with integration by parts, we find that P is a pentadiagonal matrix and S is a diagonal matrix. Repeatedly using (1.2), we can derive that

$$s_{ii} = 2(2i+3)(2i+5)(2i+7)(2i+9)a_5^{(i)}. \tag{3.25}$$

Nonzero elements of M and P can be determined accordingly using (1.2) and (1.3). Hence the linear system (3.24), under the condition of Theorem 3.1, can also be easily inverted.

4. Application to the KDV equation. There is a vast body of literature on various aspects of the KDV equation. Although most of these studies are concerned with initial value problems, the initial-boundary value problems also received considerable attention. The most natural initial-boundary value KDV equation is set in a quarter-plane (see, for instance, [28, 16, 3, 2, 12, 13, 5] and the references therein). The KDV equation on a finite spatial interval has also been considered by several authors [20, 9, 4]. Here, as an example of application to nonlinear equations, we consider the KDV equation on a finite interval:

$$\begin{aligned} &\alpha v_t + \beta v_x + \gamma v v_x + v_{xxx} = 0,\ x \in (-1,1),\ t \in (0,T], \\ &v(-1,t) = g(t),\ v(1,t) = v_x(1,t) = 0,\ \ t \in [0,T], \\ &v(x,0) = v_0(x),\ x \in (-1,1). \end{aligned} \tag{4.1}$$

The positive constants α, β, and γ are introduced to accommodate the scaling of the spatial interval. The existence and uniqueness of the solution for (4.1) can be established as in [9, 4]. Beside its own interests, (4.1) can also be viewed as a legitimate approximate model for the KDV equation on a quarter-plane before the wave reaches the right boundary.

Let us first reformulate (4.1) as an equivalent problem with homogeneous boundary conditions. To this end, let $\hat{v}(x,t) = \frac{(1-x)^2}{4}g(t)$, and write $v(x,t) = u(x,t) + \hat{v}(x,t)$. Then u satisfies the following equation with homogeneous boundary conditions:

$$\begin{aligned} &\alpha u_t + a(x,t)u + b(x,t)u_x + \gamma u u_x + u_{xxx} = f,\ x \in (-1,1),\ t \in (0,T], \\ &u(\pm 1,t) = u_x(1,t) = 0;\ t \in [0,T], \\ &u(x,0) = u_0(x) = v_0(x) - \hat{v}(x,0),\ x \in (-1,1), \end{aligned} \tag{4.2}$$

where $a(x,t) = \frac{\gamma}{2}(x-1)g(t)$, $b(x,t) = \beta + \gamma\hat{v}(x,t)$, and $f(x,t) = -\alpha\hat{v}_t(x,t)$.

For a given Δt, we set $t_k = k\Delta t$ and let $u_{\scriptscriptstyle N}^0 = \pi_{\scriptscriptstyle N} u_0$ and $u_{\scriptscriptstyle N}^1 \in V_N$ be an appropriate approximation of $u(\cdot,t_1)$, for instance; we can compute $u_{\scriptscriptstyle N}^1$ using one step of a semi-implicit first-order scheme so that for $u \in C^3(0,T; L^2_{\omega^{2,2}}(I)) \cap C(0,T; B^m_{\omega^{-2,-1}})$, we have

$$\|u_{\scriptscriptstyle N}^1 - \pi_{\scriptscriptstyle N} u(\cdot,t_1)\|_{\omega^{-1,1}} \lesssim \Delta t^2 + N^{-m}. \tag{4.3}$$

Let M be such that $|u(x,t)| \le M$ for $x \in [-1,1]$ and $t \in [0,T]$. We define a cut-off function

$$H(x) = \begin{cases} x, & |x| \le 2M, \\ 2M, & x > 2M, \\ -2M, & x < -2M. \end{cases} \tag{4.4}$$

It is easy to verify that

$$|H(x)-H(y)| \le |x-y| \ \ \forall x,\ y. \tag{4.5}$$

We consider first a modified Crank–Nicolson leap-frog dual-Petrov–Galerkin approximation:

$$\begin{aligned} &\frac{\alpha}{2\Delta t}(u_N^{k+1}-u_N^{k-1}, v_N)_{\omega^{-1,1}} + \frac{1}{2}(\partial_x(u_N^{k+1}+u_N^{k-1}), \partial_x^2(v_N\omega^{-1,1})) \\ &\quad = (f(\cdot,t_k), v_N)_{\omega^{-1,1}} + \frac{\gamma}{2}(H(u_N^k)u_N^k, \partial_x(v_N\,\omega^{-1,1})) \\ &\quad - (a\,u_N^k, v_N)_{\omega^{-1,1}} + (u_N^k, \partial_x(bv_N\,\omega^{-1,1})) \ \ \forall v_N \in V_N. \end{aligned} \tag{4.6}$$

We denote $\hat{e}_N^k = \pi_N u(\cdot,t_k) - u_N^k$, $\tilde{e}_N^k = u(\cdot,t_k) - \pi_N u(\cdot,t_k)$, and $e_N^k = u(\cdot,t_k) - u_N^k$.

THEOREM 4.1. *We assume $u \in C^3(0,T; L^2_{\omega^{2,2}}(I)) \cap C(0,T; B^m_{\omega^{-2,-1}})$ with $m > 1$. Then the scheme* (4.6) *is unconditionally stable, and the following error estimates hold:*

$$\|e_N^{n+1}\|_{\omega^{-1,1}} \lesssim \Delta t^2 + N^{-m}, \ \ 0 \le n \le [T/dt]-1,$$

$$\left(\Delta t \sum_{k=1}^{n} \|\partial_x(e_N^{k+1}+e_N^{k-1})\|^2_{\omega^{-1,0}}\right)^{\frac{1}{2}} \lesssim \Delta t^2 + N^{1-m}, \ \ 1 \le n \le [T/dt]-1.$$

Proof. Let E^k $(k=1,2,\dots)$ be the truncation error defined by

$$\begin{aligned} &\frac{\alpha}{2\Delta t}(u(\cdot,t_{k+1}) - u(\cdot,t_{k-1})) + a(\cdot,t_k)\,u(\cdot,t_k) + b(\cdot,t_k)\,u_x(\cdot,t_k) + \gamma u(\cdot,t_k)\partial_x u(\cdot,t_k) \\ &\qquad + \frac{1}{2}\partial_x^3(u(\cdot,t_{k+1}) + u(\cdot,t_{k-1})) - f(\cdot,t_k) = E^k(\cdot). \end{aligned} \tag{4.7}$$

Comparing (4.6) with (4.7) and using Lemma 2.4, we have

$$\begin{aligned} &\frac{\alpha}{2\Delta t}(\hat{e}_N^{k+1}-\hat{e}_N^{k-1}, v_N)_{\omega^{-1,1}} + \frac{1}{2}(\partial_x(\hat{e}_N^{k+1}+\hat{e}_N^{k-1}), \partial_x^2(v_N\omega^{-1,1})) \\ &\quad = \frac{\gamma}{2}(u(\cdot,t_k)^2 - H(u_N^k)u_N^k, \partial_x(v_N\,\omega^{-1,1})) \\ &\quad - (a\,(\hat{e}_N^k+\tilde{e}_N^k), v_N)_{\omega^{-1,1}} + \frac{1}{2}((\hat{e}_N^k+\tilde{e}_N^k), \partial_x(bv_N\,\omega^{-1,1})) \\ &\quad + (E^k, v_N)_{\omega^{-1,1}} - \frac{\alpha}{2\Delta t}(\tilde{e}_N^{k+1}-\tilde{e}_N^{k-1}, v_N)_{\omega^{-1,1}} \ \ \forall v_N \in V_N. \end{aligned} \tag{4.8}$$

Let $A = \max_{x\in[-1,1],\,t\in[0,T]}|a(x,t)|$ and $B = \max_{x\in[-1,1],\,t\in[0,T]}(|b(x,t)| + |\partial_x b(x,t)|)$. We now take $v_N = 2\Delta t(\hat{e}_N^{k+1}+\hat{e}_N^{k-1})$ in (4.8) and bound the right-hand side terms using repeatedly the Cauchy–Schwarz inequality and Lemma 2.2 as follows:

$$-2\Delta t(a\,(\hat{e}_N^k+\tilde{e}_N^k), \hat{e}_N^{k+1}+\hat{e}_N^{k-1})_{\omega^{-1,1}} \le \Delta t A(\|\tilde{e}_N^k\|^2_{\omega^{-1,1}} + \|\hat{e}_N^k\|^2_{\omega^{-1,1}} + \|\hat{e}_N^{k+1}+\hat{e}_N^{k-1}\|^2_{\omega^{-1,1}}),$$

$$\begin{aligned} 2\Delta t(E^k, \hat{e}_N^{k+1}+\hat{e}_N^{k-1})_{\omega^{-1,1}} &\le 2\Delta t\|E^k\|_{\omega^{2,2}}\|\hat{e}_N^{k+1}+\hat{e}_N^{k-1}\|_{\omega^{-4,0}} \\ &\le c\Delta t\|E^k\|^2_{\omega^{2,2}} + \frac{\Delta t}{36}\|\partial_x(\hat{e}_N^{k+1}+\hat{e}_N^{k-1})\|^2_{\omega^{-2,0}}. \end{aligned}$$

Let us denote $\tilde{E}_N^k = \frac{1}{2\Delta t}(\tilde{e}_N^{k+1}-\tilde{e}_N^{k-1})$. Similarly as above, we have

$$-\frac{\alpha}{2\Delta t}(\tilde{e}_N^{k+1}-\tilde{e}_N^{k-1}, \hat{e}_N^{k+1}+\hat{e}_N^{k-1})_{\omega^{-1,1}} \le c\Delta t\|\tilde{E}_N^k\|^2_{\omega^{2,2}} + \frac{\Delta t}{36}\|\partial_x(\hat{e}_N^{k+1}+\hat{e}_N^{k-1})\|^2_{\omega^{-2,0}}.$$

By the assumption $|u(x,t)| \le M$ for all x and t, we have

$$\begin{aligned}|u(\cdot,t_k)^2 - H(u_{_N}^k)u_{_N}^k| &= |u(\cdot,t_k)(H(u(\cdot,t_k)) - H(u_{_N}^k)) + H(u_{_N}^k)(u(\cdot,t_k) - u_{_N}^k)|\\ &\le 2M|u(\cdot,t_k) - u_{_N}^k| \le 2M(|\tilde{e}_{_N}^k| + |\hat{e}_{_N}^k|).\end{aligned}$$

Hence,

$$\begin{aligned}&\gamma\Delta t(u(\cdot,t_k)^2 - H(u_{_N}^k)u_{_N}^k, \partial_x(\hat{e}_{_N}^{k+1} + \hat{e}_{_N}^{k-1})\,\omega^{-1,1})\\ &\qquad\le \gamma\Delta t\|u(\cdot,t_k)^2 - H(u_{_N}^k)u_{_N}^k\|_{\omega^{0,2}}\|\partial_x(\hat{e}_{_N}^{k+1} + \hat{e}_{_N}^{k-1})\|_{\omega^{-2,0}}\\ &\qquad\le c\Delta t(\|\tilde{e}_{_N}^k\|^2_{\omega^{-2,-1}} + \|\hat{e}_{_N}^k\|^2_{\omega^{-1,1}}) + \frac{\Delta t}{36}\|\partial_x(\hat{e}_{_N}^{k+1} + \hat{e}_{_N}^{k-1})\|^2_{\omega^{-2,0}},\end{aligned}$$

To handle the remaining terms, we recall the following Hardy inequalities:

$$\int_a^b \left[\frac{1}{t-a}\int_a^t \psi(s)ds\right]^2 (t-a)^\alpha dt \le \frac{4}{1-\alpha}\int_a^b \psi^2(t)(t-a)^\alpha dt,$$

$$\int_a^b \left[\frac{1}{b-t}\int_t^b \psi(s)ds\right]^2 (b-t)^\alpha dt \le \frac{4}{1-\alpha}\int_a^b \psi^2(t)(b-t)^\alpha dt, \tag{4.9}$$

which hold for all measurable functions ϕ on (a,b) with $a<b$ and $\alpha < -1$.

Let $(a,b) = (-1,0)$, $\alpha = 0$, and $\phi(t) = \int_{-1}^t \psi(s)ds$. We find

$$\int_{-1}^0 \frac{1}{(1-t)^3(1+t)}\phi^2(t)dt \le \int_{-1}^0 \frac{1}{(t+1)^2}\phi^2(t)dt \le 4\int_{-1}^0 (\phi_t)^2 dt.$$

Let $(a,b) = (0,1)$, $\alpha = -1$, and $\phi(t) = \int_t^1 \psi(s)ds$. We find

$$\int_0^1 \frac{1}{(1-t)^3(1+t)}\phi^2(t)dt \le \int_0^1 \frac{1}{(1-t)^3}\phi^2(t)dt \le 2\int_0^1 (\phi_t)^2\frac{1}{1-t}dt.$$

Combining the above two inequalities, we obtain

$$\int_I \phi^2\omega^{-3,-1}dx \le 4\int_I (\phi_x)^2\omega^{-1,0}dx, \tag{4.10}$$

which holds for all ϕ such that $\phi(\pm1) = 0$ and $\int_I (\phi_x)^2\omega^{-1,0}dx < \infty$.

Thanks to Lemma 2.2 and (4.10),

$$\begin{aligned}&\gamma\Delta t(u(\cdot,t_k)^2 - H(u_{_N}^k)u_{_N}^k, (\hat{e}_{_N}^{k+1} + \hat{e}_{_N}^{k-1})\,\partial_x\omega^{-1,1})\\ &\qquad\le 2\gamma\Delta t\|u(\cdot,t_k)^2 - H(u_{_N}^k)u_{_N}^k\|_{\omega^{-1,1}}\|(\hat{e}_{_N}^{k+1} + \hat{e}_{_N}^{k-1})\|_{\omega^{-3,-1}}\\ &\qquad\le c\Delta t(\|\tilde{e}_{_N}^k\|^2_{\omega^{-2,-1}} + \|\hat{e}_{_N}^k\|^2_{\omega^{-1,1}}) + \frac{\Delta t}{36}\|\partial_x(\hat{e}_{_N}^{k+1} + \hat{e}_{_N}^{k-1})\|^2_{\omega^{-2,0}}.\end{aligned}$$

Similarly, we have

$$\begin{aligned}&\Delta t((\hat{e}_{_N}^k + \tilde{e}_{_N}^k), \partial_x(b(\hat{e}_{_N}^{k+1} + \hat{e}_{_N}^{k-1})\,\omega^{-1,1})) = \Delta t((\hat{e}_{_N}^k + \tilde{e}_{_N}^k), b\,\partial_x(\hat{e}_{_N}^{k+1} + \hat{e}_{_N}^{k-1})\,\omega^{-1,1})\\ &\qquad + \Delta t((\hat{e}_{_N}^k + \tilde{e}_{_N}^k), b\,(\hat{e}_{_N}^{k+1} + \hat{e}_{_N}^{k-1})\,\partial_x\omega^{-1,1}) + \Delta t(\hat{e}_{_N}^k + \tilde{e}_{_N}^k, \partial_x b\,(\hat{e}_{_N}^{k+1} + \hat{e}_{_N}^{k-1})\,\omega^{-1,1})\\ &\qquad\le cB\Delta t(\|\tilde{e}_{_N}^k\|^2_{\omega^{-2,-1}} + \|\hat{e}_{_N}^k\|^2_{\omega^{-1,1}}) + \frac{\Delta t}{36}\|\partial_x(\hat{e}_{_N}^{k+1} + \hat{e}_{_N}^{k-1})\|^2_{\omega^{-2,0}}.\end{aligned}$$

Combining the above inequalities into (4.8) and using Lemma 2.3, we obtain

$$\begin{aligned}&\alpha(\|\hat{e}_{_N}^{k+1}\|^2_{\omega^{-1,1}} - \|\hat{e}_{_N}^{k-1}\|^2_{\omega^{-1,1}}) + \frac{\Delta t}{36}\|\partial_x(\hat{e}_{_N}^{k+1} + \hat{e}_{_N}^{k-1})\|^2_{\omega^{-2,0}}\\ &\quad\le c\Delta t(\|E^k\|^2_{\omega^{2,2}} + \|\tilde{E}_N^k\|^2_{\omega^{2,2}} + \|\hat{e}_{_N}^{k+1}\|^2_{\omega^{-1,1}} + \|\hat{e}_{_N}^k\|^2_{\omega^{-1,1}} + \|\hat{e}_{_N}^{k-1}\|^2_{\omega^{-1,1}} + \|\tilde{e}_{_N}^k\|^2_{\omega^{-2,-1}}).\end{aligned}$$

We can then apply the standard discrete Gronwall lemma to the above inequality to get

$$\begin{aligned}\|\hat{e}_{_N}^{n+1}\|^2_{\omega^{-1,1}} + \frac{\Delta t}{36}\sum_{k=1}^{n}\|\partial_x(\hat{e}_{_N}^{k+1}+\hat{e}_{_N}^{k-1})\|^2_{\omega^{-2,0}} &\lesssim \|\hat{e}_{_N}^{0}\|^2_{\omega^{-1,1}} + \|\hat{e}_{_N}^{1}\|^2_{\omega^{-1,1}} \\ &+ \Delta t\sum_{k=1}^{n}(\|\tilde{e}_{_N}^{k}\|^2_{\omega^{-2,-1}} + \|E^k\|^2_{\omega^{2,2}} + \|\tilde{E}_N^k\|^2_{\omega^{2,2}}),\ \ 1\le n\le [T/\Delta t]-1.\end{aligned}$$

Using the triangular inequality, we derive that

$$\begin{aligned}\|e_{_N}^{n+1}\|^2_{\omega^{-1,1}} &\lesssim \|\hat{e}_{_N}^{0}\|^2_{\omega^{-1,1}} + \|\hat{e}_{_N}^{1}\|^2_{\omega^{-1,1}} + \|\tilde{e}_{_N}^{n+1}\|^2_{\omega^{-1,1}} \\ &+ \Delta t\sum_{k=1}^{n}(\|\tilde{e}_{_N}^{k}\|^2_{\omega^{-2,-1}} + \|E^k\|^2_{\omega^{2,2}} + \|\tilde{E}_N^k\|^2_{\omega^{2,2}})\end{aligned}$$

and

$$\begin{aligned}\frac{\Delta t}{36}\sum_{k=1}^{n}\|\partial_x(e_{_N}^{k+1}+e_{_N}^{k-1})\|^2_{\omega^{-1,0}} &\lesssim \|\hat{e}_{_N}^{0}\|^2_{\omega^{-1,1}} + \|\hat{e}_{_N}^{1}\|^2_{\omega^{-1,1}} \\ &+ \Delta t\sum_{k=1}^{n}(\|\tilde{e}_{_N}^{k}\|^2_{\omega^{-2,-1}} + \|E^k\|^2_{\omega^{2,2}} + \|\tilde{E}_N^k\|^2_{\omega^{2,2}} + \|\partial_x(\tilde{e}_{_N}^{k+1}+\tilde{e}_{_N}^{k-1})\|^2_{\omega^{-1,0}}).\end{aligned}$$

We can then conclude from the assumptions and Theorem 2.1. □

Next, we consider the standard Crank–Nicolson leap-frog weighted Galerkin approximation:

(4.11)
$$\begin{aligned}&\frac{\alpha}{2\Delta t}(u_{_N}^{k+1}-u_{_N}^{k-1},v_{_N})_{\omega^{-1,1}} + \frac12(\partial_x(u_{_N}^{k+1}+u_{_N}^{k-1}),\partial_x^2(v_{_N}\omega^{-1,1})) \\ &\quad= -(a\,u_{_N}^{k},v_{_N})_{\omega^{-1,1}} + (u_{_N}^{k},\partial_x(bv_{_N}\omega^{-1,1})) \\ &\quad+ (f(\cdot,t_k),v_{_N})_{\omega^{-1,1}} + \frac{\gamma}{2}((u_{_N}^{k})^2,\partial_x v_{_N}\,\omega^{-1,1} + v_{_N}\partial_x\omega^{-1,1})\ \ \forall v_{_N}\in V_N.\end{aligned}$$

Corollary 4.1. *Under the conditions of Theorem* 4.1*, there exists* c_0 *such that for* $\Delta t\,N\le c_0$*, the two schemes* (4.6) *and* (4.11) *are equivalent.*

Proof. We need only to show that the scheme (4.6) reduces to (4.11) under the condition that $\Delta t\,N\le c_0$. Indeed, using the estimate in Theorem 4.1, the inverse inequality $\|u\|_{L^\infty}\lesssim N^2\|u\|_{\omega^{0,1}}$ for all $u\in P_N$ (see Lemma 4.1 below), and the assumptions on u, we find that there exists c_0 such that for $\Delta t\,N\le c_0$ we have

$$\begin{aligned}\|u_{_N}^k\|_{L^\infty} &\le \|u(\cdot,t_k)\|_{L^\infty} + \|\tilde{e}_{_N}^k\|_{L^\infty} + \|\hat{e}_{_N}^k\|_{L^\infty} \\ &\le M + \|\tilde{e}_{_N}^k\|_{L^\infty} + N^2\|\hat{e}_{_N}^k\|_{\omega^{0,1}} \\ &\le M + \|\tilde{e}_{_N}^k\|_{L^\infty} + cN^2(\Delta t^2 + N^{-m}) \le 2M.\end{aligned}$$

Hence (4.6) and (4.11) are equivalent. □

The following lemma is a special case of Theorem 2.1 in [15]. For the reader's convenience, we provide an elementary proof below.

Lemma 4.1.

$$\|u\|_{L^\infty}\lesssim N^2\|u\|_{\omega^{0,1}}\ \ \forall u\in P_N.$$

Proof. Let $J_k^{0,1}$ be the kth degree Jacobi polynomial of index $(0,1)$. We recall that (cf. [27])

$$\|J_k^{0,1}\|_{L^\infty} = k+1, \;\; \|J_k^{0,1}\|_{\omega^{0,1}} = \sqrt{\frac{2}{k+1}}. \tag{4.12}$$

For any $u \in P_N$, we write $u(x) = \sum_{k=0}^N u_k J_k^{0,1}(x)$. Then

$$\begin{aligned}\|u\|_{L^\infty}^2 &\le (N+1)^2 \left(\sum_{k=0}^N |u_k|\right)^2 \le (N+1)^3 \sum_{k=0}^N |u_k|^2 \\ &\le \frac{1}{2}(N+1)^4 \sum_{k=0}^N |u_k|^2 \frac{2}{k+1} = \frac{1}{2}(N+1)^4 \|u\|_{\omega^{0,1}}^2. \qquad \square\end{aligned}$$

Remark 4.1. In practice, the nonlinear term is usually computed using the pseudospectral approach (cf. [11, 7]), which is discussed in the next section. It is not difficult to show that the results in Theorem 4.1 apply to the pseudospectral scheme (see [20] for a similar result).

Remark 4.2. We note that the result obtained here for the third-order KDV equation (4.2) can be extended to the fifth-order KDV equation, which has also attracted considerable attention (see, for instance, [18]).

5. Miscellaneous issues. We discuss in this section several extensions and practical issues related to the dual-Petrov–Galerkin method. We note, in particular, that the dual-Petrov–Galerkin method can be used with any spatial discretization method based on a variational formulation such as the finite-element method.

5.1. Other higher odd-order equations. We have discussed in detail the Legendre dual-Petrov–Galerkin method for third- and fifth-order equations. It is evident that the method can be directly applied to other higher odd-order equations of the form

$$\sum_{j=0}^{2m} a_j u^{(j)}(x) + u^{(2m+1)}(x) = 0, \;\; m \ge 3, \tag{5.1}$$

with the boundary conditions

$$u(\pm 1) = u'(\pm 1) = \cdots = u^{(m-1)}(\pm 1) = 0, \;\; u^{(m)}(1) = 0. \tag{5.2}$$

Other boundary conditions and/or variable coefficients can be treated following the discussion below.

5.2. Other boundary conditions. It must be noted that our dual-Petrov–Galerkin approach is quite flexible and can be used for other unconventional boundary conditions. For instance, Colin and Ghidaglia [9] studied the KDV equation

$$u_t + \frac{2}{L}(u_x + uu_x) + \frac{8}{L^3} u_{xxx} = 0, \;\; x \in (-1,1), \;\; t > 0, \tag{5.3}$$

with the boundary conditions

$$u(-1) = g(t), \;\; u_x(1) = u_{xx}(1) = 0. \tag{5.4}$$

Note that we have scaled the interval from $(0, L)$, which was used in [9], to $(-1, 1)$.

Let us denote

$$X_N = \{u \in P_N : u(-1) = 0,\ u_x(1) = u_{xx}(1) = 0\}. \tag{5.5}$$

Then the "dual" space is

$$X_N^* = \{v \in P_N : v_x(-1) = v(-1) = 0,\ v_{xx}(1) = 0\}. \tag{5.6}$$

There exist unique coefficients $\{a_j^{(k)},\ \tilde{a}_j^{(k)}\}$ such that

$$\begin{aligned} \phi_k &= L_k + \sum_{j=1}^{3} a_j^{(k)} L_{k+j} \in X_N,\ k = 0, 1, \dots, N-3, \\ \psi_k &= L_k + \sum_{j=1}^{3} \tilde{a}_j^{(k)} L_{k+j} \in X_N^*,\ k = 0, 1, \dots, N-3. \end{aligned} \tag{5.7}$$

Then, the Legendre dual-Petrov–Galerkin method for (5.3)–(5.4) is to find $u_{_N} = v_{_N} + \frac{(1-x)^3}{8} g(t)$ with $v_{_N} \in X_N$ such that

$$(\partial_t u_{_N}, \psi_j) + \frac{2}{L}(\partial_x u_{_N} + u_{_N} \partial_x u_{_N}, \psi_j) + (\partial_x u_{_N}, \partial_x^2 \psi_j) = 0,\ j = 0, 1, \dots, N-3. \tag{5.8}$$

One can prove results which are similar to Theorem 4.1 for this problem.

5.3. Problems with variable coefficients: Pseudospectral method in modal basis.

Let us consider, as an example, the following third-order equation:

$$\begin{aligned} &a(x)u - b(x)u_x + u_{xxx} = f,\quad x \in I = (-1, 1), \\ &u(\pm 1) = u_x(1) = 0. \end{aligned} \tag{5.9}$$

The pseudospectral dual-Petrov–Galerkin method for (5.9) is to find $u_{_N} \in V_N$ such that

$$(a(x)u_{_N}, v_{_N})_N - (b(x)u'_{_N}, v_{_N})_N + (u'_{_N}, v''_{_N})_N = (f, v_{_N})_N \quad \forall v_{_N} \in V_N^*, \tag{5.10}$$

where

$$(u, v)_N = \sum_{k=0}^{N} u(x_k) v(x_k) \omega_k \tag{5.11}$$

is the discrete inner product of u and v associated with the Legendre–Gauss–Lobatto quadrature (cf. [7]). We recall that

$$(u, v)_N = (u, v) \quad \forall uv \in P_{2N-1}. \tag{5.12}$$

Let us denote $\tilde{\psi}_i = \frac{1}{(\phi'_i, \psi''_i)} \psi_i = \frac{1}{2(2i+3)^2} \psi_i$. Then we have

$$(\phi'_j, \tilde{\psi}''_i)_N = (\phi'_j, \tilde{\psi}''_i) = \delta_{ij},\quad 0 \le i,\ j \le N-3.$$

Hence, by setting

$$
\begin{aligned}
&u_N = \sum_{k=0}^{N-3} \tilde{u}_k \phi_k, \quad \bar{u} = (\tilde{u}_0, \tilde{u}_1, \dots, \tilde{u}_{N-3})^t, \\
&\tilde{f}_k = (f, \tilde{\psi}_k)_N, \quad \bar{f} = (\tilde{f}_0, \tilde{f}_1, \dots, \tilde{f}_{N-3})^t, \\
&m_{ij} = (a(x)\phi_j, \tilde{\psi}_i)_N, \quad p_{ij} = -(b(x)\phi'_j, \tilde{\psi}_i)_N,
\end{aligned} \tag{5.13}
$$

the linear system (5.10) becomes

$$
(M + P + I)\bar{u} = \bar{f}. \tag{5.14}
$$

It is clear that the matrices M and P are full and their formation involves N^3 operations as well as the inversion of (5.14). Hence a direct approach is advisable only if one uses a small or moderate number of modes. Otherwise, an iterative method can be efficiently implemented as follows:

- Note that a conjugate gradient type iterative method does not require the explicit formation of the matrix; only the action of the matrix upon a given vector is needed at each iteration. Although the formation of M and P involves N^3 operations, their action on a given vector $\bar{u}$, i.e., $M\bar{u}$ and $P\bar{u}$, can be computed in $O(N^2)$ operations.
- The number of operations can be further reduced to a quasi-optimal $O(N\log N)$ if we use the following Chebyshev–Legendre dual-Petrov–Galerkin method (cf. [10, 24]): Find $u_N \in V_N$ such that

  $$(I_N^c(a(x)u_N), v_N)_N - (I_N^c(b(x)u'_N), v_N)_N + (u'_N, v''_N)_N = (f, v_N)_N \;\; \forall v_N \in V_N^*,$$

  where I_N^c is the interpolation operator based on the Chebyshev–Gauss–Lobatto points, while $(\cdot,\cdot)_N$ is still the discrete inner product of u and v associated with the Legendre–Gauss–Lobatto quadrature. Hence the only difference between (5.3) and (5.10) is that $a(x)u_N$ and $b(x)u'_N$ in (5.10) are replaced by $I_N^c(a(x)u_N)$ and $I_N^c(b(x)u'_N)$. Thanks to the fast Fourier transform (FFT) and the fast Chebyshev–Legendre transform [1, 24], the Legendre coefficients of $I_N^c(a(x)u_N)$ and $I_N^c(b(x)u'_N)$ can be computed in $O(N \log N)$ operations given the Legendre coefficients of u_N (see [24] for details).
- Under reasonable assumptions on $a(x)$ and $b(x)$, the matrix $M+P+I$ is well conditioned; i.e, its condition number is independent of N. We now provide a heuristic argument for this statement.
  Since $\phi_k\omega^{-1,1} \in V^*$, there exists unique $\{h_{kj}\}$ such that

  $$\phi_k \omega^{-1,1} = \sum_{j=0}^{N-3} h_{kj}\psi_j, \quad k = 0, 1, \dots, N-3.$$

  Hence we have

  $$
  \begin{aligned}
  \langle H\bar{u}, \bar{u}\rangle &= \left(\sum_{j=0}^{N-3} \tilde{u}_j \phi'_j, \sum_{k,j=0}^{N-3} \tilde{u}_j h_{kj} \psi''_j \right)_N \\
  &= \left(\sum_{j=0}^{N-3} \tilde{u}_j \phi'_j, \sum_{j=0}^{N-3} \tilde{u}_j (\phi_j \omega^{-1,1})'' \right) = (\partial_x u_N, \partial_x^2 (u_N \omega^{-1,1}))
  \end{aligned}
  $$

and

$$\langle HM\bar{u},\bar{u}\rangle = \left(a(x)\sum_{j=0}^{N-3}\tilde{u}_j\phi_j, \sum_{k,j=0}^{N-3}\tilde{u}_j h_{kj}\psi_j\right)_N$$
$$= \left(a(x)\sum_{j=0}^{N-3}\tilde{u}_j\phi_j, \sum_{j=0}^{N-3}\tilde{u}_j\phi_j\omega^{-1,1}\right)_N = (a(x)u_{_N}, u_{_N}\omega^{-1,1})_N,$$

where $\langle \bar{v},\bar{v}\rangle := \sum_{j=0}^{N-3} v_j^2$ is the inner product in l^2.

Let us assume that $0 \le a(x) \le a_1$. Then, thanks to (2.2) and (2.3), we derive from the above that there exists a constant a_2 independent of N such that

$$\langle H\bar{u},\bar{u}\rangle \;\le\; \langle H(M+I)\bar{u},\bar{u}\rangle \;\le\; a_2\langle H\bar{u},\bar{u}\rangle.$$

Hence the condition number of $M+I$, in the norm $\|\bar{v}\|_H := \langle H\bar{v},\bar{v}\rangle^{\frac{1}{2}}$, is independent of N. Under assumptions similar to those in Theorem 2.2, one can also establish that the condition number of $M+P+I$ is independent of N. This statement is confirmed by our numerical results (see the next section).

Therefore, a conjugate gradient type iterative method like BICGSTAB or CGS for (5.14) will converge in a small and fixed number (i.e., independent of N) of steps. In short, the Chebyshev–Legendre dual-Petrov–Galerkin method for (5.9) can be solved in a quasi-optimal $O(N\log N)$ operation.

Since the unknowns are coefficients of the spectral expansion, instead of the nodal values of the approximate solution at the collocation points, we refer to the above as the pseudospectral dual-Petrov–Galerkin method in *modal* basis. The modal basis presents at least three distinct advantages compared with the nodal basis:

- As demonstrated in sections 2 and 3 (see also [22, 23, 24]), for problems with constant coefficients, using an appropriate modal basis leads to sparse matrices.
- With the nodal basis, the choice of quadrature rules/collocation points plays an important role and should be made in accordance with the underlying differential equations and boundary conditions (see, for instance, [17] and the references therein). For example, the Gauss–Lobatto points are not suitable for third-order equations (cf. [21]). With the modal basis, since the use of the quadrature rule is merely to approximate the integrals in the variational formulation, the choice of quadrature rules/collocation points is not important. Therefore, for the third-order equation (5.9), we can still use the usual Gauss–Lobatto quadrature.
- Most importantly, using an appropriate modal basis leads to well-conditioned matrices as we explained above.

6. Numerical results.

We present in this section some numerical results illustrating the nice properties of our dual-Petrov–Galerkin method.

6.1. Third- and fifth-order linear equations.

Let us first look at the conditioning of our dual-Petrov–Galerkin method. We list in Table 6.1 the condition numbers of $M+P+I$ in (5.14) with various $a(x)$ and $b(x)$. Notice that in all cases, the condition numbers are small and, more importantly, independent of N.

We list in Table 6.2 the condition numbers of $\alpha M+\beta P+S$ in (3.24) scaled by the diagonal matrix S with various α and β. Once again, the condition numbers are small and independent of N.

Table 6.1
Condition numbers of (5.14).

	$a(x) = 1$	$a(x) = 10$	$a(x) = 50$	$a(x) = \sin x$	$a(x) = 10\ \exp(x)$
N	$b(x) = 0$	$b(x) = 0$	$b(x) = 0$	$b(x) = 2x - 1$	$b(x) = \cos x$
16	1.119	2.218	7.219	1.188	2.393
64	1.119	2.218	7.219	1.188	2.393
128	1.119	2.218	7.219	1.188	2.393

Table 6.2
Condition numbers of (3.24) *scaled by* S.

	$\alpha = 1$	$\alpha = 100$	$\alpha = 100$	$\alpha = 1$	$\alpha = 0$
N	$\beta = 0$	$\beta = 0$	$\beta = 100$	$\beta = -100$	$\beta = 1$
16	1.006	2.421	2.005	3.342	1.009
64	1.006	2.421	2.005	3.342	1.009
128	1.006	2.421	2.005	3.342	1.009

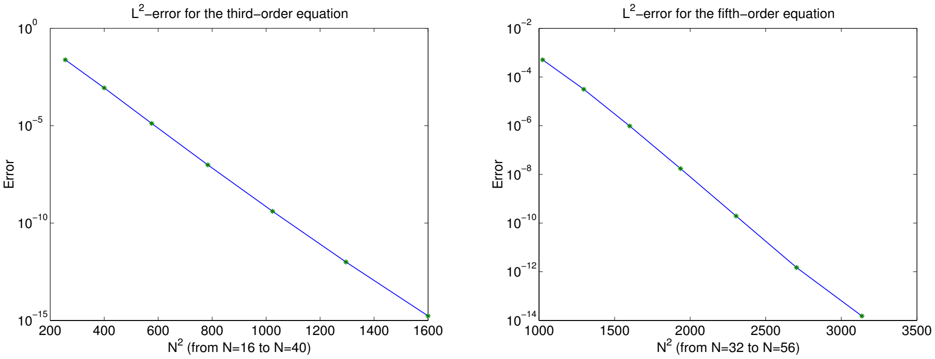


Fig. 6.1. *L^2-errors for the third- and fifth-order equations.*

Next, we look at the accuracy of our dual-Petrov–Galerkin method. We take $a(x) = \sin x$ and $b(x) = 2x-1$ in (2.4) and let the exact solution of (2.4) be $\cos(16\pi x)$. We plot in Figure 6.1 (left) the $\log_{10}$ of the L^2-error against N^2. For the fifth-order equation (3.1), we take $\alpha = \beta = 1$ and the exact solution to be $\sin^3(3\pi x)$. The $\log_{10}$ of the L^2-error against N^2 is presented in Figure 6.1 (right). The straight lines in these plots indicate that the L^2-errors converge like $\exp(-cN^2)$, a typical supergeometric convergence for analytic functions by spectral methods (cf. [6]).

6.2. KDV equation. Now, we present some numerical tests for the KDV equation. We first consider the initial value KDV problem

$$u_t + uu_x + u_{xxx} = 0, \quad u(x, 0) = u_0(x), \tag{6.1}$$

with the exact soliton solution

$$u(x, t) = 12\kappa^2 \text{sech}^2(\kappa(x - 4\kappa^2 t - x_0)).$$

Since $u(x,t)$ converges to 0 exponentially as $|x| \to \infty$, we can approximate the initial value problem (6.1) by an initial-boundary value problem for $x \in (-M, M)$ as long as the soliton does not reach the boundaries.

We take $\kappa = 0.3$, $x_0 = -20$, $M = 50$, and $\Delta t = 0.001$ so that for $N \lesssim 160$, the time discretization error is negligible compared with the spatial discretization error.

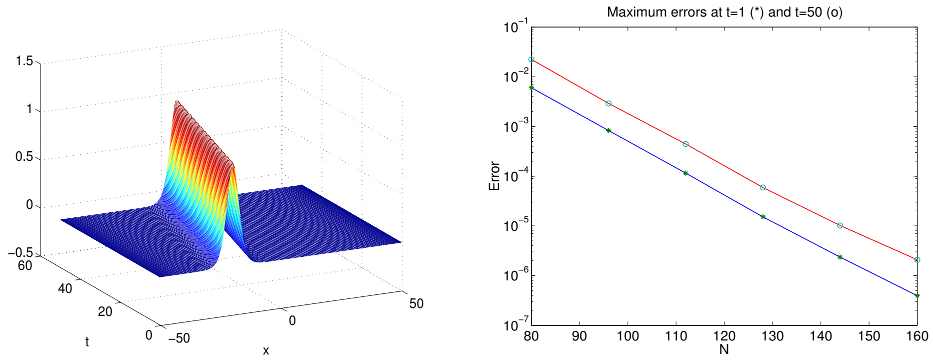


FIG. 6.2. *Exact solution for the KDV equation: Left, time evolution; right, maximum error vs. N.*

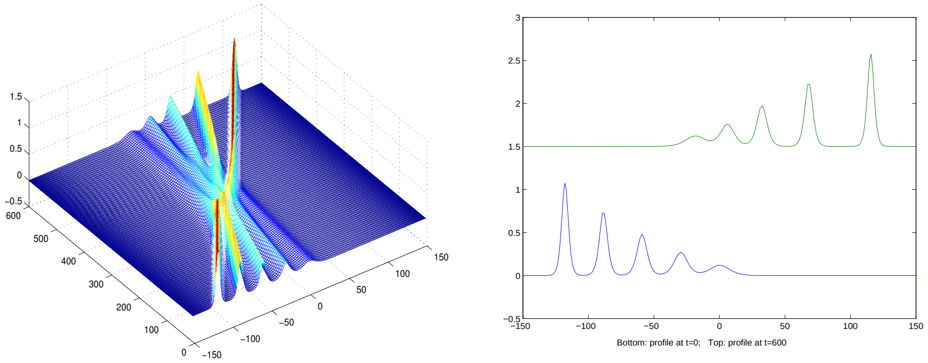


FIG. 6.3. *Interaction of five solitary waves.*

On the left of Figure 6.2, we plot the time evolution of the approximate solution, and on the right, we plot the maximum errors in the semi-log scale at $t = 1$ and $t = 50$. Note that the straight lines indicate that the errors converge like $\exp(-cN)$, which is typical for solutions that are infinitely differentiable but not analytic. The excellent accuracy for this known exact solution indicates that the KDV equation on a finite interval can be used to effectively simulate the KDV equation on a semi-infinite interval before the wave reaches the boundary.

In the following tests, we fix $M = 150$, $\Delta t = 0.02$, and $N = 256$.

Example 1: Interaction of five solitons. We start with the initial condition

$$u_0(x) = \sum_{i=1}^{5} 12\kappa_i^2 \operatorname{sech}^2(\kappa_i(x - x_i))$$

with

$$\kappa_1 = .3,\ \kappa_2 = .25,\ \kappa_3 = .2,\ \kappa_4 = .15,\ \kappa_5 = .1,$$
$$x_1 = -120,\ \ x_1 = -90,\ \ x_3 = -60,\ x_4 = -30,\ \ x_5 = 0.$$

In Figure 6.3 (left), we plot the time evolution of the solution in the (x, t) plane. We also plot the initial profile and the profile at the final step ($t = 600$) in Figure 6.3 (right). We observe that the soliton with higher amplitude travels with faster speed,

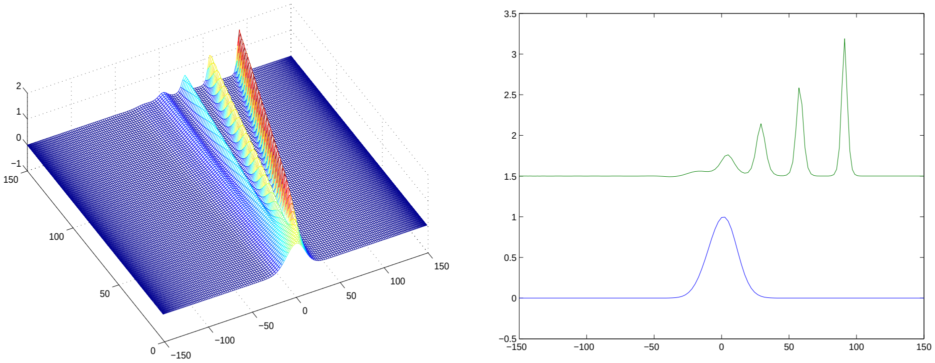


FIG. 6.4. *Solitary waves generated by an initial Gaussian profile.*

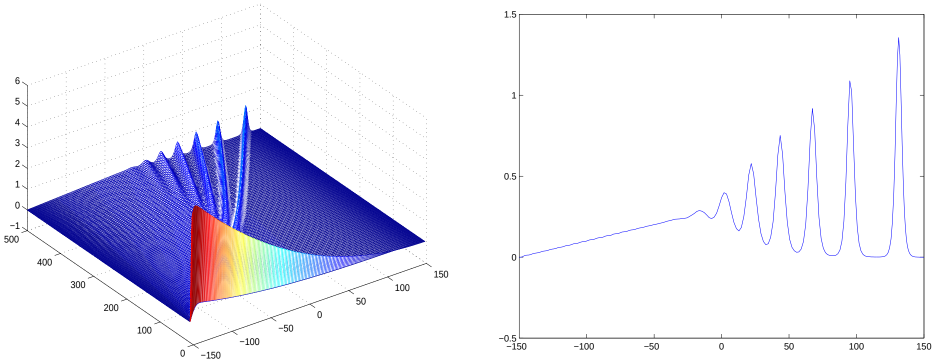


FIG. 6.5. *Solitary waves generated by a square pulse (in time) at the left boundary.*

and the amplitudes of the five solitary waves are well preserved at the final time. This indicates that our scheme has an excellent conservation property.

Example 2: Solitary waves generated by an initial Gaussian profile. We start with the initial condition $u_0(x) = \exp(-1.5x^2)$. We plot the time evolution of the solution on the left in Figure 6.4 and the profile at the final step ($t = 150$) on the right. The initial Gaussian profile has evolved into four separated solitary waves by the time $t = 150$.

Example 3: Solitary waves generated by a square pulse (in time) at the left boundary. In this example, we take $u_0(x) = 0$ but set

$$u(0,t) = \begin{cases} 5, & 0 \le t \le 5, \\ 0, & t > 5. \end{cases}$$

One may think of this situation as a dam of height five unit length that releases water for five unit time and is then shut off. We plot the time evolution of the solution on the left in Figure 6.5 and the profile at the final step ($t = 500$) on the right. The square pulse (in time) at the boundary generates a cascade of solitary waves as time evolves. This interesting phenomenon was first observed by Chu, Xiang, and Baransky in [8] (see also [12]).

7. Concluding remarks. We presented in this paper a new dual-Petrov–Galerkin method for third and higher odd-order equations. The key idea is to use test functions satisfying boundary conditions which are the "dual" of those for the trial functions. The resulting linear systems are sparse for problems with constant coefficients and well conditioned for problems with variable coefficients. By exploring the orthogonal properties of the test and trial basis functions in weighted Sobolev spaces, we were able to establish optimal error estimates for typical third-order and fifth-order linear equations and for a KDV equation on a finite interval. Obviously, the technique can be extended to other higher odd-order equations.

When combined with a Chebyshev–Legendre approach, our dual-Petrov–Galerkin method has a quasi-optimal computational complexity and is extremely accurate and efficient as illustrated by our numerical examples. Hence the method is most suitable for the study of complex dynamics of higher odd-order equations.

Finally, we note that the orthogonal polynomials $\{\phi_k\}$ and $\{\Phi_k\}$ introduced in this paper can be viewed as extensions of the Jacobi polynomials $J_k^{a,b}(x)$ with $(a, b) = (-2, -1)$ and $(a, b) = (-3, -2)$, respectively. They appear to be the most natural basis functions for, respectively, the third- and fifth-order equations (with the specified boundary conditions) considered in this paper. The extension of the Jacobi polynomials to more general (a, b) with $a,\ b < -1$ and their applications to spectral methods for other types of partial differential equations, including, in particular, hyperbolic systems, will be investigated in a forthcoming paper.

Acknowledgment. The author would like to thank Professor Benyu Guo for fruitful discussions and for carefully reading the first draft of the paper.

REFERENCES

[1] B. K. Alpert and V. Rokhlin, *A fast algorithm for the evaluation of Legendre expansions*, SIAM J. Sci. Statist. Comput., 12 (1991), pp. 158–179.

[2] J. Bona and R. Winther, *The Korteweg–de Vries equation, posed in a quarter-plane*, SIAM J. Math. Anal., 14 (1983), pp. 1056–1106.

[3] J. L. Bona, W. G. Pritchard, and L. R. Scott, *An evaluation of a model equation for water waves*, Philos. Trans. Roy. Soc. London Ser. A, 302 (1981), pp. 457–510.

[4] J. L. Bona, S. M. Sun, and B. Y. Zhang, *A non-homogeneous boundary-value problem for the Korteweg-de Vries equation posed on a finite domain*, Comm. Partial Differential Equations, 28 (2003), pp. 1391–1436.

[5] J. L. Bona, S. M. Sun, and B. Y. Zhang, *A non-homogeneous boundary-value problem for the Korteweg-de Vries equation in a quarter plane*, Trans. Amer. Math. Soc., 354 (2002), pp. 427–490.

[6] J. P. Boyd, *Chebyshev and Fourier Spectral Methods*, 2nd ed., Dover Publications, Mineola, NY, 2001.

[7] C. Canuto, M. Y. Hussaini, A. Quarteroni, and T. A. Zang, *Spectral Methods in Fluid Dynamics*, Springer-Verlag, New York, 1987.

[8] C. K. Chu, L. W. Xiang, and Y. Baransky, *Solitary waves induced by boundary motion*, Comm. Pure Appl. Math., 36 (1983), pp. 495–504.

[9] T. Colin and J.-M. Ghidaglia, *An initial-boundary value problem for the Korteweg-de Vries equation posed on a finite interval*, Adv. Differential Equations, 6 (2001), pp. 1463–1492.

[10] W. S. Don and D. Gottlieb, *The Chebyshev–Legendre method: Implementing Legendre methods on Chebyshev points*, SIAM J. Numer. Anal., 31 (1994), pp. 1519–1534.

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

[12] B.-Y. Guo, *Numerical solution of an initial-boundary value problem of the Korteweg-de Vries equation*, Acta Math. Sci. (English Ed.), 5 (1985), pp. 337–348.

[13] B.-Y. Guo and J. Shen, *On spectral approximations using modified Legendre rational functions: Application to the Korteweg-de Vries equation on the half line*, Indiana Univ. Math. J., 50 (2001), pp. 181–204. Dedicated to Professors Ciprian Foias and Roger Temam (Bloomington, IN, 2000).

[14] B.-Y. Guo, *Spectral Methods and Their Applications*, World Scientific, River Edge, NJ, 1998.

[15] B.-Y. Guo, *Jacobi approximations in certain Hilbert spaces and their applications to singular differential equations*, J. Math. Anal. Appl., 243 (2000), pp. 373–408.

[16] J. L. Hammack and H. Segur, *The Korteweg-de Vries equation and water waves.* II*, Comparison with experiments*, J. Fluid Mech., 65 (1974), pp. 289–313.

[17] W. Z. Huang and D. M. Sloan, *The pseudospectral method for third-order differential equations*, SIAM J. Numer. Anal., 29 (1992), pp. 1626–1647.

[18] S. Kichenassamy and P. J. Olver, *Existence and nonexistence of solitary wave solutions to higher-order model evolution equations*, SIAM J. Math. Anal., 23 (1992), pp. 1141–1166.

[19] H. Ma and W. Sun, *A Legendre–Petrov–Galerkin method and Chebyshev collocation method for the third-order differential equations*, SIAM J. Numer. Anal., 38 (2000), pp. 1425–1438.

[20] H. Ma and W. Sun, *Optimal error estimates of the Legendre–Petrov–Galerkin method for the Korteweg–de Vries equation*, SIAM J. Numer. Anal., 39 (2001), pp. 1380–1394.

[21] W. J. Merryfield and B. Shizgal, *Properties of collocation third-derivative operators*, J. Comput. Phys., 105 (1993), pp. 182–185.

[22] J. Shen, *Efficient spectral-Galerkin method* I*: Direct solvers of second- and fourth-order equations using Legendre polynomials*, SIAM J. Sci. Comput., 15 (1994), pp. 1489–1505.

[23] J. Shen, *Efficient spectral-Galerkin method* II*: Direct solvers of second- and fourth-order equations using Chebyshev polynomials*, SIAM J. Sci. Comput., 16 (1994), pp. 74–87.

[24] J. Shen, *Efficient Chebyshev-Legendre Galerkin methods for elliptic problems*, in Proceedings of ICOSAHOM'95, A. V. Ilin and R. Scott, eds., Houston J. Math., Houston, TX, 1996, pp. 233–240.

[25] J. Shen, *Efficient spectral-Galerkin methods* III*: Polar and cylindrical geometries*, SIAM J. Sci. Comput., 18 (1997), pp. 1583–1604.

[26] J. Shen, *Efficient spectral-Galerkin methods* IV*: Spherical geometries*, SIAM J. Sci. Comput., 20 (1999), pp. 1438–1455.

[27] G. Szegö, *Orthogonal Polynomials*, Amer. Math. Soc. Colloq. Publ. 23, AMS, New York, 1939.

[28] N. J. Zabusky and C. J. Galvin, *Shallow water waves, the Korteveg-de-Vries equation and solitons*, J. Fluid Mech., 47 (1971), pp. 811–824.

SIAM J. NUMER. ANAL.
Vol. 41, No. 5, pp. 1620–1642

THE SDFEM FOR A CONVECTION-DIFFUSION PROBLEM WITH A BOUNDARY LAYER: OPTIMAL ERROR ANALYSIS AND ENHANCEMENT OF ACCURACY*

MARTIN STYNES[†] AND LUTZ TOBISKA[‡]

Abstract. The streamline-diffusion finite element method (SDFEM) is applied to a convection-diffusion problem posed on the unit square, using a Shishkin rectangular mesh with piecewise bilinear trial functions. The hypotheses of the problem exclude interior layers but allow exponential boundary layers. An error bound is proved for $\|u^I - u^N\|_{SD}$, where u^I is the interpolant of the solution u, u^N is the SDFEM solution, and $\|\cdot\|_{SD}$ is the streamline-diffusion norm. This bound implies that $\|u - u^N\|_{L^2}$ is of optimal order, thereby settling an open question regarding the L^2-accuracy of the SDFEM on rectangular meshes. Furthermore, the bound shows that u^N is superclose to u^I, which allows the construction of a simple postprocessing that yields a more accurate solution. Enhancement of the rate of convergence by using a discrete streamline-diffusion norm is also discussed. Finally, the verification of these rates of convergence by numerical experiments is examined, and it is shown that this practice is less reliable than was previously believed.

Key words. convection-diffusion problems, streamline-diffusion method, finite elements, error estimates

AMS subject classifications. 65N30, 65N15

DOI. 10.1137/S0036142902404728

1. Introduction.

We consider the singularly perturbed boundary value problem

$$Lu := -\varepsilon\Delta u + b\cdot\nabla u + cu = f \quad \text{on } \Omega = (0,1)^2, \tag{1.1}$$

$$u = 0 \text{ on } \partial\Omega,$$

where ε is a small positive parameter, $b(x,y) = (b_1(x,y), b_2(x,y))$ with $b_1(x,y) > \beta_1 > 0$ and $b_2(x,y) > \beta_2 > 0$, $c(x,y) \geq 0$ on $\bar{\Omega}$, and

$$c(x,y) - \operatorname{div}\frac{b(x,y)}{2} \geq c_0 > 0 \text{ on } \bar{\Omega}, \tag{1.2}$$

where β_1, β_2, and c_0 are some constants. We assume that the functions b, c, and f are sufficiently smooth. These hypotheses ensure that (1.1) has a unique solution in $H_0^1(\Omega)\cap H^2(\Omega)$ for all $f \in L^2(\Omega)$. Note that for sufficiently small ε, the other hypotheses imply that (1.2) can always be ensured by the simple change of variable $v(x,y) = e^{-\sigma x}u(x,y)$ when σ is chosen suitably.

This elliptic boundary value problem, posed on a rectangular domain, is a simplification of more complex situations. In the numerical solution of models of fluid flow problems, the streamline-diffusion finite element method (SDFEM) is often used on meshes that are adapted to the flow's properties. Our aim in this paper is, on nonuniform meshes, to develop a deeper understanding of the behavior of that method that

*Received by the editors March 28, 2002; accepted for publication (in revised form) February 26, 2003; published electronically October 2, 2003. This research was supported by Enterprise Ireland International Collaboration Programme grant IC/2001/071.

http://www.siam.org/journals/sinum/41-5/40472.html

[†]National University of Ireland, Cork, Ireland (m.stynes@ucc.ie).

[‡]Institut für Analysis und Numerik, Otto-von-Guericke Universität Magdeburg, Postfach 4120, D-39016 Magdeburg, Germany (tobiska@mathematik.uni-magdeburg.de).

will be of use to others. The precise degree of accuracy that the method achieves on certain highly nonuniform meshes has not been fully understood up to now—even for problems like (1.1)—and we shall throw fresh light on this fundamental question.

Layer-adapted meshes are usually used to solve (1.1), since its solution typically has boundary layers at the sides $x = 1$ and $y = 1$ of Ω. A piecewise uniform Shishkin mesh can be chosen a priori when one has some knowledge of the structure of these layers [4, 14, 19]. This mesh has $O(N^2)$ points in a rectangular grid that is refined near the sides $x = 1$ and $y = 1$.

Shishkin meshes were originally introduced in the context of finite difference methods. The first paper to consider a finite element method on such a mesh seems to be [20], where a standard Galerkin method with piecewise bilinear trial functions was used. It was shown that $\|u^I - u^N_{Gal}\|_{1,\varepsilon} = O(N^{-1} \ln N)$, where $\|\cdot\|_{1,\varepsilon}$ is the ε-weighted energy norm defined in (2.9) below, u^I is the interpolant of the solution u from the finite element space of piecewise bilinears, and u^N_{Gal} is the computed solution. Later papers on the same method [11, 23] improved this result to $O(N^{-2} \ln^2 N)$. (In [11], a graded Bakhvalov mesh replaces the Shishkin mesh, which improves the convergence rate by removing the factor $\ln^2 N$; in [23], a discrete version of the ε-weighted energy norm is used; thus these papers have some differences from [20].) Despite these convergence results, it is computationally expensive to solve the associated discrete system of equations [10, 13], and this makes the standard Galerkin method less attractive.

The popular SDFEM, on the other hand, yields a discrete system of equations that can be solved efficiently by standard iterative methods [13]. Furthermore, in the graphical results presented in [13] for solutions of (1.1) computed on Shishkin meshes, the plots show that the solution computed by the standard Galerkin method has many small oscillations, while no oscillations are visible in the plot of the SDFEM solution. (For nonlinear problems such oscillations are highly undesirable.) Finally, Galerkin methods are much more sensitive to the choice of transition point in the Shishkin mesh than the SDFEM. Roos [17, p. 294] is strongly critical of Galerkin methods on Shishkin meshes for all these reasons, and his Figure 4 shows the undesirable oscillations that they can produce. For theoretical comparisons of the two methods that show the superiority of the SDFEM, see Remarks 4.2 and 5.1 below.

The convergence properties of the SDFEM have been widely studied [5, 15, 16, 19, 21, 24, 25], but as we discuss below, up to now no satisfactory optimal convergence result has been proved. Consequently there is an impression among researchers that, compared with the standard Galerkin method, to achieve the extra stability of the SDFEM one must pay the price of reduced accuracy in the computed solution. We shall show that, at least in the case of a rectangular Shishkin mesh, the SDFEM is no less accurate than the standard Galerkin method.

In [21], the SDFEM was applied to (1.1) using a Shishkin mesh, and it was shown that $\|u^I - u^N\|_{SD} = O(\varepsilon^{1/2} N^{-1} \ln N + N^{-3/2})$, where u^N is the computed solution, once again u^I is the interpolant of the solution u from the finite element space of piecewise bilinears, and $\|\cdot\|_{SD}$ is the streamline-diffusion norm, which will be defined in (2.7). In the present paper, we shall use sharp interpolation error identities of Lin [7] to perform a more careful and incisive analysis than was given in [21], culminating in a proof that in fact $\|u^I - u^N\|_{SD} = O(\varepsilon N^{-3/2} + N^{-2} \ln^2 N)$.

Numerical experiments in [22] seemed to indicate that the earlier result that $\|u^I - u^N\|_{SD} = O(\varepsilon^{1/2} N^{-1} \ln N + N^{-3/2})$ was optimal. In the present paper, we shall show, however, that it is easy to misinterpret the rate of convergence obtained in numerical experiments, and it now seems likely that when $\varepsilon \leq N^{-1}$, the numerical results from [22] show that $\|u^I - u^N\|_{SD} = O(N^{-2}\ln^2 N)$. This warning about the indeterminate nature of computed rates of convergence has significant implications for the inferences drawn from numerical experiments in many papers on convection-diffusion problems.

Our main result settles a long-standing conjecture regarding the accuracy of the SDFEM on regions where the solution u is smooth (i.e., away from any layers). Standard analyses (see, e.g., [19]) showed only that the rate of convergence attained in L^2 was a half-order less than optimal. (By "optimal" we mean the order of the error between the solution and its interpolant from the finite element space.) Was this apparent deficiency of the SDFEM actually observed in practice? Numerical experiments in many papers seemed to indicate that in fact one always attained optimality. Then in [24], Zhou produced an example with a smooth solution where piecewise linears on a sequence of special triangular meshes of diameter h yielded convergence of only $O(h^{3/2})$ in L^2. On the other hand, in [25] the authors showed that, when the mesh is quadrilateral, isoparametric bilinear trial functions are used, the convective flow direction b closely follows (in a precise sense) the meshlines, and the solution u is smooth on the entire domain, then one attains second-order convergence in L^2 (in fact, in L^∞). This still left open the possibility that one could construct a Zhou-type counterexample for bilinears when the convective flow was transverse to the meshlines. We show (see Remark 4.1) that this will not occur on rectangular locally uniform meshes: for (1.1) one obtains full second-order convergence (up to a logarithmic factor) in L^2 when bilinears are used. It is therefore likely that for more general problems one could use cut-off functions to prove a similar result in regions that extend downstream from an inflow boundary and on which the solution is smooth.

Furthermore, our analysis is for a problem whose solution has exponential boundary layers, and the estimates proved are valid inside those layers also with absolute constants that are independent of the data of the problem.

Our main result (Theorem 4.5) bounds the error $\|u^I - u^N\|_{SD}$. Writing $\|\cdot\|_{1,\varepsilon}$ for the ε-weighted energy norm, which is weaker than $\|\cdot\|_{SD}$, it turns out that $\|u - u^I\|_{1,\varepsilon}$ has in general a lower order of convergence than $\|u^I - u^N\|_{1,\varepsilon}$, so Theorem 4.5 does not immediately yield a satisfactory convergence result for $\|u - u^N\|_{1,\varepsilon}$.

To obtain a high order of convergence for $\|u - u^N\|$ in some norm stronger than L^2, we discuss two separate methods. First, we show that for a discrete analogue $\|\cdot\|_{SD,d}$ of the norm $\|\cdot\|_{SD}$, one has $\|u - u^N\|_{SD,d} \leq C(\varepsilon N^{-3/2} + N^{-2}\ln^2 N)$. Second, we analyze a simple local postprocessing of u^N that yields a piecewise biquadratic solution $\tilde{u}^N$ for which $\|u - \tilde{u}^N\|_{1,\varepsilon}$ has the same order as $\|u^I - u^N\|_{1,\varepsilon}$.

The techniques and results of this paper are not restricted to bilinear elements; in a forthcoming paper we shall consider higher-order elements.

The paper is organized as follows. In section 2 we describe the Shishkin mesh and the SDFEM. A decomposition of the solution u and some theoretical facts about interpolation that we shall need later are presented in section 3. The convergence of the SDFEM is analyzed in section 4. Then in section 5 the results outlined in the previous paragraph are proved. Section 6 shows that the use of numerical results to confirm theoretical rates of convergence is less reliable than was previously supposed.

Notation. Throughout the paper C will denote a generic positive constant that is independent of ε and the mesh.

We use the standard Sobolev spaces $W^{k,p}(D)$, $H^k(D) = W^{k,2}(D)$, $H_0^k(D)$, $L^p(D) = W^{0,p}(D)$ for nonnegative integers k and $1 \leq p \leq \infty$ and write $(\cdot,\cdot)_D$ for the $L^2(D)$

inner product. Here D is any measurable subset of Ω. Then $|\cdot|_{k,D}$ and $\|\cdot\|_{k,D}$ are the usual Sobolev seminorm and norm on $H^k(D)$. When $D = \Omega$, we drop D from the notation for simplicity.

2. The Shishkin mesh and the SDFEM. In this section we describe the mesh and the finite element method.

Shishkin meshes are piecewise uniform meshes, constructed a priori, that are refined inside layers. See [14, 19, 17] for a detailed discussion of their properties and uses.

Let N be an even positive integer. We let λ_x and λ_y denote two mesh transition parameters that will be used to specify where the mesh changes from coarse to fine; these are defined by

$$\lambda_x = \min\left(\frac{1}{2}, \frac{5\varepsilon}{2\beta_1}\ln N\right) \quad \text{and} \quad \lambda_y = \min\left(\frac{1}{2}, \frac{5\varepsilon}{2\beta_2}\ln N\right).$$

In these formulae, different authors make slightly different choices for the multiplier of ε/β_1 and ε/β_2. Linß [11] takes this multiplier to be 2 (as was also done in [21]), while Zhang [23] chooses 5/2, as we do in the present paper. This choice of a larger value ensures that the solution layers have decayed sufficiently on the coarse part of the Shishkin mesh.

In fact we assume that $\lambda_x = (5\varepsilon/(2\beta_1))\ln N$ and $\lambda_y = (5\varepsilon/(2\beta_2))\ln N$, as otherwise we have $N \geq \min\{e^{\beta_1/(5\varepsilon)}, e^{\beta_2/(5\varepsilon)}\}$ (which is very unlikely in practice), and one can then analyze the method using standard classical techniques.

We divide Ω as in Figure 2.1: $\bar{\Omega} = \Omega_{11} \cup \Omega_{21} \cup \Omega_{12} \cup \Omega_{22}$, where $\Omega_{11} = [0, 1 - \lambda_x] \times [0, 1 - \lambda_y], \Omega_{21} = [1 - \lambda_x, 1] \times [0, 1 - \lambda_y]$, $\Omega_{12} = [0, 1 - \lambda_x] \times [1 - \lambda_y, 1], \Omega_{22} = [1 - \lambda_x, 1] \times [1 - \lambda_y, 1]$.

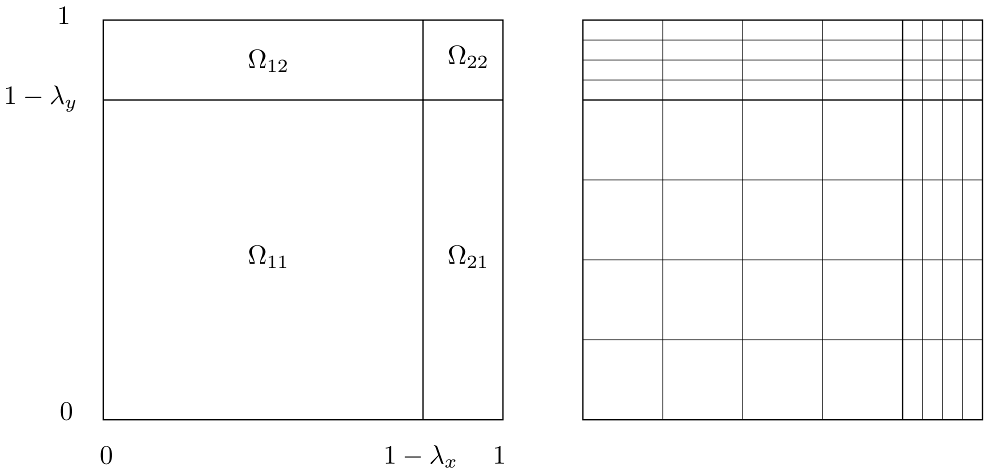


FIG. 2.1. *Division of Ω and Shishkin mesh for $\mathcal{T}_8$.*

The mesh points $\Omega^N = \{(x_i, y_j) \in \bar{\Omega} : i, j = 0, \ldots, N\}$ are the rectangular lattice defined by

$$x_i = \begin{cases} 2i(1 - \lambda_x)/N & \text{for} \quad i = 0, \ldots, N/2, \\ 1 - 2(N - i)\lambda_x/N & \text{for} \quad i = N/2 + 1, \ldots, N, \end{cases}$$

and

$$y_j = \begin{cases} 2j(1-\lambda_y)/N & \text{for} \quad j = 0, \dots, N/2, \\ 1 - 2(N-j)\lambda_y/N & \text{for} \quad j = N/2+1, \dots, N. \end{cases}$$

Our mesh is constructed by drawing lines parallel to the coordinate axes through these mesh points, so it is a tensor product of two one-dimensional piecewise uniform meshes. This divides Ω into a set $\mathcal{T}_N$ of mesh rectangles K whose sides are parallel to the axes—see Figure 2.1. The mesh is coarse on Ω_{11}, coarse/fine on $\Omega_{21} \cup \Omega_{12}$, and fine on Ω_{22}. The mesh is quasi-uniform on Ω_{11}, and its diameter d there satisfies $\sqrt{2}/N \le d \le 2\sqrt{2}/N$; on $\Omega_{12} \cup \Omega_{21}$, each mesh rectangle has dimensions $O(N^{-1})$ by $O(\varepsilon N^{-1} \ln N)$; and on Ω_{22} each rectangle is $O(\varepsilon N^{-1} \ln N)$ by $O(\varepsilon N^{-1} \ln N)$. We shall use these properties several times in our analysis. Given a mesh rectangle K, its dimensions are written as $h_{x,K}$ by $h_{y,K}$ and its barycenter is denoted by (x_K, y_K).

We now describe the SDFEM on this rectangular mesh. Our trial space V^N is the standard space of continuous piecewise bilinears that satisfy the boundary conditions of the problem:

$$V^N = \left\{ v \in C(\bar{\Omega}) : v|_{\partial\Omega} = 0 \ \text{ and } \ v|_K \in Q_1(K) \ \ \forall\, K \in \mathcal{T}_N \right\}.$$

Given any function $v(\cdot,\cdot) \in C(\bar{\Omega})$, we denote its piecewise bilinear interpolant on the mesh $\mathcal{T}_N$ by $v^I(\cdot,\cdot)$. We define the bilinear form $B_{SD}(\cdot,\cdot)$ used in the SDFEM by

$$B_{SD}(w,v) = B_{GAL}(w,v) + B_{STAB}(w,v),$$

where

$$B_{GAL}(w,v) = \varepsilon(\nabla w, \nabla v) + (b \cdot \nabla w + cw, v), \tag{2.1}$$

$$B_{STAB}(w,v) = \sum_{K \subset \Omega_{11}} \delta_K(-\varepsilon \Delta w + b \cdot \nabla w + cw, b \cdot \nabla v)_K \tag{2.2}$$

for all $(w,v) \in \tilde{H}^1(\Omega) \times H^1(\Omega)$, where $\tilde{H}^1(\Omega)$ denotes the set of functions in $H^1(\Omega)$ that lie in $H^2(K)$ for each K, and $\delta_K \ge 0$ is a user-chosen piecewise constant parameter. Now the SDFEM is defined as follows: Find $u^N \in V^N$ such that

$$B_{SD}(u^N, v^N) = (f, v^N) + \sum_{K \subset \Omega_{11}} \delta_K(f, b \cdot \nabla v^N)_K \qquad \forall v^N \in V^N. \tag{2.3}$$

The term Δu^N in $B_{STAB}(u^N, v^N)$ is zero on each K as our trial space is piecewise bilinear. We clearly have the Galerkin orthogonality property

$$B_{SD}(u - u^N, v^N) = 0 \quad \forall\, v^N \in V^N. \tag{2.4}$$

It is shown in, e.g., [19, section III.3.2.1], that if

$$0 \le \delta_K \le \frac{c_0}{(\max_K |c(x,y)|)^2} \quad \forall\, K \subset \Omega_{11}, \tag{2.5}$$

then

$$B_{SD}(v^N, v^N) \ge \frac{1}{2} \|v^N\|_{SD}^2 \quad \forall\, v^N \in V^N, \tag{2.6}$$

where we define

$$\|v\|_{SD} = \left(\varepsilon|v|_1^2 + \sum_{K\subset\Omega_{11}} \delta_K\|b\cdot\nabla v\|_{0,K}^2 + c_0\|v\|_0^2\right)^{1/2} \qquad \forall\, v\in H^1(\Omega). \tag{2.7}$$

(Here we simplified the result from [19] by observing that $\Delta v^N|_K = 0$ for each $K \in \mathcal{T}_N$.) It follows that (2.3) has a unique solution $u^N \in V^N$.

Similarly to [19, p. 233], we set

$$\delta_K = \begin{cases} N^{-1} & \text{if } K\subset\Omega_{11} \text{ and } \varepsilon\le N^{-1},\\ \varepsilon^{-1}N^{-2} & \text{if } K\subset\Omega_{11} \text{ and } \varepsilon> N^{-1},\\ 0 & \text{otherwise.}\end{cases} \tag{2.8}$$

This choice clearly satisfies (2.5) for N sufficiently large (independently of ε).

We shall also use the ε-weighted energy norm

$$\|v\|_{1,\varepsilon} = \left(\varepsilon|v|_1^2 + c_0\|v\|_0^2\right)^{1/2} \qquad \forall\, v\in H^1(\Omega). \tag{2.9}$$

This norm is clearly weaker than $\|\cdot\|_{SD}$.

3. Solution decomposition, a priori estimates. For the analysis we shall assume that the solution u can be decomposed in a way that reflects the typical behavior that is observed in solutions of (1.1) when interior layers are absent. The precise hypotheses follow.

Assumption 3.1. Assume that

$$u = S + E_{21} + E_{12} + E_{22}, \tag{3.1}$$

where there exists a constant C such that for all $(x,y)\in\Omega$ we have

$$\left|\frac{\partial^{i+j}S}{\partial x^i\partial y^j}(x,y)\right| \le C \quad \text{for } 0\le i+j\le 2, \tag{3.2}$$

$$\left|\frac{\partial^{i+j}E_{21}}{\partial x^i\partial y^j}(x,y)\right| \le C\varepsilon^{-i}e^{-\beta_1(1-x)/\varepsilon} \quad \text{for } 0\le i,j\le 2, \tag{3.3}$$

$$\left|\frac{\partial^{i+j}E_{12}}{\partial x^i\partial y^j}(x,y)\right| \le C\varepsilon^{-j}e^{-\beta_2(1-y)/\varepsilon} \quad \text{for } 0\le i,j\le 2, \tag{3.4}$$

$$\left|\frac{\partial^{i+j}E_{22}}{\partial x^i\partial y^j}(x,y)\right| \le C\varepsilon^{-(i+j)}e^{-(\beta_1(1-x)+\beta_2(1-y))/\varepsilon} \quad \text{for } 0\le i,j\le 2. \tag{3.5}$$

Furthermore, assume that $S\in H^3(\Omega)$ with

$$\|S\|_3 \le C. \tag{3.6}$$

Here S is the smooth part of u, E_{21} is an exponential boundary layer along the side $x=1$ of Ω, E_{12} is an exponential boundary layer along the side $y=1$, while E_{22} is an exponential corner layer at (1,1).

Remark 3.1. Linß [11] and Zhang [23] make assumptions close to those of Assumption 3.1. In [12] a proof is given that under certain compatibility conditions on the data f, the bounds (3.2)–(3.5) of Assumption 3.1 hold true. The extension of this result to the case $0\le i+j\le 3$ in (3.2)—which would imply (3.6)—is not trivial and furthermore would place an inordinate number of compatibility conditions on f, but it should be noted that the hypothesis (3.6) is weaker than requiring all third-order derivatives of u to be pointwise bounded.

The Shishkin mesh is highly anisotropic on $\Omega_{12} \cup \Omega_{21}$, and to obtain satisfactory interpolation error estimates on this region we invoke the sharp anisotropic interpolation analysis of [1], which includes the following result as a special case.

LEMMA 3.1. *Let K be any mesh rectangle of the Shishkin mesh $\mathcal{T}_N$. Let $v \in H^2(K)$. Then its piecewise bilinear interpolant v^I satisfies the bounds*

$$\begin{aligned}
\|v - v^I\|_{0,K} &\le C\left(h_{x,K}^2 \|v_{xx}\|_{0,K} + h_{x,K} h_{y,K} \|v_{xy}\|_{0,K} + h_{y,K}^2 \|v_{yy}\|_{0,K}\right), \\
\|(v - v^I)_x\|_{0,K} &\le C\left(h_{x,K} \|v_{xx}\|_{0,K} + h_{y,K} \|v_{xy}\|_{0,K}\right), \\
\|(v - v^I)_y\|_{0,K} &\le C\left(h_{x,K} \|v_{xy}\|_{0,K} + h_{y,K} \|v_{yy}\|_{0,K}\right).
\end{aligned}$$

The next lemma collects several results that involve the interpolants of the various terms in the decomposition (3.1).

LEMMA 3.2. *Let Assumption* 3.1 *hold true. Let S^I and E^I denote the piecewise bilinear interpolants of S and E, respectively, on the Shishkin mesh $\mathcal{T}_N$, where the function E can be any one of E_{12}, E_{21}, or E_{22}. Then there exists a constant C such that the following interpolation error estimates hold true:*

$$\|S - S^I\|_{0,\Omega} \le C\, N^{-2}, \tag{3.7}$$

$$\|E\|_{0,\Omega_{11}} \le C\, \varepsilon^{1/2} N^{-5/2}, \tag{3.8}$$

$$\varepsilon \|\Delta E\|_{L^1(\Omega_{11})} + \|\nabla E\|_{L^1(\Omega_{11})} \le C\, N^{-5/2}, \tag{3.9}$$

$$N^{-1} \|\nabla E^I\|_{0,\Omega_{11}} + \|E^I\|_{0,\Omega_{11}} \le C\left(\varepsilon^{1/2} N^{-5/2} + N^{-3}\right), \tag{3.10}$$

$$\|E - E^I\|_{0,\Omega} \le C N^{-2} \ln^2 N. \tag{3.11}$$

Proof. Inequality (3.7) is a standard classical result [2, Theorem 3.1.5]; it can also be deduced from Lemma 3.1 and (3.2).

We prove (3.8)–(3.10) only for $E = E_{21}$ since the proof for the other layer functions is similar. Applying (3.3), one gets

$$\begin{aligned}
\|E_{21}\|_{0,\Omega_{11}}^2 &\le C \int_0^{1-\lambda_y} \int_0^{1-\lambda_x} e^{-2\beta_1(1-x)/\varepsilon}\, dx\, dy \\
&\le C\varepsilon e^{-2\beta_1 \lambda_x/\varepsilon} \\
&\le C\varepsilon N^{-5},
\end{aligned}$$

which proves (3.8). For (3.9) we have similarly

$$\begin{aligned}
\varepsilon \|\Delta E_{21}\|_{L^1(\Omega_{11})} + \|\nabla E_{21}\|_{L^1(\Omega_{11})} &\le C\varepsilon^{-1} \int_0^{1-\lambda_y} \int_0^{1-\lambda_x} e^{-\beta_1(1-x)/\varepsilon}\, dx\, dy \\
&\le C e^{-\beta_1 \lambda_x/\varepsilon} \\
&\le C\, N^{-5/2}.
\end{aligned}$$

The proof of (3.10) is longer. An inverse inequality [2, Theorem 3.2.6] yields

$$N^{-1} \|\nabla E^I\|_{0,\Omega_{11}} + \|E^I\|_{0,\Omega_{11}} \le C\, \|E^I\|_{0,\Omega_{11}},$$

and it remains only to bound $\|E^I\|_{0,\Omega_{11}}$. Again invoking (3.3),

$$\|E^I_{21}\|^2_{0,\Omega_{11}} \le C \int_0^{1-\lambda_y} \sum_{i=1}^{N/2} \int_{x_{i-1}}^{x_i} e^{-2\beta_1(1-x_i)/\varepsilon}\, dx\, dy.$$

In this sum, each integral is small as a function of ε when $i < N/2$ but not when $i = N/2$, so we treat the final value of i differently. For $i = 1, \ldots, N/2-1$ and $x \in [x_{i-1}, x_i]$, we have

$$e^{-2\beta_1(1-x_i)/\varepsilon} = e^{2\beta_1(x_{N/2}-x_{N/2-1})/\varepsilon} e^{-2\beta_1(1-x_{i-1})/\varepsilon} \le e^{2\beta_1(x_{N/2}-x_{N/2-1})/\varepsilon} e^{-2\beta_1(1-x)/\varepsilon},$$

and for $i = N/2$,

$$e^{-2\beta_1(1-x_{N/2})/\varepsilon} = e^{-2\beta_1\lambda_x/\varepsilon} = N^{-5}.$$

Thus

$$\begin{aligned}
\|E^I_{21}\|^2_{0,\Omega_{11}} &\le Ce^{2\beta_1(x_{N/2}-x_{N/2-1})/\varepsilon} \int_0^{x_{N/2-1}} e^{-2\beta_1(1-x)/\varepsilon}\, dx + CN^{-6} \\
&\le C\varepsilon e^{-2\beta_1(1-x_{N/2})/\varepsilon} + CN^{-6} \\
&\le C\left(\varepsilon N^{-5} + N^{-6}\right),
\end{aligned}$$

which proves (3.10).

To get the final estimate (3.11), observe that (3.8) and (3.10) imply that

$$\|E - E^I\|_{0,\Omega_{11}} \le C(\varepsilon^{1/2}N^{-5/2} + N^{-3}) \le CN^{-2}\ln^2 N.$$

Furthermore,

$$\|E - E^I\|_{0,\Omega\setminus\Omega_{11}} \le CN^{-2}\ln^2 N.$$

To prove this inequality, apply Lemma 3.1 and the bounds (3.3)–(3.5) on each mesh rectangle, and then add the results; see [3]. □

4. Error bound for $u^I - u^N$.

The analysis of this section improves on that of [21] by invoking the sharp superconvergence results of Lin [7]. For each mesh rectangle K, we set

$$G_K(x) = \frac{1}{2}\left[(x-x_K)^2 - \left(\frac{h_{x,K}}{2}\right)^2\right], \quad F_K(y) = \frac{1}{2}\left[(y-y_K)^2 - \left(\frac{h_{y,K}}{2}\right)^2\right].$$

Denote the east, north, west, and south edges of K by $l_{i,K}$ for $i = 1, \ldots, 4$, respectively.

Lemma 4.1 (Lin identities). *Let K be a mesh rectangle. Let $w \in H^3(K)$, and let $w^I \in Q_1(K)$ be its bilinear interpolant. Then for each $v^N \in Q_1(K)$, we have*

$$\begin{aligned}
\int_K (w-w^I)_x v_x^N \,dx\,dy &= \int_K w_{xyy}\left(F_K v_x^N - \frac{1}{3}\left(F_K^2\right)' v_{xy}^N\right) dxdy,\\
\int_K (w-w^I)_x v_y^N \,dx\,dy &= \int_K \left(F_K w_{xyy}(v_y^N - G_K' v_{xy}^N) + G_K w_{xxy} v_x^N\right) dx\,dy\\
&\quad - \int_{l_{2,K}} G_K w_{xx} v_x^N \,dx + \int_{l_{4,K}} G_K w_{xx} v_x^N \,dx,\\
\int_K (w-w^I)_y v_x^N \,dx\,dy &= \int_K \left(G_K w_{xxy}(v_x^N - F_K' v_{xy}^N) + F_K w_{xyy} v_y^N\right) dx\,dy\\
&\quad - \int_{l_{1,K}} F_K w_{yy} v_y^N \,dy + \int_{l_{3,K}} F_K w_{yy} v_y^N \,dy,\\
\int_K (w-w^I)_y v_y^N \,dxdy &= \int_K w_{xxy}\left(G_K v_y^N - \frac{1}{3}\left(G_K^2\right)' v_{xy}^N\right) dx\,dy.
\end{aligned}$$

Proof. Start from the right-hand side of each identity. Since $(w^I)_{xx}$, $(w^I)_{yy}$, and all third-order derivatives of w^I vanish, these terms can be introduced at appropriate places in the right-hand side; then one integrates by parts and takes into consideration the definitions of F_K and G_K. For more details, see [7]. □

Lemma 4.2. *Let $\varphi \in W^{1,\infty}$ satisfy $\|\varphi\|_{W^{1,\infty}} \le C$ for some constant C. Let $w^I \in V^N$ be the piecewise bilinear interpolant of $w \in H^3(\Omega) \cap W^{2,\infty}(\Omega)$ on the Shishkin mesh $\mathcal{T}_N$. Then for all $v^N \in Q_1(K)$, we have*

$$\left|\int_{\Omega_{11}} \varphi(w-w^I)_x v_x^N \,dx\,dy\right| \le C\,N^{-2}\left(|w|_2 + |w|_3\right)\|v_x^N\|_0, \tag{4.1}$$

$$\begin{aligned}
\left|\int_{\Omega_{11}} \varphi(w-w^I)_x v_y^N \,dx\,dy\right| &\le C\,N^{-2}\left(|w|_{W^{2,\infty}} + \|w\|_3\right)\left(\|v_y^N\|_0 + \|v_x^N\|_0\right)\\
&\quad + C\varepsilon^{1/2}N^{-2}(\ln^{1/2} N)|w|_{W^{2,\infty}}\|v_{xy}^N\|_0,
\end{aligned}\tag{4.2}$$

$$\begin{aligned}
\left|\int_{\Omega_{11}} \varphi(w-w^I)_y v_x^N \,dx\,dy\right| &\le C\,N^{-2}\left(|w|_{W^{2,\infty}} + \|w\|_3\right)\left(\|v_x^N\|_0 + \|v_y^N\|_0\right)\\
&\quad + C\varepsilon^{1/2}N^{-2}(\ln^{1/2} N)|w|_{W^{2,\infty}}\|v_{xy}^N\|_0,
\end{aligned}\tag{4.3}$$

$$\left|\int_{\Omega_{11}} \varphi(w-w^I)_y v_y^N \,dx\,dy\right| \le C\,N^{-2}\left(|w|_2 + |w|_3\right)\|v_y^N\|_0. \tag{4.4}$$

Proof. We define a piecewise constant approximation $\bar{\varphi}$ of φ by

$$\bar{\varphi}\Big|_K = \frac{1}{\text{area K}}\int_K \varphi \,dx\,dy \quad \text{for all mesh rectangles } K.$$

Then we use the splitting

$$\begin{aligned}
&\int_{\Omega_{11}} \varphi(w-w^I)_x v_x^N \,dx\,dy\\
&\qquad = \int_{\Omega_{11}} (\varphi-\bar{\varphi})(w-w^I)_x v_x^N \,dx\,dy + \int_{\Omega_{11}} \bar{\varphi}(w-w^I)_x v_x^N \,dx\,dy.
\end{aligned}$$

The first term can be bounded using standard interpolation error estimates [2, Theorem 3.1.5] and the hypothesis $\|\varphi\|_{W^{1,\infty}} \le C$:

$$\left|\int_{\Omega_{11}} (\varphi - \bar{\varphi})(w - w^I)_x v_x^N \, dx\, dy\right| \le C\, N^{-2} |w|_2 \, \|v_x^N\|_0.$$

For the second term, we apply Lemma 4.1 and get

$$\begin{aligned}\left|\int_{\Omega_{11}} \bar{\varphi}(w - w^I)_x v_x^N \, dxdy\right| &\le C \sum_{K\subset\Omega_{11}} \|w_{xyy}\|_{0,K} \left(N^{-2}\|v_x^N\|_{0,K} + N^{-3}\|v_{xy}^N\|\right)\\ &\le C\, N^{-2}|w|_3 \, \|v_x^N\|_0\end{aligned}$$

by an inverse inequality [2, Theorem 3.2.6]. This proves (4.1). The inequality (4.4) can be shown in the same way.

To prove (4.2), we use the same technique and get

$$\begin{aligned}&\left|\int_{\Omega_{11}} \varphi(w - w^I)_x v_y^N \, dx\, dy\right|\\ &\quad\le C\, N^{-2} \left[|w|_2\|v_y^N\|_0 + |w|_3(\|v_x^N\|_0 + \|v_y^N\|_0)\right]\\ &\qquad + \left|\sum_{K\subset\Omega_{11}} \left(\int_{l_{2,K}} \bar{\varphi} G_K w_{xx} v_x^N \, dx - \int_{l_{4,K}} \bar{\varphi} G_K w_{xx} v_x^N \, dx\right)\right|.\end{aligned}$$

Set

$$\varphi_{i,j} = \bar{\varphi}(x,y) \quad \text{for } x_{i-1} < x < x_i \text{ and } y_{j-1} < y < y_j.$$

Since $v_x^N(x,0) = 0$, we have

$$\begin{aligned}&\sum_{K\subset\Omega_{11}} \left(\int_{l_{2,K}} \bar{\varphi} G_K w_{xx} v_x^N \, dx - \int_{l_{4,K}} \bar{\varphi} G_K w_{xx} v_x^N \, dx\right)\\ &\quad= \sum_{i,j=1}^{N/2} \int_{x_{i-1}}^{x_i} \varphi_{ij} \left[(G_K w_{xx} v_x^N)(x,y_j) - (G_K w_{xx} v_x^N)(x,y_{j-1})\right] dx\\ &\quad= \sum_{i=1}^{N/2} \sum_{j=1}^{N/2-1} \int_{x_{i-1}}^{x_i} (\varphi_{ij} - \varphi_{i,j+1})\, (G_K w_{xx} v_x^N)(x,y_j)\, dx\\ &\qquad\quad + \sum_{i=1}^{N/2} \int_{x_{i-1}}^{x_i} \varphi_{i,N/2} (G_K w_{xx} v_x^N)(x, y_{N/2})\, dx.\end{aligned} \tag{4.5}$$

Here we used the following observation: if K and K' are two mesh rectangles having a common horizontal edge, then v_x^N is continuous across this edge and G_K and $G_{K'}$

coincide there. Taking the first term in (4.5),

$$\begin{aligned}\left|\sum_{i=1}^{N/2}\sum_{j=1}^{N/2-1}\int_{x_{i-1}}^{x_i}(\varphi_{ij}-\varphi_{i,j+1})\,(G_K w_{xx}v_x^N)(x,y_j)\,dx\right|&\\ \le C\,N^{-3}|w|_{W^{2,\infty}}\sum_{i=1}^{N/2}\sum_{j=1}^{N/2-1}\int_{x_{i-1}}^{x_i}|v_x^N(x,y_j)|\,dx&\\ \le C\,N^{-3}|w|_{W^{2,\infty}}\sum_{i=1}^{N/2}\sum_{j=1}^{N/2-1}\sqrt{\int_{x_{i-1}}^{x_i}\int_{y_{j-1}}^{y_j}|v_x^N(x,y)|^2\,dx\,dy}&\\ \le C\,N^{-2}|w|_{W^{2,\infty}}\|v_x^N\|_0,&\end{aligned}$$

where the penultimate inequality is proved by transforming to the reference element and then scaling. To estimate the final term in (4.5), note first that $v_x^N(x,1)=0$. Starting with the representation

$$\begin{aligned}(G_K w_{xx}v_x^N)(x,y_{N/2})&=\sum_{j=N/2}^{N-1}\left[(G_K w_{xx}v_x^N)(x,y_j)-(G_K w_{xx}v_x^N)(x,y_{j+1})\right]\\ &=-\sum_{j=N/2}^{N-1}\int_{y_j}^{y_{j+1}}\frac{\partial(G_K w_{xx}v_x^N)}{\partial y}(x,y)\,dy,\end{aligned}$$

one has

$$\begin{aligned}\left|\sum_{i=1}^{N/2}\int_{x_{i-1}}^{x_i}\varphi_{i,N/2}(G_K w_{xx}v_x^N)(x,y_{N/2})\,dx\right|&\\ =\left|\sum_{i=1}^{N/2}\sum_{j=N/2}^{N-1}\int_{x_{i-1}}^{x_i}\int_{y_j}^{y_{j+1}}\varphi_{i,N/2}G_K\left(w_{xxy}v_x^N+w_{xx}v_{xy}^N\right)\,dx\,dy\right|&\\ \le CN^{-2}\sum_{K\subset\Omega_{12}}\int_K\left|w_{xxy}v_x^N+w_{xx}v_{xy}^N\right|\,dx\,dy&\\ \le CN^{-2}\left[|w|_3\|v_x^N\|_0+|w|_{W^{2,\infty}}\|v_{xy}^N\|_{L^1(\Omega_{12})}\right]&\\ \le CN^{-2}\left[|w|_3\|v_x^N\|_0+(\varepsilon\ln N)^{1/2}|w|_{W^{2,\infty}}\|v_{xy}^N\|_0\right],&\end{aligned}$$

where we used the Cauchy–Schwarz inequality to bound $\|v_{xy}^N\|_{L^1(\Omega_{12})}$. Collecting the various estimates now yields (4.2). One can similarly prove (4.3). □

Versions of the next lemma appear in both [11] and [23], but the proof of [11] contains the erroneous statement that $\|E-E^I\|_{\Omega_{22}}\le C\|E\|_{\Omega_{22}}$; Linß [9] gives a short alternative argument that fixes this mistake.

LEMMA 4.3. *Let Assumption* 3.1 *hold true. Then for all* $v^N\in V^N$, *we have*

$$\left|B_{GAL}(u-u^I,v^N)\right|\le C\big(\varepsilon N^{-3/2}+N^{-2}\ln^2 N\big)\|v^N\|_{1,\varepsilon}. \tag{4.6}$$

Proof. This result is proved in [23, section 4] under the assumption that $\|S\|_{W^{3,\infty}(\Omega)}\le C$, but this hypothesis can be relaxed. Indeed, the full regularity of S was used only in the following chain of arguments [23, (4.32)]:

$$\begin{aligned}\|S_{xxx}v^N + S_{xx}v_x^N\|_{L^1(\Omega_{21}\cup\Omega_{22})} &\le \|S\|_{W^{3,\infty}(\Omega)}\|v^N + v_x^N\|_{L^1(\Omega_{21}\cup\Omega_{22})}\\ &\le C\lambda_x^{1/2}\|v^N\|_{1,\Omega_{21}\cup\Omega_{22}}\\ &\le C(\ln^{1/2} N)\|v^N\|_{1,\varepsilon}\end{aligned}$$

for all $v^N \in V^N$. Replace this calculation by

$$\begin{aligned}\|S_{xxx}v^N + S_{xx}v_x^N\|_{L^1(\Omega_{21}\cup\Omega_{22})} &\le |S|_3\|v^N\|_0 + |S|_{W^{2,\infty}(\Omega)}\|v_x^N\|_{L^1(\Omega_{21}\cup\Omega_{22})}\\ &\le C\big[\|v^N\|_0 + (\varepsilon\ln N)^{1/2}|v^N|_1\big]\\ &\le C(\ln^{1/2} N)\|v^N\|_{1,\varepsilon},\end{aligned}$$

where we used the Cauchy–Schwarz inequality. □

LEMMA 4.4. *Let Assumption* 3.1 *hold true. Then*

$$\big|B_{STAB}(u-u^I, v^N)\big| \le CN^{-2}(\ln^{1/2} N)\|v^N\|_{SD} \qquad \forall v^N \in V^N.$$

Proof. Writing E for E_{12}, E_{21}, or E_{22}, we have

$$\begin{aligned}&\big|B_{STAB}(E-E^I, v^N)\big|\\ &\le CN^{-1}\left[\varepsilon\|\Delta E\|_{L^1(\Omega_{11})} + \|\nabla E\|_{L^1(\Omega_{11})}\right]\|b\cdot\nabla v^N\|_{L^\infty(\Omega_{11})}\\ &\quad + C\,N^{-1/2}\left(\|\nabla E^I\|_{0,\Omega_{11}} + \|E-E^I\|_{0,\Omega_{11}}\right)\|v^N\|_{SD}\\ &\le CN^{-5/2}\|b\cdot\nabla v^N\|_{0,\Omega_{11}} + CN^{-1/2}\big(\varepsilon^{1/2}N^{-3/2} + N^{-2}\ln^2 N\big)\|v^N\|_{SD}\\ &\le CN^{-2}\|v^N\|_{SD}, \end{aligned}\tag{4.7}$$

where we used Lemma 3.2 and an inverse inequality [2, Theorem 3.2.6]. Next,

$$\begin{aligned}&B_{STAB}(S-S^I, v^N)\\ &\quad = -\sum_{K\subset\Omega_{11}}\varepsilon\delta_K\left(\Delta S, b\cdot\nabla v^N\right)_K + \sum_{K\subset\Omega_{11}}\delta_K\left(b\cdot\nabla(S-S^I), b\cdot\nabla v^N\right)_K\\ &\qquad + \sum_{K\subset\Omega_{11}}\delta_K\left(c(S-S^I), b\cdot\nabla v^N\right)_K.\end{aligned}\tag{4.8}$$

For the first term in (4.8),

$$\begin{aligned}-\left(\Delta S, b\cdot\nabla v^N\right)_K &= \left(\Delta S, v^N\text{div } b\right)_K - \left(\Delta S, \text{div }(bv^N)\right)_K\\ &= \left(\Delta S, v^N\text{div } b\right)_K + \left(\nabla\Delta S, bv^N\right)_K - \int_{\partial K}\Delta S(b\cdot n_K)v^N\,d\gamma,\end{aligned}$$

after integrating by parts. Here n_K denotes the outward-pointing unit normal to the boundary ∂K of K. Now (2.8) implies that $\varepsilon\delta_K \le N^{-2}$, so

$$\begin{aligned}\left|\sum_{K\subset\Omega_{11}}\varepsilon\delta_K\left(\Delta S, b\cdot\nabla v^N\right)_K\right| &\le CN^{-2}\left(|S|_2 + |S|_3\right)\|v^N\|_0\\ &\quad + CN^{-2}\left|\int_{\partial\Omega_{11}}\Delta S(b\cdot n_K)v^N\,d\gamma\right|\end{aligned}$$

since the line integrals in the interior of Ω_{11} cancel. For the line integral over $\partial\Omega_{11}$, one can imitate our earlier analysis of the final term in (4.5) to get

$$\begin{aligned}\left|\int_{\partial\Omega_{11}} \Delta S(b\cdot n_K)v^N d\gamma\right| &= \left|-\int_{\Omega_{12}} (b_2 v^N \Delta S)_y\,dx\,dy - \int_{\Omega_{21}} (b_1 v^N \Delta S)_x\,dx\,dy\right| \\ &\le C\left[\|v^N\|_0 |S|_3 + \varepsilon^{1/2}(\ln^{1/2} N)\,|v^N|_1 |S|_{W^{2,\infty}}\right] \\ &\le C(\ln^{1/2} N)\|v^N\|_{SD}.\end{aligned}$$

Collecting these results, the first term in (4.8) is bounded by

$$\left|\sum_{K\subset\Omega_{11}} \varepsilon\delta_K \left(\Delta S, b\cdot\nabla v^N\right)_K\right| \le CN^{-2}(\ln^{1/2} N)\|v^N\|_{SD}.$$

The second term in (4.8) is handled by invoking Lemma 4.2:

$$\begin{aligned}\left|\sum_{K\subset\Omega_{11}} \delta_K \left(b\cdot\nabla(S-S^I), b\cdot\nabla v^N\right)_K\right| &\le C\,N^{-3}\left[|v^N|_1 + (\varepsilon\ln N)^{1/2}\|v^N_{xy}\|_0\right] \\ &\le C\,N^{-2}\left[|v^N|_0 + (\varepsilon\ln N)^{1/2}\|v^N\|_1\right] \\ &\le C\,N^{-2}(\ln^{1/2} N)\|v^N\|_{SD}.\end{aligned}$$

Finally, the third term in (4.8) can be bounded by

$$\left|\sum_{K\subset\Omega_{11}} \delta_K \left(c(S-S^I), b\cdot\nabla v^N\right)_K\right| \le C\,N^{-3}|v^N|_1 \le C\,N^{-2}\|v^N\|_{SD},$$

using an inverse inequality.

Each term in (4.8) has now been bounded. Recalling the estimate (4.7) for the layer part of u, the conclusion of the lemma follows. □

It is now straightforward to prove our main result.

THEOREM 4.5. *Let Assumption* 3.1 *hold true. Then the SDFEM solution* u^N *satisfies*

$$\|u^N - u^I\|_{SD} \le C\,(\varepsilon N^{-3/2} + N^{-2}\ln^2 N). \tag{4.9}$$

Proof. By (2.6) and (2.4), we have

$$\begin{aligned}\frac{1}{2}\|u^N - u^I\|_{SD}^2 &\le B_{SD}(u^N - u^I, u^N - u^I) \\ &= B_{SD}(u - u^I, u^N - u^I) \\ &= B_{GAL}(u - u^I, u^N - u^I) + B_{STAB}(u - u^I, u^N - u^I).\end{aligned}$$

Now invoke Lemmas 4.3 and 4.4 to complete the proof. □

Remark 4.1. Suppose that $\varepsilon \le N^{-1/2}\ln^2 N$, as is almost certainly true in practice. Then Theorem 4.5 implies that $\|u^I - u^N\|_0 \le CN^{-2}\ln^2 N$, which is optimal since by (3.7) and (3.11) one has $\|u - u^I\|_0 \le CN^{-2}\ln^2 N$, and this is the best possible result, as can be seen by considering $u(x,y) = e^{-\beta_1(1-x)/\varepsilon}$. Thus a triangle inequality yields $\|u - u^N\|_0 \le CN^{-2}\ln^2 N$. This optimal bound is obtained without using an Aubin–Nitsche trick.

Zhou and Rannacher [24, 25] prove a similar result on meshes that are almost uniform in the streamline direction but only when the convective flow is directed along or close to the meshlines, and they do not consider solutions that exhibit boundary layers.

Remark 4.2. Zhang [23] and Linß [11] obtain a result similar to Theorem 4.5 for the standard Galerkin method on Shishkin meshes but in the weaker norm $\|\cdot\|_{1,\varepsilon}$; compared with this, our theorem guarantees additional accuracy in the computed approximation of the streamline derivative.

Remark 4.3. It should be noted that when $\varepsilon \le N^{-1}$ (so $\delta_K = N^{-1}$ for $K \subset \Omega_{11}$), Theorem 4.5 implies, in particular, that

$$\left(\sum_{K\subset\Omega_{11}} \|b\cdot\nabla(u^I - u^N)\|_{0,K}^2 \right)^{1/2} \le CN^{-3/2}\ln^2 N,$$

although approximation theory predicts that, in general,

$$\left(\sum_{K\subset\Omega_{11}} \|b\cdot\nabla(u - u^I)\|_{0,K}^2 \right)^{1/2} \le CN^{-1}$$

is the best possible result. This phenomenon, where the order of accuracy of $\|u^I - u^N\|$ in some norm or seminorm $\|\cdot\|$ is greater than the optimal order of $\|u - u^N\|$, is called the superclose property by Lin [8], who shows that it can occur in the solution of many classical problems on rectangular meshes.

5. Error bounds for $u - u^N$.

One can use Lemma 3.1 to show that $\|u - u^I\|_{1,\varepsilon} \le CN^{-1}\ln N$ (see [3]), and this estimate is sharp. Consequently, despite the bound of Theorem 4.5, the triangle inequality yields only $\|u - u^N\|_{1,\varepsilon} \le CN^{-1}\ln N$. There are two possible ways of bypassing this barrier and proving higher-order estimates for $u - u^N$ in some H^1-type norm, as we now outline.

The first possibility is to exploit the fact that discrete norms are sometimes weaker than their continuous counterparts. For example, a discrete version of the $|\cdot|_1$ seminorm (where the H^1 integral has been replaced by a one-point quadrature rule at the barycenters of the rectangles) is used in [23]. One has superconvergence of the computed gradients at these barycenters, and this leads to an enhanced rate of convergence in the discrete norm. We follow a similar approach in section 5.1 and prove higher-order uniform convergence in a discrete version of the streamline-diffusion norm.

The second possibility is to apply to u^N a local postprocessing technique yielding a new discrete solution Pu^N for which $\|u - Pu^N\|_{1,\varepsilon} \ll \|u - u^N\|_{1,\varepsilon}$. In classical finite element computations, postprocessing is commonly used, but to the best of the authors' knowledge it has been applied only recently to singularly perturbed problems: a reaction-diffusion problem is considered in [6], while in [18] the authors show how in a convection-diffusion problem one can improve the accuracy of the computed gradient. In section 5.2, we construct a local postprocessing operator P and prove a higher-order error bound (uniformly in ε) for $\|u - Pu^N\|_{1,\varepsilon}$.

In analyzing each of these possibilities, we are forced to assume slightly more regularity of the solution than was required in previous sections, but our assumptions are no stronger than those of [18, 23].

Assumption 5.1. Assume that, for the decomposition $u = S + E_{21} + E_{12} + E_{22}$ of Assumption 3.1, there exists a constant C such that for all $(x, y) \in \Omega$ we have

$$\left|\frac{\partial^{i+j} E_{21}}{\partial x^i \partial y^j}(x,y)\right| \le C\varepsilon^{-i} e^{-\beta_1(1-x)/\varepsilon} \quad \text{for } 0 \le i+j \le 3, \tag{5.1}$$

$$\left|\frac{\partial^{i+j} E_{12}}{\partial x^i \partial y^j}(x,y)\right| \le C\varepsilon^{-j} e^{-\beta_2(1-y)/\varepsilon} \quad \text{for } 0 \le i+j \le 3, \tag{5.2}$$

$$\left|\frac{\partial^{i+j} E_{22}}{\partial x^i \partial y^j}(x,y)\right| \le C\varepsilon^{-(i+j)} e^{-(\beta_1(1-x)+\beta_2(1-y))/\varepsilon} \quad \text{for } 0 \le i+j \le 3. \tag{5.3}$$

5.1. Bound in discrete streamline-diffusion norm. Let us first introduce a discrete version $\|\cdot\|_{SD,d}$ of the streamline-diffusion norm $\|\cdot\|_{SD}$. This discrete norm is defined by replacing all integrals of derivatives in $\|\cdot\|_{SD}$ by a barycentric one-point quadrature rule. That is,

$$\begin{aligned}\|v\|_{SD,d} = \Bigg[&\sum_{K\in\mathcal{T}_N} \varepsilon\,(\operatorname{area} K)|\nabla v(x_K,y_K)|^2 \\ &+ \sum_{K\subset\Omega_{11}} \delta_K|(\operatorname{area} K)(b\cdot\nabla v)(x_K,y_K)|^2 + \|v\|_0^2\Bigg]^{1/2},\end{aligned} \tag{5.4}$$

where this norm is defined for all functions v such that $v|_K \in H^3(K)$ for all $K \in \mathcal{T}_N$.

The next result shows that the discrete streamline-diffusion norm (5.4) is weaker than the streamline-diffusion norm (2.7) on the discrete space V^N.

Lemma 5.1. *There exists a positive constant C such that*

$$\|v^N\|_{SD,d} \le C\,\|v^N\|_{SD} \quad \forall v^N \in V^N.$$

Proof. A direct calculation shows that for $K \in \mathcal{T}_N$ and $v^N \in V^N$,

$$(\operatorname{area} K)\,|(b\cdot\nabla v^N)(x_K,y_K)|^2 \le \int_K |b(x_K,y_K)\cdot\nabla v^N(x,y)|^2\,dxdy.$$

Thus

$$\sum_{K\in\mathcal{T}_N} (\operatorname{area} K)\,|\nabla v^N(x_K,y_K)|^2 \le |v^N|_1^2$$

and

$$\begin{aligned}&\sum_{K\subset\Omega_{11}} \delta_K(\operatorname{area} K)\,|(b\cdot\nabla v^N)(x_K,y_K)|^2 \\ &\qquad\le \sum_{K\subset\Omega_{11}} \int_K \delta_K|b(x_K,y_K)\cdot\nabla v^N(x,y)|^2\,dxdy.\end{aligned}$$

Using the splitting

$$|b(x_K,y_K)\cdot\nabla v^N(x,y)| \le |(b(x_K,y_K)-b(x,y))\cdot\nabla v^N(x,y)| + |(b\cdot\nabla v^N)(x,y)|$$

and an inverse inequality [2, Theorem 3.2.6], we get

$$\sum_{K\subset\Omega_{11}} \delta_K (\operatorname{area} K) |(b\cdot\nabla v^N)(x_K,y_K)|^2 \le C \sum_{K\subset\Omega_{11}} \delta_K \left(\|v^N\|_{0,K}^2 + \|b\cdot\nabla v^N\|_{0,K}^2\right).$$

As the norms $\|\cdot\|_{SD}$ and $\|\cdot\|_{SD,d}$ have identical L^2 terms, the statement of the lemma follows on putting together the above estimates. □

The next interpolation result is needed later.

LEMMA 5.2. *In addition to Assumptions* 3.1 *and* 5.1, *suppose that* $S\in W^{3,\infty}(\Omega)$ *with* $\|S\|_{W^{3,\infty}(K)} \le C$ *for some constant* C. *Then*

$$\|u-u^I\|_{SD,d} \le C\,N^{-2}\ln^2 N. \tag{5.5}$$

Proof. Combining (3.7) and (3.11), we get the estimate

$$\|u-u^I\|_0 \le C\,N^{-2}\ln^2 N. \tag{5.6}$$

At the barycenter of each rectangle $K\subset\Omega_{11}$, a Taylor expansion shows readily that the gradient of the interpolant to the smooth part of u is superconvergent:

$$|(S-S^I)_x(x_K,y_K)| \le C\,N^{-2}|S|_{3,\infty,K} \quad \forall\, K\in\mathcal{T}_N. \tag{5.7}$$

Moving on to the layer part of u, it is shown in [23, Theorem 4.1] that

$$\varepsilon \sum_{K\in\mathcal{T}_N} (\operatorname{area} K)\,|\nabla(E-E^I)(x_K,y_K)|^2 \le C\,N^{-4}\ln^4 N, \tag{5.8}$$

where the function E can be any one of E_{12}, E_{21}, or E_{22}.

Inequality (5.8) suffices to bound the ε-weighted H^1-norm component of $\|u-u^I\|_{SD,d}$, but it is inadequate for the streamline-diffusion-derivative component. Thus on Ω_{11} we shall prove a stronger result. Consider only the case $E=E_{21}$, since in the other cases the estimates are analogous. Using (3.3), for all $K\subset[x_{i-1},x_i]\times[0,1-\lambda_y]\subset\Omega_{11}$ we have

$$\begin{aligned}|\nabla E_{21}(x_K,y_K)|^2 &\le C(1+\varepsilon^{-2})e^{-2\beta_1(1-x_K)/\varepsilon}\\ &\le C\varepsilon^{-2}e^{-\beta_1 H/\varepsilon}\,e^{-2\beta_1(1-x_i)/\varepsilon},\end{aligned}$$

where we set $H=x_i-x_{i-1}$ for $1=1,2,\dots,N/2$. Hence

$$\sum_{K\subset\Omega_{11}} \delta_K(\operatorname{area} K)|\nabla E_{21}(x_K,y_K)|^2 \le C\varepsilon^{-2}e^{-\beta_1 H/\varepsilon}N^{-3}\sum_{i,j=1}^{N/2} e^{-2\beta_1(1-x_i)/\varepsilon}.$$

Now

$$\begin{aligned}\sum_{i,j=1}^{N/2} e^{-2\beta_1(1-x_i)/\varepsilon} &\le C\left(\sum_{i=1}^{N/2-1} N^2\int_{x_{i-1}}^{x_i} e^{-2\beta_1(1-x_{i-1}-H)/\varepsilon}\,dx + Ne^{-2\beta_1\lambda_x/\varepsilon}\right)\\ &\le CN^2\int_0^{1-\lambda_x-H} e^{-2\beta_1(1-x-H)/\varepsilon}\,dx + CN^{-4}\\ &\le C(\varepsilon N^{-3}+N^{-4}),\end{aligned}$$

so, since $H \geq 1/N$,

$$\sum_{K\subset\Omega_{11}} \delta_K(\operatorname{area} K)|\nabla E_{21}(x_K,y_K)|^2 \leq CN^{-5}\left[(\varepsilon N)^{-1}+(\varepsilon N)^{-2}\right]e^{-\beta_1/(\varepsilon N)}$$

$$\leq CN^{-5}, \tag{5.9}$$

as the mapping $t \mapsto (t+t^2)e^{-t}$ is bounded on $\mathbb{R}_+$. It remains only to estimate $|\nabla E_{21}^I(x_K,y_K)|$. First, (3.3) and $H \geq 1/N$ imply that

$$|\nabla E_{21}^I(x_K,y_K)| \leq CNe^{-\beta_1(1-x_i)/\varepsilon}$$

for all $K \subset [x_{i-1},x_i]\times[0,1-\lambda_y] \subset \Omega_{11}$. Then recalling the calculation above,

$$\sum_{K\subset\Omega_{11}} \delta_K(\operatorname{area} K)|\nabla E_{21}^I(x_K,y_K)|^2 \leq C\sum_{i,j=1}^{N/2} N^{-1}e^{-2\beta_1(1-x_i)/\varepsilon}$$

$$\leq C(\varepsilon N^{-4}+N^{-5}). \tag{5.10}$$

Combining (5.7)–(5.10) yields the statement of the lemma. □

THEOREM 5.3. *In addition to Assumptions* 3.1 *and* 5.1, *suppose that* $S \in W^{3,\infty}(\Omega)$ *with* $\|S\|_{W^{3,\infty}(K)} \leq C$ *for some constant* C. *Then*

$$\|u-u^N\|_{SD,d} \leq C\left(\varepsilon N^{-3/2}+N^{-2}\ln^2 N\right).$$

Proof. Lemmas 5.2 and 5.1 imply that

$$\begin{aligned}\|u-u^N\|_{SD,d} &\leq \|u-u^I\|_{SD,d}+\|u^I-u^N\|_{SD,d}\\ &\leq CN^{-2}\ln^2 N + C\,\|u^I-u^N\|_{SD}.\end{aligned}$$

Now invoke Theorem 4.5 to complete the proof. □

Remark 5.1. In practice one usually has $\varepsilon \leq N^{-1/2}\ln^2 N$, so the bound of Theorem 5.3 becomes $\|u-u^N\|_{SD,d} \leq CN^{-2}\ln^2 N$. Writing u_{Gal}^N for the numerical solution computed by the Galerkin finite element method, it is shown in [11, 23] that if $\varepsilon \leq N^{-1}$, then $\|u-u_{Gal}^N\|_{1,\varepsilon} \leq CN^{-2}\ln^2 N$, but the norm $\|\cdot\|_{1,\varepsilon}$ is weaker than $\|\cdot\|_{SD,d}$.

5.2. Postprocessing u^N.

As described at the beginning of section 5, we now show how a local postprocessing of u^N yields a piecewise biquadratic solution Pu^N for which in general $\|u-Pu^N\|_{1,\varepsilon} \ll \|u-u^N\|_{1,\varepsilon}$.

Consider a family of Shishkin meshes $\mathcal{T}_N$ with mesh points (x_i,y_j) for $i,j=0,\ldots,N$, where we require $N/2$ to be even. Then we can build a coarser mesh composed of disjoint macrorectangles M, each comprising four mesh rectangles from $\mathcal{T}_N$, where M belongs to only one of the four domains Ω_{11}, Ω_{12}, Ω_{21}, and Ω_{22}. See Figure 5.1. Associate with each macrorectangle M an interpolation operator $P_M : C(\bar{M}) \to Q_2(M)$ defined by the standard biquadratic interpolation at the barycenter, nodes, and midpoints of edges of the macrorectangle. As usual, P_M can be extended to a continuous global interpolation operator $P : C(\bar{\Omega}) \to W^N$, where W^N is the space of piecewise biquadratic finite elements, by setting

$$(Pv)|_M := P_M(v|_M) \quad \forall M.$$

Note that the macrorectangle M does not belong to $\mathcal{T}_{N/2}$ because the transition point values $1-\lambda_x$ and $1-\lambda_y$ associated with the Shishkin mesh $\mathcal{T}_N$ change when

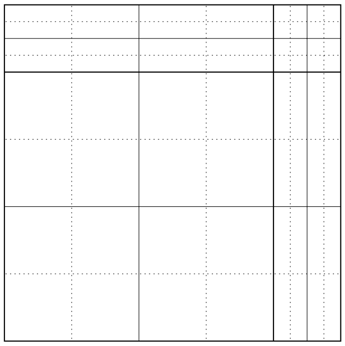


FIG. 5.1. *Macroelements built from the decomposition $\mathcal{T}_8$.*

N is replaced by $N/2$. We shall use the notation $\tilde{\mathcal{T}}_{N/2}$ for the family of macromeshes that is generated by the family of Shishkin meshes $\mathcal{T}_N$. Thus each macrorectangle $M \in \tilde{\mathcal{T}}_{N/2}$ is the union of four rectangles from $\mathcal{T}_N$.

LEMMA 5.4. *The piecewise biquadratic interpolation P has the following properties:*

$$P(v^I) = P(v) \qquad \forall v \in C(\bar{\Omega}), \tag{5.11}$$

$$\|Pv^N\|_{1,\varepsilon} \le C\|v^N\|_{1,\varepsilon} \quad \forall v^N \in V^N. \tag{5.12}$$

Proof. The identity (5.11) follows immediately from the definitions of the interpolation operators.

To prove (5.12), map the macroelement M onto the reference macroelement $\hat{M} = [-1,+1]^2$ that is the union of $M_1 = [0,1] \times [-1,0]$, $M_2 = [0,1]^2$, $M_3 = [-1,0] \times [0,1]$, and $M_4 = [-1,0]^2$. Then by scaling properties it is sufficient to show that

$$\|\hat{P}\hat{v}\|_{0,\hat{M}} \le C\|\hat{v}\|_{0,\hat{M}} \quad \text{and} \quad |\hat{P}\hat{v}|_{1,\hat{M}} \le C|\hat{v}|_{1,\hat{M}} \quad \forall \hat{v} \in Q(\hat{M}), \tag{5.13}$$

where

$$Q(\hat{M}) = \{\hat{w} \in C(\hat{M}) \;:\; \hat{w}|_{M_i} \in Q_1(M_i) \text{ for } i = 1, \dots, 4\}.$$

We have

$$\hat{P}\hat{v} = 0 \quad \Rightarrow \quad \hat{v} = 0 \qquad \forall \hat{v} \in Q(\hat{M}),$$

so the mapping $\hat{v} \mapsto \|\hat{P}\hat{v}\|_{0,\hat{M}}$ is a norm on the space $Q(\hat{M})$. Similarly

$$|\hat{P}\hat{v}|_{1,\hat{M}} = 0 \quad \Rightarrow \quad \hat{P}\hat{v} = \text{constant} \quad \Rightarrow \quad \hat{v} = \text{constant},$$

which shows that the mapping

$$\hat{v} \mapsto |\hat{P}\hat{v}|_{1,\hat{M}}$$

is a norm on the quotient space $Q(\hat{M})\backslash\mathbb{R}$. Then (5.13) follows from the equivalence of norms in finite-dimensional spaces. □

LEMMA 5.5. *Let Assumptions* 3.1 *and* 5.1 *hold true. Then*

$$\|Pu - u\|_{1,\varepsilon} \le C\big(\varepsilon N^{-3/2} + N^{-2}\ln^2 N\big). \tag{5.14}$$

Proof. When considering a family of Shishkin meshes $\mathcal{T}_N$ with $N/2$ even, the resulting family $\widetilde{\mathcal{T}}_{N/2}$ of macroelements has the same transition points as $\mathcal{T}_N$ but has $N/2$ elements in each coordinate direction. Using standard interpolation results, we get

$$\|PS - S\|_{1,\varepsilon} \le C(\varepsilon^{1/2}N^{-2} + N^{-3}) \le CN^{-2}. \tag{5.15}$$

The method of analysis of the layer parts of u varies with one's location in Ω. In order not to overburden the presentation with excessive detail, only the estimation of $\varepsilon^{1/2}\|(PE_{21} - E_{21})_x\|_0$ is described; the component $\varepsilon^{1/2}\|(PE_{21} - E_{21})_y\|_0$ of (5.14) is less badly behaved, and the treatment of E_{12} and E_{22} is broadly similar. A similar argument also shows that, analogously to (3.11),

$$\|PE - E\|_0 \le CN^{-2}\ln^2 N, \tag{5.16}$$

where E can be any one of E_{12}, E_{21}, or E_{22}.

Inside the layer (i.e., on $\Omega_{21} \cup \Omega_{22}$), one uses the following anisotropic error estimates from [1], which are analogous to those of Lemma 3.1:

$$\begin{aligned}
\|v - Pv\|_{0,M} &\le C \sum_{i+j=3} h_{x,M}^i h_{y,M}^j \left\| \frac{\partial^3 v}{\partial x^i \partial y^j} \right\|_{0,M} && \forall M \in \tilde{\mathcal{T}}_{N/2},\\
\|(v - Pv)_x\|_{0,M} &\le C \sum_{i+j=2} h_{x,M}^i h_{y,M}^j \left\| \frac{\partial^3 v}{\partial x^{i+1} \partial y^j} \right\|_{0,M} && \forall M \in \tilde{\mathcal{T}}_{N/2},\\
\|(v - Pv)_y\|_{0,M} &\le C \sum_{i+j=2} h_{x,M}^i h_{y,M}^j \left\| \frac{\partial^3 v}{\partial x^i \partial y^{j+1}} \right\|_{0,M} && \forall M \in \tilde{\mathcal{T}}_{N/2}.
\end{aligned}$$

These bounds yield

$$\begin{aligned}
&\varepsilon^{1/2}\|(E_{21} - PE_{21})_x\|_{0,\Omega_{21}\cup\Omega_{22}}\\
&\quad \le C\varepsilon^{1/2}\left[\left(\frac{\varepsilon \ln N}{N}\right)^2 \|(E_{21})_{xxx}\|_{0,\Omega_{21}\cup\Omega_{22}}\right.\\
&\qquad \left. + \left(\frac{\varepsilon \ln N}{N}\cdot\frac{1}{N}\right)\|(E_{21})_{xxy}\|_{0,\Omega_{21}\cup\Omega_{22}} + \frac{1}{N^2}\|(E_{21})_{xyy}\|_{0,\Omega_{21}\cup\Omega_{22}}\right]\\
&\quad \le CN^{-2}\ln^2 N,
\end{aligned} \tag{5.17}$$

by (5.1).

Outside the layer region $\Omega_{21} \cup \Omega_{22}$, it is best to apply the triangle inequality:

$$\varepsilon^{1/2}\|(E_{21} - PE_{21})_x\|_{0,\Omega_{12}\cup\Omega_{11}} \le \varepsilon^{1/2}\|(E_{21})_x\|_{0,\Omega_{12}\cup\Omega_{11}} + \varepsilon^{1/2}\|(PE_{21})_x\|_{0,\Omega_{12}\cup\Omega_{11}}.$$

Then (5.1) implies that

$$\varepsilon^{1/2}\|(E_{21})_x\|_{0,\Omega_{12}\cup\Omega_{11}} \le CN^{-5/2}. \tag{5.18}$$

For the other term $\varepsilon^{1/2}\|(PE_{21})_x\|_{0,\Omega_{12}\cup\Omega_{11}}$, on each $K \subset \Omega_{12}$ use the easily verified inequality $\|(PE_{21})_x\|_{L^\infty(K)} \le C\|(E_{21})_x\|_{L^\infty(K)}$, and then exploit the fact that the area of Ω_{12} is at most $C(\varepsilon \ln N)^{1/2}$, while on Ω_{11} simply invoke an inverse inequality. These techniques yield (compare with the derivation of (3.10))

$$\begin{aligned}\varepsilon\|(PE_{21})_x\|_{0,\Omega_{12}}^2 &\leq C\varepsilon^{-1}\int_{1-\lambda_y}^{1}\sum_{i=1}^{N/4}\int_{x_{2i-2}}^{x_{2i}} \exp\left(-\frac{2\beta_1(1-x_{2i})}{\varepsilon}\right) dx\, dy \\ &\leq C(\varepsilon N^{-5}+N^{-6})\ln N \qquad (5.19)\end{aligned}$$

and

$$\begin{aligned}\varepsilon^{1/2}\|(PE_{21})_x\|_{0,\Omega_{11}} &\leq C\varepsilon^{1/2}N\|PE_{21}\|_{0,\Omega_{11}} \\ &\leq C\varepsilon^{1/2}N\int_{0}^{1-\lambda_y}\sum_{i=1}^{N/4}\int_{x_{2i-2}}^{x_{2i}} \exp\left(-\frac{2\beta_1(1-x_{2i})}{\varepsilon}\right) dx\, dy \\ &\leq C\varepsilon^{1/2}(\varepsilon^{1/2}N^{-3/2}+N^{-2}). \qquad (5.20)\end{aligned}$$

Now combine (5.17)–(5.20) to get

$$\varepsilon^{1/2}\|(PE_{21}-E_{21})_x\|_0 \leq C\big(\varepsilon N^{-3/2}+N^{-2}\ln^2 N\big).$$

Recalling (5.15) and (5.16), we are done. □

Theorem 5.6. *Let Assumptions* 3.1 *and* 5.1 *hold true. Then after postprocessing by P, the numerical solution u^N generated by the SDFEM satisfies*

$$\|u-Pu^N\|_{1,\varepsilon} \leq C\big(\varepsilon N^{-3/2}+N^{-2}\ln^2 N\big).$$

Proof. The triangle inequality and Lemmas 5.4 and 5.5 yield

$$\begin{aligned}\|u-Pu^N\|_{1,\varepsilon} &\leq \|u-Pu\|_{1,\varepsilon}+\|P(u^I-u^N)\|_{1,\varepsilon} \\ &\leq C\big(\varepsilon N^{-3/2}+N^{-2}\ln^2 N+\|u^I-u^N\|_{1,\varepsilon}\big) \\ &\leq C(\varepsilon N^{-3/2}+N^{-2}\ln^2 N),\end{aligned}$$

where we invoked Theorem 4.5. □

Remark 5.2. In practice we can assume that $\varepsilon \leq N^{-1/2}\ln^2 N$ and obtain (aside from a logarithmic factor) second-order convergence in Theorem 5.6.

Remark 5.3. Theorem 5.6 implies that

$$\varepsilon^{1/2}\|\nabla u-\nabla Pu^N\|_0 \leq C(\varepsilon N^{-3/2}+N^{-2}\ln^2 N).$$

In [18] a recovery technique is applied to the solution u_{Gal}^N computed by the standard Galerkin method on the Shishkin mesh $\mathcal{T}_N$, producing a recovered gradient Ru_{Gal}^N for which $\varepsilon^{1/2}\|\nabla u - Ru_{Gal}^N\|_0 \leq CN^{-2}\ln^{5/2}N$ when $\varepsilon \leq N^{-1}$. Thus our postprocessed solution Pu^N includes a recovery of the gradient that is accurate to a slightly higher order. The two methods are not the same; ∇Pu^N is piecewise linear and discontinuous, while Ru_{Gal}^N is piecewise linear and continuous.

6. What rates of convergence are observed? In this section we give numerical results that appear to support our theoretical results, as is standard practice. However, we shall also reveal that in analyzing the numerical results of singularly perturbed problems solved on Shishkin meshes, it is easy to mistake one rate of convergence for another.

In Figure 6.1 is shown the experimental data from [22, Figure 2], which is for (1.1) with $b = (2 + x, 1 + y)$, $c = 2$, and the right-hand side f chosen so that

$$u(x, y) = 2(\sin x) \left(1 - \exp\left(-\frac{3(1 - x)}{\varepsilon}\right)\right) y^2 \left(1 - \exp\left(-\frac{2(1 - y)}{\varepsilon}\right)\right).$$

This is a typical problem of type (1.1). In [22], where only the comparison curve for $O(N^{-3/2})$ was drawn, the numerical results were interpreted as showing that experimentally one observes convergence of order $N^{-3/2}$. Of course Theorem 4.5 of the present paper predicts instead convergence of order $N^{-2} \ln^2 N$, so in Figure 6.1 we also draw this comparison curve. One can then see that indeed Theorem 4.5 does appear to predict the correct rate of convergence, but one can also see that the two comparison curves are quite similar, so it is easy to understand the misinterpretation made in [22].

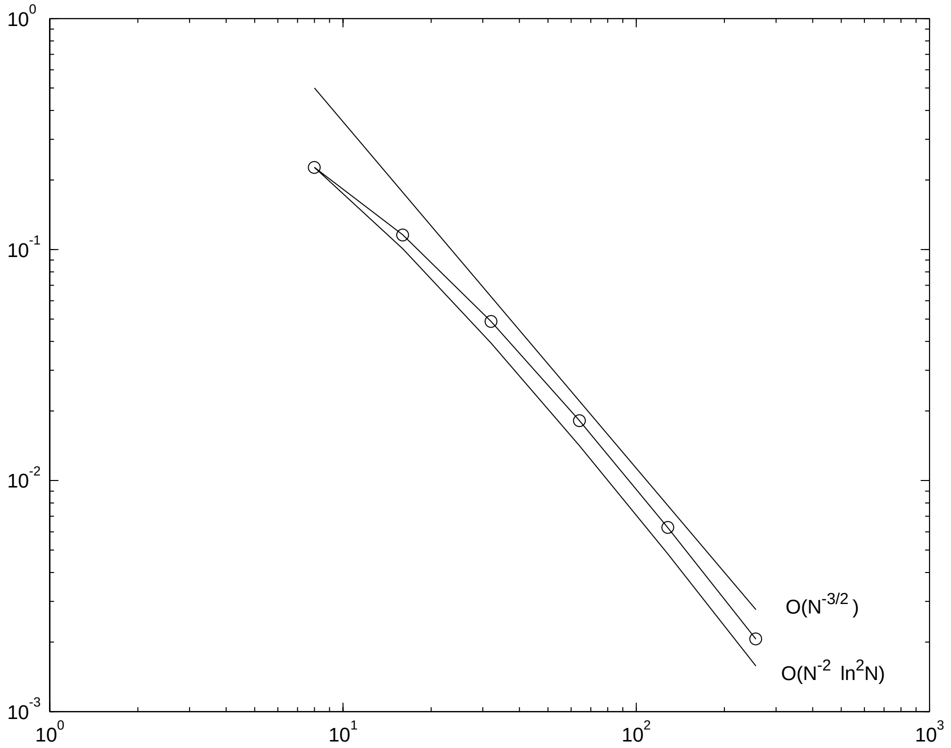


FIG. 6.1. *Error in* $\|u^I - u^N\|_{SD}$.

Can we be sure from Figure 6.1 that the experimental order of convergence is $N^{-2} \ln^2 N$? No. It is possible that a comparison curve for $O(N^{-\alpha} \ln^{\beta} N)$ (for some constants α and β) would look slightly more convincing.

Table 6.1 presents this data in another way. It gives the computed rates of convergence based on the errors in Figure 6.1 when one assumes that the error has the form $CN^{-\alpha}$ or $C(N^{-1} \ln N)^{-\alpha}$ for some constant α. Once again it is difficult to draw a firm conclusion about the actual rate that is observed.

By now, the reader may be of the opinion that these difficulties in interpretation of experimental results stem from the use of values of N that are too small. Surely

Table 6.1
Computed rates of convergence for errors from Figure 6.1.

N	16	32	64	128	256
Rate $N^{-\alpha}$; α value	0.9694	1.2455	1.4254	1.5365	1.6080
Rate $(N^{-1}\ln N)^{\alpha}$; α value	1.6572	1.8369	1.9341	1.9760	1.9917

Table 6.2
Convergence rates for $E(N) = CN^{-2}\ln^2 N$; rate of $N^{-\alpha}$ is assumed.

N	32	64	128	256	512	1024	2048	4096
α	1.3561	1.4739	1.5552	1.6147	1.6601	1.6960	1.7250	1.7489

for N sufficiently large there can be no doubt about the actual rate of convergence observed?

To throw some light on the answer to this question, Table 6.2 gives the rates of convergence, computed in the standard way when one assumes convergence of order $N^{-\alpha}$ for some constant α, from a function of the form $E(N) = CN^{-2}\ln^2 N$. That is, these numbers are not obtained by solving a boundary value problem but are computed by assuming that the numerical error function $E(N)$ has the exact form $CN^{-2}\ln^2 N$ for some constant C. If this table were computed from successive numerical solutions to a problem like (1.1) and one did not know a priori that the rates represent $O(N^{-2}\ln^2 N)$ convergence, one might easily interpret them as $O(N^{-7/4})$. Taking larger values of N would certainly reveal the misinterpretation, but already the table has reached $N = 4096$, which for our two-dimensional problem means that one has about 16 million unknowns. Thus taking a value of N sufficiently large to confirm beyond doubt a rate of convergence when solving (1.1) is in general impractical. It seems that Figure 6.1 is the best we can do.

REFERENCES

[1] T. Apel and M. Dobrowolski, *Anisotropic interpolation with applications to the finite element method*, Computing, 47 (1992), pp. 277–293.

[2] P. Ciarlet, *The Finite Element Method for Elliptic Problems*, Classics Appl. Math. 40, SIAM, Philadelphia, 2002.

[3] M. Dobrowolski and H.-G. Roos, *A priori estimates for the solution of convection-diffusion problems and interpolation on Shishkin meshes*, Z. Anal. Anwendungen, 16 (1997), pp. 1001–1012.

[4] P. Farrell, A. Hegarty, J. Miller, E. O'Riordan, and G. Shishkin, *Robust Computational Techniques for Boundary Layers*, Chapman & Hall/CRC, Boca Raton, FL, 2000.

[5] C. Johnson, A. Schatz, and L. Wahlbin, *Crosswind smear and pointwise errors in the streamline diffusion finite element methods*, Math. Comp., 49 (1987), pp. 25–38.

[6] J. Li and M. F. Wheeler, *Uniform convergence and superconvergence of mixed finite element methods on anisotropically refined grids*, SIAM J. Numer. Anal., 38 (2000), pp. 770–798.

[7] Q. Lin, *A rectangle test for finite element analysis*, in Proceedings of Systems Science and Systems Engineering, Great Wall Culture Publishing, Hong Kong, 1991, pp. 213–216.

[8] Q. Lin, *Superclose FE-theory becomes a table of integrals*, in Finite Element Methods, Lecture Notes in Pure and Appl. Math. 196, Marcel Dekker, New York, 1998, pp. 217–226.

[9] T. Linß, *private communication*, Institute for Numerical Mathematics, Technischen Universität Dresden, Dresden, Germany, 2000.

[10] T. Linß, *Analysis of a Galerkin finite element method on a Bakhvalov-Shishkin mesh for a linear convection-diffusion problem*, IMA J. Numer. Anal., 20 (2000), pp. 621–632.

[11] T. Linß, *Uniform superconvergence of a Galerkin finite element method on Shishkin-type meshes*, Numer. Methods Partial Differential Equations, 16 (2000), pp. 426–440.

[12] T. Linß and M. Stynes, *Asymptotic analysis and Shishkin-type decomposition for an elliptic convection-diffusion problem*, J. Math. Anal. Appl., 261 (2001), pp. 604–632.

[13] T. Linß and M. Stynes, *Numerical methods on Shishkin meshes for linear convection-diffusion problems*, Comput. Methods Appl. Mech. Engrg., 190 (2001), pp. 3527–3542.

[14] J. Miller, E. O'Riordan, and G. Shishkin, *Fitted Numerical Methods for Singular Perturbation Problems*, World Scientific, Singapore, 1996.

[15] U. Nävert, *A Finite Element Method for Convection-Diffusion Problems*, Ph.D. thesis, Chalmers University of Technology, Göteborg, Sweden, 1982.

[16] K. Niijima, *Pointwise error estimates for a streamline diffusion finite element scheme*, Numer. Math., 56 (1990), pp. 707–719.

[17] H.-G. Roos, *Layer-adapted grids for singular perturbation problems*, ZAMM Z. Angew. Math. Mech., 78 (1998), pp. 291–309.

[18] H.-G. Roos and T. Linss, *Gradient recovery for singularly perturbed boundary value problems* II: *Two-dimensional convection-diffusion*, Math. Models Methods Appl. Sci., 11 (2001), pp. 1169–1179.

[19] H.-G. Roos, M. Stynes, and L. Tobiska, *Numerical Methods for Singularly Perturbed Differential Equations*, Springer-Verlag, Berlin, Heidelberg New York, 1996.

[20] M. Stynes and E. O'Riordan, *A uniformly convergent Galerkin method on a Shishkin mesh for a convection-diffusion problem*, J. Math. Anal. Appl., 214 (1997), pp. 36–54.

[21] M. Stynes and L. Tobiska, *Analysis of the streamline–diffusion finite element method for a convection-diffusion problem with exponential layers*, East-West J. Numer. Math., 9 (2001), pp. 59–76.

[22] L. Tobiska, G. Matthies, and M. Stynes, *Convergence properties of the streamline-diffusion finite element method on a Shishkin mesh for singularly perturbed elliptic equations with exponential layers*, in Analytical and Numerical Methods for Convection-Dominated and Singularly Perturbed Problems, L. Vulkov, J. Miller, and G. Shishkin, eds., Nova Science Publishers, New York, 2000, pp. 123–132.

[23] Z. Zhang, *Finite element superconvergence on Shishkin mesh for* 2*-d convection-diffusion problems*, Math. Comp., 72 (2003), pp. 1147–1177.

[24] G. Zhou, *How accurate is the streamline diffusion finite element method?*, Math. Comp., 66 (1997), pp. 31–44.

[25] G. Zhou and R. Rannacher, *Pointwise superconvergence of the streamline diffusion finite element method*, Numer. Methods Partial Differential Equations, 12 (1996), pp. 123–145.

SIAM J. NUMER. ANAL.
Vol. 41, No. 5, pp. 1643–1681

OPTIMAL SCHWARZ WAVEFORM RELAXATION FOR THE ONE DIMENSIONAL WAVE EQUATION*

MARTIN J. GANDER†, LAURENCE HALPERN‡, AND FRÉDÉRIC NATAF§

Abstract. We introduce a nonoverlapping variant of the Schwarz waveform relaxation algorithm for wave propagation problems with variable coefficients in one spatial dimension. We derive transmission conditions which lead to convergence of the algorithm in a number of iterations equal to the number of subdomains, independently of the length of the time interval. These optimal transmission conditions are in general nonlocal, but we show that the nonlocality depends on the time interval under consideration, and we introduce time windows to obtain optimal performance of the algorithm with local transmission conditions in the case of piecewise constant wave speed. We show that convergence in two iterations can be achieved independently of the number of subdomains in that case. The algorithm thus scales optimally with the number of subdomains, provided the time windows are chosen appropriately. For continuously varying coefficients we prove convergence of the algorithm with local transmission conditions using energy estimates. We then introduce a finite volume discretization which permits computations on nonmatching grids, and we prove convergence of the fully discrete Schwarz waveform relaxation algorithm. We finally illustrate our analysis with numerical experiments.

Key words. domain decomposition, waveform relaxation, Schwarz methods

AMS subject classifications. 65F10, 65N22

DOI. 10.1137/S003614290139559X

1. Introduction.

Domain decomposition methods have been mainly developed and analyzed for elliptic coercive problems, and their convergence theory is well understood; see [42, 8, 38, 37] and references therein. When treating evolution problems, the classical approach consists of discretizing the time dimension uniformly on the whole domain by an implicit scheme and then treating the obtained problems at each time step by a classical domain decomposition method for steady problems. For parabolic problems, see, for example, [6, 30, 7], and for hyperbolic problems see [2, 41].

This approach has two disadvantages. First, one needs to impose a uniform time discretization for all subdomains, and thus one loses one of the main features of domain decomposition algorithms, namely, to adapt the solution process to the physical properties of the subdomain. It is still possible to refine in space, but for evolution problems this is not sufficient, since the space and time discretization are linked in general by stability constraints and conditions on the dispersion of the numerical scheme. Second, the algorithm needs to communicate small amounts of information at each time step. Each communication involves in addition to the cost for the data transmitted a startup cost independent of the amount of data transmitted. It can thus be of interest to communicate larger packages of data at once over several time steps instead of many small packages to save communication time. This factor can be-

*Received by the editors September 25, 2001; accepted for publication (in revised form) February 25, 2003; published electronically October 28, 2003.
http://www.siam.org/journals/sinum/41-5/39559.html

†Department of Mathematics and Statistics, McGill University, Montreal, Canada (mgander@math.mcgill.ca). Part of this work was performed when this author was visiting the CMAP with financial support from the Swiss National Science Foundation.

‡Département de Mathématiques, Université Paris XIII, 93430 Villetaneuse, France and CMAP, Ecole Polytechnique, 91128 Palaiseau, France (halpern@math.univ-paris13.fr).

§CMAP, CNRS UMR 7641, Ecole Polytechnique, 91128 Palaiseau, France (nataf@cmap.polytechnique.fr).

come important if the algorithm runs on an existing network of workstations without special high performance links.

To avoid the above disadvantages, we propose in this paper an approach different from the classical one. We decompose the original domain into subdomains as in the classical case, but we do not discretize the time dimension. Instead we solve time dependent subproblems on each subdomain. This approach is related to waveform relaxation algorithms for ordinary differential equations and has first been considered for partial differential equations by Bjørhus in [5, 4], where first order hyperbolic problems were analyzed, in which case only incoming characteristic information can be imposed on subdomain interfaces. An overlapping Schwarz algorithm of this type has been analyzed for the heat equation in [13, 19, 18], and for more general parabolic problems in [20, 14], which led to a new understanding of the performance of the waveform relaxation algorithm when applied to parabolic partial differential equations; in particular, a new and faster asymptotic convergence rate is obtained with overlapping subdomain splitting compared to the classical waveform relaxation rate for Jacobi splittings. For overlapping and nonoverlapping Schwarz waveform relaxation methods for the wave equation and convection reaction diffusion equation, see [15].

We are focusing in this paper on wave propagation phenomena in the presence of variable and discontinuous coefficients. We first perform an analysis at the continuous level and derive transmission conditions for nonoverlapping Schwarz waveform relaxation algorithms which lead to optimal convergence. The optimal transmission conditions involve linear operators $\mathcal{S}_j$ related to the Dirichlet to Neumann maps at the artificial interfaces. For elliptic problems, results of this type have been studied in [9, 34, 33, 12, 17]. These optimal transmission conditions are nonlocal in general, but we show that the nonlocality depends on the time interval under consideration in the wave equation case. We introduce then time windows to obtain optimal performance of the algorithm with local transmission conditions for piecewise constant wave speed. We show that convergence in two iterations can be achieved independently of the number of subdomains in that case. The algorithm thus scales optimally with the number of subdomains, without any additional mechanism like a coarse grid. For continuously varying coefficients, we prove convergence of the algorithm with local transmission conditions using energy estimates. We then introduce a finite volume discretization and analyze the fully discrete Schwarz waveform relaxation algorithm. This algorithm allows us to use nonmatching grids both in space and time on different subdomains so that the resolution can be adapted to the underlying physical properties of the problem. For piecewise constant wave speed, we analyze the convergence of the algorithm using discrete Laplace transforms, a tool introduced for the continuous analysis of waveform relaxation algorithms by Miekkala and Nevanlinna in [31] and later used by Nevanlinna in [35, 36]. For an analysis of waveform relaxation algorithms discretized in time, see also Janssen and Vandewalle [22] and references therein. For continuously varying wave speed, we prove stability of the subdomain problems and convergence on nonmatching grids using energy estimates. Our approach is an alternative to the mortar method for nonmatching grids [3]; see also [1]. We finally illustrate the analysis with numerical experiments for model problems and a simulation for a typical underwater sound speed profile from an application. For a different approach of a space time decomposition for evolution problems using virtual controls, see [28], and for other ways of grid refinement in space and time, see [2, 10, 25].

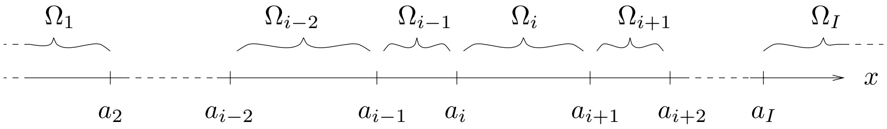


FIG. 2.1. *Domain decomposition into I nonoverlapping subdomains.*

2. The optimal Schwarz waveform relaxation algorithm.

We consider the second order wave equation with variable wave speed in one dimension,

$$\mathcal{L}(u) = \frac{1}{c^2(x)}\frac{\partial^2 u}{\partial t^2} - \frac{\partial^2 u}{\partial x^2} = f, \tag{2.1}$$

on the domain $\mathbb{R} \times (0, T)$ with initial conditions

$$u(x, 0) = p(x), \quad \frac{\partial u}{\partial t}(x, 0) = q(x).$$

For $0 < \underline{c} \leq c(x) \leq \bar{c} < \infty$ there exists a unique weak solution u of (2.1) on any bounded time interval $t \in [0, T]$; see [26, 27].

2.1. A general nonoverlapping Schwarz waveform relaxation algorithm.

We decompose the domain $\mathbb{R}$ into I nonoverlapping subdomains $\Omega_i = (a_i, a_{i+1})$, $a_j < a_i$, for $j < i$ and $a_1 = -\infty$, $a_{I+1} = +\infty$ as given in Figure 2.1. We introduce a general nonoverlapping Schwarz waveform relaxation algorithm

$$\begin{array}{rcll} \mathcal{L}(u_i^{k+1}) & = & f, & \text{in } \Omega_i \times (0, T), \\ \mathcal{B}_i^-(u_i^{k+1})(a_i, t) & = & \mathcal{B}_i^-(u_{i-1}^k)(a_i, t), & t \in (0, T), \\ \mathcal{B}_i^+(u_i^{k+1})(a_{i+1}, t) & = & \mathcal{B}_i^+(u_{i+1}^k)(a_{i+1}, t), & t \in (0, T), \\ u_i^{k+1}(x, 0) & = & p(x), & x \in \Omega_i, \\ \frac{\partial u_i^{k+1}}{\partial t}(x, 0) & = & q(x), & x \in \Omega_i, \end{array} \tag{2.2}$$

where $\mathcal{B}_i^{\pm}$ are linear transmission operators which we will determine to get optimal performance of the algorithm. For ease of notation, we define here

$$u_0^k := 0, \quad u_{I+1}^k := 0 \tag{2.3}$$

so that the index i in (2.2) ranges from $i = 1, 2, \ldots, I$. Note that we call this algorithm a waveform relaxation algorithm because time dependent problems are solved on subdomains as in the waveform relaxation algorithm for large systems of ordinary differential equations [24]. The algorithm we consider here is a Jacobi-type or additive Schwarz algorithm, since all the subdomains are treated in parallel. A Gauss–Seidel or multiplicative Schwarz algorithm could be considered as well. But since the analysis would be similar, we focus in this paper on the additive version of the algorithm only.

2.2. Transmission conditions for optimal convergence.

For elliptic problems, the Dirichlet to Neumann map has been used in [32] to define optimal transmission conditions. For wave propagation, it is more convenient to introduce the linear operator $\mathcal{S}_1(x_0)$ defined by

$$\mathcal{S}_1(x_0) : g(t) \mapsto \frac{\partial v}{\partial x}(x_0, t), \tag{2.4}$$

where $v(x,t)$ is the solution of

$$\begin{array}{rcll} \mathcal{L}(v) & = & 0, & \text{in } (-\infty, x_0) \times (0,T), \\ \frac{\partial v}{\partial t}(x_0,t) & = & g(t), & t \in (0,T), \\ v(x,0) = \frac{\partial v}{\partial t}(x,0) & = & 0, & x \in (-\infty, x_0), \end{array} \tag{2.5}$$

and the linear operator $\mathcal{S}_2(x_0)$ is defined by

$$\mathcal{S}_2(x_0) : g(t) \mapsto \frac{\partial v}{\partial x}(x_0,t), \tag{2.6}$$

where $v(x,t)$ is the solution of

$$\begin{array}{rcll} \mathcal{L}(v) & = & 0, & \text{in } (x_0, \infty) \times (0,T), \\ \frac{\partial v}{\partial t}(x_0,t) & = & g(t), & t \in (0,T), \\ v(x,0) = \frac{\partial v}{\partial t}(x,0) & = & 0, & x \in (x_0, \infty). \end{array} \tag{2.7}$$

The operators $\mathcal{S}_j$ defined in (2.5) and (2.7) are the key ingredients in obtaining an optimal Schwarz waveform relaxation algorithm. This algorithm is obtained by choosing the transmission operators $\mathcal{B}_i^{\pm}$ in the algorithm (2.2) to be

$$\mathcal{B}_i^- := \mathcal{S}_1(a_i)\partial_t - \partial_x, \quad \mathcal{B}_i^+ := \mathcal{S}_2(a_{i+1})\partial_t - \partial_x. \tag{2.8}$$

This choice is not arbitrary. The absorption property of these transmission operators allows the algorithm to compute subdomain solutions which do not see the interfaces and hence are exact; for the steady convection diffusion case, see [34].

Theorem 2.1 (convergence in I steps). *The nonoverlapping Schwarz waveform relaxation algorithm* (2.2) *with transmission operators defined by* (2.8) *converges in* I *iterations, where I denotes the number of subdomains.*

Proof. First note that convergence in less than I iterations, where I denotes the number of subdomains, is not possible over long time intervals, since the solution on each subdomain depends on the data on all the other subdomains and information is propagated only locally to neighboring subdomains. To show that the algorithm with the transmission operators (2.8) achieves convergence in I iterations and therefore is optimal, we rewrite the algorithm (2.2) in substructured form on the interfaces only. By linearity it suffices to consider the homogeneous case (the error equations) only $(f = p = q = 0)$ and to prove convergence to zero. We denote the interface values, which subdomain Ω_i obtains from its neighbors Ω_{i-1} and Ω_{i+1}, by

$$g_i^{k^-}(t) := \mathcal{B}_i^-(u_{i-1}^k)(a_i,t), \quad g_i^{k^+}(t) := \mathcal{B}_i^+(u_{i+1}^k)(a_{i+1},t) \quad \forall t \in (0,T)$$

and put all these values $g_i^{k^-}(t)$ and $g_i^{k^+}(t)$ together into the vector-valued function $g^k := (g_1^{k^+}, g_2^{k^-}, g_2^{k^+}, \ldots, g_I^{k^-})$. Note that there is only one element in g^k for the leftmost and rightmost subdomain, since they both extend to infinity. One step of the Schwarz waveform relaxation algorithm (2.2) can now be seen as a linear map taking a vector-valued function g^k as input and producing a new vector-valued function g^{k+1} as output. On each subdomain Ω_i for interior subdomains, $i = 2, \ldots I-1$, there are two linear mappings, both taking as input arguments the values of the neighboring subdomains and one producing a new value on the left boundary

$$A_i^- : (g_i^{k^-}, g_i^{k^+}) \mapsto \mathcal{B}_{i-1}^+\left(u_i^{k+1}\right)(a_i, \cdot)$$

and the other one a new value on the right boundary

$$A_i^+ : (g_i^{k^-}, g_i^{k^+}) \mapsto \mathcal{B}_{i+1}^- \left(u_i^{k+1}\right)(a_{i+1}, \cdot).$$

For the outermost subdomains there is only one linear map each, taking one input argument only,

$$A_1^+ : g_1^{k^+} \mapsto \mathcal{B}_2^- \left(u_1^{k+1}\right)(a_2, \cdot) \quad \text{and} \quad A_I^- : g_I^{k^-} \mapsto \mathcal{B}_{I-1}^+ \left(u_I^{k+1}\right)(a_I, \cdot).$$

Note that by the definition of the operators $\mathcal{S}_j$, both A_1^+ and A_I^- map any argument to zero: for any function $g(t)$ we have

$$A_1^+(g) \equiv A_I^-(g) \equiv 0. \tag{2.9}$$

This can be seen for A_1^+, for example, by

$$A_1^+(g) = \mathcal{B}_2^- (v)(a_2, \cdot) = (\mathcal{S}_1(a_2)\partial_t - \partial_x)v = v_x(a_2) - v_x(a_2) = 0,$$

where we have used that by the definition of $\mathcal{S}_1$ the function v is a solution of (2.5). Similarly for interior subdomains, we have for any function $g(t)$

$$A_i^-(g, 0) \equiv A_i^+(0, g) \equiv 0. \tag{2.10}$$

However, this implies by linearity that $A_i^+(g, h)$ depends only on g and $A_i^-(g, h)$ depends only on h, since

$$\begin{aligned} A_i^+(g, h) &= A_i^+(g, 0) + A_i^+(0, h) =: \widetilde{A}_i^+(g), \\ A_i^-(g, h) &= A_i^-(g, 0) + A_i^-(0, h) =: \widetilde{A}_i^-(h). \end{aligned} \tag{2.11}$$

Using these linear mappings on each subdomain, a complete step of the nonoverlapping Schwarz waveform relaxation algorithm can be described by the linear map $\mathcal{A} : g^k \mapsto g^{k+1}$, where $\mathcal{A}g^k$ is defined by

$$\begin{aligned} \mathcal{A}g^k &= \mathcal{A}(g_1^{k^+}, g_2^{k^-}, g_2^{k^+}, \dots, g_I^{k^-}) \\ &= (A_2^-(g_2^{k^-}, g_2^{k^+}), A_1^+(g_1^{k^+}), A_3^-(g_3^{k^-}, g_3^{k^+}), A_2^+(g_2^{k^-}, g_2^{k^+}), \dots, A_I^-(g_I^{k^-}), A_{I-1}^+(g_{I-1}^{k^-}, g_{I-1}^{k^+})) \\ &= (\widetilde{A}_2^-(g_2^{k^+}), 0, \widetilde{A}_3^-(g_3^{k^+}), \widetilde{A}_2^+(g_2^{k^-}), \dots, 0, \widetilde{A}_{I-1}^+(g_{I-1}^{k^-})) \end{aligned}$$

or, written in matrix form,

$$\mathcal{A} = \left[\begin{array}{ccccccccc} 0 & 0 & \widetilde{A}_2^- & & & & & & \\ 0 & 0 & 0 & 0 & & & & & \\ 0 & 0 & 0 & 0 & \widetilde{A}_3^- & & & & \\ & \widetilde{A}_2^+ & 0 & 0 & 0 & \ddots & & & \\ & & 0 & 0 & 0 & & 0 & & \\ & & & \widetilde{A}_3^+ & 0 & & 0 & \widetilde{A}_{I-1}^- & \\ & & & & \ddots & & 0 & 0 & 0 \\ & & & & & 0 & 0 & 0 & 0 \\ & & & & & & \widetilde{A}_{I-1}^+ & 0 & 0 \end{array}\right].$$

Now proving that the nonoverlapping Schwarz waveform relaxation algorithm (2.2) converges in I iterations is equivalent to showing that the Schwarz iteration map satisfies $\mathcal{A}^{I-1} = 0$ since then after $I-1$ iterations all the interface values are zero, and thus after one more iteration the solution will be converged to zero everywhere. We prove this by showing that $\mathcal{A}^{I-1}$ applied to an arbitrary vector-valued function $e(t) = (e_1(t), e_2(t), \ldots, e_{2I-2}(t))$ equals zero. The first application of $\mathcal{A}$ will delete the second and second to last entry in e. The second application therefore will delete the fourth and fourth to last entries in e because of the structure of $\mathcal{A}$. This process continues until the $I-1$ application of $\mathcal{A}$ deleted the $2I-2$, and the $2I-2$ to last entry, which is the first entry in e. Thus now $\mathcal{A}^{I-1}e$ is the zero vector, and in the next step the solution is zero everywhere. □

Note that the proof uses only the fundamental property of the linear operators $\mathcal{S}_j$ leading to the transparent transmission conditions and no special properties of the wave equation. The result is therefore valid for other partial differential equations as well where the appropriate operators $\mathcal{S}_j$ can be defined.

3. Optimal convergence with local transmission conditions for piecewise constant wave speed.

We now consider the wave equation (2.1) with piecewise constant wave speed to investigate the optimal transmission operators further. This will lead to the interesting result of convergence in fewer iterations than the number of subdomains on certain bounded time intervals. We consider first the wave equation (2.1) with two physical domains

$$O_1 = \mathbb{R}^- \text{ with } c(x) = c_1 \text{ and } O_2 = \mathbb{R}^+ \text{ with } c(x) = c_2. \tag{3.1}$$

In this case we can compute the linear operators $\mathcal{S}_j$ explicitly, and from them we gain more insight into the optimal Schwarz waveform relaxation algorithm. In subsection 3.3, we generalize the results to an arbitrary number of discontinuities.

3.1. Identification of the optimal nonlocal transmission conditions.

We define the ratio r by

$$r := \frac{c_2 - c_1}{c_2 + c_1}. \tag{3.2}$$

Lemma 3.1. *In the case of piecewise constant wave speed* (3.1), *the linear operators* $\mathcal{S}_j$ *in* (2.4), (2.6) *are given by*

$$(\mathcal{S}_1(x_0))\, g(t) = \begin{cases} \frac{1}{c_1} g(t), & x_0 \in O_1, \\ \frac{1}{c_2}\Big(g(t) + 2\sum_{k=1}^{\lfloor \frac{c_2 t}{2x_0} \rfloor} r^k g(t - 2kx_0/c_2)\Big), & x_0 \in O_2, \end{cases} \tag{3.3}$$

and

$$(\mathcal{S}_2(x_0))\, g(t) = \begin{cases} \frac{1}{c_1}\Big(g(t) + 2\sum_{k=1}^{\lfloor \frac{c_1 t}{2x_0} \rfloor} r^k g(t + 2kx_0/c_1)\Big), & x_0 \in O_1, \\ \frac{1}{c_2} g(t), & x_0 \in O_2. \end{cases} \tag{3.4}$$

Proof. This result can be obtained by explicitly computing the solutions; for details, see [16]. □

In this special case, one can see why the operators $\mathcal{S}_j$ are nonlocal in general: they have to include reflections stemming from the discontinuity in the wave speed between the two different physical domains. Using Theorem 2.1, the nonoverlapping Schwarz waveform relaxation algorithm for the wave equation with a discontinuity in the wave speed converges in I steps, where I denotes the number of subdomains. However, an implementation of these nonlocal transmission conditions is rather complicated, and we do not recommend this, especially if several discontinuities occur in the physical domain. However, the result indicates a better approach already, and we develop it in the next subsection.

3.2. Local transmission conditions using the time evolution. The optimal transmission operators (2.8) depend on the time interval under consideration, since the linear operators $\mathcal{S}_j$ in (3.3), (3.4) contain a sum with a number of terms proportional to the length of the time interval. In particular, for a given time interval $[0, T]$ we have at most

$$\max\left(\left\lfloor\frac{c_1 T}{2x_0}\right\rfloor, \left\lfloor\frac{c_2 T}{2x_0}\right\rfloor\right)$$

terms in the sum to obtain optimal convergence. We thus obtain the following important corollary of Theorem 2.1.

COROLLARY 3.2. *In the case of piecewise constant wave speed* (3.1), *if the discontinuity at $x = 0$ lies within the subdomain Ω_l, $a_l < 0 < a_{l+1}$, then the nonoverlapping Schwarz waveform relaxation algorithm* (2.2) *with local transmission operators*

$$\mathcal{B}_i^- := \frac{1}{c(a_i)}\partial_t - \partial_x, \quad \mathcal{B}_i^+ := \frac{1}{c(a_{i+1})}\partial_t + \partial_x \tag{3.5}$$

converges in I iterations, where I denotes the number of subdomains, if the computation is restricted to the time interval $t \in [0, T]$ with

$$T \leq T_1 = 2\min\left(\frac{|a_l|}{c(a_l)}, \frac{|a_{l+1}|}{c(a_{l+1})}\right). \tag{3.6}$$

Proof. If we choose T such that

$$\max_{1<j\leq I}\left\lfloor\frac{c(a_j)T}{2|a_j|}\right\rfloor \equiv 0, \tag{3.7}$$

then there are no terms left in the sum of the operators $\mathcal{S}_j$ in (3.3), (3.4), and thus the optimal transmission operators become the local operators (3.5). However, the maximum in condition (3.7) can only be attained for either $j = l$ or $j = l + 1$ because the discontinuity lies in subdomain Ω_l, and thus (3.7) is equivalent to the condition (3.6). □

This corollary suggests avoiding the costly nonlocal transmission conditions by cutting the given time domain $[0, T]$ into time windows of length T_1 given in (3.6). Then the algorithm can employ local transmission conditions and will still converge in at most I iterations.

However, condition (3.6) can impose very small time windows if a_l or a_{l+1} is very close to the discontinuity at $x = 0$. At first glance, this suggests that it is best to place the subdomains so that the discontinuities lie inside the subdomains, away from its boundaries. The optimal location for $a_l < 0 < a_{l+1}$ would be such that

$$\frac{|a_l|}{c(a_l)} = \frac{|a_{l+1}|}{c(a_{l+1})} \tag{3.8}$$

to maximize the time interval (3.6) one can use with the algorithm and local transmission conditions. There is, however, a better choice: taking the limit of the operators

$\mathcal{S}_j$ in (3.3,3.4) as x_0 goes to zero, we find

$$(\mathcal{S}_1(0))g(t) = \frac{1}{c_1}g(t), \quad (\mathcal{S}_2(0))g(t) = \frac{1}{c_2}g(t), \tag{3.9}$$

and thus the operators $\mathcal{S}_j$ become local operators in that case. This suggests that aligning physical domains with computational ones is an advantage for the transmission conditions. Defining

$$c(x^-) := c(x-0), \quad c(x^+) := c(x+0)$$

to include the correct limits when the discontinuity lies exactly at an interface between two subdomains, we obtain the following corollary.

COROLLARY 3.3. *In the case of piecewise constant wave speed* (3.1), *if the discontinuity lies on the interface between the two subdomains Ω_l and Ω_{l+1}, $a_{l+1} = 0$, then the nonoverlapping Schwarz waveform relaxation algorithm* (2.2) *with local transmission operators*

$$\mathcal{B}_i^- := \frac{1}{c(a_i^-)}\partial_t - \partial_x, \quad \mathcal{B}_i^+ := \frac{1}{c(a_{i+1}^+)}\partial_t + \partial_x \tag{3.10}$$

converges in I iterations, where I denotes the number of subdomains, if the computation is restricted to the time interval $t \in [0, T]$ with

$$T \leq T_2 = 2\min\left(\frac{|a_l|}{c(a_l)}, \frac{|a_{l+2}|}{c(a_{l+2})}\right). \tag{3.11}$$

Proof. If we place the discontinuity directly between the two subdomains Ω_l and Ω_{l+1}, then the optimal transmission conditions between Ω_l and Ω_{l+1} are local, as seen in (3.9). Therefore, the largest time interval we can choose for local transmission conditions depends now only on the total width of Ω_l and Ω_{l+1}, which leads to condition (3.11). □

Hence, with the discontinuity at $x = 0$ aligned with a subdomain boundary, say, at $a_{l+1} = 0$, one would choose the subdomain boundaries a_l and a_{l+2} such that

$$\frac{|a_l|}{c(a_l)} = \frac{|a_{l+2}|}{c(a_{l+2})} \tag{3.12}$$

to maximize the possible time interval in (3.11), where the algorithm can be used with local transmission conditions. This choice leads to a longer time interval than the choice with the discontinuity within one subdomain (3.8) since $a_l < a_{l+1} < a_{l+2}$.

3.3. Convergence in two iterations independent of the number of subdomains.

In more realistic situations, there will be more than one discontinuity in the computational domain, which seems to complicate the situation because for the global optimal transmission conditions of the type (3.3), (3.4), one would need to track more and more reflections from the various discontinuities in the wave speed. However, due to the finite speed of propagation in the wave equation, the previous analysis can be applied locally using time windows again. In addition, with time windows, not every subdomain solution depends on the solution on all the other subdomains if the time interval is short enough. Neighboring information suffices in that case, and it is thus possible to reduce the number of iterations below I for I subdomains.

Suppose we have J physical domains $O_j = (d_j, d_{j+1})$ with constant wave speed per physical domain, $c(x) = c_j$ for $d_j < x < d_{j+1}$, $j = 1, \ldots, J$, $d_1 = -\infty$, and

$d_{J+1} = \infty$. We decompose the physical domain $\mathbb{R}$ into I computational subdomains $\Omega_i = (a_i, a_{i+1})$ as before. We denote by n_i the number of discontinuities within each subdomain Ω_i, and we exclude for the moment the case where a discontinuity is aligned precisely between two computational subdomains. We also denote by m_i the index of the first physical domain O_{m_i} which intersects the computational subdomain Ω_i. We define the transmission time t_i of a signal across subdomain Ω_i by

$$t_i := \begin{cases} \dfrac{a_{i+1} - a_i}{c_{m_i}} & \text{if } n_i = 0, \\ \dfrac{d_{m_i+1} - a_i}{c_{m_i}} + \displaystyle\sum_{k=1}^{n_i-1} \dfrac{d_{m_i+k+1} - d_{m_i+k}}{c_{m_i+k}} + \dfrac{a_{i+1} - d_{m_i+n_i}}{c_{m_i+n_i}} & \text{if } n_i > 0. \end{cases} \tag{3.13}$$

We also define the reflection time at each interface a_i of the computational subdomains by

$$\tau_i := 2 \min_j \frac{|a_i - d_j|}{c(a_i)}. \tag{3.14}$$

These two time constants allow us to formulate conditions for convergence in less than I steps.

Theorem 3.4. *The overlapping Schwarz waveform relaxation algorithm* (2.2) *with local transmission conditions* (3.10) *and any discontinuities strictly in the interior of the computational subdomains converges in two iterations independently of the number of subdomains if the time interval* $[0, T]$ *is chosen such that*

$$T \leq T_3 = \min\left(\min_i t_i, \min_i \tau_i\right), \tag{3.15}$$

where t_i *is defined in* (3.13) *and* τ_i *is defined in* (3.14).

Proof. Consider one of the computational domains Ω_i. The solution on that domain depends only on the solution of the neighboring domains Ω_{i+1} and Ω_{i-1} determined by their initial conditions, because the time interval $[0, T]$ given by (3.15) is too short for any signal to reach domain Ω_i across the neighboring subdomains due to condition (3.15). So after one iteration, the exact boundary conditions for domain Ω_i are available if the transmission conditions employed at the boundary of Ω_i are exact absorbing boundary conditions. However, this is ensured by condition (3.15) as well because T is smaller than any reflection time τ_i so that the local transmission conditions (3.10) are indeed exactly absorbing. Thus the second iteration produces the exact solution on subdomain Ω_i. Since this argument holds for all computational subdomains, the result is established. □

As in Corollary 3.2, this result can require very small time intervals, since the reflection times τ_i can be very small when a discontinuity approaches a subdomain boundary. This can however be avoided as before by aligning physical discontinuities with the boundaries of the subdomains—a natural approach for domain decomposition. Doing this for all discontinuities, the minimal transmission time $\min_i t_i$ becomes necessarily smaller than the minimal reflection time $\min_i \tau_i$ because the reflection time requires the signal to go across a subdomain twice. (There are no discontinuities within subdomains anymore.) In addition, the transmission times formula (3.13) simplifies greatly, becoming

$$t_i = \frac{a_{i+1} - a_i}{c(a_i^+)}.$$

We therefore obtain the following theorem.

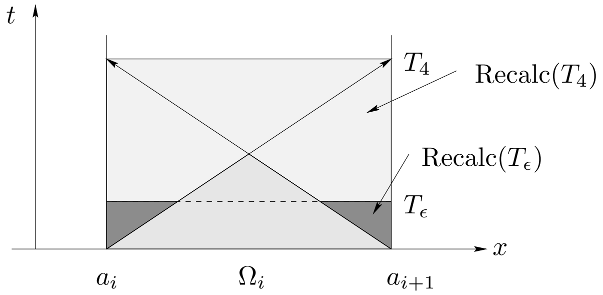


FIG. 3.1. *Toward the optimal algorithm: Unknowns to recompute for the time windows T_4 and T_ϵ, respectively.*

THEOREM 3.5. *The overlapping Schwarz waveform relaxation algorithm* (2.2) *with local transmission conditions* (3.10) *and any discontinuities aligned with the computational subdomain boundaries converges in two iterations independently of the number of subdomains if the time interval* $[0, T]$ *is chosen such that*

$$T \leq T_4 = \min_i \frac{a_{i+1} - a_i}{c(a_i^+)}. \tag{3.16}$$

Proof. The argument is the same as in the previous theorem. ☐

Further computation can be saved by noting that only values above the characteristics in each subdomain need to be recomputed during the second iteration, as shown in Figure 3.1. If the time window is chosen to be $[0, T_4]$ from Theorem 3.5, then convergence will be achieved in two iterations, and in the second iteration only the variables in the region denoted by Recalc(T_4) need to be recalculated. If we choose however an even smaller time window T_ϵ, then far fewer variables need to be recalculated in the second iteration, namely, the ones denoted by Recalc(T_ϵ). Thus, with our algorithm, the solution of the wave equation can be optimally parallelized: the parallel algorithm run on a sequential machine will run at a cost $1 + \epsilon$ of the optimal sequential code, provided the cost is linear in the number of unknowns. The optimal choice of T_ϵ depends on the latency time of the network linking the computational nodes. If the latency time is small, then a short T_ϵ will lead to the best performance, since almost no values need to be recomputed. If the latency time is important, however, it is better to communicate larger amounts of data each time a communication needs to be done. This can be achieved by choosing a larger T_ϵ and will lead to faster solution times even if more values need to be recomputed in the second iteration.

4. Convergence with local transmission conditions for continuous wave speed.

Discontinuous wave speeds allowed us to use local transmission conditions in the Schwarz waveform relaxation algorithm and still get optimal performance. If the wave speed is varying continuously, such a result cannot hold anymore, because reflections become relevant immediately. Nevertheless, the algorithm is well defined with local transmission conditions, and we prove that it converges using energy estimates. Energy estimates are useful tools for proving well-posedness of boundary or initial boundary value problems, in particular for variable coefficients. They have been used to analyze the convergence of Schwarz algorithms in the stationary case before; see, for example, [29], [11], or [33]. We extend these techniques here to time dependent problems.

Let u be a solution of the wave equation in the interval $[a,b]$ for $t \geq 0$,

$$\frac{1}{c^2(x)}\frac{\partial^2 u}{\partial t^2} - \frac{\partial^2 u}{\partial x^2} = 0, \tag{4.1}$$

where $c(x)$ is now any continuous function. We define the kinetic and potential energies by

$$E_K(u)(t) := \frac{1}{2}\int_a^b \frac{1}{c^2(x)}\left(\frac{\partial u}{\partial t}(x,t)\right)^2 dx,\ E_P(u)(t) := \frac{1}{2}\int_a^b \left(\frac{\partial u}{\partial x}(x,t)\right)^2 dx \tag{4.2}$$

and the total energy by the sum $E := E_K + E_P$. Multiplying (4.1) by $\frac{\partial u}{\partial t}$ and integrating on the interval $[a,b]$ yield the following theorem.

Theorem 4.1 (continuous energy identity). *The energy identity*

$$\frac{d}{dt}[E(u)(t)] + \frac{\partial u}{\partial t}(a,t)\frac{\partial u}{\partial x}(a,t) - \frac{\partial u}{\partial t}(b,t)\frac{\partial u}{\partial x}(b,t) = 0 \tag{4.3}$$

holds for any positive time t.

4.1. Well-posedness of the continuous subdomain problems. Introducing the general progressive and regressive transport operators

$$\mathcal{T}_\alpha^+ = \frac{1}{\alpha}\frac{\partial}{\partial t} + \frac{\partial}{\partial x}, \quad \mathcal{T}_\alpha^- = \frac{1}{\alpha}\frac{\partial}{\partial t} - \frac{\partial}{\partial x}, \tag{4.4}$$

where α is a positive real number, we can rewrite (4.3) for any positive α and β as

$$\frac{d}{dt}[E(u)(t)] + \frac{\alpha}{4}\left[\mathcal{T}_\alpha^+ u(a,t)\right]^2 + \frac{\beta}{4}\left[\mathcal{T}_\beta^- u(b,t)\right]^2 = \frac{\alpha}{4}\left[\mathcal{T}_\alpha^- u(a,t)\right]^2 + \frac{\beta}{4}\left[\mathcal{T}_\beta^+ u(b,t)\right]^2. \tag{4.5}$$

Suppose that the boundary conditions (4.4) are given by

$$\mathcal{T}_\alpha^- u(a,t) = g^-(t), \quad \mathcal{T}_\beta^+ u(b,t) = g^+(t). \tag{4.6}$$

Then we get a bound on the energy on any finite time interval.

Theorem 4.2. *For the wave equation* (4.1) *on* $[a,b]$ *with boundary conditions* (4.6), *the energy* $E(u)(t)$ *on* $[a,b]$ *stays bounded for all finite time* t,

$$E(u)(t) \leq E(u)(0) + \int_0^t \left[\frac{\alpha}{4}|g^-(\tau)|^2 + \frac{\beta}{4}|g^+(\tau)|^2\right] d\tau. \tag{4.7}$$

By standard techniques (see, for example, [26]), the well-posedness is then established.

4.2. Convergence with local transmission conditions. Consider now the domain decomposition algorithm (2.2) for continuously variable wave speed $c(x)$. By linearity, it suffices to consider the homogeneous wave equation with homogeneous initial conditions and prove convergence to zero. The local transmission operators (3.10) can be expressed in terms of the transport operators (4.4),

$$\mathcal{B}_i^- = \mathcal{T}_{c(a_i)}^-, \quad \mathcal{B}_i^+ = \mathcal{T}_{c(a_{i+1})}^+, \tag{4.8}$$

where $i = 1, \ldots, I$.

THEOREM 4.3. *Suppose the velocity is continuous on the interfaces a_i. Then on any time interval $[0,T]$, the nonoverlapping Schwarz waveform relaxation algorithm, with local transmission conditions*

$$\begin{aligned} \mathcal{T}^-_{c(a_i)}(u_i^{k+1})(a_i,\cdot) &= \mathcal{T}^-_{c(a_i)}(u_{i-1}^{k})(a_i,\cdot) && \text{on } (0,T),\\ \mathcal{T}^+_{c(a_{i+1})}(u_i^{k+1})(a_{i+1},\cdot) &= \mathcal{T}^+_{c(a_{i+1})}(u_{i+1}^{k})(a_{i+1},\cdot) && \text{on } (0,T), \end{aligned} \tag{4.9}$$

converges in the energy norm,

$$\sum_{i=1}^{I} E(u_i^k)(T) \longrightarrow 0 \text{ as } k \longrightarrow \infty.$$

Proof. We can write (4.5) on the interval $[a_i, a_{i+1}]$, with $\alpha = c(a_i)$ and $\beta = c(a_{i+1})$, which gives

$$\begin{aligned} \tfrac{d}{dt}[E(u_i^{k+1})(\cdot)] &+ \tfrac{c(a_i)}{4}\left[\mathcal{T}^+_{c(a_i)}(u_i^{k+1})(a_i,\cdot)\right]^2 + \tfrac{c(a_{i+1})}{4}\left[\mathcal{T}^-_{c(a_{i+1})}(u_i^{k+1})(a_{i+1},\cdot)\right]^2\\ &= \tfrac{c(a_i)}{4}\left[\mathcal{T}^-_{c(a_i)}(u_i^{k+1})(a_i,\cdot)\right]^2 + \tfrac{c(a_{i+1})}{4}\left[\mathcal{T}^+_{c(a_{i+1})}(u_i^{k+1})(a_{i+1},\cdot)\right]^2 \end{aligned} \tag{4.10}$$

for $1 \le i \le I$, with the convention

$$\mathcal{T}^+_{c(a_1)}(u_1^{k+1})(a_1,\cdot) = 0, \quad \mathcal{T}^-_{c(a_{I+1})}(u_I^{k+1})(a_{I+1},\cdot) = 0, \tag{4.11}$$

and by using the boundary conditions, we obtain

$$\begin{aligned} \tfrac{d}{dt}[E(u_i^{k+1})(\cdot)] &+ \tfrac{c(a_i)}{4}\left[\mathcal{T}^+_{c(a_i)}(u_i^{k+1})(a_i,\cdot)\right]^2 + \tfrac{c(a_{i+1})}{4}\left[\mathcal{T}^-_{c(a_{i+1})}(u_i^{k+1})(a_{i+1},\cdot)\right]^2\\ &= \tfrac{c(a_i)}{4}\left[\mathcal{T}^-_{c(a_i)}(u_{i-1}^{k})(a_i,\cdot)\right]^2 + \tfrac{c(a_{i+1})}{4}\left[\mathcal{T}^+_{c(a_{i+1})}(u_{i+1}^{k})(a_{i+1},\cdot)\right]^2. \end{aligned} \tag{4.12}$$

Summing these equations for $1 \le i \le I$ and shifting the indices of the two sums on the right-hand side, we find

$$\begin{aligned} \sum_{i=1}^{I}\frac{d}{dt}[E(u_i^{k+1})(\cdot)] + \sum_{i=1}^{I}\frac{c(a_i)}{4}\left[\mathcal{T}^+_{c(a_i)}(u_i^{k+1})(a_i,\cdot)\right]^2 + \sum_{i=1}^{I}\frac{c(a_{i+1})}{4}\left[\mathcal{T}^-_{c(a_{i+1})}(u_i^{k+1})(a_{i+1},\cdot)\right]^2\\ = \sum_{i=0}^{I-1}\frac{c(a_{i+1})}{4}\left[\mathcal{T}^-_{c(a_{i+1})}(u_i^{k})(a_{i+1},\cdot)\right]^2 + \sum_{i=2}^{I+1}\frac{c(a_i)}{4}\left[\mathcal{T}^+_{c(a_i)}(u_i^{k})(a_i,\cdot)\right]^2. \end{aligned} \tag{4.13}$$

Now note that by (2.3) we have $u_{I+1}^k = u_0^k = 0$ and thus $\mathcal{T}^+_{c(a_{I+1})}(u_{I+1}^k)(a_{I+1},\cdot) = \mathcal{T}^-_{c(a_1)}(u_0^k)(a_1,\cdot) = 0$. Together with (4.11), we obtain the energy equality

$$\begin{aligned} \sum_{i=1}^{I}\frac{d}{dt}[E(u_i^{k+1})(\cdot)] + \sum_{i=2}^{I}\frac{c(a_i)}{4}\left[\mathcal{T}^+_{c(a_i)}(u_i^{k+1})(a_i,\cdot)\right]^2 + \sum_{i=1}^{I-1}\frac{c(a_{i+1})}{4}\left[\mathcal{T}^-_{c(a_{i+1})}(u_i^{k+1})(a_{i+1},\cdot)\right]^2\\ = \sum_{i=2}^{I}\frac{c(a_i)}{4}\left[\mathcal{T}^+_{c(a_i)}(u_i^{k})(a_i,\cdot)\right]^2 + \sum_{i=1}^{I-1}\frac{c(a_{i+1})}{4}\left[\mathcal{T}^-_{c(a_{i+1})}(u_i^{k})(a_{i+1},\cdot)\right]^2. \end{aligned} \tag{4.14}$$

Now we have the same terms on the boundary, on the left for iteration step $k+1$ and on the right for iteration step k. Defining

$$\hat{E}^k(t) := \sum_{i=1}^{I} E(u_i^{k+1})(t),$$

$$\hat{E}_B^k(t) := \sum_{i=2}^{I} \frac{c(a_i)}{4} \left[\mathcal{T}_{c(a_i)}^{+}(u_i^k, \cdot)(a_i)\right]^2 + \sum_{i=1}^{I-1} \frac{c(a_{i+1})}{4} \left[\mathcal{T}_{c(a_{i+1})}^{-}(u_i^k)(a_{i+1}, \cdot)\right]^2,$$

we find the energy equality

$$\frac{d}{dt}\hat{E}^{k+1} + \hat{E}_B^{k+1} = \hat{E}_B^k \quad \text{on } (0,T),$$

and thus summing up over all iteration steps $k = 0, \ldots, K$ and denoting the sum of the energies $\hat{E}^k$ at each step by

$$E^K := \sum_{k=0}^{K} \hat{E}^k,$$

we find by cancellation of the $\hat{E}_B^k$ terms

$$\frac{d}{dt}E^K + \hat{E}_B^K = \hat{E}_B^0 \quad \text{on } (0,T).$$

Since $\hat{E}_B^K \geq 0$, we obtain

$$\frac{d}{dt}E^K \leq \hat{E}_B^0 \quad \text{on } (0,T).$$

Now integrating over $(0,T)$ and noting that $E^K(0) = 0$, we find

$$E^K(T) \leq \int_0^T \hat{E}_B^0(t)dt,$$

and thus the total energy E^K is uniformly bounded independently of the number of iterations K. Hence the energy at each iteration must go to zero, and the algorithm converges. □

5. A finite volume discretization. We discretize the wave equation (2.1) on each subdomain $\Omega_i \times (0,T)$, $i = 1, \ldots, I$, separately, using a finite volume discretization on rectangular grids. For simplicity we set $f = 0$. We allow nonmatching grids on different subdomains, with $J_i + 2$ points in space numbered from 0 up to $J_i + 1$ and $\Delta x_i = (a_{i+1} - a_i)/(J_i + 1)$ and $N_i + 1$ grid points in time with $\Delta t_i = T/N_i$ numbered from 0 up to N_i. Note that we chose for the exposition here uniform spacing in time per subdomain, but the techniques developed are not limited to this special case. We denote the numerical approximation to $u_i^k(a_i + j\Delta x_i, n\Delta t_i)$ on Ω_i at iteration step k by $U_i^k(j,n)$. To simplify the notation, we omit the index i on quantities depending on the subdomain and the index k referring to the iteration as long as we are discussing one subdomain only.

5.1. Discretization of the subdomain problem.

5.1.1. Interior points. Denoting by D the volume around a grid point ($x = a_i + j\Delta x_i, t = n\Delta t_i$) in the interior of subdomain $\Omega_i \times (0,T)$, as shown in Figure 5.1 on the left, we obtain the finite volume scheme by integrating the equation over the

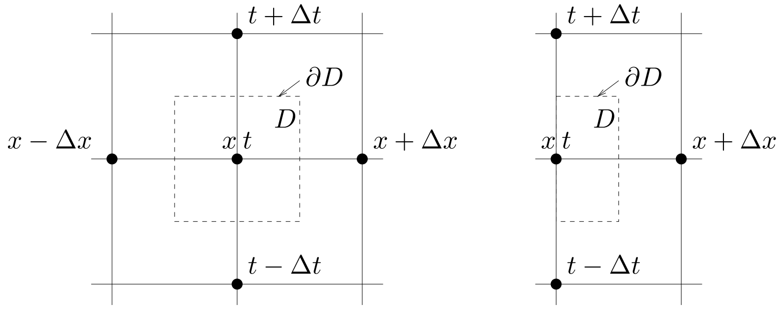


FIG. 5.1. *Control volume of an interior grid point and a boundary grid point.*

volume D and applying the divergence theorem,

$$\begin{aligned} 0 &= \int_{x-\Delta x/2}^{x+\Delta x/2}\int_{t-\Delta t/2}^{t+\Delta t/2}\left[\frac{1}{c^2(\xi)}\frac{\partial^2 u}{\partial t^2}(\xi,\tau)-\frac{\partial^2 u}{\partial x^2}(\xi,\tau)\right]d\tau d\xi \\ &= \int_{x-\Delta x/2}^{x+\Delta x/2}\frac{1}{c^2(\xi)}\frac{\partial u}{\partial t}(\xi,t+\Delta t/2)d\xi-\int_{x-\Delta x/2}^{x+\Delta x/2}\frac{1}{c^2(\xi)}\frac{\partial u}{\partial t}(\xi,t-\Delta t/2)d\xi \qquad (5.1)\\ &\quad -\int_{t-\Delta t/2}^{t+\Delta t/2}\frac{\partial u}{\partial x}(x+\Delta x/2,\tau)d\tau+\int_{t-\Delta t/2}^{t+\Delta t/2}\frac{\partial u}{\partial x}(x-\Delta x/2,\tau)d\tau. \end{aligned}$$

Now we approximate the remaining derivatives by finite differences on the grid:

$$\begin{aligned} D_t^+(U)(j,n) &:= \tfrac{U(j,n+1)-U(j,n)}{\Delta t} \approx \tfrac{\partial u}{\partial t}(\xi,t+\Delta t/2), \\ D_t^-(U)(j,n) &:= \tfrac{U(j,n)-U(j,n-1)}{\Delta t} \approx \tfrac{\partial u}{\partial t}(\xi,t-\Delta t/2), \qquad x-\tfrac{\Delta x}{2}\le\xi\le x+\tfrac{\Delta x}{2}, \\ D_x^+(U)(j,n) &:= \tfrac{U(j+1,n)-U(j,n)}{\Delta x} \approx \tfrac{\partial u}{\partial x}(x+\Delta x/2,\tau), \qquad (5.2)\\ D_x^-(U)(j,n) &:= \tfrac{U(j,n)-U(j-1,n)}{\Delta x} \approx \tfrac{\partial u}{\partial x}(x-\Delta x/2,\tau), \qquad t-\tfrac{\Delta t}{2}\le\tau\le t+\tfrac{\Delta t}{2}. \end{aligned}$$

We introduce a discrete speed function $C_i(j)$ which approximates $c(a_i+j\Delta x_i)$ through the integral relation

$$\frac{\Delta x_i}{C_i^2(j)} := \int_{x-\Delta x_i/2}^{x+\Delta x_i/2}\frac{1}{c^2(\xi)}d\xi, \qquad (5.3)$$

and we will omit the subdomain index i as long as we are on one subdomain. We thus obtain from (5.1) the discrete scheme

$$0=\frac{\Delta x}{C^2(j)}(D_t^+-D_t^-)(U)(j,n)-\Delta t(D_x^+-D_x^-)(U)(j,n),$$

which yields on using the identities $\Delta t D_t^+D_t^- = D_t^+-D_t^-$ and $\Delta x D_x^+D_x^- = D_x^+-D_x^-$ the well-known finite difference scheme

$$\left(\frac{1}{C^2(j)}D_t^+D_t^- - D_x^+D_x^-\right)(U)(j,n)=0, \quad 1\le j\le J, \qquad (5.4)$$

for points in the interior of subdomains.

5.1.2. Boundary points. So far the finite volume scheme led to a similar discretization as a finite difference scheme. On the boundary, however, the finite volume scheme leads automatically to a consistent discretization of the transmission conditions, whereas a finite difference discretization would require a special treatment. In addition, the finite volume scheme leads naturally to the correct transmission operators when using nonmatching grids in different subdomains.

Suppose the point $(x = a_i, t = n\Delta t_i)$ is on the left boundary of subdomain $\Omega_i \times (0, T)$, as shown in Figure 5.1 on the right. Then we have only half a volume D to integrate over. Proceeding as before, we obtain

$$\begin{aligned} 0 = &\int_{x}^{x+\Delta x/2} \frac{1}{c^2(\xi)} \frac{\partial u}{\partial t}(\xi, t + \Delta t/2) d\xi - \int_{x}^{x+\Delta x/2} \frac{1}{c^2(\xi)} \frac{\partial u}{\partial t}(\xi, t - \Delta t/2) d\xi \\ &- \int_{t-\Delta t/2}^{t+\Delta t/2} \frac{\partial u}{\partial x}(x + \Delta x/2, \tau) d\tau + \int_{t-\Delta t/2}^{t+\Delta t/2} \frac{\partial u}{\partial x}(x, \tau) d\tau. \end{aligned}$$

Again we can approximate $\frac{\partial u}{\partial t}$ and $\frac{\partial u}{\partial x}$ by the finite differences given in (5.2) except on the left side of the control volume where we cannot approximate $\frac{\partial u}{\partial x}(x, \tau)$ by a finite difference, since we are on the boundary and the point at $x - \Delta x$ is not available. We approximate only on the three other sides by finite differences and obtain

$$0 = \left(\frac{\Delta x}{2C^2(0)} (D_t^+ - D_t^-) - \Delta t D_x^+ \right) (U)(0, n) + \int_{t-\Delta t/2}^{t+\Delta t/2} \frac{\partial u}{\partial x}(x, \tau) d\tau. \tag{5.5}$$

Note that this equation defines the spatial derivative along the boundary, once all the grid values are known. However, to compute the grid values, we need to use the transmission condition imposed on the left boundary which also defines the spatial derivative at the boundary, since it is of the form

$$\mathcal{B}^-(u)(x, t) = \left(\frac{1}{c(x^-)} \frac{\partial u}{\partial t} - \frac{\partial u}{\partial x} \right) (x, t) = g^-(t), \tag{5.6}$$

where $g^-(t)$ is a given boundary condition. Solving for $\frac{\partial u}{\partial x}$ and integrating, we find

$$\int_{t-\Delta t/2}^{t+\Delta t/2} \frac{\partial u}{\partial x}(x, \tau) d\tau = \int_{t-\Delta t/2}^{t+\Delta t/2} \frac{1}{c(x^-)} \frac{\partial u}{\partial t}(x, \tau) d\tau - \int_{t-\Delta t/2}^{t+\Delta t/2} g^-(\tau) d\tau, \tag{5.7}$$

which gives us the missing expression for the spatial derivative in the discrete scheme (5.5). The newly introduced time derivative on the right can be approximated again by finite differences as in (5.2), on the upper part of the integral by D_t^+ and on the lower part by D_t^-. Summing those contributions, we obtain a centered finite difference,

$$D_t^0(U)(j, n) := \frac{U(j, n+1) - U(j, n-1)}{2\Delta t}, \tag{5.8}$$

and we get on denoting by $C^- := c(x^-)$ for the integral of $\frac{\partial u}{\partial x}$

$$\int_{t-\Delta t/2}^{t+\Delta t/2} \frac{\partial u}{\partial x}(x, \tau) d\tau = \frac{\Delta t}{C^-} D_t^0(U)(0, n) - \int_{t-\Delta t/2}^{t+\Delta t/2} g^-(\tau) d\tau, \tag{5.9}$$

which we insert into our scheme (5.5). Denoting the integral over the boundary condition $g^-(t)$ by

$$\Delta t G^-(n) := \int_{t-\Delta t/2}^{t+\Delta t/2} g^-(\tau)d\tau, \quad t = n\Delta t,$$

we obtain the discretization

$$0 = \left(\frac{\Delta x}{2C^2(0)} D_t^+ D_t^- - D_x^+ + \frac{1}{C^-} D_t^0\right)(U)(0,n) - G^-(n). \tag{5.10}$$

This result also defines the discrete transmission operator B^-. Comparing this with (5.6), we find (where we add now the subdomain index i for completeness)

$$B_i^-(U_i)(0,n) := \left(\frac{\Delta x_i}{2C_i^2(0)} D_t^+ D_t^- - D_x^+ + \frac{1}{C_{i-1}(J_{i-1}+1)} D_t^0\right)(U_i)(0,n), \tag{5.11}$$

where we used the fact that $C^- = C_{i-1}(J_{i-1}+1)$. Similarly, for a point (x,t) on the right boundary of a subdomain with imposed transmission condition

$$\mathcal{B}^+(u)(x,t) = \left(\frac{1}{c(x^+)}\frac{\partial u}{\partial t} + \frac{\partial u}{\partial x}\right)(x,t) = g^+(t), \tag{5.12}$$

one obtains on defining

$$\Delta t G^+(n) := \int_{t-\Delta t/2}^{t+\Delta t/2} g^+(\tau)d\tau, \quad t = n\Delta t,$$

and $C^+ := c(x^+)$ the discrete scheme

$$0 = \left(\frac{\Delta x}{2C^2(J+1)} D_t^+ D_t^- + D_x^- + \frac{1}{C^+} D_t^0\right)(U)(J+1,n) - G^+(n) \tag{5.13}$$

and thus the definition of the discrete transmission operator for subdomain i

$$B_i^+(U_i)(J_i+1,n) := \left(\frac{\Delta x_i}{2C_i^2(J_i+1)} D_t^+ D_t^- + D_x^- + \frac{1}{C_{i+1}(0)} D_t^0\right)(U_i)(J_i+1,n), \tag{5.14}$$

where we used that $C^+ = C_{i+1}(0)$.

5.1.3. Points on the initial line. Suppose $(x = a_i + j\Delta x_i, 0)$ is a grid point on the interior of the initial line of subdomain $\Omega_i \times (0,T)$. We have again half a volume D to integrate over, as shown in Figure 5.2 on the left. Integrating as before, we obtain

$$\begin{aligned} 0 = &\int_{x-\Delta x/2}^{x+\Delta x/2} \frac{1}{c^2(\xi)}\frac{\partial u}{\partial t}(\xi, \Delta t/2)d\xi - \int_{x-\Delta x/2}^{x+\Delta x/2} \frac{1}{c^2(\xi)}\frac{\partial u}{\partial t}(\xi, 0)d\xi \\ &- \int_0^{\Delta t/2} \frac{\partial u}{\partial x}(x+\Delta x/2, \tau)d\tau + \int_0^{\Delta t/2} \frac{\partial u}{\partial x}(x-\Delta x/2, \tau)d\tau. \end{aligned}$$

Now the remaining derivatives can be approximated by finite differences (5.2) except

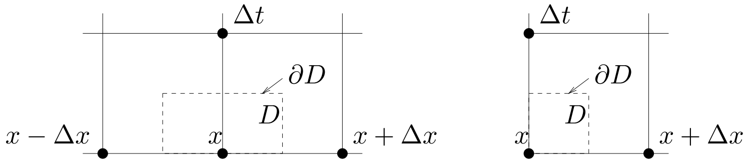


FIG. 5.2. *Control volume of a grid point on the initial line and in a corner.*

$\frac{\partial u}{\partial t}(\xi,0)$. However, this derivative is given explicitly by the initial condition, and approximating it on one grid cell by

$$\Delta x \frac{U_t(j)}{C^2(j)} := \int_{x-\Delta x/2}^{x+\Delta x/2} \frac{1}{c^2(\xi)} \frac{\partial u}{\partial t}(\xi,0)d\xi,$$

we obtain the scheme

$$\left(\frac{1}{C^2(j)}D_t^+ - \frac{\Delta t}{2}D_x^+D_x^-\right)(U)(j,0) - \frac{1}{C^2(j)}U_t(j) = 0. \tag{5.15}$$

5.1.4. Corner points. For the corner points on the initial line, there is only a quarter of the original finite volume left to integrate over. For example, on the left corner we obtain according to Figure 5.2

$$\begin{aligned} 0 = &\int_0^{\Delta x/2} \frac{1}{c^2(\xi)} \frac{\partial u}{\partial t}(\xi,\Delta t/2)d\xi - \int_0^{\Delta x/2} \frac{1}{c^2(\xi)} \frac{\partial u}{\partial t}(\xi,0)d\xi \\ &- \int_0^{\Delta t/2} \frac{\partial u}{\partial x}(\Delta x/2,\tau)d\tau + \int_0^{\Delta t/2} \frac{\partial u}{\partial x}(0,\tau)d\tau. \end{aligned}$$

Here two of the remaining derivatives can be approximated by the finite differences (5.2), whereas $\frac{\partial u}{\partial t}(\xi,0)$ is given by the initial condition and $\frac{\partial u}{\partial x}(0,\tau)$ has to be obtained from the transmission condition by proceeding as before along the boundary. We obtain the discrete scheme

$$0 = \left(\frac{\Delta x}{2C^2(0)}D_t^+ - \frac{\Delta t}{2}D_x^+ + \frac{\Delta t}{2C^-}D_t^+\right)(U)(0,0) - \frac{\Delta x}{2C^2(0)}U_t(0) - \frac{\Delta t}{2}G^-(0),$$

and thus the discrete transmission operator B^- on the initial line on the left is obtained by dividing through by $\Delta t/2$:

$$B_i^-(U_i)(0,0) = \left(\frac{\Delta x_i}{\Delta t C_i^2(0)}D_t^+ - D_x^+ + \frac{1}{C_{i-1}(J_{i-1}+1)}D_t^+\right)(U_i)(0,0) - \frac{\Delta x_i}{\Delta t C_i^2(0)}U_{t,i}(0). \tag{5.16}$$

Similarly, for the corner point on the right, we get

$$0 = \left(\frac{\Delta x}{2C^2(J+1)}D_t^+ + \frac{\Delta t}{2}D_x^- + \frac{\Delta t}{2C^+}D_t^+\right)(U)(J+1,0) - \frac{\Delta x}{2C^2(J+1)}U_t(J+1) - \frac{\Delta t}{2}G^+(0),$$

and thus the discrete transmission operator B^+ on the initial line on the right is

$$\begin{aligned} B_i^+(U_i)(J_i+1,0) = &\left(\frac{\Delta x_i}{\Delta t C_i^2(J_i+1)}D_t^+ + D_x^- + \frac{1}{C_{i+1}(0)}D_t^+\right)(U_i)(J_i+1,0) \\ &- \frac{\Delta x_i}{\Delta t C_i^2(J_i+1)}U_{t,i}(J_i+1). \end{aligned} \tag{5.17}$$

For given discrete transmission conditions $G^-(n)$ and $G^+(n)$, $n = 0, \dots, N$, the above discrete scheme describes a numerical method to solve one subproblem on one subdomain.

5.2. Extraction of the transmission conditions from neighboring subdomains. Now the boundary values $g_i^-(t)$ and $g_i^+(t)$ imposed through the transmission conditions on subdomain Ω_i have to come from the neighboring subdomains Ω_{i-1} and Ω_{i+1}. We thus need to calculate from our discrete scheme above on the neighboring subdomains the integrals

$$(5.18)\quad \int_{t_i-\Delta t_i/2}^{t_i+\Delta t_i/2} g_i^-(\tau)d\tau = \int_{t_i-\Delta t_i/2}^{t_i+\Delta t_i/2} \left[\frac{1}{c(x^-)}\frac{\partial u_{i-1}}{\partial t}(x,\tau) - \frac{\partial u_{i-1}}{\partial x}(x,\tau)\right] d\tau,$$

where (x,t) is on the right of subdomain Ω_{i-1} and similarly

$$(5.19)\quad \int_{t_i-\Delta t_i/2}^{t_i+\Delta t_i/2} g_i^+(\tau)d\tau = \int_{t_i-\Delta t/2}^{t_i+\Delta t_i/2} \left[\frac{1}{c(x^+)}\frac{\partial u_{i+1}}{\partial t}(x,\tau) + \frac{\partial u_{i+1}}{\partial x}(x,\tau)\right] d\tau,$$

where (x,t) is on the left of the subdomain Ω_{i+1}. Let us take, for example, the subdomain to the right, Ω_{i+1}. To perform the integration (5.19) we note that the numerical approximation to $\frac{\partial u_{i+1}}{\partial t}$ in the finite volume scheme is piecewise constant and according to (5.2) given by $D_t^+(U_{i+1})(0,n)$ for $t \in [n\Delta t_{i+1}, (n+1)\Delta t_{i+1})$. Similarly, the numerical approximation to $\frac{\partial u_{i+1}}{\partial x}$ is piecewise constant in the finite volume scheme. According to (5.5), it is given for $t \in [(n-\frac{1}{2})\Delta t_{i+1}, (n+\frac{1}{2})\Delta t_{i+1})$ by

$$\left(-\frac{\Delta x_{i+1}}{2\Delta t_{i+1}C_{i+1}^2(0)}(D_t^+ - D_t^-) + D_x^+\right)(U_{i+1})(0,n).$$

Inserting these two numerical approximations into (5.19), we obtain on one grid cell of Ω_{i+1}

$$\int_{(n-1/2)\Delta t_{i+1}}^{(n+1/2)\Delta t_{i+1}} g_i^+(\tau)d\tau = \left(-\frac{\Delta x_{i+1}\Delta t_{i+1}}{2C_{i+1}^2(0)}D_t^+D_t^- + \Delta t_{i+1}D_x^+ + \frac{\Delta t_{i+1}}{C_{i+1}(0)}D_t^0\right)(U_{i+1})(0,n),$$

and thus the definition of the discrete transmission operator $\widetilde{B}_i^+$ is

$$\widetilde{B}_i^+(U_{i+1})(0,n) := \left(-\frac{\Delta x_{i+1}}{2C_{i+1}^2(0)}D_t^+D_t^- + D_x^+ + \frac{1}{C_{i+1}(0)}D_t^0\right)(U_{i+1})(0,n) = \widetilde{G}_i^+(n).$$
(5.20)

Similarly, we find on the left subdomain Ω_{i-1} the discrete transmission operator $\widetilde{B}_i^-$ to be

$$\begin{aligned}\widetilde{B}_i^-(U_{i-1})(J_{i-1}+1,n) &:= \left(-\frac{\Delta x_{i-1}}{2C_{i-1}^2(J_{i-1}+1)}D_t^+D_t^- - D_x^- + \frac{1}{C_{i-1}(J_{i-1}+1)}D_t^0\right)(U_{i-1})(J_{i-1}+1,n)\\ &= \widetilde{G}_i^-(n).\end{aligned}$$
(5.21)

Note that in the discrete case $B_i^\pm$ and $\widetilde{B}_i^\pm$ are different operators, whereas in the continuous case we found the identical operator $\mathcal{B}_i^\pm$. On the initial line we find accordingly

$$\begin{aligned}
\widetilde{B}_i^+(U_{i+1})(0,0) &:= \left(-\tfrac{\Delta x_{i+1}}{\Delta t_{i+1} C_{i+1}^2(0)} D_t^+ + D_x^+ + \tfrac{1}{C_{i+1}(0)} D_t^+\right)(U_{i+1})(0,0) \\
&\quad + \tfrac{\Delta x_{i+1}}{\Delta t_{i+1} C_{i+1}^2(0)} U_t(0), \\
\widetilde{B}_i^-(U_{i-1})(J_{i-1}+1,0) &:= \left(-\tfrac{\Delta x_{i-1}}{\Delta t_{i-1} C_{i-1}^2(J_{i-1}+1)} D_t^+ - D_x^- + \tfrac{1}{C_{i-1}(J_{i-1}+1)} D_t^+\right)(U_{i-1})(J_{i-1}+1,0) \\
&\quad + \tfrac{\Delta x_{i-1}}{\Delta t_{i-1} C_{i-1}^2(J_{i-1}+1)} U_t(J_{i-1}+1),
\end{aligned} \tag{5.22}$$

and we have

$$\widetilde{B}_i^+(U_{i+1})(0,0) = \widetilde{G}_i^+(0), \quad \widetilde{B}_i^-(U_{i-1})(J_{i-1}+1,0) = \widetilde{G}_i^-(0).$$

5.3. Projections for different grids. If different grids are used on different subdomains, the extracted transmission condition $\widetilde{G}_i^+$ is a vector in $\mathbb{R}^{N_{i+1}+1}$ and $\widetilde{G}_i^-$ is a vector in $\mathbb{R}^{N_{i-1}+1}$ which both represent step functions on their corresponding grids, and what we need to impose on the boundary on Ω_i are vectors $G_i^\pm$ in $\mathbb{R}^{N_i+1}$. We thus need to introduce a projection operation to transfer the boundary values onto the grid of Ω_i. Suppose we are given a vector $\boldsymbol{v} = (v_0, \ldots, v_N) \in \mathbb{R}^{N+1}$ which represents the values of a step function on the corresponding intervals $I_n = (t_n, t_{n+1})$, where $t_0 = 0$, $t_{N+1} = T$, and $\cup_{n=0}^N \overline{I_n} = [0,T]$, and the intervals do not overlap. Then we define the scalar product on $\mathbb{R}^{N+1}$ by

$$(\boldsymbol{v}, \boldsymbol{w})_{N+1} := \sum_{n=0}^{N} |I_n| v_n w_n,$$

where $|I_n|$ denotes the length of the interval I_n. We thus obtain the induced norm on $\mathbb{R}^{N+1}$

$$||\boldsymbol{v}||_{N+1}^2 := (\boldsymbol{v}, \boldsymbol{v})_{N+1}.$$

We first define the operator $\mathbb{F} : \mathbb{R}^{N+1} \longrightarrow L^2(0,T)$, which constructs a piecewise constant function on the intervals I_n from the vector $\boldsymbol{v}$,

$$\mathbb{F} : \boldsymbol{v} \longmapsto f(t) := v_n, \quad t \in I_n.$$

Then we define the operator $\mathbb{E} : L^2(0,T) \longrightarrow \mathbb{R}^{N+1}$ which projects a given function $f(t)$ onto a vector $\boldsymbol{v} \in \mathbb{R}^{N+1}$ corresponding to a piecewise constant function in the intervals I_n:

$$\mathbb{E} : f(t) \longmapsto v_n := \frac{1}{|I_n|} \int_{I_n} f(t)dt.$$

Denoting by $\mathbb{F}_i$ and $\mathbb{E}_i$ the corresponding operators using the grid of Ω_i, we define the operator $\mathbb{P}_{i,j} : \mathbb{R}^{N_i+1} \longrightarrow \mathbb{R}^{N_j+1}$ by

$$\mathbb{P}_{i,j} := \mathbb{E}_j \circ \mathbb{F}_i. \tag{5.23}$$

A direct calculation shows that for any u in $\mathbb{R}^{N_i+1}$ we have

$$||\mathbb{P}_{i,j} u||_{N_j+1} \leq ||u||_{N_i+1}, \tag{5.24}$$

which is a natural consequence of the L^2 projection on piecewise constant functions. To perform the projection $\mathbb{P}_{i,j}$ between arbitrary grids is a nontrivial task, since one needs to find the intersections of corresponding arbitrary grid cells. For one dimension however, there is a short, concise algorithm; see the appendix.

5.4. The discrete Schwarz waveform relaxation algorithm. We obtain the discrete Schwarz waveform relaxation algorithm on subdomains Ω_i, $i = 1, \dots, I$, with nonmatching grids

$$\begin{aligned}
\left(\tfrac{1}{C_i^2(j)}D_t^+D_t^- - D_x^+D_x^-\right)(U_i^{k+1})(j,n) &= 0, \quad 1 \le j \le J_i,\ 1 \le n \le N_i,\\
\left(\tfrac{1}{C_i^2(j)}D_t^+ - \tfrac{\Delta t_i}{2}D_x^+D_x^-\right)(U_i^{k+1})(j,0) - \tfrac{1}{C_i^2(j)}U_{t,i}(j) &= 0, \quad 1 \le j \le J_i,\\
B_i^-(U_i^{k+1})(0,\cdot) &= \mathbb{P}_{i-1,i}\widetilde{B}_i^-(U_{i-1}^k)(J_{i-1}+1,\cdot),\\
B_i^+(U_i^{k+1})(J_i+1,\cdot) &= \mathbb{P}_{i+1,i}\widetilde{B}_i^+(U_{i+1}^k)(0,\cdot),
\end{aligned} \tag{5.25}$$

where the operators $\mathbb{P}_{i\pm1,i}$ are defined in (5.23), the discrete transmission operators $B_i^\pm$ for $n \ge 1$ are given in (5.11), (5.14), and the extraction operators $\widetilde{B}_i^\pm$ are for $n \ge 1$ given in (5.20), (5.21). For $n = 0$ the corresponding operators are given in (5.16), (5.17), and (5.22). For conforming grids with these transmission conditions, the solution obtained at convergence satisfies the finite volume discretization scheme without decomposition, as one can see by taking the difference of $B_i^\pm$ and $\widetilde{B}_i^\pm$. For example, for constant wave speed across the interface we obtain

$$B_i^+ - \widetilde{B}_i^+ = \frac{\Delta x}{C^2}D_t^+D_t^- + D_x^- - D_x^+ = \Delta x\left(\frac{1}{C^2}D_t^+D_t^- - D_x^-D_x^+\right),$$

which is the discretized wave operator (5.4). For nonconforming grids, (5.25) defines the solution when converged. For a different definition of a solution on nonmatching grids, see [10].

6. Normal mode analysis and convergence proof for piecewise constant wave speed and two subdomains.

We consider two subdomains $\Omega_i, i = 1, 2$, with piecewise constant velocity c_i per subdomain. We discretize the problem on each subdomain in space with spatial discretization parameter Δx_i, and we keep the time discretization uniform across the subdomains with discretization parameter Δt. Then there is no projection in the transmission operators in (5.25). We denote by γ_i the Courant–Friedrichs–Lewy (CFL) number in the corresponding subdomain Ω_i,

$$\gamma_i = c_i\frac{\Delta t}{\Delta x_i}. \tag{6.1}$$

For the stability of the Cauchy problem, we suppose that $\gamma_i < 1$ [39]. By linearity it suffices to analyze algorithm (5.25) for homogeneous initial conditions and to prove convergence to zero. To avoid the special case of the interface conditions for $n = 0$ in the analysis, we set $U(j,0) = U(j,1) = 0$, which corresponds to initial conditions $u(x,0) = u_t(x,0) = 0$.

6.1. Discrete Laplace transforms.

The discrete Laplace transform of a grid function $v = \{v_n\}_{n\ge0}$ on a regular grid with time step Δt is defined for $\eta > 0$ by [39]

$$\mathcal{L}v(s) = \hat{v}(s) = \frac{1}{\sqrt{2\pi}}\Delta t\sum_{n\ge0} e^{-sn\Delta t}v_n, \quad s = \eta + i\tau,\ |\tau| \le \frac{\pi}{\Delta t}, \tag{6.2}$$

and the inversion formula is given by

$$v_n = \frac{1}{\sqrt{2\pi}}\int_{-\frac{\pi}{\Delta t}}^{\frac{\pi}{\Delta t}} e^{sn\Delta t}\hat{v}(s)d\tau = -\frac{i}{\sqrt{2\pi}}\int_{|z|=e^{\eta\Delta t}} z^{n-1}\hat{v}(z)dz.$$

The corresponding norms are

$$(6.3)\qquad ||v||_{\eta,\Delta t} = \left(\Delta t \sum_{n\geq 0} e^{-2\eta n\Delta t}|v_n|^2\right)^{\frac{1}{2}}, \qquad ||\hat{v}||_\eta = \left(\int_{-\frac{\pi}{\Delta t}}^{\frac{\pi}{\Delta t}} |\hat{v}(\eta + i\tau)|^2 d\tau\right)^{\frac{1}{2}},$$

and we have Parseval's equality

$$(6.4)\qquad ||v||_{\eta,\Delta t} = ||\hat{v}||_\eta.$$

Suppose $U(j,n)$ is a solution of the difference equation

$$(6.5)\qquad \left(\frac{1}{C^2}D_t^+D_t^- - D_x^+D_x^-\right)(U)(j,n) = 0$$

with the initial condition $U(j,0) = U(j,1) = 0$. We denote by $\hat{U}(j,s)$ the discrete Laplace transform in time of $U(j,n)$. Equation (6.5) becomes the difference equation

$$(6.6)\qquad \gamma^2\hat{U}(j-1,s) - 2(\gamma^2 + h(z))\hat{U}(j,s) + \gamma^2\hat{U}(j+1,s) = 0,$$

with $z = e^{s\Delta t}$, $h(z) = \frac{1}{2}(z + \frac{1}{z}) - 1$, and $\gamma = c\Delta t/\Delta x$. The solutions of (6.6) are formed by powers of the roots of the second order equation

$$(6.7)\qquad \gamma^2r^2 - 2(\gamma^2 + h(z))r + \gamma^2 = 0.$$

We need several technical lemmas about these roots.

LEMMA 6.1. *For $\eta = 0$ and $\tau = 0$, (6.7) has one double root $r_\pm = 1$. For $\eta = 0$ and $|\sin(\frac{\tau\Delta t}{2})| = \gamma$, (6.7) has one double root $r_\pm = -1$.*

Proof. One can do the analysis on a case by case basis. □

LEMMA 6.2. *For $|z| > 1$ (i.e., $\eta > 0$), (6.7) has one root r_- whose modulus is strictly less than 1 and one root r_+ whose modulus is strictly bigger than 1.*

Proof. The discriminant of (6.7) is

$$(6.8)\qquad \Delta = h(z)(2\gamma^2 + h(z)),$$

and for it to vanish we have the sequence of necessary and sufficient conditions

$$\begin{aligned} \Delta = 0 \quad &\Longleftrightarrow \quad h(z) = 0 \quad \text{or} \quad 2\gamma^2 + h(z) = 0 \\ &\Longleftrightarrow \quad z = 1 \quad \text{or} \quad z + \tfrac{1}{z} - 2 + 4\gamma^2 = 0 \\ &\Longleftrightarrow \quad z = 1 \quad \text{or} \quad z = 1 - 2\gamma^2 \pm 2i\gamma\sqrt{1-\gamma^2}. \end{aligned}$$

In both cases $|z| = 1$, as a short computation in the second case shows, and $|z| = 1$ is excluded in this lemma and treated in Lemma 6.3. Hence for $|z| > 1$ there are two distinct roots whose product equals 1. Therefore, either they are a complex conjugate of modulus 1 or one is of modulus strictly bigger than 1 whereas the other is of modulus strictly less than 1. It thus remains to exclude the complex conjugate case. We find

$$\bar{r} = \frac{1}{r} \iff r + \bar{r} = r + \frac{1}{r} \iff \frac{\gamma^2 + h(z)}{\gamma^2} \in \mathbb{R} \iff h(z) \in \mathbb{R},$$

$$\bar{r} = \frac{1}{r} \iff z + \frac{1}{z} \in \mathbb{R} \iff \eta\Delta t = 0 \text{ or } \tau\Delta t = 0, \pm\pi.$$

If $\eta = 0$, we have $|z| = 1$, which is again excluded by the conditions of the lemma. If, on the other hand, $\tau\Delta t = 0$, we compute the real part of the root r and obtain

$$\Re(r) = 1 + \frac{h(z)}{\gamma^2} = 1 + \frac{2}{\gamma^2}\sinh^2\left(\frac{\eta\Delta t}{2}\right) \geq 1,$$

and if $\tau\Delta t = \pm\pi$, we have

$$\Re(r) = 1 - \frac{2}{\gamma^2}\cosh^2\left(\frac{\eta\Delta t}{2}\right) \leq -1,$$

and in both cases $|r| > 1$, a contradiction which excludes this case as well and hence proves the lemma. ◻

LEMMA 6.3. *For $|z| = 1$ (i.e., $\eta = 0$) and $|\sin(\frac{\tau\Delta t}{2})|$ different from 0 and γ, (6.7) has two distinct roots r_- and r_+,*

$$r_\pm = \begin{cases} \frac{1}{\gamma^2}\left[\gamma^2 - 2\sin^2(\frac{\tau\Delta t}{2}) \pm 2i\sin(\frac{\tau\Delta t}{2})\sqrt{\gamma^2 - \sin^2(\frac{\tau\Delta t}{2})}\right] & \textit{if } |\sin(\frac{\tau\Delta t}{2})| < \gamma, \\ \frac{1}{\gamma^2}\left[\gamma^2 - 2\sin^2(\frac{\tau\Delta t}{2}) \mp 2\,|\sin(\frac{\tau\Delta t}{2})|\sqrt{-\gamma^2 + \sin^2(\frac{\tau\Delta t}{2})}\right] & \textit{if } |\sin(\frac{\tau\Delta t}{2})| > \gamma. \end{cases}$$

Proof. For $\eta = 0$, one can compute the roots directly. The only difficulty is the determination of the signs in front of the square root. In the case $|\sin(\frac{\tau\Delta t}{2})| < \gamma$, the roots are complex conjugate. We compute the roots $r_\pm(z)$ for $z = (1+\epsilon)z_0$ with $z_0 = e^{i\theta}$ and let ϵ tend to zero. The sign is then defined by continuity. For $|\sin(\frac{\tau\Delta t}{2})| > \gamma$, there are two real roots, and the sign is determined by the fact that $|r_+| > 1$ and $|r_-| < 1$. ◻

LEMMA 6.4. *For z real positive, which corresponds to $\tau\Delta t = 0$, we have*

$$r_\pm = \frac{1}{\gamma^2}\left[\gamma^2 + 2\sinh^2\left(\frac{\eta\Delta t}{2}\right) \pm 2\sinh\left(\frac{\eta\Delta t}{2}\right)\sqrt{\gamma^2 + \sinh^2\left(\frac{\eta\Delta t}{2}\right)}\right],$$

and $0 < r_- < 1$, $r_+ > 1$.

For z real negative, which corresponds to $\tau\Delta t = \pi$, we have

$$r_\pm = \frac{1}{\gamma^2}\left[\gamma^2 - 2\cosh^2\left(\frac{\eta\Delta t}{2}\right) \mp 2\cosh\left(\frac{\eta\Delta t}{2}\right)\sqrt{-\gamma^2 + \cosh^2\left(\frac{\eta\Delta t}{2}\right)}\right],$$

and $-1 < r_- < 0$, $r_+ < -1$.

Proof. For $\tau\Delta t = 0, \pm\pi$, which means z real, one can do the analysis on a case by case basis. ◻

In all cases except for Lemma 6.1, there are functions $a_+(s)$ and $a_-(s)$ such that for all j the solution of (6.6) is given by

$$\hat{U}(j,s) = a_+(s)r_+^j + a_-(s)r_-^j. \tag{6.9}$$

In the case of Lemma 6.1 there exist functions $a(s)$ and $b(s)$ such that for all j

$$\hat{U}(j,s) = (a(s)j + b(s))r_\pm^j.$$

6.2. The discrete homogeneous subdomain problem. We consider, for example, the problem posed in Ω_1, with a nonhomogeneous boundary condition of type B_1^+,

$$\begin{aligned}
\left(\frac{1}{C^2}D_t^+D_t^- - D_x^+D_x^-\right)(U)(j,n) &= 0, && -\infty < j < 0,\ n \geq 1,\\
U(j,0) &= U(j,1) = 0, && -\infty < j < 0,\\
\left(\frac{\Delta x}{2C^2}D_t^+D_t^- + D_x^- + \frac{1}{\alpha C}D_t^0\right)(U)(0,n) &= g(n), && n \geq 1,
\end{aligned} \tag{6.10}$$

where α is a given, strictly positive real number. Applying the discrete Laplace transform, we obtain with the results of the previous subsection that every solution bounded in space is of the form

$$\hat{U}(j,s) = a_+(s)r_+^j.$$

Applying the discrete Laplace transform to the boundary condition, we get

$$\left(\frac{\Delta x}{C^2\Delta t^2}h(z) + \frac{1}{\Delta x}(1-r_-) + \frac{1}{\alpha C\Delta t}\left(z - \frac{1}{z}\right)\right)\hat{U}(0,s) = \hat{g}(s),$$

and introducing the notation

$$k(z) = \frac{1}{2}\left(z - \frac{1}{z}\right), \quad E(z,\gamma,\alpha) = \frac{1}{\gamma^2}h(z) + 1 - r_- + \frac{1}{\alpha\gamma}k(z), \tag{6.11}$$

the boundary condition becomes $E(z,\gamma,\alpha)a_+(s) = \Delta x\hat{g}(s)$. We call the problem well-posed in the sense of Gustafsson, Kreiss, and Sundstrom (GKS) if the preceding equation is invertible for all z with $|z| \geq 1$. If z_0 is such that $E(z_0,\gamma,\alpha) = 0$, we call z_0 a generalized eigenvalue [39].

THEOREM 6.5. *If $\gamma < 1$ and $|z| \geq 1$ with $z \neq 1$, then for any strictly positive α, $E(z,\gamma,\alpha) \neq 0$. The only generalized eigenvalues are $z = 1$ and if $\gamma = 1$, $z = -1$.*

Proof. Using the relation $r_+ + r_- = \frac{2}{\gamma^2}(\gamma^2 + h(z))$ satisfied by the roots of (6.7), we find

$$E(z,\gamma,\alpha) = \frac{1}{2}(r_+ - r_-) + \frac{k(z)}{\alpha\gamma}. \tag{6.12}$$

For $z = 1$ we obtain by Lemma 6.1 that $E(1,\gamma,\alpha) = 0$. If $\gamma = 1$, we also get for $z = -1$ by Lemma 6.1 that $E(-1,\gamma,\alpha) = 0$. We have to show now that there are no other generalized eigenvalues. For any generalized eigenvalue z, we must have $E(z,\gamma,\alpha) = 0$, which means

$$\frac{r_+ - r_-}{2} = -\frac{k(z)}{\alpha\gamma}. \tag{6.13}$$

Squaring both sides and using the relations of the roots r_+ and r_- of the quadratic equation (6.7) to the coefficients of that equation, we obtain

$$\frac{k^2(z)}{\alpha^2\gamma^2} = \frac{1}{4}(r_+ - r_-)^2 = \frac{1}{4}\left((r_+ + r_-)^2 - 4r_+r_-\right) = \frac{(\gamma^2 + h(z))^2}{\gamma^4} - 1 = \frac{h(z)\cdot(2\gamma^2 + h(z))}{\gamma^4}.$$

Inserting the definitions of $h(z)$ from (6.6) and $k(z)$ from (6.11) and factoring, we obtain that in order to be a generalized eigenvalue, z has to satisfy the following equation:

$$(z-1)^2\left((\gamma^2 - \alpha^2)z^2 + 2(\alpha^2 - 2\alpha^2\gamma^2 + \gamma^2)z + \gamma^2 - \alpha^2\right) = 0. \tag{6.14}$$

The first factor contains the generalized eigenvalue $z = 1$ we have found earlier. For $\gamma = 1$ the second factor contains the generalized eigenvalue $z = -1$, and for $\alpha = 1$ it contains again the generalized eigenvalue $z = 1$. It remains to show that for $\gamma < 1$ and $\alpha \neq 1$ the solutions of the second factor are introduced by the squaring, and they are not solutions to the original equation (6.13) and therefore not generalized eigenvalues. To do this, we perform the change of variables

$$a^2 = \frac{1}{\gamma^2} - 1, \quad \epsilon b^2 = \frac{1}{\alpha^2} - 1$$

with $\epsilon = \pm 1$, $\epsilon(\frac{1}{\alpha^2} - 1) > 0$, $a > 0$, $b > 0$ in the second factor of (6.14). Note that we exclude the case $b = 0 \Leftrightarrow \alpha = 1$ because then the only solution is $z = 1$. We obtain after the change of variables for the second factor of (6.14)

$$(\epsilon b^2 - a^2)z^2 + 2(\epsilon b^2 + a^2)z + \epsilon b^2 - a^2 = 0$$

or, equivalently,

$$\epsilon b^2(z+1)^2 = a^2(z-1)^2.$$

Now if $\epsilon = +1$, the only root with modulus greater than or equal to 1 is

$$z_1 = \frac{a+b}{a-b}.$$

Using Lemma 6.4, we see that the signs of $(r_+ - r_-)(z_1)$ and $k(z_1)$ are equal, which contradicts (6.13), and thus z_1 is not a generalized eigenvalue; $E(z_1) \neq 0$. If $\epsilon = -1$, there are two complex conjugate roots of modulus 1,

$$z_1 = \frac{a - ib}{a + ib} = e^{i\tau\Delta t}, \quad z_2 = \frac{a + ib}{a - ib} = e^{-i\tau\Delta t}.$$

To apply Lemma 6.3 we need to check that $|\sin(\tau\Delta t/2)|$ is different from 0 and γ. To do so, note that the real part of both z_1 and z_2 is given by $\cos\tau\Delta t = \frac{a^2-b^2}{a^2+b^2}$, and thus we obtain for $|\sin(\tau\Delta t/2)|$

$$\sin^2\frac{\tau\Delta t}{2} = \frac{1}{2} - \frac{1}{2}\cos(\tau\Delta t) = \frac{b^2}{a^2+b^2} > 0.$$

Now since $\epsilon = -1$, we have $b^2 < 1$, and therefore

$$\sin^2\frac{\tau\Delta t}{2} = \frac{b^2}{a^2+b^2} < \frac{1}{a^2+1} = \gamma^2.$$

Hence the first case of Lemma 6.3 applies and we obtain

$$r_+ - r_- = \frac{4i}{\gamma^2}\sin\left(\frac{\tau\Delta t}{2}\right)\sqrt{\gamma^2 - \sin^2\frac{\tau\Delta t}{2}}, \quad k(z) = 2i\sin\tau\Delta t,$$

which again contradicts (6.13) because of the sign. The results for z_2 are the same with a sign change. □

The values $z = 1$ (and $z = -1$ if $\gamma = 1$) are the generalized eigenvalues in the sense of GKS. Following the analysis of Trefethen [40], they correspond to stationary solutions which propagate at the same time toward the left and the right. We will see that this does not affect the convergence of the Schwarz method.

6.3. Convergence rate. We denote by $\hat{U}_i^k(j,s)$, $i=1,2$ the discrete Laplace transforms in time of the iterates $U_i^k(j,n)$, $i=1,2$, in algorithm (5.25). We obtain with the results from subsection 6.1

$$\hat{U}_1^k(j,s)=\hat{U}_1^k(0,s)r_{1,+}^j, \qquad \hat{U}_2^k(j,s)=\hat{U}_2^k(0,s)r_{2,-}^j, \tag{6.15}$$

where $r_{i,+}$and $r_{i,-}$ are the roots of the second order equation (6.7) defined in each subdomain,

$$\gamma_i^2 r^2 - 2(\gamma_i^2+h(z))r+\gamma_i^2=0, \quad \text{with } \gamma_i = c_i\tfrac{\Delta t}{\Delta x_i}.$$

The coefficients $\hat{U}_1^k(0,s)$ and $\hat{U}_2^k(0,s)$ are determined iteratively by the Laplace transform of the transmission conditions in (5.25). The discrete transmission operators in the Laplace transformed domain are given by

$$\begin{array}{rcccl} b_1^+(z) &=& E(z,\gamma_1,c_2/c_1) &=& \dfrac{1}{\gamma_1^2}h(z)+1-r_{1,-}+\dfrac{c_1}{c_2\gamma_1}k(z),\\ \widetilde{b}_1^+(z) &=& -E(z,\gamma_2,1) &=& -\dfrac{1}{\gamma_2^2}h(z)-1+r_{2,-}+\dfrac{1}{\gamma_2}k(z),\\ b_2^-(z) &=& E(z,\gamma_2,c_1/c_2) &=& \dfrac{1}{\gamma_2^2}h(z)+1-r_{2,-}+\dfrac{c_2}{c_1\gamma_2}k(z),\\ \widetilde{b}_2^-(z) &=& -E(z,\gamma_1,1) &=& -\dfrac{1}{\gamma_1^2}h(z)-1+r_{1,-}+\dfrac{1}{\gamma_1}k(z). \end{array} \tag{6.16}$$

The transmission conditions impose therefore

$$\frac{1}{\Delta x_1}b_1^+(z)\hat{U}_1^{k+1}(0,s)=\frac{1}{\Delta x_2}\widetilde{b}_1^+(z)\hat{U}_2^k(0,s),$$
$$\frac{1}{\Delta x_2}b_2^-(z)\hat{U}_2^{k+1}(0,s)=\frac{1}{\Delta x_1}\widetilde{b}_2^-(z)\hat{U}_1^k(0,s).$$

Inserting the second equation at iteration k into the first one, we find

$$\hat{U}_1^{k+1}(0,s)=\frac{\widetilde{b}_1^+(z)}{b_1^+(z)}\frac{\widetilde{b}_2^-(z)}{b_2^-(z)}\hat{U}_1^{k-1}(0,s)$$

and a similar relation for $\hat{U}_2^{k+1}(0,s)$. Defining

$$\sigma(z,\gamma)=\frac{1}{\gamma^2}h(z)+1-r_-=\frac{1}{2}(r_+-r_-), \quad \rho(z,\gamma,q)=\frac{-\sigma(z,\gamma)+\frac{1}{\gamma}k(z)}{\sigma(z,\gamma)+\frac{q}{\gamma}k(z)}, \tag{6.17}$$

we obtain for the convergence rate of the discrete Schwarz waveform relaxation algorithm

$$R(z,\gamma_1,\gamma_2,c_1/c_2):=\frac{\widetilde{b}_1^+(z)}{b_1^+(z)}\frac{\widetilde{b}_2^-(z)}{b_2^-(z)}=\rho(z,\gamma_2,c_2/c_1)\rho(z,\gamma_1,c_1/c_2), \tag{6.18}$$

and by induction we find

$$\hat{U}_i^{2k}(0,s)=R^k\hat{U}_i^0(0,s), \quad i=1,2. \tag{6.19}$$

LEMMA 6.6. *The convergence rate $R(z,\gamma_1,\gamma_2,c_1/c_2)$ is an analytic function of z for $|z|\geq 1$.*

Proof. By Theorem 6.5, for $\gamma < 1$, $z = 1$ is the only root of $E(z, \gamma, \alpha)$ such that $|z| \geq 1$. Furthermore, a Taylor expansion shows that it is a simple root; for z close to 1 we have

$$E(z, \gamma, \alpha) \approx \frac{\alpha + 1}{\alpha\gamma}(z - 1).$$

With (6.16) and (6.18) we see that $z = 1$ is only an apparent pole for R, which concludes the proof. □

Since R is analytic for $|z| \geq 1$, which corresponds to $\eta \geq 0$, R satisfies a maximum principle for $\eta \geq 0$ and hence attains its maximum on the boundary $\eta = 0$. It therefore suffices to study the behavior of R for $\eta = 0$, and we do this by studying the factors $\rho(z, \gamma, q)$ for $\eta = 0$. Setting $\omega := \frac{\tau\Delta t}{2}$, we consider ω varying between 0 and $\frac{\pi}{2}$. (The same computations apply for negative ω.) For $\eta = 0$ we have the explicit formulas

$$\rho(z, \gamma, q) = \begin{cases} \dfrac{-\sqrt{\gamma^2 - \sin^2\omega} + \gamma\cos\omega}{\sqrt{\gamma^2 - \sin^2\omega} + q\gamma\cos\omega} & \text{if } \sin\omega < \gamma, \\[2ex] \dfrac{\sqrt{\sin^2\omega - \gamma^2} + i\gamma\cos\omega}{-\sqrt{\sin^2\omega - \gamma^2} + iq\gamma\cos\omega} & \text{if } \sin\omega > \gamma, \\[2ex] \dfrac{1}{q} & \text{if } \sin\omega = \gamma. \end{cases}$$

To find a first necessary condition for convergence of the Schwarz method, we choose ω_1 such that $\sin\omega_1 = \gamma_1$, $\omega_1 \in (0, \frac{\pi}{2})$, and $q = \dfrac{c_1}{c_2}$. We obtain for the convergence rate at $\eta = 0$

$$R = \frac{1}{q}\begin{cases} \dfrac{\gamma_2\cos\omega_1 - \sqrt{\gamma_2^2 - \gamma_1^2}}{\gamma_2\cos\omega_1 + q\sqrt{\gamma_2^2 - \gamma_1^2}} & \text{if } \gamma_1 < \gamma_2, \\[2ex] \dfrac{i\gamma_2\cos\omega_1 + \sqrt{\gamma_2^2 - \gamma_1^2}}{i\gamma_2\cos\omega_1 - q\sqrt{\gamma_2^2 - \gamma_1^2}} & \text{if } \gamma_1 > \gamma_2. \end{cases}$$

In the first case, R is a real number strictly between 0 and 1, and in the second case, if $q < 1$, $|R| > 1$. We therefore have the following theorem.

THEOREM 6.7. *If the convergence rate R given in* (6.18) *of the discrete Schwarz waveform relaxation algorithm is bounded by* 1 *for all z of modulus larger than or equal to* 1, *then*

$$(c_1 - c_2)(\gamma_1 - \gamma_2) \geq 0; \tag{6.20}$$

in other words, $c_1 > c_2$ implies $\gamma_1 \geq \gamma_2$.

We now study the variations of $|\rho(z, \gamma, q)|$ for $\eta = 0$ as a function of ω. An example is shown in Figure 6.1 for the case $\gamma_1 < \gamma_2$ and $q = c_1/c_2 < 1$. The complete results are obtained by explicitly computing the derivatives and are summarized in Table 6.1. They rely on

$$\frac{d}{dz}\rho(z, \gamma, q) = \frac{(q + 1)(\gamma^2 - 1)}{\gamma^5}\frac{h(z)^2}{z\sigma(z)(\sigma(z) + \frac{q}{\gamma}k(z))^2}, \tag{6.21}$$

where h is given in (6.6), k in (6.11), and σ in (6.17). For $\omega \leq \arcsin(\gamma_2)$, we have

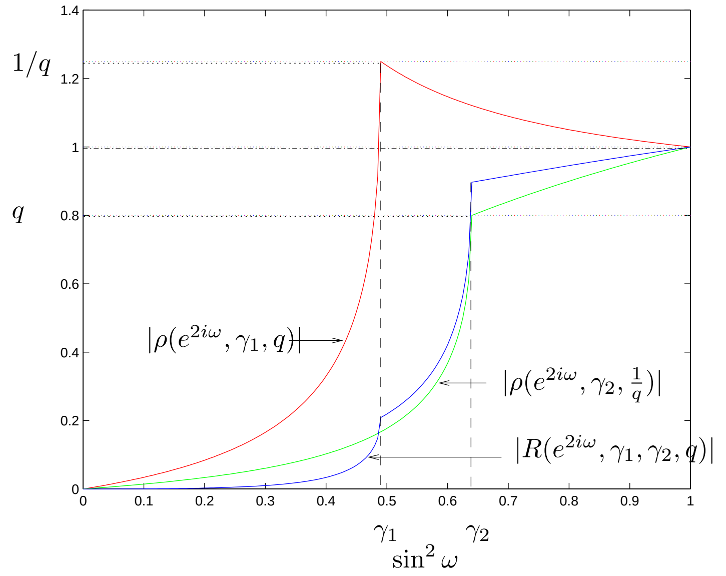


FIG. 6.1. *Example of the dependence of* $|\rho(e^{2i\omega},\gamma,q)|$ *as a function of* ω *at* $\eta=0$*,* $q=c_1/c_2$*.*

TABLE 6.1
Behavior of $|\rho(e^{2i\omega},\gamma,q)|$ *as a function of* $\omega>0$ *for* $\eta=0$*.*

	$\omega=0$	$0<\sin(\omega)<\gamma$	$\sin(\omega)=\gamma$	$\gamma<\sin(\omega)<1$	$\sin(\omega)=1$
$q<1$	0	$\nearrow$	$1/q$	$\searrow$	1
$q>1$	0	$\nearrow$	$1/q$	$\nearrow$	1

$|\rho(e^{2i\omega},\gamma_1,q)|\le 1/q$ and $|\rho(e^{2i\omega},\gamma_2,\frac{1}{q})|\le q$. By (6.21), a final explicit computation shows that the modulus of R is an increasing function of ω for $\omega\ge\arcsin(\gamma_2)$ and

$$\sup_{\omega\in[0,\frac{\pi}{2}]}|R(e^{2i\omega},\gamma_1,\gamma_2,q)|=|R(-1,\gamma_1,\gamma_2,q)|=1. \tag{6.22}$$

We therefore have the following theorem.

THEOREM 6.8. *For* $(c_1-c_2)(\gamma_1-\gamma_2)\ge 0$ *the convergence rate* $R(z,\gamma_1,\gamma_2,c_1/c_2)$ *satisfies*

$$\sup_{|z|=1}|R(z,\gamma_1,\gamma_2,c_1/c_2)|=1.$$

For purely propagating modes, $\eta=0$*, the convergence rate equals* 1*.*

For $\eta>0$, however, we have the following convergence result.

THEOREM 6.9. *For* $(c_1-c_2)(\gamma_1-\gamma_2)\ge 0$ *and* $\eta>0$ *fixed, there exists a constant* K *strictly positive such that, for* $\eta\Delta t$ *sufficiently small but nonzero, the convergence rate satisfies*

$$\sup_{|z|=e^{\eta\Delta t}}|R(z,\gamma_1,\gamma_2,c_1/c_2)|\le 1-K\eta\Delta t.$$

Proof. By (6.21) we can calculate the derivative of $|\rho|$ with respect to $\eta\Delta t$ and get

$$\frac{d}{d(\eta\Delta t)}(|\rho(z,\gamma,q)|^2)\bigg|_{\eta\Delta t=0}=2\frac{(q+1)(\gamma^2-1)}{\gamma^5}\Re\frac{h(z)^2\bar{\rho}(z)}{\sigma(z)(\sigma(z)+\frac{q}{\gamma}k(z))^2}.$$

For $\eta\Delta t = 0$, h is real, k is purely imaginary, and by Lemma 6.3 we have the explicit value of σ which is purely imaginary for $\sin(\omega) < \gamma$ and real negative for $\sin(\omega) > \gamma$. We thus have

$$\begin{cases} \frac{d}{d(\eta\Delta t)}(|\rho(z,\gamma,q)|^2)\Big|_{\eta\Delta t=0} = 0 & \text{if } \sin(\omega) < \gamma, \\ \frac{d}{d(\eta\Delta t)}(|\rho(z,\gamma,q)|^2)\Big|_{\eta\Delta t=0} < 0 & \text{if } \sin(\omega) > \gamma, \end{cases}$$

which together with (6.22) gives the desired result. □

6.4. Convergence of the discrete Schwarz waveform relaxation algorithm.

We introduce the discrete norms in space and time

$$||U||_{\Omega_i,\eta,\Delta t} = \left(\Delta t\Delta x_i \sum_{j\in\Omega_i}\sum_{n\geq 0} e^{-2\eta n\Delta t}|U(j,n)|^2\right)^{\frac{1}{2}}. \tag{6.23}$$

Theorem 6.10. *Let U_i^p be the iterates of algorithm* (5.25). *For* $(c_1-c_2)(\gamma_1-\gamma_2) \geq 0$ *there exists a positive constant K such that for $\eta\Delta t$ sufficiently small but nonzero, we have*

$$||U_i^p||_{\Omega_i,\eta,\Delta t} \leq (1-K\eta\Delta t)^{\lfloor \frac{p}{2}\rfloor} \max_{i=1,2} ||U_i^0||_{\Omega_i,\eta,\Delta t}.$$

Proof. By (6.15) and (6.19), we have $\hat{U}_i^{2k}(j,s) = R^k\hat{U}_i^0(j,s)$ for any j,s and therefore

$$\begin{aligned} ||U_i^{2k}||^2_{\Omega_i,\eta,\Delta t} &= \int_{|z|=e^{\eta\Delta t}} |R(z)|^{2k}||\hat{U}_i^0(z)||^2_{\Omega_i}dz \leq \sup_{|z|=e^{\eta\Delta t}}|R(z)|^{2k}\int_{|z|=e^{\eta\Delta t}}||\hat{U}^0(z)||^2_{\Omega_1}dz \\ &\leq \sup_{|z|=e^{\eta\Delta t}}|R(z)|^{2k}||U_i^0||^2_{\Omega_i,\eta,\Delta t} \ \leq\ (1-K\eta\Delta t)^{2k}||U_i^0||^2_{\Omega_i,\eta,\Delta t}. \end{aligned}$$

A similar argument holds for U_i^{2k+1}. □

7. Energy estimates and convergence proof for continuous wave speed.

We consider here the case of I subdomains, with a continuous velocity, and nonuniform grids in space and time. We use the same approach as in the continuous case to prove convergence of the discrete domain decomposition algorithm. Such estimates have been used in [21] in the context of discrete absorbing boundary conditions for the wave equation and in [10] to prove stability for a nonuniform scheme.

7.1. Stability for the discrete subdomain problem.

Let $U(j,n)$ for $0 \leq j \leq J+1$ and $0 \leq n \leq N$ solve the leap-frog scheme

$$\frac{1}{C^2(j)}D_t^+D_t^-(U)(j,n) - D_x^+D_x^-(U)(j,n) = 0, \quad 1 \leq j \leq J. \tag{7.1}$$

Here n stands for the discrete time variable and j for the discrete space variable. We define a discrete energy. First, we denote by $V = \{V(j)\}_{0\leq j\leq J+1}$ a sequence in $\mathbb{R}^{J+2}$, and we define for $V, W \in \mathbb{R}^{J+2}$ a bilinear form on $\mathbb{R}^{J+2}$ by

$$a_h(V,W) = \frac{\Delta x}{2}\sum_{j=1}^{J+1} D_x^-(V)(j)\cdot D_x^-(W)(j). \tag{7.2}$$

Accordingly, for any positive n, $V(n)$ stands for the sequence $\{V(j,n)\}_{0\leq j\leq J+1}$. The discrete energy E_n at time step n, global in space, is defined as the sum of a discrete kinetic energy $E_{K,n}$ and a discrete potential energy $E_{P,n}$ given by

$$\begin{aligned} E_{K,n} &= \frac{\Delta x}{2}\left[\frac{1}{2C^2(0)}(D_t^-(V)(0,n))^2 + \sum_{j=1}^{J}\frac{1}{C^2(j)}(D_t^-(V)(j,n))^2 \right. \\ &\qquad\qquad \left. + \frac{1}{2C^2(J+1)}(D_t^-(V)(J+1,n))^2\right], \\ E_{P,n} &= a_h(V(n),V(n-1)), \\ E_n &= E_{K,n} + E_{P,n}. \end{aligned} \tag{7.3}$$

The quantity $E_{K,n}$ is clearly a discrete kinetic energy. It is less evident to identify E_n as an energy. The following lemma gives a lower bound for E_n under a CFL condition and hence shows that E_n is indeed an energy.

LEMMA 7.1. *For any $n \geq 1$, we have*

$$E_n \geq \left(1 - \left(C\frac{\Delta t}{\Delta x}\right)^2\right)E_{K,n}, \tag{7.4}$$

where C is defined by $C = \sup_{1\leq j\leq J+1} C(j)$. Hence, under the CFL condition

$$C\frac{\Delta t}{\Delta x} < 1, \tag{7.5}$$

E_n is bounded from below by an energy.

Proof. For any $V, W \in \mathbb{R}^{J+2}$ we have

$$a_h(V,W) = \frac{1}{4}A_h(V+W) - \frac{1}{4}A_h(V-W) \tag{7.6}$$

with A_h defined by $A_h(V) = a_h(V,V)$. Since a_h is a positive bilinear form, the first term on the right-hand side is positive, which gives a first lower bound on E_n,

$$E_n \geq E_{K,n} - \frac{1}{4}A_h(V(n)-V(n-1)). \tag{7.7}$$

It remains to estimate the second term on the right-hand side. Using the well-known inequality

$$(a+b)^2 \leq 2(a^2+b^2), \tag{7.8}$$

we obtain

$$\begin{aligned} A_h(V(n)-V(n-1)) &= \frac{\Delta x}{2}\sum_{j=1}^{J+1}\left[D_x^-(V)(j,n) - D_x^-(V)(j,n-1)\right]^2 \\ &= \frac{\Delta x}{2}\sum_{j=1}^{J+1}\frac{\Delta t^2}{\Delta x^2}\left[D_t^-(V)(j,n) - D_t^-(V)(j-1,n)\right]^2 \\ &\leq C^2\frac{\Delta t^2}{\Delta x^2}\left[\Delta x\sum_{j=1}^{J+1}\frac{1}{C^2(j)}(D_t^-(V)(j,n))^2 \right. \\ &\qquad\qquad \left. + \Delta x\sum_{j=0}^{J}\frac{1}{C^2(j)}(D_t^-(V)(j,n))^2\right]. \end{aligned}$$

Thus

$$A_h(V(n) - V(n-1)) \leq 4C^2 \frac{\Delta t^2}{\Delta x^2} E_{K,n}, \tag{7.9}$$

which, together with (7.7), gives the desired result (7.4). □

Having defined a discrete energy, we obtain a discrete energy identity as stated in the following theorem.

THEOREM 7.2 (discrete energy identity). *For any $n \geq 1$, if $U(j,n)$ is a solution of* (7.1), *we have the energy identity*

$$\begin{aligned} E_{n+1} - E_n \ &+\ \Delta t D_t^0(U)(0,n) \cdot \left(D_x^+ - \frac{\Delta x}{2C^2(0)} D_t^+ D_t^- \right)(U)(0,n) \\ &= \Delta t D_t^0(U)(J+1,n) \cdot \left(D_x^- + \frac{\Delta x}{2C^2(J+1)} D_t^+ D_t^- \right)(U)(J+1,n). \end{aligned} \tag{7.10}$$

Furthermore, if $U(j,0)$ is a solution of (5.15), *we have the energy identity*

$$\begin{aligned} E_{K,1} + E_1 \quad &+ \quad \tfrac{\Delta t}{2} D_t^+(U)(0,0) \cdot \left(D_x^+ - \frac{1}{C^2(0)} D_t^+ \right)(U)(0,0) \\ &= \quad \frac{\Delta t}{2} D_t^+(U)(J+1,0) \cdot \left(D_x^- + \frac{1}{C^2(J+1)} D_t^+ \right)(U)(J+1,0) \\ &+ \quad \Delta x \sum_{j=1}^{J} \frac{1}{C^2(j)} U_t(j) D_t^+(U)(j,0) + \frac{\Delta x}{2} \sum_{j=1}^{J+1} (D_x^-(U)(j,0))^2. \end{aligned} \tag{7.11}$$

Proof. The proof is the discrete analogue to the proof in the continuous case. The problem here is that there is no canonical translation of the derivatives and the integrals. For $n \geq 1$, the appropriate choice is to multiply (7.1) by the centered finite differences $D_t^0(U)(j,n)$. Then we sum up for $1 \leq j \leq J$. We obtain for the derivatives in time denoted by I_1

$$\begin{aligned} I_1 &= \sum_{j=1}^{J} \frac{1}{C^2(j)} \left(D_t^+ D_t^-(U)(j,n) \right) \left(D_t^0(U)(j,n) \right) \\ &= \frac{1}{2\Delta t} \sum_{j=1}^{J} \frac{1}{C^2(j)} \left((D_t^+ - D_t^-)(U)(j,n) \right) \left((D_t^+ + D_t^-)(U)(j,n) \right) \\ &= \frac{1}{2\Delta t} \sum_{j=1}^{J} \frac{1}{C^2(j)} \left[\left(D_t^+(U)(j,n) \right)^2 - \left(D_t^+(U)(j,n-1) \right)^2 \right], \end{aligned}$$

where we used $D_t^-(U)(j,n) = D_t^+(U(j,n-1))$, and for the derivatives in space denoted by I_2

$$\begin{aligned} I_2 &= \sum_{j=1}^{J} D_x^+ D_x^-(U)(j,n) \cdot D_t^0(U)(j,n) \\ &= \frac{1}{\Delta x} \left[\sum_{j=1}^{J} D_x^+(U)(j,n) \cdot D_t^0(U)(j,n) - \sum_{j=1}^{J} D_x^-(U)(j,n) \cdot D_t^0(U)(j,n) \right]. \end{aligned}$$

By a translation of indices in the first sum of I_2 using $D_x^+(U)(j,n) = D_x^-(U(j+1,n))$, we get

$$\begin{aligned}
I_2 = {} & \frac{1}{\Delta x}\sum_{j=1}^{J+1} D_x^-(U)(j,n)\cdot\left(D_t^0(U)(j-1,n) - D_t^0(U)(j,n)\right)\\
& +\frac{1}{\Delta x}\left(-D_x^+(U)(0,n)\cdot D_t^0(U)(0,n) + D_x^-(U)(J+1,n))\cdot D_t^0(U)(J+1,n)\right)\\
= {} & \frac{1}{2\Delta t}\left[-\sum_{j=1}^{J+1} D_x^-(U)(j,n{+}1)\cdot D_x^-(U)(j,n) + \sum_{j=1}^{J+1} D_x^-(U)(j,n)\cdot D_x^-(U)(j,n-1)\right]\\
& +\frac{1}{\Delta x}\left(-D_x^+(U(0,n))\cdot D_t^0(U)(0,n) + D_x^-(U(J+1,n))\cdot D_t^0(U)(J+1,n)\right).
\end{aligned}$$

We now compute the difference $I_1 - I_2$ and find

$$\begin{aligned}
0 = {} & \frac{1}{\Delta t}\left[\frac{1}{2}\sum_{j=1}^{J}\frac{1}{C^2(j)}(D_t^-(U)(j,n+1))^2 - \frac{1}{2}\sum_{j=1}^{J}\frac{1}{C^2(j)}(D_t^-(U)(j,n))^2\right]\\
& +\frac{1}{\Delta t\Delta x}\left(a_h(U(n+1),U(n)) - a_h(U(n),U(n-1))\right)\\
& +\frac{1}{\Delta x}\left(D_x^+(U)(0,n)\cdot D_t^0(U)(0,n) - D_x^-(U)(J+1,n)\cdot D_t^0(U)(J+1,n)\right).
\end{aligned}$$

Using the definition of E_n, we finally obtain

$$\begin{aligned}
0 = {} & \frac{1}{\Delta t\Delta x}(E_{n+1} - E_n) + \frac{1}{4C^2(0)\Delta t}\left[-(D_t^+(U)(0,n))^2 + (D_t^-(U)(0,n))^2\right]\\
& +\frac{D_x^+(U)(0,n)\cdot D_t^0(U)(0,n)}{\Delta x} + \frac{-(D_t^+(U)(J{+}1,n))^2 + (D_t^-(U)(J{+}1,n))^2}{4C^2(J{+}1)\Delta t}\\
& -\frac{1}{\Delta x}D_x^-(U)(J+1,n)\cdot D_t^0(U)(J+1,n),
\end{aligned}$$

which gives (7.10) using the identities $D_t^+ - D_t^- = \Delta t D_t^+ D_t^-$ and $D_t^+ + D_t^- = 2D_t^0$. For $n = 0$, the appropriate choice is to multiply (5.15) by the forward finite difference $D_t^+(U)(j,0)$ and to perform the same computations. □

We define the discrete boundary operators

$$\begin{aligned}
T_{\alpha,C}^- &:= \frac{1}{\alpha}D_t^0 - D_x^+ + \frac{\Delta x}{2C^2}D_t^+D_t^-, & \widetilde{T}_{\alpha,C}^- &:= \frac{1}{\alpha}D_t^0 - D_x^- - \frac{\Delta x}{2C^2}D_t^+D_t^-,\\
T_{\alpha,C}^+ &:= \frac{1}{\alpha}D_t^0 + D_x^- + \frac{\Delta x}{2C^2}D_t^+D_t^-, & \widetilde{T}_{\alpha,C}^+ &:= \frac{1}{\alpha}D_t^0 + D_x^+ - \frac{\Delta x}{2C^2}D_t^+D_t^-,
\end{aligned}\tag{7.12}$$

to be applied to $U(j,n)$ for $n \geq 1$, where α is a positive real number. For $n = 0$, $D_t^+D_t^-/2$ above is replaced by $D_t^+/\Delta t$, and D_t^0 is replaced by D_t^+. Using the identity $ab = \frac{\alpha}{4}\left((\frac{1}{\alpha}a + b)^2 - (\frac{1}{\alpha}a - b)^2\right)$ for $\alpha > 0$, the energy identities (7.10), (7.11) can be rewritten for any positive α and β as

$$\begin{aligned}
E_{n+1} - E_n &+ \frac{\Delta t}{4}\Big(\alpha\big(\widetilde{T}_{\alpha,C(0)}^+(U)(0,n)\big)^2 + \beta\big(\widetilde{T}_{\beta,C(J+1)}^-(U)(J+1,n)\big)^2\Big)\\
&= \frac{\Delta t}{4}\Big(\alpha\big(T_{\alpha,C(0)}^-(U)(0,n)\big)^2 + \beta\big(T_{\beta,C(J+1)}^+(U)(J+1,n)\big)^2\Big),
\end{aligned}\tag{7.13}$$

$$
\begin{aligned}
2E_{K,1}+2E_1 \quad &+ \quad \frac{\Delta t}{4}\Big(\alpha\big(\widetilde{T}^{+}_{\alpha,C(0)}(U)(0,0)\big)^2+\beta\big(\widetilde{T}^{-}_{\beta,C(J+1)}(U)(J+1,0)\big)^2\Big)\\
&= \quad \frac{\Delta t}{4}\Big(\alpha\big(T^{-}_{\alpha,C(0)}(U)(0,0)\big)^2+\beta\big(T^{+}_{\beta,C(J+1)}(U)(J+1,0)\big)^2\Big)\\
&\qquad +2\Delta x\sum_{j=1}^{J}\frac{1}{C^2(j)}U_t(j)D_t^{+}(U)(j,0)+\Delta x\sum_{j=1}^{J+1}(D_x^{-}(U)(j,0))^2.
\end{aligned}
\tag{7.14}
$$

Suppose now the discrete boundary conditions are to be given for $n \geq 0$ by

$$
T^{-}_{\alpha,C(0)}(U)(0,n)=G^{-}(n),\quad T^{+}_{\beta,C(J+1)}(U)(J+1,n)=G^{+}(n)
\tag{7.15}
$$

and the initial conditions are to be given by $\{U(j,0)\}$, $\{U_t(j)\}$. Summing (7.13) in time and adding (7.14), we get

$$
\begin{aligned}
E_{n+1}+2E_{K,1}+E_1 \leq\ &\frac{1}{4}\Delta t\sum_{p=0}^{n}\big(\alpha(G^{-}(p))^2+\beta(G^{+}(p))^2\big)\\
&+2\Delta x\sum_{j=1}^{J}\frac{1}{C^2(j)}U_t(j)D_t^{+}(U)(j,0)+\Delta x\sum_{j=1}^{J+1}(D_x^{-}(U)(j,0))^2.
\end{aligned}
$$

Using the discrete Cauchy–Schwarz inequality on the right-hand side, we get stability for the numerical scheme.

THEOREM 7.3 (stability). *Suppose $U(j,n)$ is a solution of* (7.1), *together with initial conditions and boundary conditions* (7.15), *with α and β positive. For any positive time step n one has*

$$
\begin{aligned}
E_{n+1}+E_1 \leq\ &\frac{1}{4}\Delta t\sum_{p=0}^{n}(\alpha(G^{-}(p))^2+\beta(G^{+}(p))^2)\\
&+\ \Delta x\sum_{j=1}^{J+1}(D_x^{-}(U)(j,0))^2+\Delta x\sum_{j=1}^{J}\frac{1}{C^2(j)}(U_t(j))^2.
\end{aligned}
\tag{7.16}
$$

Thus, under the CFL condition $\sup_{1\leq j\leq J}C(j)\frac{\Delta t}{\Delta x}<1$ required in Lemma 7.1, *the scheme is stable.*

7.2. Convergence of the discrete Schwarz waveform relaxation algorithm. Corresponding to the continuous analysis, we take the velocity to be continuous at the interfaces. To shorten the notation we denote by T_i^{-} the operator $T^{-}_{C(a_i),C(a_i)}$ and the others accordingly. To analyze convergence of the discretized domain decomposition algorithm (5.25), it suffices to consider homogeneous initial conditions in (5.25) and to prove convergence to zero.

THEOREM 7.4. *Assume that the velocity is continuous on the interfaces a_i. If the CFL condition* (7.5) *is satisfied by the discretization in each subdomain, then the nonoverlapping discrete Schwarz waveform relaxation algorithm* (5.25) *with homogeneous initial condition converges to zero on any time interval $[0,T]$ in the energy norm*

$$
\sum_{i=1}^{I}E_{N_i}(U_i^k)\to 0 \text{ as } k\to\infty.
$$

Proof. The energy estimates (7.13) and (7.14) can be rewritten as

$$\begin{aligned} E_{n+1}^{k+1} - E_n^{k+1} + \frac{\Delta t_i}{4}\Big(c(a_i)\big(\widetilde{T}_i^+(U_i^{k+1})(0,n)\big)^2 + c(a_{i+1})\big(\widetilde{T}_{i+1}^-(U_i^{k+1})(J_i+1,n)\big)^2\Big) \\ = \frac{\Delta t_i}{4}\Big(c(a_i)\big(T_i^-(U_i^{k+1})(0,n)\big)^2 + c(a_{i+1})\big(T_{i+1}^+(U_i^{k+1})(J_i+1,n)\big)^2\Big), \end{aligned} \tag{7.17}$$

$$\begin{aligned} 2E_{K,1} + 2E_1 + \frac{\Delta t_i}{4}\Big(c(a_i)\big(\widetilde{T}_i^+(U_i^{k+1})(0,0)\big)^2 + c(a_{i+1})\big(\widetilde{T}_{i+1}^-(U_i^{k+1})(J_i+1,0)\big)^2\Big) \\ = \frac{\Delta t_i}{4}\Big(c(a_i)\big(T_i^-(U_i^{k+1})(0,0)\big)^2 + c(a_{i+1})\big(T_{i+1}^+(U_i^{k+1})(J_i+1,0)\big)^2\Big). \end{aligned} \tag{7.18}$$

We define the boundary energies

$$\begin{aligned} &\widetilde{F}_{i,n}^{k,+} = \tfrac{\Delta t_i}{4} c(a_i)(\widetilde{T}_i^+(U_i^k)(0,n))^2, && \widetilde{F}_{i,n}^{k,-} = \tfrac{\Delta t_{i-1}}{4} c(a_i)(\widetilde{T}_i^-(U_{i-1}^k)(J_{i-1}+1,n))^2, \\ &F_{i,n}^{k,+} = \tfrac{\Delta t_{i-1}}{4} c(a_i)(T_i^+(U_{i-1}^k)(J_{i-1}+1,n))^2, && F_{i,n}^{k,-} = \tfrac{\Delta t_i}{4} c(a_i)(T_i^-(U_i^k)(0,n))^2 \end{aligned} \tag{7.19}$$

and rewrite (7.17) and (7.18) as

$$\begin{aligned} [E_{n+1} - E_n](U_i^{k+1}) + \widetilde{F}_{i,n}^{k+1,+} + \widetilde{F}_{i+1,n}^{k+1,-} &= F_{i+1,n}^{k+1,+} + F_{i,n}^{k+1,-}, \\ [2E_{K,1} + 2E_1]\,(U_i^{k+1}) + \widetilde{F}_{i,0}^{k+1,+} + \widetilde{F}_{i+1,0}^{k+1,-} &= F_{i+1,0}^{k+1,+} + F_{i,0}^{k+1,-}. \end{aligned} \tag{7.20}$$

Summing these equations in every subdomain for $1 \le n \le N_i$, we find

$$\begin{aligned} [E_{N_i+1} + 2E_{K,1} + E_1](U_i^{k+1}) + \sum_{n=0}^{N_i} \widetilde{F}_{i,n}^{k+1,+} + \sum_{n=0}^{N_i} \widetilde{F}_{i,n}^{k+1,-} \\ = \sum_{n=0}^{N_i} F_{i,n}^{k+1,+} + \sum_{n=0}^{N_i} F_{i,n}^{k+1,-}. \end{aligned} \tag{7.21}$$

Using now the transmission conditions and the fact that the projection is a contraction in L^2, we get

$$\begin{aligned} [E_{N_i+1} + 2E_{K,1} + E_1](U_i^{k+1}) + \sum_{n=0}^{N_i} \widetilde{F}_{i,n}^{k+1,+} + \sum_{n=0}^{N_i} \widetilde{F}_{i,n}^{k+1,-} \\ \le \sum_{n=0}^{N_{i+1}} \widetilde{F}_{i+1,n}^{k,+} + \sum_{n=0}^{N_{i-1}} \widetilde{F}_{i-1,n}^{k,-}. \end{aligned} \tag{7.22}$$

Note now that by definition we have as in the continuous case $F_1^{k,\pm} = F_{I+1}^{k,\pm} = 0$. Thus, defining the total boundary energy at iteration k by

$$\widetilde{F}^k = \sum_{i=1}^{I} \sum_{n=0}^{N_i} [\widetilde{F}_{i,n}^{k,-} + \widetilde{F}_{i,n}^{k,+}],$$

we have, by summing in i and shifting the indices, the inequality

$$\sum_{i=1}^{I} [E_{N_i+1} + 2E_{K,1} + E_1](U_i^{k+1}) + \widetilde{F}^{k+1} \le \widetilde{F}^k. \tag{7.23}$$

Thus the same arguments as in the continuous case prove that $\sum_{i=1}^{I} E_{N_i}(U_i^k) \to 0$ as $k \to \infty$. □

8. Numerical experiments.

We perform the numerical experiments on the wave equation

$$\frac{\partial^2 u}{\partial t^2} = c^2(x)\frac{\partial^2 u}{\partial x^2}, \quad 0 < x < L,\ 0 < t < T, \tag{8.1}$$

where we truncate the spatial domain at 0 and L using absorbing boundary conditions so that the results obtained correspond to the analysis on an infinite domain. We discretize the wave equation and the optimal transmission conditions using the finite volume scheme presented in section 5.

8.1. Optimal global transmission conditions.

First we test the convergence result proved in Theorem 2.1 for a two subdomain problem with $L = 2$, $T = 2$, and a constant wave speed $c(x) = 1$. The domain is partitioned at $x = 1$, and the optimal transmission conditions at the continuous level are local. We use the initial conditions

$$\begin{aligned} u(x,0) &= 0, \\ \frac{\partial u}{\partial t}(x,0) &= -100(0.5-x)e^{-50(0.5-x)^2}, \end{aligned}$$

and we start the iteration with the initial guess zero. Table 8.1 shows for various mesh parameters the difference of the domain decomposition algorithm result after two iterations and the numerical solution on the whole domain and compares this value to the truncation error, the difference between the numerical solution on the whole domain, and the exact solution. One can see that the discretization of the optimal local transmission conditions leads to an algorithm which converges in two iterations to well below the accuracy of the numerical scheme.

Table 8.1
Convergence in two iterations to below the accuracy of the discretization.

Grid	Error after 2 iterations	Discretization error
50 x 50	2.6128e-04	2.1515e-02
100 x 100	2.7305e-05	4.9472e-03
200 x 200	3.2361e-06	1.2218e-03
400 x 400	3.9852e-07	3.0321e-04
800 x 800	4.9548e-08	7.5567e-05

For the next model problem, we choose $L = 6$, $T = 8$, and a speed function $c(x)$ with a discontinuity at $x = 1$,

$$c(x) = \begin{cases} 1 & 1 < x < 6, \\ 2 & 0 < x < 1. \end{cases}$$

We decompose the domain into three subdomains, $\Omega_1 = [0,2]\times[0,6]$, $\Omega_2 = [2,4]\times[0,6]$, and $\Omega_3 = [4,6]\times[0,6]$, and we use the initial conditions

$$\begin{aligned} u(x,0) &= 0, \\ \frac{\partial u}{\partial t}(x,0) &= -20(5-x)e^{-10(5-x)^2}. \end{aligned}$$

We use again a discretization of the optimal transmission conditions, which are nonlocal in this case. We start the Schwarz waveform relaxation with a zero initial guess. Table 8.2 shows that the algorithm converges at the third iteration to the discretization error level and illustrates Theorem 2.1, which states that we should find convergence in a number of iterations identical to the number of subdomains. More accuracy is achieved as the iteration progresses further. The nonlocal transmission conditions (2.8) require this value of T to include three terms of the sum of the operators $\mathcal{S}_j$ in (3.3), (3.4) on the middle subdomain and one term on the right subdomain. We computed the solution on a uniform grid with $\Delta x = 1/20$ and $\Delta t = 1/40$ for this example.

Table 8.2

Convergence for a 3 *subdomain problem and a discontinuity in the left subdomain.*

Iteration	0	1	2	3	4	5
$\lvert\lvert u - u_k\rvert\rvert_\infty$	5.0230e-01	5.0164e-01	5.0065e-01	4.0289e-03	3.9800e-03	3.9535e-05

8.2. Optimal local transmission conditions. Now we introduce time windows for the same example to be able to use optimal local transmission conditions. We cut the time domain into four equal pieces $[0, 2]$, $[2, 4]$, $[4, 6]$, and $[6, 8]$ such that the condition (3.15) according to Theorem 3.4 is satisfied. We solve the problem consecutively on the four time subdomains, and in each time window we expect the algorithm to converge in two iterations. We show the convergence results for the first time window only, $[0, T = 2]$. Table 8.3 shows the error in the infinity norm over five iterations for the same mesh parameters as before. The algorithm converges now already at the second iteration as predicted by Theorem 3.4 to the discretization error level, and more accuracy is achieved as the iteration progresses.

Table 8.3

Convergence with local transmission conditions over a shorter time interval.

Iteration	0	1	2	3	4	5
$\lvert\lvert u - u_k\rvert\rvert_\infty$	5.0200e-01	5.0135e-01	1.2089e-03	9.0654e-06	1.4775e-06	1.2551e-06

8.3. Nonconforming grids. As an illustration of Theorem 3.5 on nonconforming grids, we consider a problem with a layered medium of six layers, and we decompose the domain into six subdomains corresponding to the different layers, $\Omega_i = [i - 1, i]$ with corresponding wave speeds $c_i \in \{1, 2/3, 1/2, 3/4, 1, 4/5\}$. We discretize each subdomain with a grid in space using $\Delta x_i = 1/50$ and in time using an appropriate time step satisfying the CFL condition $c_i \frac{\Delta t_i}{\Delta x_i} < 1$ but close to 1, which is important for accuracy in the propagation properties of the solution, so different time steps are essential in different subdomains. This leads to nonconforming grids between subdomains. Since we have no algorithm that computes the entire solution over nonconforming grids to compare with, we choose to compute the zero solution to the homogeneous problem with zero initial conditions. We start with a nonzero initial guess on the artificial interfaces, $g^{\pm}(t) = 1$. According to Theorem 3.5 the algorithm will converge in two iterations if $T \leq 1$. Table 8.4 shows that this is also observed numerically. After two iterations the Schwarz waveform relaxation has converged to the precision of the numerical scheme. Figure 8.1 shows a solution computed on nonmatching grids with the optimal Schwarz waveform relaxation algorithm.

8.4. Variable wave speed and local transmission conditions. Now we consider a variable propagation speed $c(x)$ for which convergence of the Schwarz waveform relaxation algorithm with local transmission conditions was proved in Theorem 4.3

Table 8.4

Convergence with local transmission conditions in 2 iterations to the level of the truncation error for a problem with five discontinuities and six subdomains aligned with the discontinuities.

Iteration	0	1	2
$\|\|u-u_k\|\|_\infty$	5.0234e+00	5.0234e+00	1.1738e-02

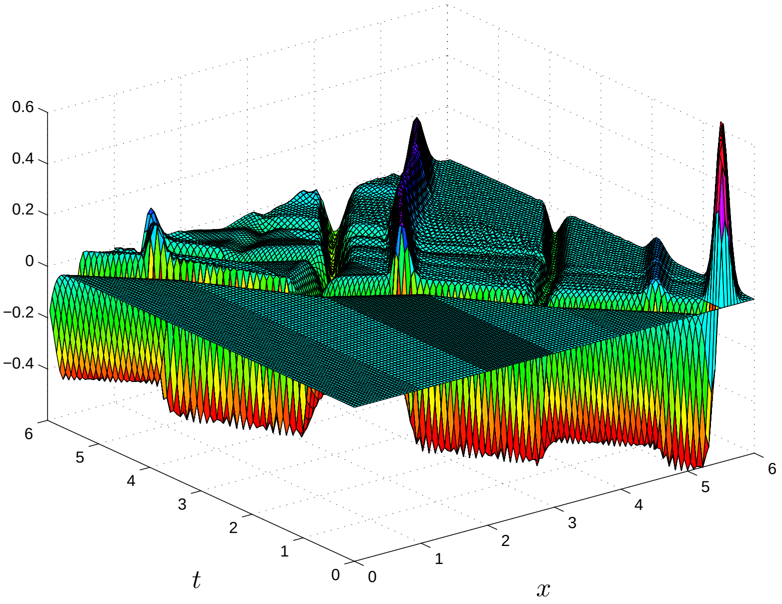


Fig. 8.1. *Computation with nonmatching grids on a layered medium with five discontinuities, a pulse created in the surface layer and propagating downward.*

using energy estimates. The speed profile, which is a typical underwater profile, was obtained from [23], and it is given as a function of depth by

$$c(x) = \begin{cases} 300, & \frac{m}{s}, \quad x < 0 \text{ (above ground)}, \\ 1500 - x/12, & \frac{m}{s}, \quad 0 < x < 120, \\ 1480 + x/12, & \frac{m}{s}, \quad 120 < x < 240, \\ 1505, & \frac{m}{s}, \quad x > 240. \end{cases}$$

We decompose the domain into two subdomains $\Omega_1 = [0, 300]$, and $\Omega_2 = [300, 600]$, and we apply the domain decomposition algorithm with the local transmission conditions (3.10), which would be exact if the sound speed was identically constant over both subdomains and equal to the sound speed at the artificial interface at $x = 300$. Table 8.5 shows the convergence of the algorithm for the variable sound speed for a time interval $[0, 1/2]$. The algorithm converges again to the accuracy of the scheme in two iterations, even though the sound speed is variable in this example. This is because the variation is small in scale, and thus the local approximations to the transmission conditions are sufficiently accurate to lead to the convergence in two steps. Note also that continuing the iteration, the error is further reduced.

Table 8.5

Convergence behavior of the algorithm for the variable sound speed profile from an application.

Iteration	0	1	2	3
$\|\|u - u_k\|\|_\infty$	5.0316e+00	5.0316e+00	9.9024e-03	6.2439e-04

9. Conclusions. We have presented and analyzed a nonoverlapping Schwarz waveform relaxation algorithm for the one dimensional wave equation with variable coefficients, both at the continuous and the discrete level. The algorithm permits the use of grid CFL conditions adapted to the local wave speeds and nonmatching grids on different subdomains, and it has optimal scalability when implemented on a parallel computer. The formulation of the algorithm is quite general; it can be applied to the wave equation in higher dimensions and even to other types of evolution equations. The convergence result with the optimal transmission conditions also holds in these more general situations; specific results for the wave equation in higher dimensions will be presented elsewhere. The convergence analysis for the discretized algorithm with approximate transmission conditions, however, is specific to the one dimensional wave equation with variable coefficients. Although the ideas can be generalized to higher dimensions, the discrete energy estimates present a real challenge. The advantage of the continuous analysis is, however, that convergence results similar to the continuous ones can be expected to hold when the algorithm is consistently discretized.

Appendix. A projection algorithm for nonmatching grids.

The projection operation between nonconforming grids as given in (5.23) is not an easy task in an algorithm, since one cell can intersect with an arbitrary number of neighboring ones or even not intersect at all if it is fully contained in the neighboring one. In one dimension, the following short algorithm in Matlab performs this task in an efficient manner:

```
function b=transfer(a,ta,tb);
% TRANSFER integrates a stepfunction between given intervals
%   b=transfer(a,ta,tb); computes the integral of the
%   stepfunction with values a(j) in [ta(j),ta(j+1)] in the
%   intervals [tb(i),tb(i+1)] and stores the result in b(i).
%   Note that the first and last entry in ta and tb must be equal.

n=length(tb);           % n-1 is the length of b
ta(length(ta))=tb(n);   % numerical equality for proper termination
j=1;
for i=1:n-1,
b(i)=0;
m=ta(j+1)-tb(i);
while ta(j+1)<tb(i+1),
b(i)=b(i)+m*a(j);
j=j+1;
m=ta(j+1)-ta(j);
end;
m=m-(ta(j+1)-tb(i+1));
b(i)=b(i)+m*a(j);
end;
```

Given a vector $\texttt{ta} = [0, t_a(1), \ldots, T]$ of arbitrary grid points in time and a piecewise constant function on the intervals between the grid points ta whose values are given in the vector a, the algorithm computes the integrals of a on the intervals between the grid points of a second grid given in the vector $\texttt{tb} = [0, t_b(1), \ldots, T]$. The algorithm does a single pass without any special cases using the fact that the grid points are sorted in time. It advances automatically on whatever side the next cell boundary is coming and handles any possible cases of nonmatching grids at a one dimensional interface.

Acknowledgment. We would like to thank the unknown referee for his detailed comments and suggestions, which were very helpful in the preparation of the final version of the manuscript.

REFERENCES

[1] Y. Achdou, C. Japhet, Y. Maday, and F. Nataf, *A new cement to glue non-conforming grids with Robin interface conditions: The finite volume case*, Numer. Math., 92 (2002), pp. 593–620.

[2] A. Bamberger, R. Glowinski, and Q. H. Tran, *A domain decomposition method for the acoustic wave equation with discontinuous coefficients and grid change*, SIAM J. Numer. Anal., 34 (1997), pp. 603–639.

[3] C. Bernardi, Y. Maday, and A. T. Patera, *Domain decomposition by the mortar element method*, in Asymptotic and Numerical Methods for Partial Differential Equations with Critical Parameters, H. G. Kaper, M. Garby, and G. W. Pieper, eds., NATO Adv. Sci. Inst. Ser. C Math. Phys. Sci. 384, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1993, pp. 269–286.

[4] M. Bjørhus, *A note on the convergence of discretized dynamic iteration*, BIT, 35 (1995), pp. 291–296.

[5] M. Bjørhus, *On Domain Decomposition, Subdomain Iteration and Waveform Relaxation*, Ph.D. thesis, University of Trondheim, Trondheim, Norway, 1995.

[6] X.-C. Cai, *Additive Schwarz algorithms for parabolic convection-diffusion equations*, Numer. Math., 60 (1991), pp. 41–61.

[7] X.-C. Cai, *Multiplicative Schwarz methods for parabolic problems*, SIAM J. Sci. Comput., 15 (1994), pp. 587–603.

[8] T. F. Chan and T. P. Mathew, *Domain decomposition algorithms*, in Acta Numerica 1994, Acta Numer., Cambridge University Press, Cambridge, UK, 1994, pp. 61–143.

[9] P. Charton, F. Nataf, and F. Rogier, *Méthode de décomposition de domaine pour l'équation d'advection-diffusion*, C. R. Acad. Sci. Paris Sèr. I Math., 313 (1991), pp. 623–626.

[10] F. Collino, T. Fouquet, and P. Joly, *Une méthode de raffinement de maillage espace-temps pour le système de Maxwell* 1*-d*, C. R. Acad. Sci. Paris Sèr. I Math., 328 (1999), pp. 263–268.

[11] B. Després, P. Joly, and J. E. Roberts, *A domain decomposition method for the harmonic Maxwell equations*, in Iterative Methods in Linear Algebra (Brussels, 1991), North–Holland, Amsterdam, 1992, pp. 475–484.

[12] B. Engquist and H.-K. Zhao, *Absorbing boundary conditions for domain decomposition*, Appl. Numer. Math., 27 (1998), pp. 341–365.

[13] M. J. Gander, *Overlapping Schwarz for parabolic problems*, in Proceedings of the Ninth International Conference on Domain Decomposition Methods, P. E. Bjørstad, M. Espedal, and D. Keyes, eds., Bergen, Norway, 1997, pp. 97–104.

[14] M. J. Gander, *Overlapping Schwarz waveform relaxation for parabolic problems*, in Tenth International Conference on Domain Decomposition Methods, Contemp. Math. 218, J. Mandel, C. Farhat, and X.-C. Cai, eds., AMS, Providence, RI, 1998, pp. 425–431.

[15] M. J. Gander, L. Halpern, and F. Nataf, *Optimal convergence for overlapping and non-overlapping Schwarz waveform relaxation*, in Eleventh International Conference of Domain Decomposition Methods, C.-H. Lai, P. Bjørstad, M. Cross, and O. Widlund, eds., DDM.org, Augsburg, 1999, pp. 27–36.

[16] M. J. Gander, L. Halpern, and F. Nataf, *Optimal Schwarz Waveform Relaxation for the One Dimensional Wave Equation*, Tech. rep. 469, CMAP, Ecole Polytechnique, Palaiseau, France, 2001.

[17] M. J. Gander, L. Halpern, and F. Nataf, *Optimized Schwarz methods*, in Twelfth International Conference on Domain Decomposition Methods (Chiba, Japan), T. Chan, T. Kako, H. Kawarada, and O. Pironneau, eds., Domain Decomposition Press, Bergen, 2001, pp. 15–28.

[18] M. J. Gander and A. M. Stuart, *Space-time continuous analysis of waveform relaxation for the heat equation*, SIAM J. Sci. Comput., 19 (1998), pp. 2014–2031.

[19] M. J. Gander and H. Zhao, *Overlapping Schwarz waveform relaxation for parabolic problems in higher dimension*, in Proceedings of Algoritmy 14, A. Handlovičová, M. Komorníkova, and K. Mikula, eds., Slovak Technical University, 1997, pp. 42–51.

[20] E. Giladi and H. Keller, *Space time domain decomposition for parabolic problems*, Numer. Math., 93 (2002), pp. 279–313.

[21] L. Halpern, *Absorbing boundary conditions for the discretization schemes of the one-dimensional wave equation*, Math. Comp., 38 (1982), pp. 415–429.

[22] J. Janssen and S. Vandewalle, *Multigrid waveform relaxation on spatial finite element meshes: The discrete-time case*, SIAM J. Sci. Comput., 17 (1996), pp. 133–155.

[23] D. Lee, G. Botseas, and J. S. Papadakis, *Finite difference solution to the parabolic wave equation*, J. Acoustic Society of America, 70 (1981), pp. 798–799.

[24] E. Lelarasmee, A. E. Ruehli, and A. L. Sangiovanni-Vincentelli, *The waveform relaxation method for time-domain analysis of large scale integrated circuits*, IEEE Trans. Computer-Aided Design of Integrated Circuits and Systems, 1 (1982), pp. 131–145.

[25] J.-L. Lions, Y. Maday, and G. Turinici, *A parareal in time discretization of PDE's*, C.R. Acad. Sci. Paris Sèr. I Math., 332 (2001), pp. 661–668.

[26] J.-L. Lions and E. Magenes, *Problèmes aux limites non homogènes et applications*, *Vol.* 1, Travaux et Recherches Mathématiques 17, Dunod, Paris, 1968.

[27] J.-L. Lions and E. Magenes, *Problèmes aux limites non homogènes et applications*, *Vol.* 2, Travaux et Recherches Mathématiques 18, Dunod, Paris, 1968.

[28] J.-L. Lions and O. Pironneau, *Non-overlapping domain decomposition for evolution operators*, C. R. Acad. Sci. Paris Sèr. I Math., 330 (2000), pp. 943–950.

[29] P.-L. Lions, *On the Schwarz alternating method III: A variant for nonoverlapping subdomains*, in Third International Symposium on Domain Decomposition Methods for Partial Differential Equations, Proceedings in Applied Mathematics 50, T. F. Chan, R. Glowinski, J. Périaux, and O. B. Widlund, eds., SIAM, Philadelphia, 1990, pp. 202–223.

[30] G. A. Meurant, *Numerical experiments with a domain decomposition method for parabolic problems on parallel computers*, in Fourth International Symposium on Domain Decomposition Methods for Partial Differential Equations, Proceedings in Applied Mathematics 51, R. Glowinski, Y. A. Kuznetsov, G. A. Meurant, J. Périaux, and O. B. Widlund, eds., SIAM, Philadelphia, 1991, pp. 394–408.

[31] U. Miekkala and O. Nevanlinna, *Convergence of dynamic iteration methods for initial value problems*, SIAM J. Sci. Statist. Comput., 8 (1987), pp. 459–482.

[32] F. Nataf and F. Nier, *Convergence rate of some domain decomposition methods for overlapping and nonoverlapping subdomains*, Numer. Math., 75 (1997), pp. 357–377.

[33] F. Nataf and F. Rogier, *Factorization of the convection-diffusion operator and the Schwarz algorithm*, Math. Models Methods Appl. Sci., 5 (1995), pp. 67–93.

[34] F. Nataf, F. Rogier, and E. de Sturler, *Optimal Interface Conditions for Domain Decomposition Methods*, Tech. rep. 301, CMAP, Ecole Polytechnique, Palaiseau, France, 1994.

[35] O. Nevanlinna, *Remarks on Picard-Lindelöf iterations* I, BIT, 29 (1989), pp. 328–346.

[36] O. Nevanlinna, *Remarks on Picard-Lindelöf iterations* II, BIT, 29 (1989), pp. 535–562.

[37] A. Quarteroni and A. Valli, *Domain Decomposition Methods for Partial Differential Equations*, Oxford Science Publications, Oxford University Press, New York, 1999.

[38] B. F. Smith, P. E. Bjørstad, and W. Gropp, *Domain Decomposition: Parallel Multilevel Methods for Elliptic Partial Differential Equations*, Cambridge University Press, Cambridge, UK, 1996.

[39] J. C. Strikwerda, *Finite Difference Schemes and Partial Differential Equations*, Wadsworth & Brooks/Cole Advanced Books & Software, Pacific Grove, CA, 1989.

[40] L. N. Trefethen, *Instability of difference models for hyperbolic initial-boundary value problems*, Comm. Pure Appl. Math., 37 (1984), pp. 329–367.

[41] Y. Wu, X.-C. Cai, and D. E. Keyes, *Additive Schwarz methods for hyperbolic equations*, in Tenth International Conference on Domain Decomposition Methods, J. Mandel, C. Farhat, and X.-C. Cai, eds., Contemp. Math. 218, AMS, Providence, RI, 1998, pp. 513–521.

[42] J. Xu, *Iterative methods by space decomposition and subspace correction*, SIAM Rev., 34 (1992), pp. 581–613.

SIAM J. NUMER. ANAL.
Vol. 41, No. 5, pp. 1682–1708

A NONOVERLAPPING DOMAIN DECOMPOSITION METHOD FOR MAXWELL'S EQUATIONS IN THREE DIMENSIONS[*]

QIYA HU[†] AND JUN ZOU[‡]

Abstract. In this paper, we propose a nonoverlapping domain decomposition method for solving the three-dimensional Maxwell equations, based on the edge element discretization. For the Schur complement system on the interface, we construct an efficient preconditioner by introducing two special coarse subspaces defined on the nonoverlapping subdomains. It is shown that the condition number of the preconditioned system grows only polylogarithmically with the ratio between the subdomain diameter and the finite element mesh size but possibly depends on the jumps of the coefficients.

Key words. Maxwell's equations, Nédélec finite elements, nonoverlapping domain decomposition, condition numbers

AMS subject classifications. 65N30, 65N55

DOI. 10.1137/S0036142901396909

1. Introduction.

In the numerical solution of the Maxwell equations, one needs to repeatedly solve the following system [9], [12], [17], [21], [28], [30]:

$$\nabla\times(\alpha\,\nabla\times\mathbf{u})+\beta\mathbf{u}=\mathbf{f}\quad\text{in}\quad\Omega, \tag{1.1}$$

where Ω is an open polyhedral domain in $\mathbf{R}^3$ and the coefficients $\alpha(x)$ and $\beta(x)$ are two positive bounded functions in Ω. Among various boundary conditions for (1.1), we shall consider the perfect conductor condition

$$\mathbf{u}\times\mathbf{n}=0\quad\text{on}\quad\partial\Omega, \tag{1.2}$$

where $\mathbf{n}$ is the unit outward normal vector on $\partial\Omega$.

Both the nodal and edge finite element methods have been widely used for solving the system (1.1)–(1.2); see, for example, [5], [10], [11], [12], [22], [24]. However, the algebraic systems arising from the discretization by the edge element methods are very different from the ones arising from the discretization by the standard nodal finite element methods. So the nonoverlapping domain decomposition theory for the nodal element systems, which has been well developed for second order elliptic problems in the past two decades (see the survey articles [13] [33]), does not work for the edge element systems in general, especially in three dimensions. During the last five years, there has been a rapidly growing interest in domain decomposition methods (DDMs) for solving the system (1.1)–(1.2). Some substructuring DDMs were studied for two-dimensional Maxwell equations in [29], [30] and for a different three dimensional model problem in [31]. Overlapping Schwarz methods were investigated in

[*] Received by the editors October 23, 2001; accepted for publication (in revised form) March 17, 2003; published electronically October 28, 2003.

http://www.siam.org/journals/sinum/41-5/39690.html

[†] Institute of Computational Mathematics and Scientific/Engineering Computing, Academy of Mathematical and System Sciences, The Chinese Academy of Sciences, Beijing 100080, China (hqy@lsec.cc.ac.cn). The work of this author was supported by Special Funds for Major State Basic Research Projects of China G1999032804.

[‡] Department of Mathematics, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong (zou@math.cuhk.edu.hk). The work of this author was completely supported by Hong Kong RGC grants (Projects CUHK4048/02P and 403403).

[15], [28], [16] for three-dimensional Maxwell equations. As far as the nonoverlapping DDMs are concerned, very few works can be found in the literature. A nonoverlapping DDM with two subdomains was proposed in [3] for Maxwell equations in three dimensions. The current work represents some initial efforts in the construction of efficient nonoverlapping DDMs for the case with general multiple subdomains. As we shall see, not only the construction of the coarse subspaces but also the estimates of the condition numbers of the preconditioned systems for the three-dimensional case with multiple nonoverlapping subdomains are much more difficult and tricky than in the two-dimensional case or the three-dimensional case with overlapping subdomains.

In this paper, we will propose an efficient preconditioner for the Schur complement system arising from the nonoverlapping DDM based on the edge element discretization. For the analysis of our new method, some important inequalities will be established for discrete functions in edge element spaces. We believe these inequalities should also be useful to the future developments in the field. It will be shown that the resulting preconditioned system has a nearly optimal condition number; namely, the condition number grows only polylogarithmically with the ratio between the subdomain diameter and the finite element mesh size. Unlike the optimal nonoverlapping domain decomposition preconditioners for elliptic problems [13], [25], [33], we are still unable to conclude whether the condition number of the preconditioned system generated by our nonoverlapping DDM is independent of the jumps of the coefficients. This is an important problem that we are currently working on.

The paper is arranged as follows. The edge element discretization of the system (1.1)–(1.2) and some basic formulae and definitions will be described in section 2. The construction of nonoverlapping domain decomposition preconditioners and the main results of the paper are discussed in section 3. Section 4 presents some auxiliary lemmas, which are needed in section 5 to deal with the technical difficulties in the estimates of the condition numbers.

2. Domain decompositions and discretizations. This section is devoted to the introduction of the nonoverlapping domain decomposition and the weak form and the edge element discretization of the system (1.1)–(1.2) as well as some discrete operators.

Domain decomposition. We decompose the physical domain Ω into N nonoverlapping tetrahedral subdomains $\{\Omega_i\}_i^N$, with each Ω_i of size d (see [7], [33]). The faces and vertices of the subdomains are always denoted by $\textsc{f}$ and $\textsc{v}$, while the common (open) face of the subdomains Ω_i and Ω_j are denoted by Γ_{ij}, and the union of all such common faces is denoted by Γ, i.e., $\Gamma = \cup\bar{\Gamma}_{ij}$. Γ will be called *the interface.* By Γ_i we denote the intersection of Γ with the boundary of the subdomain Ω_i. So we have $\Gamma_i = \partial\Omega_i$ if Ω_i is an interior subdomain of Ω.

Finite element triangulation. Further, we divide each subdomain Ω_i into smaller tetrahedral elements of size h so that elements from the neighboring two subdomains have an intersection which is either empty or a single nodal point or an edge or a face on the interface Γ. The resulting triangulation of the domain Ω is denoted by $\mathcal{T}_h$, which is assumed to be quasi-uniform (cf. [33]), while the set of edges and the set of nodes in $\mathcal{T}_h$ are denoted by $\mathcal{E}_h$ and $\mathcal{N}_h$, respectively.

Weak formulation. The primary goal of this paper is to construct an efficient nonoverlapping DDM for solving the discrete system arising from the edge element discretization of (1.1). For this, we first introduce its weak form and then the edge element discretization of the weak form. Let $H(\mathbf{curl};\Omega)$ be the Sobolev space consisting of all square integrable functions whose $\mathbf{curl}$'s are also square integrable in Ω, and let $H_0(\mathbf{curl};\Omega)$ be a subspace of $H(\mathbf{curl};\Omega)$ with all functions whose tangential components vanish on $\partial\Omega$, i.e., $\mathbf{v}\times\mathbf{n}=0$ on $\partial\Omega$ for all $\mathbf{v}\in H_0(\mathbf{curl};\Omega)$. Then, by integration by parts, one derives immediately the variational problem associated with the system (1.1)–(1.2).

Find $\mathbf{u}\in H_0(\mathbf{curl};\Omega)$ such that

$$A(\mathbf{u},\mathbf{v})=(\mathbf{f},\mathbf{v})\quad\forall\mathbf{v}\in H_0(\mathbf{curl};\Omega),\tag{2.1}$$

where $A(\cdot,\cdot)$ is a bilinear form given by

$$A(\mathbf{u},\mathbf{v})=(\alpha\nabla\times\mathbf{u},\nabla\times\mathbf{v})+(\beta\mathbf{u},\mathbf{v}),\quad \mathbf{u},\mathbf{v}\in H(\mathbf{curl};\Omega).$$

Here and in what follows, $(\cdot,\cdot)$ denotes the scalar product in $L^2(\Omega)$ or $L^2(\Omega)^3$.

Edge element discretization. The Nédélec edge element space, of the lowest order, is a subspace of piecewise linear polynomials defined on $\mathcal{T}_h$ (cf. [14] and [23]):

$$V_h(\Omega)=\Big\{\mathbf{v}\in H_0(\mathbf{curl};\Omega);\ \mathbf{v}\,|_K\in R(K)\ \forall K\in\mathcal{T}_h\Big\},$$

where $R(K)$ is a subset of all linear polynomials on the element K of the form

$$R(K)=\Big\{\mathbf{a}+\mathbf{b}\times\mathbf{x};\ \mathbf{a},\mathbf{b}\in\mathbf{R}^3,\ \mathbf{x}\in K\Big\}.$$

It is known [14], [23] that the tangential components of any edge element function $\mathbf{v}$ of $V_h(\Omega)$ are continuous on all edges of every element in the triangulation $\mathcal{T}_h$, and $\mathbf{v}$ is uniquely determined by its moments on edges of $\mathcal{T}_h$:

$$\left\{\lambda_e(\mathbf{v})=\int_e\mathbf{v}\cdot\mathbf{t}_e ds;\ e\in\mathcal{E}_h\right\},$$

where $\mathbf{t}_e$ denotes the unit vector on the edge e. Let $\{L_e;\ e\in\mathcal{E}_h\}$ be the edge element basis functions of $V_h(\Omega)$ satisfying

$$\lambda_{e'}(L_e)=\begin{cases}1 & \text{if } e'=e\,,\\ 0 & \text{if } e'\neq e\,;\end{cases}$$

then the edge element basis function L_e associated with the edge e has the representation

$$L_e=c_e\left(\lambda_1^e\nabla\lambda_2^e-\lambda_2^e\nabla\lambda_1^e\right),\tag{2.2}$$

where c_e is a constant independent of h, and λ_1^e and λ_2^e are two barycentric basis functions at the two endpoints of e. Furthermore, each function $\mathbf{v}$ of $V_h(\Omega)$ can be expressed as

$$\mathbf{v}(\mathbf{x})=\sum_{e\in\mathcal{E}_h}\lambda_e(\mathbf{v})L_e(\mathbf{x}),\quad \mathbf{x}\in\Omega\,.$$

With the above notation, the edge element approximation to the variational problem (2.1) can be formulated as follows: Find $\mathbf{u}_h\in V_h(\Omega)$ such that

$$A(\mathbf{u}_h,\mathbf{v}_h)=(\mathbf{f},\mathbf{v}_h)\quad\forall\mathbf{v}_h\in V_h(\Omega),\tag{2.3}$$

where $A_h(\cdot,\cdot)$ is a bilinear form given by

$$A_h(\mathbf{u}_h,\mathbf{v}_h)=\sum_{i=1}^{N}A_i(\mathbf{u}_h,\mathbf{v}_h)$$

with each $A_i(\cdot,\cdot)$ defined only on the subdomain Ω_i:

$$A_i(\mathbf{u},\mathbf{v})=(\alpha\,\nabla\times\mathbf{u},\nabla\times\mathbf{v})_{\Omega_i}+(\beta\,\mathbf{u},\mathbf{v})_{\Omega_i},\quad i=1,2,\ldots,N.$$

Some edge element subspaces. In section 3, we will formulate our DDM for solving the edge element system (2.3). Before doing so, we need to introduce more notation, subspaces, and discrete operation tools.

We will often use G to represent a subset of Γ, which may be the entire interface Γ or the local interface Γ_i or a face F of Γ_i. The notation e, with $e\subset G$, always means that e is an edge of $\mathcal{T}_h$ and lies on G. By restricting $V_h(\Omega)$ on G, we generate a subspace of $L^2(G)^3$:

$$V_h(G)=\Big\{\psi\in L^2(G)^3;\ \psi=\mathbf{v}\times\mathbf{n}\ \text{ on }\ G\ \text{ for some }\ \mathbf{v}\in V_h(\Omega)\Big\}.$$

By $V_h(\Omega_i)$ we denote the restriction of $V_h(\Omega)$ on the subdomain Ω_i. The following two local subspaces of $V_h(\Omega_i)$ and $V_h(\mathrm{F})$ will be important to our subsequent analysis:

$$V_h^0(\Omega_i)=\Big\{\mathbf{v}\in V_h(\Omega_i);\ \mathbf{v}\times\mathbf{n}=0\ \text{ on }\ \Gamma_i\Big\},$$

$$V_h^0(\mathrm{F})=\Big\{\Phi=\mathbf{v}\times\mathbf{n}\in V_h(\mathrm{F});\ \lambda_e(\mathbf{v})=0\ \ \forall\ e\subset\partial\mathrm{F}\cap\mathcal{E}_h\Big\}.$$

Discrete operators. We will often use the natural restriction operator from $V_h(\Gamma)$ onto $V_h(G)$, denoted by $\mathbf{I}_G$, and the natural zero extension operator from $V_h(G)$ into $L^2(\Gamma)^3$, denoted by $\mathbf{I}_G^t$. By definition it is clear that for a face F, $\mathbf{I}_\mathrm{F}^t\mathbf{v}\in V_h(\Gamma)$ if and only if $\mathbf{v}\in V_h^0(\mathrm{F})$, and $\mathbf{I}_G$ and $\mathbf{I}_G^t$ satisfy

$$\langle\mathbf{I}_G\Psi,\Phi\rangle_G=\langle\Psi,\mathbf{I}_G^t\Phi\rangle\quad\forall\,\Psi\in V_h(\Gamma),\ \Phi\in V_h(G),$$

where $\langle\cdot,\cdot\rangle_G$ stands for the L^2-inner product in $L^2(G)$ or $L^2(G)^3$, and the subscript G will be dropped when $G=\Gamma$. Also, we shall write $\mathbf{I}_i=\mathbf{I}_{\Gamma_i}$ and $\mathbf{I}_{ij}^t=\mathbf{I}_{\Gamma_{ij}}^t$.

For any face F of Ω_i, we use F_b to denote the union of all $\mathcal{T}_h$-induced (closed) triangles on F which have at least one edge lying on $\partial\mathrm{F}$ and F_∂ to denote the open set $\mathrm{F}\backslash\mathrm{F}_b$.

By definition, for any $\Phi\in V_h(\Gamma_i)$, there exists a $\mathbf{v}\in V_h(\Omega_i)$ such that $\Phi=\mathbf{v}\times\mathbf{n}$ on Γ_i. So Φ has the representation of the form

$$\Phi(\mathbf{x})=\sum_{e\subset\Gamma_i}\lambda_e(\mathbf{v})(L_e\times\mathbf{n})(\mathbf{x}),\quad \mathbf{x}\in\ \Gamma_i. \tag{2.4}$$

For any open face F on Γ_i, we define an operator $\mathbf{I}_{\mathrm{F}_\partial}^0:V_h(\Gamma_i)\rightarrow\mathbf{I}_\mathrm{F}^tV_h^0(\mathrm{F})$ by

$$(\mathbf{I}_{\mathrm{F}_\partial}^0\Phi)(\mathbf{x})=\sum_{e\subset\mathrm{F}_\partial}\lambda_e(\mathbf{v})(L_e\times\mathbf{n})(\mathbf{x}),\quad \mathbf{x}\in\Gamma_i, \tag{2.5}$$

and an operator $\mathbf{I}^0_{\mathrm{F}_b}$ by

$$(\mathbf{I}^0_{\mathrm{F}_b}\Phi)(\mathbf{x}) = \sum_{e\subset \mathrm{F}_b} \lambda_e(\mathbf{v})\,\mathbf{I}^t_{\mathrm{F}}(L_e \times \mathbf{n})(\mathbf{x}), \quad \mathbf{x}\in\Gamma_i.$$

Some nodal element spaces. From time to time, we shall also need some nodal element spaces in the analyses—for example, the continuous piecewise linear finite element space $Z_h(\Omega)$ of $H^1_0(\Omega)$, its restriction $Z_h(\Gamma)$ on Γ and $Z_h(\Omega_i)$ on any subdomain Ω_i, and the restriction $Z_h(\Gamma_i)$ of $Z_h(\Omega_i)$ on the local interface Γ_i and $Z_h(\mathrm{F})$ on a face F.

The operator $\mathrm{I}^t_{\mathrm{F}} : Z_h(\mathrm{F}) \to L^2(\Gamma)$ is defined similarly to $\mathbf{I}^t_{\mathrm{F}}$.

For a subset G of Γ_i, we introduce a "local" subspace

$$Z^0_h(G) = \{v\in Z_h(\Gamma_i);\ v = 0 \text{ at all nodes on } \Gamma_i\backslash G\}.$$

For any open face $\mathrm{F}\subset\Gamma_i$, we will use $\mathrm{I}^0_{\mathrm{F}} : Z_h(\Gamma_i)\to Z^0_h(\mathrm{F})$ and $\mathrm{I}^0_{\partial\mathrm{F}} : Z_h(\Gamma_i)\to Z^0_h(\partial\mathrm{F})$ to denote the natural restriction operators (see [33]).

curl- and harmonic extension operators. The next two extension operators will play an important role in the subsequent analysis. The first is the discrete **curl-**extension operator $\mathbf{R}^i_h : V_h(\Gamma_i)\to V_h(\Omega_i)$ defined as follows: For any $\Phi\in V_h(\Gamma_i)$, $\mathbf{R}^i_h\Phi\in V_h(\Omega_i)$ satisfies $\mathbf{R}^i_h\Phi\times\mathbf{n} = \Phi$ on Γ_i and solves

$$A_i(\mathbf{R}^i_h\Phi, \mathbf{v}_h) = 0 \quad \forall \mathbf{v}_h\in V^0_h(\Omega_i).$$

The second is the discrete harmonic extension operator $R^i_h : Z_h(\Gamma_i)\to Z_h(\Omega_i)$ defined as follows: For any $v_h\in Z_h(\Gamma_i)$, $R^i_h v_h\in Z_h(\Omega_i)$ satisfies $R^i_h v_h = v_h$ on Ω_i and

$$(\nabla R^i_h v_h, \nabla w_h) = 0 \quad \forall\, w_h\in Z_h(\Omega_i)\cap H^1_0(\Omega_i)\,.$$

3. Nonoverlapping DDMs.

In this section, we propose a nonoverlapping DDM for solving the edge element system (2.3). The notation $\langle\cdot,\cdot\rangle_{\Gamma_i}$ and $(\cdot,\cdot)_{\Omega_i}$ shall be used for the scalar products in $L^2(\Gamma_i)$ and $L^2(\Omega_i)$, respectively.

3.1. The interface equation.

For the solution $\mathbf{u}_h$ to the system (2.3), we write $\mathbf{u}_{hi} = \mathbf{u}_h|_{\Omega_i}$. It follows from (2.3) that

$$A_i(\mathbf{u}_{hi}, \mathbf{v}_h) = (\mathbf{f}, \mathbf{v}_h)_{\Omega_i} \quad \forall \mathbf{v}_h\in V^0_h(\Omega_i). \tag{3.1}$$

This indicates that if the tangential components $\mathbf{u}_{hi}\times\mathbf{n}_i$ are known on Γ_i the "local" unknown $\mathbf{u}_{hi}$ can be obtained by solving the local equation (3.1).

Next, we will establish an equation for the interface quantity $\Phi = \mathbf{u}_h\times\mathbf{n}$ on Γ. To do so, we introduce a "local" interface operator $\mathbf{S}_i : V_h(\Gamma_i)\to V_h(\Gamma_i)^*$ by

$$\langle \mathbf{S}_i\Phi_i, \Psi_i\rangle_{\Gamma_i} = A_i(\mathbf{R}^i_h\Phi_i, \mathbf{R}^i_h\Psi_i) \quad \forall \Psi_i, \Phi_i\in V_h(\Gamma_i).$$

Using the obvious decomposition

$$\mathbf{u}_{hi} = \mathbf{u}^0_{hi} + \mathbf{R}^i_h(\mathbf{u}_{hi}\times\mathbf{n}_i)$$

with $\mathbf{u}^0_{hi}\in V^0_h(\Omega_i)$, solving (3.1), (2.3) reduces to the interface equation (cf. [27])

$$\sum_{i=1}^N \langle \mathbf{S}_i\mathbf{I}_i\Phi, \mathbf{I}_i\Psi\rangle_{\Gamma_i} = \sum_{i=1}^N (\mathbf{f}, \mathbf{R}^i_h\mathbf{I}_i\Psi)_{\Omega_i} \quad \forall \Psi\in V_h(\Gamma). \tag{3.2}$$

Let $\mathbf{g} \in V_h(\Gamma)^*$ be defined by

$$\langle \mathbf{g}, \Psi \rangle_\Gamma = \sum_{i=1}^{N} (\mathbf{f}, \mathbf{R}_h^i \mathbf{I}_i \Psi)_{\Omega_i} \quad \forall \Psi \in V_h(\Gamma),$$

and let $\mathbf{S} = \sum_{i=1}^{N} \mathbf{I}_i^t \mathbf{S}_i \mathbf{I}_i$; then (3.2) may be written as

$$\langle \mathbf{S}\Phi, \Psi \rangle = \langle \mathbf{g}, \Psi \rangle \quad \forall \Psi \in V_h(\Gamma). \tag{3.3}$$

With $\Phi = \mathbf{u}_h \times \mathbf{n}$ available on Γ, the solution of (2.3) can be obtained by solving one subproblem, (3.1), on each subdomain Ω_i. Therefore, the solution of (2.3) reduces to the one of the interface problem (3.3). However, it is very expensive to solve this interface equation directly. Instead, we will construct an efficient preconditioner for $\mathbf{S}$; then (3.3) can be solved by the preconditioned CG method.

3.2. Preconditioners for the interface operator S. We now start to construct a preconditioner for $\mathbf{S}$. As usual, a good preconditioner should involve both local solvers and global coarse solvers.

First, the local solvers can be constructed on each local face Γ_{ij}. For each Γ_{ij}, we define a "local" operator $\mathbf{S}_{ij} : V_h^0(\Gamma_{ij}) \to V_h^0(\Gamma_{ij})^*$ by

$$\begin{gathered}\langle \mathbf{S}_{ij}\Phi_{ij}, \Psi_{ij} \rangle_{\Gamma_{ij}} = A_i(\mathbf{R}_h^i \mathbf{I}_{ij}^t \Phi_{ij}, \mathbf{R}_h^i \mathbf{I}_{ij}^t \Psi_{ij}) + A_j(\mathbf{R}_h^j \mathbf{I}_{ij}^t \Phi_{ij}, \mathbf{R}_h^j \mathbf{I}_{ij}^t \Psi_{ij}) \\ \forall \Phi_{ij}, \Psi_{ij} \in V_h^0(\Gamma_{ij}),\end{gathered}$$

and $\mathbf{S}_{ij}^{-1}$ will be our desired local solvers. The construction of the global coarse solvers is much more tricky and technical. Before doing this, we would like to illustrate our main idea about the construction. The essential difficulty in the construction of a coarse solver lies in two facts: (1) The edge element space $V_h(\Omega)$, different from the nodal element space, is not a subspace of $H^1(\Omega)^3$; (2) for any $\mathbf{v}_h \in V_h(\Omega)$, its tangential components are continuous on all *cross-edges*, namely, the edges which are shared by more than two fine elements (tangential components make no sense at the *cross-points* in two dimensions), but the moments on the *cross-edges* are not sufficient to determine the values of the tangential trace $\mathbf{v}_h \times \mathbf{n}$ on these edges. As one will see, we have the Helmholtz decomposition

$$V_h(\Omega) = \mathbf{grad}\, Z_h(\Omega) + \tilde{V}_h(\Omega),$$

where $\tilde{V}_h(\Omega)$ corresponds to the divergence-free part and is closely related to the space $H^1(\Omega)^3$. Thus it seems necessary to construct two coarse subspaces and coarse solvers, corresponding to the **curl**-free and divergence-free subspaces $\nabla Z_h(\Omega)$ and $\tilde{V}_h(\Omega)$, respectively.

For the construction of the coarse subspaces, we introduce some more notation below. For any subdomain Ω_i, by $\mathcal{W}_i$ we denote the set of the edges of Ω_i, which belong to at least two other local interfaces Γ_j, $j \neq i$. On each $\mathcal{W}_i$, we define the discrete L^2-scalar product

$$\langle \varphi, \psi \rangle_{h,\mathcal{W}_i} = h \sum_{\mathbf{x} \in \mathcal{N}_h \cap \mathcal{W}_i} \varphi(\mathbf{x})\psi(\mathbf{x}) \quad \forall\, \varphi, \psi \in Z_h(\Gamma_i)\,;$$

the corresponding norm is denoted by $\|\cdot\|_{h,\mathcal{W}_i}$. Let

$$\Delta_i = \bigcup_{\mathrm{F} \subset \Gamma_i} \mathrm{F}_b, \quad i = 1, \ldots, N.$$

We introduce a norm $\|\cdot\|_{*,\Delta_i}$ that is induced from the following inner product in $L^2(\Delta_i)^3$:

$$\langle \mathbf{v}\times\mathbf{n}, \mathbf{w}\times\mathbf{n}\rangle_{*,\Delta_i} = \sum_{K\subset\Delta_i} \langle \mathbf{v}\times\mathbf{n}, \mathbf{w}\times\mathbf{n}\rangle_{\partial K} \quad \forall\, \mathbf{v}\times\mathbf{n}, \mathbf{w}\times\mathbf{n} \in V_h(\Gamma_i),$$

where the summation is over all triangles K in Δ_i.

For any given subset G of Ω and function φ in $L^2(G)$, we use $\gamma_{_G}(\varphi)$ for the average value of φ on G. Similarly, for a vector $\mathbf{v} = (v_1, v_2, v_3)$ in $L^2(G)^3$, we use $\Upsilon_G(\mathbf{v})$ for the constant vector with three average values $\gamma_{_G}(v_1)$, $\gamma_{_G}(v_2)$, and $\gamma_{_G}(v_3)$ as its components.

Now we define two discrete operators in $Z_h(\Gamma)$ and $V_h(\Gamma)$ which will generate two coarse subspaces. For any $\varphi \in Z_h(\Gamma)$, we define $\pi_0\varphi \in Z_h(\Gamma)$ by

$$\pi_0\varphi(\mathbf{x}) = \begin{cases} \varphi(\mathbf{x}) & \text{for } \mathbf{x}\in \mathcal{W}_i \cap \mathcal{N}_h \ (i=1,\dots,N), \\ \gamma_{_{\partial \mathrm{F}}}(\varphi) & \text{for } \mathbf{x}\in \mathrm{F}\cap\mathcal{N}_h \ (\mathrm{F}\subset\Gamma). \end{cases} \tag{3.4}$$

Similarly, for each $\mathbf{v}\times\mathbf{n} \in V_h(\Gamma)$, we define $\Pi_0\mathbf{v}\times\mathbf{n} \in V_h(\Gamma)$ such that

$$\lambda_e(\Pi_0\mathbf{v}) = \begin{cases} \lambda_e(\mathbf{v}) & \text{for } e\subset \Delta_i\cup\Omega_i \ (i=1,\dots,N), \\ \lambda_e(\Upsilon_{\partial\mathrm{F}}(\mathbf{v})) & \text{for } e\subset \mathrm{F}_\partial \ (\mathrm{F}\subset\Gamma). \end{cases}$$

Note that although $\Pi_0\mathbf{v}$ involves the degrees of freedom inside Ω_i, $\Pi_0\mathbf{v}\times\mathbf{n}$ is determined on Γ uniquely by the moments $\lambda_e(\mathbf{v})$ for all $e\subset\Gamma$. Thus $\Pi_0\mathbf{v}\times\mathbf{n}\in V_h(\Gamma)$ can also be defined directly by

$$\Pi_0\mathbf{v}\times\mathbf{n} = \begin{cases} \mathbf{v}\times\mathbf{n} & \text{on } \Delta_i \ (i=1,\dots,N), \\ \Upsilon_{\partial\mathrm{F}}(\mathbf{v}\times\mathbf{n}) & \text{on } \mathrm{F}_\partial \ (\mathrm{F}\subset\Gamma), \end{cases}$$

where we have used the fact that the normal vector $\mathbf{n}$ is constant on any face $\mathrm{F}\subset\Gamma$ and

$$\Upsilon_{\partial\mathrm{F}}(\mathbf{v})\times\mathbf{n}\big|_{\mathrm{F}} = \Upsilon_{\partial\mathrm{F}}(\mathbf{v}\times\mathbf{n}).$$

Now, we can define the two coarse subspaces:

$$V_h^{01}(\Gamma) = \Big\{\Phi_0 \in V_h(\Gamma);\ \mathbf{I}_i\Phi_0 = \mathbf{grad}(R_0^i \mathrm{I}_i \pi_0\varphi)\times\mathbf{n} \text{ on } \Gamma_i \text{ for some } \varphi\in Z_h(\Gamma)\Big\},$$

$$V_h^{02}(\Gamma) = \Big\{\mathbf{v}_0\times\mathbf{n} \in V_h(\Gamma);\ \mathbf{v}_0 = \Pi_0\mathbf{v} \text{ for some } \mathbf{v}\times\mathbf{n}\in V_h(\Gamma)\Big\}.$$

The operator R_0^i used in $V_h^{01}(\Gamma)$ is the zero extension into the interior of Ω_i; namely, for any $v_h\in Z_h(\Gamma_i)$, $R_0^i v_h \in Z_h(\Omega_i)$ takes the same values as v_h on Γ_i and vanishes at all interior nodes of Ω_i. We can define two coarse solvers $\mathbf{S}_{0k} : V_h^{0k}(\Gamma)\to V_h^{0k}(\Gamma)^*$, $k=1,2$, associated with these coarse subspaces. For any $\Phi_0, \Psi_0 \in V_h^{01}(\Gamma)$, there exist $\varphi,\ \psi\in Z_h(\Gamma)$ such that on Γ_i,

$$\mathbf{I}_i\Phi_0 = \mathbf{grad}(R_0^i\mathrm{I}_i\pi_0\varphi)\times\mathbf{n}, \quad \mathbf{I}_i\Psi_0 = \mathbf{grad}(R_0^i\mathrm{I}_i\pi_0\psi)\times\mathbf{n}\,.$$

Then $\mathbf{S}_{01}$ is defined by

$$\langle \mathbf{S}_{01}\Phi_0, \Psi_0\rangle = [1+\log(d/h)]\sum_{i=1}^N \langle \pi_0\varphi - \gamma_{_{\mathcal{W}_i}}(\pi_0\varphi), \pi_0\psi - \gamma_{_{\mathcal{W}_i}}(\pi_0\psi)\rangle_{h,\mathcal{W}_i}.$$

Similarly, for any $\Phi_0, \Psi_0 \in V_h^{02}(\Gamma)$, there exist $\mathbf{v}$, $\mathbf{w} \in V_h(\Omega)$ such that on Γ_i,

$$\mathbf{I}_i\Phi_0 = \Pi_0\mathbf{v} \times \mathbf{n}, \quad \mathbf{I}_i\Psi_0 = \Pi_0\mathbf{w} \times \mathbf{n}.$$

Then $\mathbf{S}_{02}$ is defined by

$$\langle \mathbf{S}_{02}\Phi_0, \Psi_0\rangle = [1 + \log(d/h)] \sum_{i=1}^{N} \langle \Phi_0 - \Upsilon_{\Delta_i}(\mathbf{v}) \times \mathbf{n}, \Psi_0 - \Upsilon_{\Delta_i}(\mathbf{w}) \times \mathbf{n}\rangle_{*,\Delta_i} \\ + d^2\langle \Phi_0, \Psi_0\rangle_{*,\Delta_i}.$$

Hereafter, $\Upsilon_{\Delta_i}(\mathbf{v})$ is the constant vector satisfying

$$\|\Phi_0 - \Upsilon_{\Delta_i}(\mathbf{v}) \times \mathbf{n}\|^2_{*,\Delta_i} = \min_{C_{\Delta_i} \in \mathcal{R}^3} \|\Phi_0 - C_{\Delta_i} \times \mathbf{n}\|^2_{*,\Delta_i},$$

which can be viewed as some average of Φ_0 on Δ_i. And the average is well defined.

Finally, the preconditioner for the interface operator $\mathbf{S}$ can be defined as follows:

$$\mathbf{M}^{-1} = \mathbf{S}_{01}^{-1} + \mathbf{S}_{02}^{-1} + \sum_{\Gamma_{ij}} \mathbf{I}_{ij}^t \mathbf{S}_{ij}^{-1} \mathbf{I}_{ij}. \tag{3.5}$$

For this preconditioner, we have the following theorem.

THEOREM 3.1. *The condition number of the preconditioned system can be estimated by*

$$\text{cond}(\mathbf{M}^{-1}\mathbf{S}) \le C[1 + \log(d/h)]^3. \tag{3.6}$$

Remark 3.1. A simple algorithm to implement the coarse solver $\mathbf{S}_{01}$ can be found in [33]. By the minimum property of the average $\Upsilon_{\Delta_i}(\Phi_0)$, we can also derive a simple algorithm for implementing the coarse solver $\mathbf{S}_{02}$, which is similar to the one in [33]. Note that one may also use the inner product $h^{-1}\langle\cdot,\cdot\rangle_{\Delta_i}$ in the definition of $\mathbf{S}_{02}$ instead of the inner product $\langle\cdot,\cdot\rangle_{*,\Delta_i}$. Furthermore, one may use the discrete $L^2(\Delta_i)^3$-inner product

$$\langle\langle \mathbf{v} \times \mathbf{n}, \mathbf{w} \times \mathbf{n}\rangle\rangle_{h,\Delta_i} = \sum_{e \subset \Delta_i} \lambda_e(\mathbf{v})\lambda_e(\mathbf{w}) \quad \forall \mathbf{v} \times \mathbf{n}, \mathbf{w} \times \mathbf{n} \in V_h(\Gamma_i),$$

to define the coarse solver $\mathbf{S}_{02}$, but we do not know yet how to verify the existence of the corresponding average.

Remark 3.2. The "local" operator $\mathbf{S}_{ij}$ may be replaced by any other spectrally equivalent operator, for example, the operator defined by

$$\langle \mathbf{S}_{ij}^i \Phi_{ij}, \Psi_{ij}\rangle_{\Gamma_{ij}} = A_i(\mathbf{R}_h^i \mathbf{I}_{ij}^t \Phi_{ij}, \mathbf{R}_h^i \mathbf{I}_{ij}^t \Psi_{ij}) \quad \forall \Psi_{ij} \in V_h^0(\Gamma_{ij}).$$

$\mathbf{S}_{ij}^i$ is easier to implement than $\mathbf{S}_{ij}$, but it loses the symmetry with respect to the face Γ_{ij}.

Remark 3.3. Based on our current analysis in section 5, the constant C in the condition number estimate (3.6) may have a factor $\gamma_{\max}/\gamma_{\min}$ related to the coefficients in (1.1), where $\gamma_{\max}$ is the supremum of $\beta(x)$ and $\alpha^2(x)$ over $\bar{\Omega}$, and $\gamma_{\min}$ is the infimum of $\beta(x)$ and $\alpha^2(x)$ over $\bar{\Omega}$. It is possible to improve such dependence on the coefficients if a more localized and sharper analysis can be found.

Remark 3.4. The nodal element coarse interpolant π_0 is widely used in nonoverlapping DDMs for second order elliptic problems [13], [33]. The new edge element coarse interpolant Π_0 is very similar to π_0 but with some essential differences. For a $H^1(\Omega)^3$ vector-valued function $\mathbf{v}$, there is no trace on the wirebasket set $\mathcal{W}_i$, and the coarse interpolants $\pi_0\mathbf{v}$ and $\Pi_0\mathbf{v}$ make no sense. However, it is known that π_0 is stable in the nodal element space $Z_h(\Gamma_i)$ [13], [33]. Likewise, we shall show in section 4 that Π_0 is stable in the edge element space $V_h(\Gamma_i)$, with the stability constants growing only polylogarithmically with d/h. This explains somewhat why we can achieve a logarithmical bound (3.6) on the condition number.

4. Some auxiliary lemmas.

As we shall see, the estimate (3.6) of the condition number $\text{cond}(\mathbf{M}^{-1}\mathbf{S})$ for the preconditioned system is rather technical. This section presents some basic properties of Sobolev spaces and auxiliary lemmas, which are needed to deal with the technical difficulties in the estimate of the condition number. The proofs will be provided in the appendix. The constant C will be used often in what follows for the generic constant that may take different values at different occasions.

4.1. The scaled norms.

A large part of the condition number estimate will be carried out on the subdomains, for which we need some scaled norms. For the space $H^1(\Omega_i)^3$, we define a scaled norm by

$$\|\mathbf{v}\|_{1,\Omega_i} = (|\mathbf{v}|^2_{1,\Omega_i} + d^{-2}\|\mathbf{v}\|^2_{0,\Omega_i})^{\frac{1}{2}} \quad \forall\, \mathbf{v} \in H^1(\Omega_i)^3,$$

while for the space $H(\mathbf{curl};\Omega_i)$, the restriction of $H_0(\mathbf{curl};\Omega)$ on the subdomain Ω_i, and the interface space $H^{-\frac{1}{2}}(\Gamma_i)$, we define their scaled norms by

$$\|\mathbf{v}\|_{\mathbf{curl};\Omega_i} = \left(\|\mathbf{curl}\,\mathbf{v}\|^2_{0,\Omega_i} + d^{-2}\|\mathbf{v}\|^2_{0,\Omega_i}\right)^{\frac{1}{2}} \quad \forall\, \mathbf{v} \in H(\mathbf{curl};\Omega_i)\,,$$

$$\|\lambda\|_{-\frac{1}{2},\Gamma_i} = \sup_{v\in H^{\frac{1}{2}}(\Gamma_i)} \frac{|\langle\lambda, v\rangle_{\Gamma_i}|}{\|v\|_{\frac{1}{2},\Gamma_i}} \quad \forall\, \lambda \in H^{-\frac{1}{2}}(\Gamma_i),$$

where

$$\|v\|_{\frac{1}{2},\Gamma_i} = (|v|^2_{\frac{1}{2},\Gamma_i} + d^{-1}\|v\|^2_{0,\Gamma_i})^{\frac{1}{2}}.$$

For any $\Phi \in V_h(\Gamma_i)$, we use $\text{div}_\tau\Phi$ to denote the tangential divergence of Φ; see [2] and [3] for the definition of $\text{div}_\tau\Phi$. It is known that $\text{div}_\tau\Phi \in H^{-\frac{1}{2}}(\Gamma_i)$, so it makes sense to define the norm

$$\|\Phi\|_{\mathcal{X}_{\Gamma_i}} = d^{-1}\|\Phi\|_{-\frac{1}{2},\Gamma_i} + \|\text{div}_\tau\Phi\|_{-\frac{1}{2},\Gamma_i}.$$

The next two estimates on this norm $\|\cdot\|_{\mathcal{X}_{\Gamma_i}}$ can be found in [3].

LEMMA 4.1. *The discrete* **curl**-*extension* $\mathbf{R}^i_h\Phi \in V_h(\Omega_i)$ *satisfies*

$$\|R^i_h\Phi\|_{\mathbf{curl};\Omega_i} \le C\|\Phi\|_{\mathcal{X}_{\Gamma_i}}. \tag{4.1}$$

LEMMA 4.2. *Let* $\mathbf{u} \in V_h(\Omega_i)$, *which satisfies* $\mathbf{u}\times\mathbf{n} = \Phi$ *on* Γ_i. *Then*

$$\|\Phi\|_{\mathcal{X}_{\Gamma_i}} \le C\|\mathbf{u}\|_{\mathbf{curl};\Omega_i}. \tag{4.2}$$

4.2. Estimates with the norm $\|\cdot\|_{1/2,\Gamma_i}$ and the edge element interpolant. The results in Lemma 4.3 can be found in [7] and [33].

LEMMA 4.3. *For any $\varphi \in Z_h(\Gamma)$, we have*

$$C\,|\pi_0\varphi|^2_{\frac{1}{2},\Gamma_i} \le [1+\log(d/h)]\|\varphi - \gamma_{\mathcal{W}_i}(\varphi)\|^2_{h,\mathcal{W}_i} \le C[1+\log(d/h)]^2|\varphi|^2_{\frac{1}{2},\Gamma_i} \tag{4.3}$$

and for any face F $\subset \Gamma_i$,

$$\|\mathrm{I}^0_{\mathrm{F}}(\varphi - \pi_0\varphi)\|^2_{\frac{1}{2},\Gamma_i} \le C[1+\log(d/h)]^2|\varphi|^2_{\frac{1}{2},\Gamma_i}\,. \tag{4.4}$$

Now we define an interpolation operator $\mathbf{r}_h$ associated with the space $V_h(\Omega)$. For any appropriately smooth $\mathbf{v}$, $\mathbf{r}_h\mathbf{v} \in V_h(\Omega)$ is a function in $V_h(\Omega)$ which has the same moments on the edges of $\mathcal{T}_h$ as $\mathbf{v}$, namely,

$$\int_e \mathbf{r}_h\mathbf{v}\cdot\mathbf{t}_e ds = \int_e \mathbf{v}\cdot\mathbf{t}_e ds \quad \forall\, \mathbf{v}\in H^1(\Omega) \text{ and } \ e\in\mathcal{E}_h.$$

The interpolant $\mathbf{r}_h\mathbf{v}$ is well defined on each element K for all $\mathbf{v}$ lying in the space

$$\Big\{\mathbf{w}\in L^p(K)^3;\ \mathbf{curl}\,\mathbf{v}\in L^p(K)^3 \text{ and } \mathbf{v}\times\mathbf{n}\in L^p(\partial K)^3\Big\}$$

with $p > 2$; see Lemma 4.7 in [4]. From this we immediately know that $\mathbf{r}_h\mathbf{v}$ is well defined for all $\mathbf{v}$ in $H^1(\Omega)^3$ whose **curl** is in $L^p(K)^3$.

The following three lemmas present some estimates on the interpolation operator r_h. The proof of the first lemma below is quite similar to the proofs of Lemma 4.7 in [4] and Lemma 3.2 in [12], and details can be found in [20].

LEMMA 4.4. *Let $\mathbf{w}\in H^1(\Omega_i)^3$ and its interpolant $\mathbf{r}_h\mathbf{w}$ be well defined in $V_h(\Omega_i)$. Also, we assume that* $\mathbf{curl}\,\mathbf{w} = \mathbf{curl}\,\mathbf{v}_h$ *for some $\mathbf{v}_h\in V_h(\Omega_i)$. Then*

$$\|\mathbf{r}_h\mathbf{w}-\mathbf{w}\|_{0,\Omega_i} \le Ch(|\mathbf{w}|^2_{1,\Omega_i} + \|\mathbf{curl}\,\mathbf{v}_h\|^2_{0,\Omega_i})^{\frac{1}{2}}. \tag{4.5}$$

LEMMA 4.5. *Under the same assumptions as in Lemma* 4.4, *for any face* F *of* Γ_i *we have*

$$\|(\mathbf{r}_h\mathbf{w})\times\mathbf{n}\|_{*,\mathrm{F}_b} \le C[1+\log(d/h)]^{\frac{1}{2}}(\|\mathbf{w}\|^2_{1,\Omega_i} + \|\mathbf{curl}\,\mathbf{v}_h\|^2_{0,\Omega_i})^{\frac{1}{2}}. \tag{4.6}$$

LEMMA 4.6. *Under the same assumptions as in Lemma* 4.4, *for any face* F *of* Γ_i *we have*

$$d^{-2}\|\mathbf{r}_h\mathbf{w} - \Upsilon_{\partial\mathrm{F}}(\mathbf{r}_h\mathbf{w})\|^2_{0,\Omega_i} \le C[1+\log(d/h)](|\mathbf{w}|^2_{1,\Omega_i} + \|\mathbf{curl}\,\mathbf{v}_h\|^2_{0,\Omega_i}), \tag{4.7}$$

$$d^{-2}\|\mathbf{w} - \Upsilon_{\partial\mathrm{F}}(\mathbf{r}_h\mathbf{w})\|_{0,\Omega_i} \le C[1+\log(d/h)](|\mathbf{w}|^2_{1,\Omega_i} + \|\mathbf{curl}\,\mathbf{v}_h\|^2_{0,\Omega_i}). \tag{4.8}$$

4.3. Some estimates with the norm$\|\cdot\|_{\mathcal{X}_{\Gamma_i}}$.

LEMMA 4.7. *Let $\mathbf{w}$ and $\mathbf{v}_h$ be the same as specified in Lemma* 4.4, *and $\Phi = \mathbf{r}_h\mathbf{w}\times\mathbf{n}$ on Γ_i. Then for any face* F $\subset\Gamma_i$ *we have*

$$\|\mathbf{I}^0_{\mathrm{F}_\partial}\Phi\|_{\mathcal{X}_{\Gamma_i}} \le C[1+\log(d/h)](\|\Phi\|_{\mathcal{X}_{\Gamma_i}} + \|\mathbf{w}\|_{1,\Omega_i} + \|\mathbf{curl}\ \mathbf{v}_h\|_{0,\Omega_i}). \tag{4.9}$$

LEMMA 4.8. *Let $\Phi = \mathbf{v}\times\mathbf{n}\in V_h(\Gamma_i)$ on Γ_i, and*

$$\mathbf{I}^0_{\Delta_i}\Phi(\mathbf{x}) = \sum_{e\subset\Delta_i}\lambda_e(\mathbf{v})(L_e\times\mathbf{n}_i)(\mathbf{x}), \quad \mathbf{x}\in\Gamma_i.$$

We have

$$\|\mathbf{I}^0_{\Delta_i}\Phi\|_{\mathcal{X}_{\Gamma_i}} \le C[1+\log(d/h)]^{\frac{1}{2}}\|\Phi\|_{*,\Delta_i}. \tag{4.10}$$

LEMMA 4.9. *Assume that $\mathbf{v}\in V_h(\Omega)$ and* F $\subset\Gamma_k$. *Then*

$$\|\mathbf{I}^0_{\mathrm{F}_\partial}(\Upsilon_{\partial\mathrm{F}}(\Pi_0\mathbf{v})\times\mathbf{n})\|^2_{\mathcal{X}_{\Gamma_k}} \le C[1+\log(d/h)]\|(\Pi_0\mathbf{v})\times\mathbf{n}\|^2_{*,\mathrm{F}_b}. \tag{4.11}$$

5. The estimate of condition number. This section is devoted to the estimate (3.6) of the condition number of the preconditioned system $\mathbf{M}^{-1}\mathbf{S}$. The estimation will be done by using the following additive Schwarz framework [26], [32], whose proof is standard (cf. [18] and [27]).

LEMMA 5.1. *Assume that the following two conditions hold:*

(i) *For any* $\Phi \in V_h(\Gamma)$ *there is a decomposition* $\Phi = \Phi_{01} + \Phi_{02} + \sum_{i<j} \mathrm{I}_{ij}^t \Phi_{ij}$, *with* $\Phi_{0k} \in V_h^{0k}(\Gamma)$ $(k = 1,\ 2)$ *and* $\Phi_{ij} \in V_h^0(\Gamma_{ij})$, *such that*

$$\langle \mathbf{S}_{01}\Phi_{01}, \Phi_{01}\rangle + \langle \mathbf{S}_{02}\Phi_{02}, \Phi_{02}\rangle + \sum_{i<j} \langle \mathbf{S}_{ij}\Phi_{ij}, \Phi_{ij}\rangle_{\Gamma_{ij}} \le C_1 \langle \mathbf{S}\Phi, \Phi\rangle; \tag{5.1}$$

(ii) *For any* $\Psi_{0k} \in V_h^{0k}(\Gamma)$ $(k = 1, 2)$ *and* $\Psi_{ij} \in V_h^0(\Gamma_{ij})$, *we have*

$$\left\langle \mathbf{S}\left(\sum_{i<j} \mathbf{I}_{ij}^t \Psi_{ij} + \Psi_{01} + \Psi_{02}\right),\ \sum_{i<j} \mathbf{I}_{ij}^t \Psi_{ij} + \Psi_{01} + \Psi_{02}\right\rangle \tag{5.2}$$
$$\le C_2 \left\{ \sum_{i<j} \langle \mathbf{S}_{ij}\Psi_{ij}, \Psi_{ij}\rangle_{\Gamma_{ij}} + \langle \mathbf{S}_{01}\Psi_{01}, \Psi_{01}\rangle + \langle \mathbf{S}_{02}\Psi_{02}, \Psi_{02}\rangle \right\}.$$

Then we have $\mathrm{cond}(\mathbf{M}^{-1}\mathbf{S}) \le C_1 C_2$.

The rest of this section applies Lemma 5.1 to show Theorem 3.1, the main result of this paper. First, we construct the important decomposition required in the lemma. For this, we will make use of the so-called regular decomposition instead of the usual $L^2(\Omega)$-orthogonal Helmholtz decomposition [14].

For any $\mathbf{v} \in H_0(\mathbf{curl};\Omega)$, there exist some $\mathbf{w} \in H_0^1(\Omega)^3$ and $p \in H_0^1(\Omega)$ such that the following regular decomposition holds (cf. [6], [16]):

$$\mathbf{v} = \nabla p + \mathbf{w} \tag{5.3}$$

with the estimates

$$\|w\|_{0,\Omega} + \|p\|_{1,\Omega} \le C\, \|\mathbf{v}\|_{0,\Omega}\,, \quad |\mathbf{v}|_{1,\Omega} \le C \|\mathbf{curl}\,\mathbf{v}\|_{0,\Omega}\,. \tag{5.4}$$

We remark that the use of Helmholtz-type or regular decompositions is a fundamental technique for the analysis of preconditioners for $H(\mathbf{curl};\Omega)$- and $H(\mathrm{div};\Omega)$-elliptic problems [1], [15], [17], [16], [28].

Now, for any $\Phi \in V_h(\Gamma)$, we define a $\mathbf{v}_h \in V_h(\Omega)$ such that $\mathbf{v}_h = \mathbf{R}_h^i \mathbf{I}_i \Phi$ in each subdomain Ω_i. By the regular decomposition (5.3), there exist $p \in H_0^1(\Omega)$ and $\mathbf{w} \in H_0^1(\Omega)^3$ such that

$$\mathbf{v}_h = \mathbf{grad}\ p + \mathbf{w}\,. \tag{5.5}$$

As $\mathbf{w} \in H_0^1(\Omega)^3$ and $\mathbf{curl}\,\mathbf{w} = \mathbf{curl}\,\mathbf{v}_h$, so $\mathbf{r}_h\mathbf{w}$ is well defined (see subsection 4.2). This, with (5.5), implies

$$\mathbf{v}_h = \mathbf{r}_h \mathbf{grad}\ p + \mathbf{r}_h \mathbf{w}\,.$$

By Lemma 5.10 in [14], there exists a function $p_h \in Z_h(\Omega)$ such that

$$\mathbf{v}_h = \mathbf{grad}\ p_h + \mathbf{r}_h \mathbf{w} = \mathbf{grad} p_h + \mathbf{w}_h \tag{5.6}$$

with $\mathbf{w}_h = \mathbf{r}_h \mathbf{w} \in V_h(\Omega)$. By (5.5) and (5.6), we know

$$\mathbf{curl}\ \mathbf{w}_h = \mathbf{curl}\ \mathbf{w} = \mathbf{curl}\ \mathbf{v}_h. \tag{5.7}$$

Now we are ready to show Theorem 3.1 using Lemma 5.1. We divide the proof into four steps.

Step 1. Establish a suitable decomposition for $\Phi \in V_h(\Gamma)$. For ease of notation, we introduce $p_h^0 \in Z_h(\Omega)$ and Φ_{01} by

$$\begin{aligned} p_h^0 &= R_h^i \mathrm{I}_i \pi_0 (p_h|_\Gamma) \quad \text{in} \quad \Omega_i, \quad i = 1, \ldots, N, \\ \Phi_{01}(\mathbf{x}) &= (\mathbf{grad}\ (p_h^0|_{\Omega_i}) \times \mathbf{n})(\mathbf{x}), \quad \mathbf{x} \in \Gamma_i,\ i = 1, 2, \ldots, N. \end{aligned}$$

By direct checking, we can also write

$$\Phi_{01}(\mathbf{x}) = (\mathbf{grad}\ (\tilde{p}_h^0|_{\Omega_i}) \times \mathbf{n})(\mathbf{x}), \quad \mathbf{x} \in \Gamma_i,$$

with $\tilde{p}_h^0 = R_0^i \mathrm{I}_i \pi_0 (p_h|_\Gamma)$. So we know $\Phi_{01}(\mathbf{x}) \in V_h^{01}(\Gamma)$. Next, we choose $\mathbf{w}_{02} = \Pi_0 \mathbf{w}_h \in V_h(\Omega)$ and let

$$\Phi_{02} = (\mathbf{w}_{02} \times \mathbf{n})|_\Gamma \in V_h^{02}(\Gamma).$$

Define $\Phi_{ij} \in V_h(\Gamma_{ij})$ by

$$\begin{aligned} \Phi_{ij} &= \mathbf{I}_{ij}((\mathbf{grad}\ p_h + \mathbf{w}_h) \times \mathbf{n}) - \mathbf{I}_{ij}(\Phi_{01} + \Phi_{02}) \\ &= \mathbf{I}_{ij}(\mathbf{grad}\ (p_h - p_h^0) \times \mathbf{n}) + \mathbf{I}_{ij}(\mathbf{w}_h \times \mathbf{n} - \Phi_{02}) \\ &= \mathbf{I}_{ij}(\mathbf{grad}\ (p_h - p_h^0) \times \mathbf{n}) + \mathbf{I}_{ij}((\mathbf{w}_h - \mathbf{w}_{02}) \times \mathbf{n}). \end{aligned}$$

Noting the fact that $p_h^0 - p_h$ vanishes on the wirebasket set $\mathcal{W}_i$, we can easily verify that $\lambda_e(\mathbf{grad}\ (p_h - p_h^0)) = 0$ for any $e \in \mathcal{E}_h \cap \mathcal{W}_i$. Also, we have $\lambda_e(\mathbf{w}_h - \mathbf{w}_{02}) = 0$ for any face e on Δ_i. Thus $\Phi_{ij} \in V_h^0(\Gamma_{ij})$, and the following decomposition holds:

$$\Phi = \Phi_{01} + \Phi_{02} + \sum_{\Gamma_{ij}} \mathbf{I}_{ij}^t \Phi_{ij}. \tag{5.8}$$

Step 2. Prove the estimate

$$\sum_{\Gamma_{ij}} \langle \mathbf{S}_{ij} \Phi_{ij}, \Phi_{ij} \rangle_{\Gamma_{ij}} \le C[1 + \log(d/h)]^3 \langle \mathbf{S}\Phi, \Phi \rangle. \tag{5.9}$$

For any face Γ_{ij} of Γ_i, we define

$$\begin{aligned} p_{ij}^i &= R_h^i \mathrm{I}_{ij}^t [(p_h - p_h^0)|_{\Gamma_{ij}}] \in Z_h(\Omega_i), \\ \mathbf{w}_{ij}^i &= \mathbf{R}_h^i \mathbf{I}_{ij}^t [((\mathbf{w}_h - \mathbf{w}_{02}) \times \mathbf{n})|_{\Gamma_{ij}}] \in V_h(\Omega_i), \\ \mathbf{v}_{ij}^i &= \mathbf{grad}\ p_{ij}^i + \mathbf{w}_{ij}^i \in V_h(\Omega_i). \end{aligned}$$

Using the fact that

$$\mathbf{R}_h^i \mathbf{I}_{ij}^t \Phi_{ij} \times \mathbf{n} = \mathbf{I}_{ij}^t \Phi_{ij} = \mathbf{v}_{ij}^i \times \mathbf{n} \quad \text{on} \quad \Gamma_i,$$

we obtain by the minimum **curl**-energy property of the discrete **curl**-extension that

$$\begin{aligned} A_i(\mathbf{R}_h^i \mathbf{I}_{ij}^t \Phi_{ij}, \mathbf{R}_h^i \mathbf{I}_{ij}^t \Phi_{ij}) \le A_i(\mathbf{v}_{ij}^i, \mathbf{v}_{ij}^i) &= \|\alpha^{\frac{1}{2}} \mathbf{curl}\ \mathbf{w}_{ij}^i\|_{0,\Omega_i}^2 + \|\beta^{\frac{1}{2}} \mathbf{v}_{ij}^i\|_{0,\Omega_i}^2 \\ &\le C(\|\mathbf{grad}\ p_{ij}^i\|_{0,\Omega_i}^2 + \|\mathbf{w}_{ij}^i\|_{\mathbf{curl},\Omega_i}^2). \end{aligned} \tag{5.10}$$

As $p_h^0 = \pi_0(p_h|_\Gamma)$ on Γ, we have

$$\mathrm{I}_{ij}^t[(p_h - p_h^0)|_{\Gamma_{ij}}] = \mathrm{I}_{ij}^0(p_h|_\Gamma - \pi_0(p_h|_\Gamma)).$$

Thus, using (4.4) and the trace theorem, we obtain

$$\begin{aligned}(5.11)\qquad \|\mathbf{grad}\ p_{ij}^i\|_{0,\Omega_i}^2 &= |p_{ij}^i|_{1,\Omega_i}^2 \le C|\mathrm{I}_{ij}^t[(p_h - p_h^0)|_{\Gamma_{ij}}]|_{\frac{1}{2},\Gamma_i}^2\\ &\le C[1+\log(d/h)]^2|p_h|_{\frac{1}{2},\Gamma_i}^2\\ &\le C[1+\log(d/h)]^2|p_h|_{1,\Omega_i}^2.\end{aligned}$$

We next estimate $\mathbf{w}_{ij}^i$. For each (open) common face $\mathrm{F} = \Gamma_{ij}$ shared by Ω_i and Ω_j, it follows from the definition of Π_0 that

$$\lambda_e(\mathbf{w}_h - \mathbf{w}_{02}) = \begin{cases} 0 & \text{if } e \subset \mathrm{F}_b,\\ \lambda_e(\mathbf{w}_h - \Upsilon_{\partial \mathrm{F}}(\mathbf{w}_h)) & \text{if } e \subset \mathrm{F}_\partial\,. \end{cases}$$

Then we derive by using (5.7) and Lemmas 4.1 and 4.7 that

$$\begin{aligned}(5.12)\qquad \|\mathbf{w}_{ij}^i\|_{\mathbf{curl},\Omega_i}^2 &\le C\|\mathbf{I}_{ij}^t[((\mathbf{w}_h - \mathbf{w}_{02})\times\mathbf{n})|_{\Gamma_{ij}}]\|_{\mathcal{X}_{\Gamma_i}}^2\\ &= C\|\mathbf{I}_{\mathrm{F}_\partial}^0[(\mathbf{w}_h - \Upsilon_{\partial\Gamma_{ij}}(\mathbf{w}_h))\times\mathbf{n}]\|_{\mathcal{X}_{\Gamma_i}}^2\\ &\le C[1+\log(d/h)]^2(\|(\mathbf{w}_h - \Upsilon_{\partial\Gamma_{ij}}(\mathbf{w}_h))\times\mathbf{n}\|_{\mathcal{X}_{\Gamma_i}}^2\\ &\quad +\|\mathbf{w} - \Upsilon_{\partial\Gamma_{ij}}(\mathbf{w}_h)\|_{1,\Omega_i}^2 + \|\mathbf{curl}\ \mathbf{v}_h\|_{0,\Omega_i}^2).\end{aligned}$$

On the other hand, for the term $(\mathbf{w}_h - \Upsilon_{\partial\Gamma_{ij}}(\mathbf{w}_h))\times\mathbf{n}$ we have by Lemma 4.2 and (5.5) that

$$\begin{aligned}&\|(\mathbf{w}_h - \Upsilon_{\partial\Gamma_{ij}}(\mathbf{w}_h))\times\mathbf{n}\|_{\mathcal{X}_{\Gamma_i}}^2\\ &\le C\|\mathbf{w}_h - \Upsilon_{\partial\Gamma_{ij}}(\mathbf{w}_h)\|_{\mathbf{curl};\Omega_i}^2\\ &= C(\|\mathbf{curl}\ \mathbf{w}_h\|_{0;\Omega_i}^2 + d^{-2}\|\mathbf{w}_h - \Upsilon_{\partial\Gamma_{ij}}(\mathbf{w}_h)\|_{0;\Omega_i}^2)\\ &= C(\|\mathbf{curl}\ \mathbf{v}_h\|_{0;\Omega_i}^2 + d^{-2}\|\mathbf{w}_h - \Upsilon_{\partial\Gamma_{ij}}(\mathbf{w}_h)\|_{0;\Omega_i}^2).\end{aligned}$$

Combining this with (5.12) and using Lemma 4.6 give

$$\|\mathbf{w}_{ij}^i\|_{\mathbf{curl},\Omega_i}^2 \le C[1+\log(d/h)]^3(|\mathbf{w}|_{1,\Omega_i}^2 + \|\mathbf{curl}\ \mathbf{v}_h\|_{0,\Omega_i}^2).$$

With this estimate, (5.10), and (5.11), we come to

$$(5.13)\quad A_i(\mathbf{R}_h^i\mathbf{I}_{ij}^t\Phi_{ij}, \mathbf{R}_h^i\mathbf{I}_{ij}^t\Phi_{ij}) \le C[1+\log(d/h)]^3(|p_h|_{1,\Omega_i}^2 + |\mathbf{w}|_{1,\Omega_i}^2 + \|\mathbf{curl}\ \mathbf{v}_h\|_{0,\Omega_i}^2).$$

Similarly, we have

$$A_j(\mathbf{R}_h^j\mathbf{I}_{ij}^t\Phi_{ij}, \mathbf{R}_h^j\mathbf{I}_{ij}^t\Phi_{ij}) \le C[1+\log(d/h)]^3(|p_h|_{1,\Omega_j}^2 + |\mathbf{w}|_{1,\Omega_j}^2 + \|\mathbf{curl}\ \mathbf{v}_h\|_{0,\Omega_j}^2).$$

So we have proved

$$\begin{aligned}\langle S_{ij}\Phi_{ij}, \Phi_{ij}\rangle_{\Gamma_{ij}} \le C[1+\log(d/h)]^3(&|p_h|_{1,\Omega_i}^2 + |p_h|_{1,\Omega_j}^2 + |\mathbf{w}|_{1,\Omega_i}^2\\ &+|\mathbf{w}|_{1,\Omega_j}^2 + \|\mathbf{curl}\ \mathbf{v}_h\|_{0,\Omega_i}^2 + \|\mathbf{curl}\ \mathbf{v}_h\|_{0,\Omega_j}^2),\end{aligned}$$

or

$$\begin{aligned}(5.14)\quad \sum_{\Gamma_{ij}}\langle \mathbf{S}_{ij}\Phi_{ij},\Phi_{ij}\rangle_{\Gamma_{ij}} &\le C[1+\log(d/h)]^3\sum_{i=1}^{N}(|p_h|^2_{1,\Omega_i}+|\mathbf{w}|^2_{1,\Omega_i}+\|\mathbf{curl}\ \mathbf{v}_h\|^2_{0,\Omega_i})\\ &= C[1+\log(d/h)]^3\Bigg(|p_h|^2_{1,\Omega}+|\mathbf{w}|^2_{1,\Omega}\\ &\qquad + \sum_{i=1}^{N}\|\mathbf{curl}\ \mathbf{v}_h\|^2_{0,\Omega_i}\Bigg).\end{aligned}$$

To prove (5.9), it suffices to show

$$(5.15)\qquad |p_h|^2_{1,\Omega}+|\mathbf{w}|^2_{1,\Omega}\le C(\|\mathbf{v}_h\|^2_{0,\Omega}+\|\mathbf{curl}\ \mathbf{v}_h\|^2_{0,\Omega}),$$

as this, with (5.14), implies

$$\begin{aligned}\sum_{\Gamma_{ij}}\langle \mathbf{S}_{ij}\Phi_{ij},\Phi_{ij}\rangle_{\Gamma_{ij}} &\le C[1+\log(d/h)]^3\sum_{i=1}^{N}(\|\mathbf{curl}\ \mathbf{v}_h\|^2_{0,\Omega_i}+\|\mathbf{v}_h\|^2_{0,\Omega_i})\\ &\le C[1+\log(d/h)]^3\sum_{i=1}^{N}A_i(\mathbf{R}_h^i\mathbf{I}_i\Phi,\mathbf{R}_h^i\mathbf{I}_i\Phi).\end{aligned}$$

Next we show (5.15). It follows from (5.4) and (5.7) that

$$(5.16)\qquad |\mathbf{w}|^2_{1,\Omega}\le C\|\mathbf{curl}\ \mathbf{w}\|^2_{0,\Omega}=C\|\mathbf{curl}\ \mathbf{v}_h\|^2_{0,\Omega}\,.$$

However, by Lemma 4.4 and (5.4) we obtain that

$$\|\mathbf{r}_h\mathbf{w}\|^2_{0,\Omega}\le C\,(h^2\,\|\mathbf{curl}\,\mathbf{v}_h\|^2_{0,\Omega}+h^2\,|\mathbf{w}|^2_{1,\Omega}+\|\mathbf{w}\|^2_{0,\Omega})\le C\,\|\mathbf{curl}\,\mathbf{v}_h\|^2_{0,\Omega}.$$

Inequality (5.15) is then a consequence of this estimate, (5.16), and the triangle inequality

$$\|\nabla p_h\|_{0,\Omega}\le\|\mathbf{v}_h\|_{0,\Omega}+\|\mathbf{r}_h\mathbf{w}\|_{0,\Omega}\,.$$

Step 3. Derive the estimate

$$(5.17)\qquad \langle\mathbf{S}_{01}\Phi_{01},\Phi_{01}\rangle+\langle\mathbf{S}_{02}\Phi_{02},\Phi_{02}\rangle\le C[1+\log(d/h)]^2\langle\mathbf{S}\Phi,\Phi\rangle.$$

It follows from the definitions of $\mathbf{S}_{01}$ and Φ_{01} that

$$\langle\mathbf{S}_{01}\Phi_{01},\Phi_{01}\rangle=[1+\log(d/h)]\sum_{i=1}^{N}\|p_h^0-\gamma_{\Delta_i}p_h^0\|^2_{h,\Delta_i}.$$

Thus, by (4.3) and the trace theorem, we have

$$\begin{aligned}(5.18)\qquad \langle\mathbf{S}_{01}\Phi_{01},\Phi_{01}\rangle &\le C[1+\log(d/h)]^2\sum_{i=1}^{N}|p_h|^2_{\frac{1}{2},\Gamma_i}\\ &\le C[1+\log(d/h)]^2\sum_{i=1}^{N}|p_h|^2_{1,\Omega_i}\end{aligned}$$

$$\leq C[1+\log(d/h)]^2|p_h|^2_{1,\Omega}.$$

By the definitions of $\mathbf{S}_{02}$ and Φ_{02}, we know

$$\langle \mathbf{S}_{02}\Phi_{02},\Phi_{02}\rangle = [1+\log(d/h)]\sum_{i=1}^{N}(\|(\mathbf{w}_h-\Upsilon_{\Delta_i}(\mathbf{w}_h))\times\mathbf{n}\|^2_{*,\Delta_i}+d^2\|\mathbf{w}_h\times\mathbf{n}\|^2_{*,\Delta_i}). \tag{5.19}$$

From the definition of $\Upsilon_{\Delta_i}(\mathbf{w}_h)$, we have

$$\|(\mathbf{w}_h-\Upsilon_{\Delta_i}(\mathbf{w}_h))\times\mathbf{n}\|^2_{*,\Delta_i}\leq\|(\mathbf{w}_h-\Upsilon_{\Gamma_i}(\mathbf{w}))\times\mathbf{n}\|^2_{*,\Delta_i}.$$

This, with Lemma 4.5 and the Poincaré inequality, gives

$$\begin{aligned}\|(\mathbf{w}_h-\Upsilon_{\Delta_i}(\mathbf{w}_h))\times\mathbf{n}\|^2_{*,\Delta_i} &\leq \sum_{\mathrm{F}\subset\Gamma_i}\|(\mathbf{w}_h-\Upsilon_{\Gamma_i}(\mathbf{w}))\times\mathbf{n}\|^2_{*,\mathrm{F}_b}\\ &\leq C[1+\log(d/h)](\|\mathbf{w}-\Upsilon_{\Gamma_i}(\mathbf{w})\|^2_{1,\Omega_i}\\ &\quad+\|\mathbf{curl}\ (\mathbf{v}_h-\Upsilon_{\Gamma_i}(\mathbf{w}))\|^2_{0,\Omega_i})\\ &\leq [1+\log(d/h)](|\mathbf{w}|^2_{1,\Omega_i}+\|\mathbf{curl}\ \mathbf{v}_h\|^2_{0,\Omega_i}).\end{aligned} \tag{5.20}$$

The other terms in (5.19) are estimated by Lemma 4.5 and (5.5) as follows:

$$\begin{aligned}d^2\|\mathbf{w}_h\times\mathbf{n}\|^2_{*,\Delta_i} &= d^2\sum_{\mathrm{F}\subset\Gamma_i}\|\mathbf{w}_h\times\mathbf{n}\|^2_{*,\mathrm{F}_b}\\ &\leq Cd^2[1+\log(d/h)](\|\mathbf{w}\|^2_{1,\Omega_i}+\|\mathbf{curl}\,\mathbf{v}_h\|^2_{0,\Omega_i})\\ &= C[1+\log(d/h)](d^2|\mathbf{w}|^2_{1,\Omega_i}+\|\mathbf{w}\|^2_{0,\Omega_i}+d^2\|\mathbf{curl}\,\mathbf{v}_h\|^2_{0,\Omega_i})\\ &\leq C[1+\log(d/h)](|\mathbf{w}|^2_{1,\Omega_i}+\|\mathbf{v}_h\|^2_{0,\Omega_i}+\|\mathbf{curl}\,\mathbf{v}_h\|^2_{0,\Omega_i}).\end{aligned}$$

So we have proved by (5.19) that

$$\langle \mathbf{S}_{02}\Phi_{02},\Phi_{02}\rangle\leq C[1+\log(d/h)]^2(|\mathbf{w}|^2_{1,\Omega}+\|\mathbf{v}_h\|^2_{0,\Omega}+\|\mathbf{curl}\,\mathbf{v}_h\|^2_{0,\Omega}),$$

which, together with (5.18), yields

$$\begin{aligned}&\langle \mathbf{S}_{01}\Phi_{01},\Phi_{01}\rangle+\langle \mathbf{S}_{02}\Phi_{02},\Phi_{02}\rangle\\ &\leq C[1+\log(d/h)]^2(|p_h|^2_{1,\Omega}+|\mathbf{w}|^2_{1,\Omega}+\|\mathbf{v}_h\|^2_{0,\Omega}+\|\mathbf{curl}\ \mathbf{v}_h\|^2_{0,\Omega})\\ &\leq C[1+\log(d/h)]^2(|p_h|^2_{1,\Omega}+|\mathbf{curl}\ \mathbf{w}|^2_{1,\Omega}+\|\mathbf{v}_h\|^2_{0,\Omega}+\|\mathbf{curl}\ \mathbf{v}_h\|^2_{0,\Omega})\\ &\leq C[1+\log(d/h)]^2(|p_h|^2_{1,\Omega}+\|\mathbf{v}_h\|^2_{0,\Omega}+\|\mathbf{curl}\ \mathbf{v}_h\|^2_{0,\Omega})\\ &\leq C[1+\log(d/h)]^2(\|\mathbf{v}_h\|^2_{0,\Omega}+\|\mathbf{curl}\ \mathbf{v}_h\|^2_{0,\Omega})\\ &\leq C[1+\log(d/h)]^2\langle\mathbf{S}\Phi,\Phi\rangle.\end{aligned}$$

The estimates (5.9) and (5.17) indicate that the constant C_1 in (5.1) can be bounded by $C[1+\log(d/h)]^3$.

Step 4. Estimate the constant C_2 in (5.2). It is easy to see that

$$\mathbf{I}_k\left(\sum_{\Gamma_{ij}}\mathbf{I}^t_{ij}\Psi_{ij}+\Psi_{01}+\Psi_{02}\right)=\sum_{\Gamma_{ij}\subset\Gamma_k}\mathbf{I}^t_{ij}\Phi_{ij}+\mathbf{I}_k\Psi_{01}+\mathbf{I}_k\Psi_{02}.$$

Hence

(5.21)

$$\begin{aligned}&\left\langle \mathbf{S}\left(\sum_{\Gamma_{ij}} \mathbf{I}_{ij}^t\Psi_{ij}+\Psi_{01}+\Psi_{02}\right),\ \sum_{\Gamma_{ij}} \mathbf{I}_{ij}^t\Psi_{ij}+\Psi_{01}+\Psi_{02}\right\rangle\\ &\le C\sum_{k=1}^N\left\{\sum_{\Gamma_{ij}\subset\Gamma_k}\langle \mathbf{S}_k\mathbf{I}_{ij}^t\Psi_{ij},\mathbf{I}_{ij}^t\Phi_{ij}\rangle_{\Gamma_k}+\langle \mathbf{S}_k\mathbf{I}_k\Psi_{01},\mathbf{I}_k\Psi_{01}\rangle_{\Gamma_k}+\langle \mathbf{S}_k\mathbf{I}_k\Psi_{02},\mathbf{I}_k\Psi_{02}\rangle_{\Gamma_k}\right\}\\ &\le C\sum_{k=1}^N\left\{\sum_{\Gamma_{ij}\subset\Gamma_k}\langle \mathbf{S}_{ij}\Psi_{ij},\Psi_{ij}\rangle_{\Gamma_{ij}}+A_k(\mathbf{R}_h^k\mathbf{I}_k\Psi_{01},\mathbf{R}_h^k\mathbf{I}_k\Psi_{01})+A_k(\mathbf{R}_h^k\mathbf{I}_k\Psi_{02},\mathbf{R}_h^k\mathbf{I}_k\Psi_{02})\right\}.\end{aligned}$$

As each face Γ_{ij} is shared only by two subdomains Ω_i and Ω_j, we have

$$\sum_{k=1}^N\sum_{\Gamma_{ij}\subset\Gamma_k}\langle \mathbf{S}_{ij}\Psi_{ij},\Psi_{ij}\rangle_{\Gamma_{ij}}\le C\sum_{\Gamma_{ij}}\langle \mathbf{S}_{ij}\Psi_{ij},\Psi_{ij}\rangle_{\Gamma_{ij}}. \tag{5.22}$$

Note that $\Psi_{01}\in V_h^{01}(\Gamma)$ can be written as

$$\mathbf{I}_k\Psi_{01}=\mathbf{grad}(R_h^k\mathrm{I}_k\pi_0\psi)\times\mathbf{n}\quad\text{on}\ \ \Gamma_k$$

for some $\psi\in Z_h(\Gamma)$, so we have

$$\begin{aligned}A_k(\mathbf{R}_h^k\mathbf{I}_k\Psi_{01},\mathbf{R}_h^k\mathbf{I}_k\Psi_{01})&\le A_k(\mathbf{grad}(R_h^k\mathrm{I}_k\pi_0\psi),\mathbf{grad}(R_h^k\mathrm{I}_k\pi_0\psi))\\ &=|\beta^{\frac12}R_h^k\mathrm{I}_k\pi_0\psi|_{1,\Omega_k}^2\le C|\pi_0\psi|_{\frac12,\Gamma_k}^2.\end{aligned}$$

Then it follows from (4.3) that

$$A_k(\mathbf{R}_h^k\mathbf{I}_k\Psi_{01},\mathbf{R}_h^k\mathbf{I}_k\Psi_{01})\le C[1+\log(d/h)]\|\pi_0\psi-\gamma_{\mathcal{W}_k}(\pi_0\psi)\|_{h,\mathcal{W}_k}^2.$$

This, with the definition of $\mathbf{S}_{01}$, shows

$$\sum_{k=1}^N A_k(\mathbf{R}_h^k\mathbf{I}_k\Psi_{01},\mathbf{R}_h^k\mathbf{I}_k\Psi_{01})\le C\langle \mathbf{S}_{01}\Psi_{01},\Psi_{01}\rangle. \tag{5.23}$$

We next estimate the last term in (5.21). We can write $\Psi_{02}\in V_h^{02}(\Gamma)$ as follows:

$$\mathbf{I}_k\Psi_{02}=\Pi_0\mathbf{v}\times\mathbf{n}=[\Pi_0\mathbf{v}-\Upsilon_{\Delta_k}(\Pi_0\mathbf{v})]\times\mathbf{n}+\Upsilon_{\Delta_k}(\Pi_0\mathbf{v})\times\mathbf{n}\quad\text{on}\ \ \Gamma_i$$

for some $\mathbf{v}\in V_h(\Gamma)$. Then, by the triangle inequality, we obtain

$$\begin{aligned}&A_k(\mathbf{R}_h^k\mathbf{I}_k\Psi_{02},\mathbf{R}_h^k\mathbf{I}_k\Psi_{02})\\ &\le 2A_k(\mathbf{R}_h^k\mathbf{I}_k[\Pi_0\mathbf{v}-\Upsilon_{\Delta_k}(\Pi_0\mathbf{v})]\times\mathbf{n},\mathbf{R}_h^k\mathbf{I}_k[\Pi_0\mathbf{v}-\Upsilon_{\Delta_k}(\Pi_0\mathbf{v})\times\mathbf{n}])\\ &\quad+A_k(\mathbf{R}_h^k\mathbf{I}_k[\Upsilon_{\Delta_k}(\Pi_0\mathbf{v})\times\mathbf{n}],\mathbf{R}_h^k\mathbf{I}_k[\Upsilon_{\Delta_k}(\Pi_0\mathbf{v})\times\mathbf{n}])).\end{aligned}$$

Furthermore, using Lemma 4.1 and the minimum **curl**-energy property of the discrete **curl**-extension, we obtain (note that $\Upsilon_{\Delta_k}(\Pi_0\mathbf{v})$ is a constant vector)

$$\begin{aligned}&A_k(\mathbf{R}_h^k\mathbf{I}_k\Psi_{02},\mathbf{R}_h^k\mathbf{I}_k\Psi_{02})\\ &\le C(\|[\Pi_0\mathbf{v}-\Upsilon_{\mathcal{W}_k}(\Pi_0\mathbf{v})]\times\mathbf{n}\|_{\mathcal{X}_{\Gamma_i}}^2+A_k(\Upsilon_{\Delta_k}(\Pi_0\mathbf{v}),\Upsilon_{\Delta_k}(\Pi_0\mathbf{v})))\\ &=C(\|[\Pi_0\mathbf{v}-\Upsilon_{\Delta_k}(\Pi_0\mathbf{v})]\times\mathbf{n}\|_{\mathcal{X}_{\Gamma_k}}^2+\|\Upsilon_{\Delta_k}(\Pi_0\mathbf{v})\|_{0,\Omega_k}^2),\end{aligned} \tag{5.24}$$

where the last term can be estimated using the Hölder inequality and direct computation:

$$\|\Upsilon_{\Delta_k}(\Pi_0\mathbf{v})\|^2_{0,\Omega_k} = d^3|\Upsilon_{\Delta_k}(\Pi_0\mathbf{v})|^2 \leq Cd^2\|\Upsilon_{\Delta_k}(\Pi_0\mathbf{v})\|^2_{*,\Delta_k} \leq Cd^2\|(\Pi_0\mathbf{v})\times\mathbf{n}\|^2_{*,\Delta_k}. \tag{5.25}$$

Next, we show that the first term in (5.28) has the following bound:

$$\|[\Pi_0\mathbf{v} - \Upsilon_{\Delta_k}(\Pi_0\mathbf{v})]\times\mathbf{n}\|^2_{\mathcal{X}_{\Gamma_k}} \leq C[1+\log(d/h)]\|[\Pi_0\mathbf{v} - \Upsilon_{\Delta_k}(\Pi_0\mathbf{v})]\times\mathbf{n}\|^2_{*,\Delta_k}. \tag{5.26}$$

For this, it suffices to prove

$$\|(\Pi_0\mathbf{v})\times\mathbf{n}\|^2_{\mathcal{X}_{\Gamma_k}} \leq C[1+\log(d/h)]\|(\Pi_0\mathbf{v})\times\mathbf{n}\|^2_{*,\Delta_k} \quad \forall \mathbf{v}\in V_h(\Omega). \tag{5.27}$$

To see this, using the relation

$$\mathbf{I}_k[(\Pi_0\mathbf{v})\times\mathbf{n}] = \mathbf{I}^0_{\Delta_k}(\Pi_0\mathbf{v}\times\mathbf{n}) + \sum_{\mathrm{F}\subset\Gamma_k}\mathbf{I}^0_{\mathrm{F}_\partial}(\Upsilon_{\partial\mathrm{F}}(\Pi_0\mathbf{v})\times\mathbf{n}),$$

we have

$$\|(\Pi_0\mathbf{v})\times\mathbf{n}\|^2_{\mathcal{X}_{\Gamma_k}} \leq C\left(\|\mathbf{I}^0_{\Delta_i}(\Pi_0\mathbf{v}\times\mathbf{n})\|^2_{\mathcal{X}_{\Gamma_k}} + \sum_{\mathrm{F}\subset\Gamma_k}\|\mathbf{I}^0_{\mathrm{F}_\partial}(\Upsilon_{\partial\mathrm{F}}(\Pi_0\mathbf{v})\times\mathbf{n})\|^2_{\mathcal{X}_{\Gamma_k}}\right).$$

This, together with Lemmas 4.8 and 4.9, yields (5.27).

Finally, we obtain using (5.24), (5.25), and (5.26) that

$$A_k(\mathbf{R}^k_h\mathbf{I}_k\Psi_{02}, \mathbf{R}^k_h\mathbf{I}_k\Psi_{02}) \leq C([1+\log(d/h)]\|[\mathbf{v}-\Upsilon_{\Delta_k}(\mathbf{v})]\times\mathbf{n}\|^2_{*,\Delta_k} + d^2\|\mathbf{v}\times\mathbf{n}\|^2_{*,\Delta_k}),$$

which implies

$$\sum_{k=1}^N A_k(\mathbf{R}^k_h\mathbf{I}_k\Psi_{02}, \mathbf{R}^k_h\mathbf{I}_k\Psi_{02}) \leq C\langle\mathbf{S}_{02}\Psi_{02}, \Psi_{02}\rangle.$$

This estimate with (5.22)–(5.23) indicates that the constant C_2 in (5.2) is bounded by a constant independent of h and d.

6. Appendix.

This appendix provides the technical proofs for the auxiliary lemmas in Section 4.

6.1. Proofs of Lemmas 4.5 and 4.6.

In this subsection we shall prove Lemmata 4.5 and 4.6. For this, we first give some auxiliary results. The first lemma can be found in [7], [33].

LEMMA 6.1. *Let $v_h \in Z_h(\Gamma_i)$. Then, for any $\mathrm{F}\subset\Gamma_i$, we have*

$$\|v_h\|_{0,\partial\mathrm{F}} \leq C[1+\log(d/h)]^{\frac{1}{2}}\|v_h\|_{\frac{1}{2},\Gamma_i}, \tag{6.1}$$

$$\|\mathrm{I}^0_{\mathrm{F}}v_h\|_{\frac{1}{2},\Gamma_i} \leq C[1+\log(d/h)]\|v_h\|_{\frac{1}{2},\Gamma_i}, \tag{6.2}$$

$$|\mathrm{I}^0_{\partial\mathrm{F}}v_h|_{\frac{1}{2},\mathrm{F}} \leq C[1+\log(d/h)]^{\frac{1}{2}}\|v_h\|_{\frac{1}{2},\Gamma_i}. \tag{6.3}$$

LEMMA 6.2. *Assume that $\mathbf{v}_h \in Z_h(\Omega_i)^3$. Then, for any face F of Γ_i we have*

$$d^{-2}\|\mathbf{v}_h - \Upsilon_{\partial\mathrm{F}}(\mathbf{v}_h)\|^2_{0,\Omega_i} \leq C[1+\log(d/h)]|\mathbf{v}_h|^2_{1,\Omega_i}. \tag{6.4}$$

Proof. Since $\Upsilon_{\partial\mathrm{F}}(\cdot)$ is invariant with constant vectors, we have

$$\begin{aligned} & d^{-2}\|\mathbf{v}_h - \Upsilon_{\partial\mathrm{F}}(\mathbf{v}_h)\|^2_{0,\Omega_i} = d^{-2}\|\mathbf{v}_h - \Upsilon_{\mathrm{F}}(\mathbf{v}_h) - \Upsilon_{\partial\mathrm{F}}(\mathbf{v}_h - \Upsilon_{\mathrm{F}}(\mathbf{v}_h))\|^2_{0,\Omega_i} \\ (6.5)\quad & \le 2d^{-2}(\|\mathbf{v}_h - \Upsilon_{\mathrm{F}}(\mathbf{v}_h)\|^2_{0,\Omega_i} + \|\Upsilon_{\partial\mathrm{F}}(\mathbf{v}_h - \Upsilon_{\mathrm{F}}(\mathbf{v}_h))\|^2_{0,\Omega_i}). \end{aligned}$$

It can be verified, by the Hölder inequality, that

$$\|\Upsilon_{\partial\mathrm{F}}(\mathbf{v}_h - \Upsilon_{\mathrm{F}}(\mathbf{v}_h))\|^2_{0,\Omega_i} \le Cd^3|\Upsilon_{\partial\mathrm{F}}(\mathbf{v}_h - \Upsilon_{\mathrm{F}}(\mathbf{v}_h))|^2 \le Cd^2\|\mathbf{v}_h - \Upsilon_{\mathrm{F}}(\mathbf{v}_h)\|^2_{0,\partial\mathrm{F}}.$$

This, together with (6.1) and the trace theorem, yields

$$\begin{aligned} d^{-2}\|\Upsilon_{\partial\mathrm{F}}(\mathbf{v}_h - \Upsilon_{\mathrm{F}}(\mathbf{v}_h))\|^2_{0,\Omega_i} &\le C[1+\log(d/h)]\|\mathbf{v}_h - \Upsilon_{\mathrm{F}}(\mathbf{v}_h)\|^2_{\frac{1}{2},\Gamma_i} \\ &\le C[1+\log(d/h)]\|\mathbf{v}_h - \Upsilon_{\mathrm{F}}(\mathbf{v}_h)\|^2_{1,\Omega_i}. \end{aligned}$$

Now (6.4) follows from this, (6.5), and the Friedrich's inequality. □

For any face F of Γ_i, we define a quantity (not a norm) on F_b as follows:

$$\|\mathbf{v}\|_{*,\mathrm{F}_b} = \left(\sum_{K\in\mathrm{F}_b}\|\mathbf{v}\|^2_{0,\partial K}\right)^{\frac{1}{2}} \quad \forall\, \mathbf{v}\in Z_h(\Gamma_i)^3 \quad \text{or} \quad \mathbf{v}\in V_h(\Gamma_i).$$

Lemma 6.3. *Assume that* $\mathbf{v}_h \in Z_h(\Gamma_i)^3$. *Then*

$$\|\mathbf{v}_h\|_{*,\mathrm{F}_b} \le C[1+\log(d/h)]^{\frac{1}{2}}\|\mathbf{v}_h\|_{\frac{1}{2},\Gamma_i}. \tag{6.6}$$

Proof. Consider a triangle $K\in\mathrm{F}_b$, and let e be one of its edges lying on $\partial\mathrm{F}$. Then we have

$$\|\mathbf{v}_h\|^2_{0,\partial K} \le 2(\|\mathbf{v}_h - \Upsilon_e(\mathbf{v}_h)\|^2_{0,\partial K} + \|\Upsilon_e(\mathbf{v}_h)\|^2_{0,\partial K}). \tag{6.7}$$

By the Poincaré inequality we obtain

$$h^{-1}\|\mathbf{v}_h - \Upsilon_e(\mathbf{v}_h)\|^2_{0,\partial K} \le h^{-2}\|\mathbf{v}_h - \Upsilon_e(\mathbf{v}_h)\|^2_{0,K} \le C|\mathbf{v}_h|^2_{1,K}.$$

Thus

$$\|\mathbf{v}_h - \Upsilon_e(\mathbf{v}_h)\|^2_{0,\partial K} \le Ch|\mathbf{v}_h|^2_{1,K}. \tag{6.8}$$

On the other hand, it can be verified directly that

$$\|\Upsilon_e(\mathbf{v}_h)\|^2_{0,\partial K} \le Ch|\Upsilon_e(\mathbf{v}_h)|^2 \le C\|\mathbf{v}_h\|^2_{0,e}.$$

Substituting this and (6.8) into (6.7) and then summing over all the edges e on K yield

$$\|\mathbf{v}_h\|^2_{0,\partial K} \le C(h|\mathbf{v}_h|^2_{1,\mathrm{F}} + \|\mathbf{v}_h\|^2_{0,\partial\mathrm{F}}) \le C(|\mathbf{v}_h|^2_{1/2,\mathrm{F}} + \|\mathbf{v}_h\|^2_{0,\partial\mathrm{F}}).$$

Now, (6.6) follows from (6.1). □

Proof of Lemma 4.5. Let $\mathbf{P}_h$: $L^2(\Omega_i)^3 \to Z_h(\Omega_i)^3$ be the L^2-projection operator, which is known to have the following H^s-stability (with $0\le s\le 1$) and estimate [8]:

$$\|\mathbf{P}_h\mathbf{w}\|_{s,\Omega_i} \le C\|\mathbf{w}\|_{s,\Omega_i}, \quad \|\mathbf{w} - \mathbf{P}_h\mathbf{w}\|_{0,\Omega_i} \le C\,h\,|\mathbf{w}|_{1,\Omega_i}\,. \tag{6.9}$$

It is easy to verify that

$$\|(\mathbf{r}_h\mathbf{w})\times\mathbf{n}\|_{*,\mathrm{F}_b}\le C\|\mathbf{r}_h\mathbf{w}\|^2_{*,\mathrm{F}_b}\le C\sum_{e\subset\mathrm{F}_b}(\|\mathbf{P}_h\mathbf{w}\|^2_{0,e}+\|\mathbf{r}_h\mathbf{w}-\mathbf{P}_h\mathbf{w}\|^2_{0,e}). \tag{6.10}$$

Let $K_e\in\mathcal{T}_h$ be an element in Ω_i with e being one of its edges, and $\{\lambda_i\}_{i=1}^4$ the barycentric basis functions at the four vertices of K_e, λ_1, and λ_2, correspond to two end-points of e. By the expression (2.2) of the edge element basis functions, it is easy to verify that $(\mathbf{r}_h\mathbf{w}-\mathbf{P}_h\mathbf{w})$ can be written, in the element K_e, as

$$\mathbf{r}_h\mathbf{w}-\mathbf{P}_h\mathbf{w}=\left(\sum_{i=1}^4 a_i\lambda_i\,,\ \sum_{i=1}^4 b_i\lambda_i\,,\ \sum_{i=1}^4 c_i\lambda_i\right)^T\,,$$

where $a_i,\ b_i$, and c_i $(i=1,2,3,4)$ are constants which may depend on h. By the standard scaling argument, we obtain

$$\|\mathbf{r}_h\mathbf{w}-\mathbf{P}_h\mathbf{w}\|^2_{0,K_e}\ge\bar{C}h^3\sum_{i=1}^4(a_i^2+b_i^2+c_i^2)\,,$$
$$\|\mathbf{r}_h\mathbf{w}-\mathbf{P}_h\mathbf{w}\|^2_{0,e}\le\tilde{C}h\sum_{i=1}^2(a_i^2+b_i^2+c_i^2).$$

This implies

$$\|\mathbf{r}_h\mathbf{w}-\mathbf{P}_h\mathbf{w}\|^2_{0,e}\le Ch^{-2}\|\mathbf{r}_h\mathbf{w}-\mathbf{P}_h\mathbf{w}\|^2_{0,K_e}\,,$$

and so we have

$$\sum_{e\subset\mathrm{F}_b}\|\mathbf{r}_h\mathbf{w}–\mathbf{P}_h\mathbf{w}\|^2_{0,e}\le Ch^{-2}\sum_{e\subset\mathrm{F}_b}\|\mathbf{r}_h\mathbf{w}–\mathbf{P}_h\mathbf{w}\|^2_{0,K_e}\le Ch^{-2}\|\mathbf{r}_h\mathbf{w}–\mathbf{P}_h\mathbf{w}\|^2_{0,\Omega_i}. \tag{6.11}$$

This with (6.10) leads to

$$\|\mathbf{r}_h\mathbf{w}\|^2_{*,\mathrm{F}_b}\le C(\|\mathbf{P}_h\mathbf{w}\|^2_{*,\mathrm{F}_b}+h^{-2}\|\mathbf{r}_h\mathbf{w}-\mathbf{P}_h\mathbf{w}\|^2_{0,\Omega_i}). \tag{6.12}$$

On the other hand, by (6.6), the trace theorem, and (6.9), we obtain

$$\begin{aligned}\|\mathbf{P}_h\mathbf{w}\|_{*,\mathrm{F}_b}&\le C[1+\log(d/h)]^{\frac12}\|\mathbf{P}_h\mathbf{w}\|_{\frac12,\Gamma_i}\\&\le C[1+\log(d/h)]^{\frac12}\|\mathbf{P}_h\mathbf{w}\|_{1,\Omega_i}\\&\le C[1+\log(d/h)]^{\frac12}\|\mathbf{w}\|_{1,\Omega_i},\end{aligned} \tag{6.13}$$

while by the triangle inequality, (4.5), and (6.9), we deduce

$$\begin{aligned}h^{-1}\|\mathbf{r}_h\mathbf{w}-\mathbf{P}_h\mathbf{w}\|_{0,\Omega_i}&\le h^{-1}(\|\mathbf{r}_h\mathbf{w}-\mathbf{w}\|_{0,\Omega_i}+\|\mathbf{P}_h\mathbf{w}-\mathbf{w}\|_{0,\Omega_i})\\&\le C(|\mathbf{w}|^2_{1,\Omega_i}+\|\mathbf{curl}\,\mathbf{v}_h\|^2_{0,\Omega_i})^{\frac12}.\end{aligned} \tag{6.14}$$

Now, (4.6) follows readily from (6.12)–(6.14). □

Proof of Lemma 4.6. We can write

$$\mathbf{r}_h\mathbf{w}-\Upsilon_{\partial\mathrm{F}}(\mathbf{r}_h\mathbf{w})=(\mathbf{r}_h\mathbf{w}-\mathbf{P}_h\mathbf{w})+(\mathbf{P}_h\mathbf{w}-\Upsilon_{\partial\mathrm{F}}(\mathbf{P}_h\mathbf{w}))+\Upsilon_{\partial\mathrm{F}}(\mathbf{P}_h\mathbf{w}-\mathbf{r}_h\mathbf{w});$$

then, by the triangle inequality,

$$\|\mathbf{r}_h\mathbf{w} - \Upsilon_{\partial \mathrm{F}}(\mathbf{r}_h\mathbf{w})\|^2_{0,\Omega_i} \le 3(\|\mathbf{P}_h\mathbf{w} - \Upsilon_{\partial \mathrm{F}}(\mathbf{P}_h\mathbf{w})\|^2_{0,\Omega_i} + \|\mathbf{r}_h\mathbf{w} - \mathbf{P}_h\mathbf{w}\|^2_{0,\Omega_i} + \|\Upsilon_{\partial \mathrm{F}}(\mathbf{P}_h\mathbf{w} - \mathbf{r}_h\mathbf{w})\|^2_{0,\Omega_i}). \tag{6.15}$$

Using (6.4) and (6.9), we know

$$\|\mathbf{P}_h\mathbf{w} - \Upsilon_{\partial \mathrm{F}}(\mathbf{P}_h\mathbf{w})\|^2_{0,\Omega_i} \le Cd^2[1+\log(d/h)]|\mathbf{P}_h\mathbf{w}|^2_{1,\Omega_i} \le Cd^2[1+\log(d/h)]|\mathbf{w}|^2_{1,\Omega_i}. \tag{6.16}$$

On the other hand, by the definition of $\Upsilon_{\partial \mathrm{F}}$, one can verify directly that

$$\|\Upsilon_{\partial \mathrm{F}}(\mathbf{P}_h\mathbf{w} - \mathbf{r}_h\mathbf{w})\|^2_{0,\Omega_i} \le Cd^3|\Upsilon_{\partial \mathrm{F}}(\mathbf{P}_h\mathbf{w} - \mathbf{r}_h\mathbf{w})|^2 \le C\, d^2\|\mathbf{P}_h\mathbf{w} - \mathbf{r}_h\mathbf{w}\|^2_{0,\mathrm{F}_b}.$$

This with (6.11) gives

$$\|\Upsilon_{\partial \mathrm{F}}(\mathbf{P}_h\mathbf{w} - \mathbf{r}_h\mathbf{w})\|^2_{0,\Omega_i} \le Cd^2h^{-2}\,\|\mathbf{r}_h\mathbf{w} - \mathbf{P}_h\mathbf{w}\|^2_{0,\Omega_i},$$

and so we obtain by (6.14) that

$$\begin{aligned}&\|\mathbf{r}_h\mathbf{w} - \mathbf{P}_h\mathbf{w}\|^2_{0,\Omega_i} + \|\Upsilon_{\partial \mathrm{F}}(\mathbf{P}_h\mathbf{w} - \mathbf{r}_h\mathbf{w})\|^2_{0,\Omega_i}\\ &\le C(1+d^2h^{-2})\|\mathbf{r}_h\mathbf{w} - \mathbf{P}_h\mathbf{w}\|^2_{0,\Omega_i}\\ &\le C(h^2+d^2)(|\mathbf{w}|^2_{1,\Omega_i} + \|\mathbf{curl}\,\mathbf{v}_h\|^2_{0,\Omega_i}).\end{aligned}$$

Now (4.7) follows from this, (6.15), and (6.16).

Finally, the relation

$$\mathbf{w} - \Upsilon_{\partial \mathrm{F}}(\mathbf{r}_h\mathbf{w}) = (\mathbf{w} - \mathbf{r}_h\mathbf{w}) + (\mathbf{r}_h\mathbf{w} - \Upsilon_{\partial \mathrm{F}}(\mathbf{r}_h\mathbf{w})),$$

with (4.7) and Lemma 4.4, leads to (4.8) directly. □

6.2. Proofs of Lemmas 4.7, 4.8, and 4.9. The proofs of these lemmas are rather technical, and we will start with some auxiliary results.

LEMMA 6.4. *For any $\Phi \in V_h(\Gamma_i)$, we have*

$$\|\Phi\|_{0,\Gamma_i} \le Ch^{-\frac{1}{2}}\|\Phi\|_{-\frac{1}{2},\Gamma_i}, \quad \|\mathbf{I}^0_{\mathrm{F}_b}\Phi\|_{0,\mathrm{F}} \le Ch^{\frac{1}{2}}\|\Phi\|_{*,\mathrm{F}_b}. \tag{6.17}$$

Proof. The first estimate was proved in [3]. We prove only the second inequality in (6.17). For any $\Phi \in V_h(\Gamma_i)$, we can write $\Phi = \mathbf{v} \times \mathbf{n}$ on Γ_i for some $\mathbf{v} \in V_h(\Omega_i)$. Using the definitions of $\mathbf{I}^0_{\mathrm{F}_b}$, we deduce

$$\|\mathbf{I}^0_{\mathrm{F}_b}\Phi\|^2_{0,\mathrm{F}} \le C\sum_{e\subset \mathrm{F}_b}\lambda^2_e(\mathbf{v})\|L_e \times \mathbf{n}_i\|^2_{0,\mathrm{F}}. \tag{6.18}$$

It follows by (2.2) that $\|L_e \times \mathbf{n}\|^2_{0,\mathrm{F}} \le C$. This, together with (6.18), yields

$$\|\mathbf{I}^0_{\mathrm{F}_b}\Phi\|^2_{0,\mathrm{F}} \le C\sum_{e\subset \mathrm{F}_b}\lambda^2_e(\mathbf{v}).$$

Now we need only to prove

$$\lambda^2_e(\mathbf{v}) \le Ch\|\Phi\|^2_{0,e} \quad \forall\, e \subset \mathrm{F}_b \subset \Gamma_i. \tag{6.19}$$

Noting the fact that $\mathbf{v} = (\mathbf{v} \cdot \mathbf{n})\mathbf{n} + \mathbf{n} \times \mathbf{v} \times \mathbf{n}$ on F, for any $e \subset \mathrm{F}$ we have

$$\mathbf{v}|_{\mathrm{F}} \cdot \mathbf{t}_e = (\mathbf{n} \times \mathbf{v} \times \mathbf{n})|_{\mathrm{F}} \cdot \mathbf{t}_e.$$

Thus (6.19) comes readily from the following:

$$\lambda_e^2(\mathbf{v}) = \left|\int_e \mathbf{v} \cdot \mathbf{t}_e ds\right|^2 \leq \int_e |\mathbf{n} \times \mathbf{v} \times \mathbf{n}|^2 ds \int_e |\mathbf{t}_e|^2 ds \leq Ch \int_e |\mathbf{n} \times \mathbf{v}|^2 ds. \quad \square \tag{6.20}$$

LEMMA 6.5. *Let $\Phi \in V_h(\Gamma_i)$, and let $\mathbf{I}^0_{\mathrm{F}_\partial}\Phi$ be defined as in* (2.5). *Then*

$$\|\mathbf{I}^0_{\mathrm{F}_\partial}\Phi\|_{-\frac{1}{2},\Gamma_i} \leq C([1 + \log(d/h)]\|\Phi\|_{-\frac{1}{2},\Gamma_i} + h^{\frac{1}{2}}\|\Phi\|_{*,\mathrm{F}_b}). \tag{6.21}$$

Proof. The proof is similar to that of Lemma 6 in [19]. However, for the reader's convenience, we still give a complete proof below.

For any $\mathbf{v} \in H^{1/2}(\Gamma_i)^3$, let $\mathbf{v}_h \in Z_h(\Gamma_i)^3$ be the $L^2(\Gamma_i)$-projection of $\mathbf{v}$. Then

$$|\langle \mathbf{I}^0_{\mathrm{F}_\partial}\Phi, \mathbf{v}\rangle_{\Gamma_i}| \leq |\langle \mathbf{I}^0_{\mathrm{F}_\partial}\Phi, \mathbf{v} - \mathbf{v}_h\rangle_{\Gamma_i}| + |\langle \mathbf{I}^0_{\mathrm{F}_\partial}\Phi, \mathbf{v}_h\rangle_{\Gamma_i}|. \tag{6.22}$$

It is known that

$$\|\mathbf{v}_h - \mathbf{v}\|_{0,\Gamma_i} \leq Ch^{\frac{1}{2}}\|\mathbf{v}\|_{\frac{1}{2},\Gamma_i}, \quad \|\mathbf{v}_h\|_{\frac{1}{2},\Gamma_i} \leq C\|\mathbf{v}\|_{\frac{1}{2},\Gamma_i}. \tag{6.23}$$

This, together with (6.17), leads to

$$\begin{aligned} |\langle \mathbf{I}^0_{\mathrm{F}_\partial}\Phi, \mathbf{v} - \mathbf{v}_h\rangle_{\Gamma_i}| &\leq \|\mathbf{I}^0_{\mathrm{F}_\partial}\Phi\|_{0,\Gamma_i}\, \|\mathbf{v} - \mathbf{v}_h\|_{0,\Gamma_i} \\ &\leq Ch^{1/2}\|\Phi\|_{0,\Gamma_i}\|\mathbf{v}\|_{\frac{1}{2},\Gamma_i} \leq C\|\Phi\|_{-\frac{1}{2},\Gamma_i}\, \|\mathbf{v}\|_{\frac{1}{2},\Gamma_i}. \end{aligned} \tag{6.24}$$

On the other hand, from the definitions of the operators $\mathbf{I}^0_{\mathrm{F}_\partial}$ and $\mathbf{I}^0_{\mathrm{F}_b}$, we have $\mathbf{I}^0_{\mathrm{F}_\partial}\Phi = \Phi - \mathbf{I}^0_{\mathrm{F}_b}\Phi$ on F. Then

$$|\langle \mathbf{I}^0_{\mathrm{F}_\partial}\Phi, \mathbf{v}_h\rangle_{\Gamma_i}| = |\langle \mathbf{I}^0_{\mathrm{F}_\partial}\Phi, \mathbf{v}_h\rangle_{\mathrm{F}}| \leq |\langle \Phi, \mathbf{v}_h\rangle_{\mathrm{F}}| + |\langle \mathbf{I}^0_{\mathrm{F}_b}\Phi, \mathbf{v}_h\rangle_{\mathrm{F}}|. \tag{6.25}$$

It follows from (6.17) that

$$|\langle \mathbf{I}^0_{\mathrm{F}_b}\Phi, \mathbf{v}_h\rangle_{\mathrm{F}}| \leq \|\mathbf{I}^0_{\mathrm{F}_b}\Phi\|_{0,\mathrm{F}}\, \|\mathbf{v}_h\|_{0,\mathrm{F}} \leq Ch^{\frac{1}{2}}\|\Phi\|_{*,\mathrm{F}_b}\, \|\mathbf{v}_h\|_{\frac{1}{2},\Gamma_i}. \tag{6.26}$$

For the term $\langle \Phi, \mathbf{v}_h\rangle_{\mathrm{F}}$ in (6.25), we use the simple decomposition

$$\mathbf{v}_h(\mathbf{x}) = \mathrm{I}^0_{\mathrm{F}}\mathbf{v}_h(\mathbf{x}) + \mathrm{I}^0_{\partial\mathrm{F}}\mathbf{v}_h(\mathbf{x}) \quad \forall\, \mathbf{x} \in \mathrm{F} \tag{6.27}$$

to derive (note that $\mathrm{I}^0_{\mathrm{F}}\mathbf{v}_h(\mathbf{x}) = \mathbf{0}$ on $\Gamma_i \backslash \mathrm{F}$)

$$\begin{aligned} |\langle \Phi, \mathbf{v}_h\rangle_{\mathrm{F}}| &\leq |\langle \Phi, \mathrm{I}^0_{\mathrm{F}}\mathbf{v}_h\rangle_{\mathrm{F}}| + |\langle \Phi, \mathrm{I}^0_{\partial\mathrm{F}}\mathbf{v}_h\rangle_{\mathrm{F}}| \\ &\leq |\langle \Phi, \mathrm{I}^0_{\mathrm{F}}\mathbf{v}_h\rangle_{\Gamma_i}| + \|\Phi\|_{0,\mathrm{F}}\, \|\mathrm{I}^0_{\partial\mathrm{F}}\mathbf{v}_h\|_{0,\mathrm{F}} \\ &\leq \|\Phi\|_{-\frac{1}{2},\Gamma_i}\, \|\mathrm{I}^0_{\mathrm{F}}\mathbf{v}_h\|_{\frac{1}{2},\Gamma_i} + Ch^{\frac{1}{2}}\|\Phi\|_{0,\Gamma_i}\, \|\mathbf{v}_h\|_{0,\partial\mathrm{F}}, \end{aligned}$$

where a direct computation is used to bound the term $\|\mathrm{I}^0_{\partial\mathrm{F}}\mathbf{v}_h\|_{0,\mathrm{F}}$ by $h^{1/2}\|\mathbf{v}_h\|_{0,\partial\mathrm{F}}$ using the discrete L^2-norm. This with (6.17), (6.2), and (6.1) yields

$$|\langle \Phi, \mathbf{v}_h\rangle_{\mathrm{F}}| \leq C[1 + \log(d/h)]\|\Phi\|_{-\frac{1}{2},\Gamma_i}\, \|\mathbf{v}_h\|_{\frac{1}{2},\Gamma_i}.$$

Substituting it and (6.26) into (6.25) yields

$$|\langle \mathbf{I}^0_{\mathrm{F}_\partial}\Phi, \mathbf{v}_h\rangle_{\mathrm{F}}| \leq C([1+\log(d/h)]\|\Phi\|_{-\frac{1}{2},\Gamma_i} + h^{\frac{1}{2}}\|\Phi\|_{*,\mathrm{F}_b})\|\mathbf{v}_h\|_{\frac{1}{2},\Gamma_i},$$

which, along with (6.22) and (6.24), leads to

$$|\langle \mathbf{I}^0_{\mathrm{F}_\partial}\Phi, \mathbf{v}\rangle_{\Gamma_i}| \leq C([1+\log(d/h)]\|\Phi\|_{-\frac{1}{2},\Gamma_i} + h^{\frac{1}{2}}\|\Phi\|_{*,\mathrm{F}_b})\|\mathbf{v}\|_{\frac{1}{2},\Gamma_i}.$$

Now (6.21) follows directly from the definition of the norm $\|\cdot\|_{-1/2,\Gamma_i}$. □

Next, we are going to prove Lemma 6.10 on the estimate of $\|\mathrm{div}_\tau(\mathbf{I}^0_{\mathrm{F}}\Phi)\|_{-\frac{1}{2},\Gamma_i}$ for all $\Phi \in V_h(\Gamma_i)$. To do so, we have to present some auxiliary results first (Lemmas 6.6–6.9).

LEMMA 6.6. *Let $\varphi \in L^2(\Gamma_i)$ be piecewise constant with respect to the $\mathcal{T}_h$-induced triangulation $\mathcal{T}_{h,i}$ on Γ_i. Then*

$$\|\varphi\|_{0,\Gamma_i} \leq Ch^{-\frac{1}{2}}\|\varphi\|_{-\frac{1}{2},\Gamma_i}. \tag{6.28}$$

Proof. By definition,

$$\|\varphi\|_{-\frac{1}{2},\Gamma_i} = \sup_{\psi\in H^{1/2}(\Gamma_i)} \frac{|\langle\varphi,\psi\rangle_{\Gamma_i}|}{\|\psi\|_{\frac{1}{2},\Gamma_i}}.$$

The inequality (6.28) then follows if we can construct a function $\psi_0 \in H^{\frac{1}{2}}(\Gamma_i)$ such that

$$|\langle\varphi,\psi_0\rangle_{\Gamma_i}| \geq C\|\varphi\|_{0,\Gamma_i}\|\psi_0\|_{0,\Gamma_i}, \quad \|\psi_0\|_{\frac{1}{2},\Gamma_i} \leq Ch^{-\frac{1}{2}}\|\psi_0\|_{0,\Gamma_i}. \tag{6.29}$$

To construct the function ψ_0 for each triangle $K \in \mathcal{T}_{h,i}$ and lying on Γ_i, with O_K being its barycenter, we refine K by connecting O_K with three vertices of K. Let a_K denote the (constant) value of φ on the triangle K, and let ψ_0 be a piecewise linear function on K with respect to this subdivision such that ψ_0 equals a_K at O_K and vanishes on the edges of K. It is clear that such a function ψ_0 is in $H^{1/2}(\Gamma_i)$. As ψ_0 is piecewise linear on the entire boundary Γ_i with respect to the subdivision of $\mathcal{T}_h$, the second inequality in (6.29) follows directly from the inverse inequality. Moreover, by the equivalent discrete L^2-norms we have

$$\|\psi_0\|^2_{0,\Gamma_i} \leq Ch^2 \sum_{K\in\mathcal{T}_{h,i}} |a_K|^2. \tag{6.30}$$

Let S_K be the area of the triangle K. We have

$$\begin{aligned}|\langle\varphi,\psi_0\rangle_{\Gamma_i}| &= \left|\sum_{K\in\mathcal{T}_{h,i}} \langle\varphi,\psi_0\rangle_K\right| = \left|\sum_{K\in\mathcal{T}_{h,i}} a_K\langle 1,\psi_0\rangle_K\right| \\ &= \frac{1}{3}\left|\sum_{K\in\mathcal{T}_{h,i}} a_K^2 S_K\right| \geq Ch^2\sum_{K\in\mathcal{T}_{h,i}}|a_K^2|.\end{aligned}$$

Now the first inequality of (6.29) follows readily from this and (6.30). □

The next lemma can be shown similarly as Lemma 6.5 by using Lemma 6.6.

LEMMA 6.7. *Let φ be the same as in Lemma 6.6; then*

$$\|\mathrm{I}^t_{\mathrm{F}}(\varphi|_{\mathrm{F}})\|_{-\frac{1}{2},\Gamma_i} \leq C[1+\log(d/h)]\|\varphi\|_{-\frac{1}{2},\Gamma_i}. \tag{6.31}$$

For the proof, we introduce some new functions. For any $\Phi = \mathbf{v}\times\mathbf{n} \in V_h(\Gamma_i)$ and any face $\mathrm{F} \subset \Gamma_i$, we define a function in $L^2(\Gamma_i)$ as follows:

$$\varphi_{\mathrm{F}_b}(\mathbf{x}) = \sum_{e\subset \mathrm{F}_b} \lambda_e(\mathbf{v})(\mathbf{n}_i\cdot \mathbf{curl}\, L_e)(\mathbf{x}), \quad \mathbf{x}\in \bar{\mathrm{F}};\quad \varphi_{\mathrm{F}_b}(\mathbf{x}) = 0, \quad \mathbf{x}\in \Gamma_i\backslash\bar{\mathrm{F}}, \tag{6.32}$$

where $\{L_e; e\in \mathcal{E}_h\}$ are the edge element basis functions defined in (2.2). One can see that φ_{F_b} is piecewise constant on Γ_i, and it vanishes everywhere except in those triangles which are in F and have a vertex on ∂F at least. We now present two estimates for $\varphi_{\mathrm{F}_b}(\mathbf{x})$ below.

Lemma 6.8. *For any $\Phi\in V_h(\Gamma_i)$ and any face* F *of Γ_i, we have*

$$\|\varphi_{\mathrm{F}_b}\|_{-\frac{1}{2},\Gamma_i} \le Ch^{\frac{1}{2}}[1+\log(d/h)]^{\frac{1}{2}}\|\varphi_{\mathrm{F}_b}\|_{0,\mathrm{F}}. \tag{6.33}$$

Proof. For any $v\in H^{1/2}(\Gamma_i)$, let $v_h {\in} Z_h(\Gamma_i)$ be the $L^2(\Gamma_i)$-projection of v. We see directly from (6.27), (6.23), and (6.1) that

$$\begin{aligned}|\langle\varphi_{\mathrm{F}_b}, v\rangle_{\Gamma_i}| &\le |\langle\varphi_{\mathrm{F}_b}, v-v_h\rangle_{\Gamma_i} + |\langle\varphi_{\mathrm{F}_b}, \mathrm{I}^0_{\partial\mathrm{F}}v_h\rangle_{\Gamma_i}| + |\langle\varphi_{\mathrm{F}_b}, \mathrm{I}^0_{\mathrm{F}}v_h\rangle_{\Gamma_i}|\\ &\le Ch^{\frac{1}{2}}[1+\log(d/h)]^{\frac{1}{2}}\|\varphi_{\mathrm{F}_b}\|_{0,\mathrm{F}}\,\|v\|_{\frac{1}{2},\Gamma_i} + |\langle\varphi_{\mathrm{F}_b}, \mathrm{I}^0_{\mathrm{F}}v_h\rangle_{\mathrm{F}}|,\end{aligned}$$

where we have used the fact that $\varphi_{\mathrm{F}_b} = 0$ on $\Gamma_i\backslash$F. It remains to show that

$$|\langle\varphi_{\mathrm{F}_b}, \mathrm{I}^0_{\mathrm{F}}v_h\rangle_{\mathrm{F}}| \le Ch^{\frac{1}{2}}[1+\log(d/h)]^{\frac{1}{2}}\|\varphi_{\mathrm{F}_b}\|_{0,\mathrm{F}}\,\|\mathbf{v}\|_{\frac{1}{2},\Gamma_i}. \tag{6.34}$$

Let F_c denote the union of all triangles that have at least one of their vertices lying on ∂F. We regroup the triangles in F_c such that $\mathrm{F}_c = \cup K$, with each K being one triangle or a union of two triangles and having at least one of its edges lying on ∂F. Then by the definition of φ_{F_b} and the Hölder inequality, we have

$$\begin{aligned}|\langle\varphi_{\mathrm{F}_b}, \mathrm{I}^0_{\mathrm{F}}v_h\rangle_{\mathrm{F}}| = |\langle\varphi_{\mathrm{F}_b}, \mathrm{I}^0_{\mathrm{F}}v_h\rangle_{\mathrm{F}_c}| &= \left|\sum_K\langle\varphi_{\mathrm{F}_b}, \mathrm{I}^0_{\mathrm{F}}v_h\rangle_K\right|\\ &\le \sum_K\|\varphi_{\mathrm{F}_b}\|_{0,K}\,\|\mathrm{I}^0_{\mathrm{F}}v_h\|_{0,K}.\end{aligned} \tag{6.35}$$

As each $K\in \mathrm{F}_c$ has an edge lying on ∂F, $\mathrm{I}^0_{\mathrm{F}}v_h$ vanishes on the edge. Then by Friedrich's inequality we obtain

$$\|\mathrm{I}^0_{\mathrm{F}}v_h\|_{0,K} \le Ch^{\frac{1}{2}}|\mathrm{I}^0_{\mathrm{F}}v_h|_{\frac{1}{2},K}.$$

Plugging this in (6.35) and using the Cauchy–Schwarz inequality, we derive

$$\begin{aligned}|\langle\varphi_{\mathrm{F}_b}, \mathrm{I}^0_{\mathrm{F}}v_h\rangle_{\mathrm{F}}| &\le Ch^{\frac{1}{2}}\left\{\sum_K\|\varphi_{\mathrm{F}_b}\|^2_{0,K}\right\}^{\frac{1}{2}}\left\{\sum_K|\mathrm{I}^0_{\mathrm{F}}v_h|^2_{\frac{1}{2},K}\right\}^{\frac{1}{2}}\\ &= Ch^{\frac{1}{2}}\{\|\varphi_{\mathrm{F}_b}\|^2_{0,\mathrm{F}_c}\}^{\frac{1}{2}}\,\{|\mathrm{I}^0_{\mathrm{F}}v_h|^2_{\frac{1}{2},\mathrm{F}_c}\}^{\frac{1}{2}}\\ &\le Ch^{\frac{1}{2}}\|\varphi_{\mathrm{F}_b}\|_{0,\mathrm{F}}\;\;|\mathrm{I}^0_{\mathrm{F}}v_h|_{\frac{1}{2},\mathrm{F}}.\end{aligned} \tag{6.36}$$

On the other hand, it follows from (6.27) and (6.3) that

$$|\mathrm{I}^0_{\mathrm{F}}v_h|_{\frac{1}{2},\mathrm{F}} = |v_h - \mathrm{I}^0_{\partial\mathrm{F}}v_h|_{\frac{1}{2},\mathrm{F}} \le |v_h|_{\frac{1}{2},\mathrm{F}} + |\mathrm{I}^0_{\partial\mathrm{F}}v_h|_{\frac{1}{2},\mathrm{F}} \le C[1+\log(d/h)]^{\frac{1}{2}}\|v_h\|_{\frac{1}{2},\Gamma_i}.$$

This, together with (6.36), gives (6.34). □

LEMMA 6.9. *Assume that $\Phi = \mathbf{v} \times \mathbf{n} \in V_h(\Gamma_i)$. Then*

$$\|\varphi_{\mathrm{F}_b}\|_{0,\mathrm{F}} \le Ch^{-\frac{1}{2}}\|\Phi\|_{*,\mathrm{F}_b}. \tag{6.37}$$

Proof. We have by the definitions of φ_{F_b} that

$$\|\varphi_{\mathrm{F}_b}\|_{0,\mathrm{F}}^2 \le C \sum_{e\subset \mathrm{F}_b} \lambda_e^2(\mathbf{v})\|\mathbf{n}_i \cdot \mathbf{curl}\ L_e\|_{0,\mathrm{F}}^2. \tag{6.38}$$

It follows from (2.2) that $\mathbf{curl}\, L_e = c_e \nabla\lambda_1^e \times \nabla\lambda_2^e$, which gives $\|\mathbf{n}_i \cdot \mathbf{curl}\, L_e\|_{0,\mathrm{F}}^2 \le Ch^{-2}$. Then we derive from (6.38) that

$$\|\varphi_{\mathrm{F}_b}\|_{0,\mathrm{F}}^2 \le Ch^{-2} \sum_{e\subset \mathrm{F}_b} \lambda_e^2(\mathbf{v}).$$

This, together with (6.20), gives the desired results. □

LEMMA 6.10. *For any $\Phi = \mathbf{v} \times \mathbf{n} \in V_h(\Gamma_i)$, we have*

$$\|\mathrm{div}_\tau(\mathbf{I}_\mathrm{F}^0\Phi)\|_{-\frac{1}{2},\Gamma_i} \le C[1+\log(d/h)]\|\mathrm{div}_\tau\Phi\|_{-\frac{1}{2},\Gamma_i} + C[1+\log(d/h)]^{\frac{1}{2}}\|\Phi\|_{*,\mathrm{F}_b}. \tag{6.39}$$

Proof. We use Lemmas 6.7, 6.8, and 6.9 to estimate $\mathrm{div}_\tau(\mathbf{I}_\mathrm{F}^0\Phi)$. By Green's formula and the definition of $\mathrm{div}_\tau\Phi$, one can verify (cf. [2]) that

$$\mathrm{div}_\tau\Phi = \mathrm{div}_\tau(\mathbf{v}\times\mathbf{n})|_{\Gamma_i} = -(\mathbf{n}_i \cdot \mathbf{curl}\ \mathbf{v})|_{\Gamma_i} \quad \text{in} \ \ H^{-\frac{1}{2}}(\Gamma_i).$$

Thus $\mathrm{div}_\tau\Phi$ is a piecewise constant function on Γ_i. It suffices to prove that

$$\mathrm{div}_\tau(\mathbf{I}_{\mathrm{F}_\partial}^0\Phi) = \mathrm{I}_\mathrm{F}^t(\mathrm{div}_\tau\Phi|_\mathrm{F}) + \varphi_{\mathrm{F}_b} \quad \text{in} \ \ H^{-\frac{1}{2}}(\Gamma_i). \tag{6.40}$$

As $\mathrm{div}_\tau(\mathbf{I}_{\mathrm{F}_\partial}^0\Phi) = 0$ on $\Gamma_i\backslash\bar{\mathrm{F}}$, the inequality (6.40) is valid in $\Gamma_i\backslash\bar{\mathrm{F}}$. However, on the face $\bar{\mathrm{F}}$, we have by (2.4) and (2.5) that

$$\mathrm{div}_\tau\Phi = \sum_{e\subset\overline{\mathrm{F}}} \lambda_e(\mathbf{v})\mathrm{div}_\tau(L_e\times\mathbf{n}_i)\,, \quad \mathrm{div}_\tau(\mathbf{I}_{\mathrm{F}_\partial}^0\Phi) = \sum_{e\subset\mathrm{F}_\partial} \lambda_e(\mathbf{v})\mathrm{div}_\tau(L_e\times\mathbf{n}_i)\,.$$

Hence

$$\mathrm{div}_\tau\Phi - \mathrm{div}_\tau(\mathbf{I}_{\mathrm{F}_\partial}^0\Phi) = \sum_{e\subset\mathrm{F}_b} \lambda_e(\mathbf{v})\mathrm{div}_\tau(L_e\times\mathbf{n}_i) \quad \text{on} \ \ \bar{\mathrm{F}}. \tag{6.41}$$

Noting that (see (2.10) in [2])

$$\mathrm{div}_\tau(L_e\times\mathbf{n}_i)|_{\Gamma_i} = -(\mathbf{n}_i\cdot\mathbf{curl}\ L_e)|_{\Gamma_i} \quad \text{in} \ \ H^{-\frac{1}{2}}(\Gamma_i),$$

we see that (6.40) holds also on $\bar{\mathrm{F}}$, using (6.41) and (6.32). □

The following result can be proved in an analogous way as Lemma 6.6.

LEMMA 6.11. *For any $\Phi \in V_h(\Gamma_i)$ and any face* F *of Γ_i, we have*

$$\|\mathbf{I}_{\mathrm{F}_b}^0\Phi\|_{-\frac{1}{2},\mathrm{F}} \le Ch^{\frac{1}{2}}[1+\log(d/h)]^{\frac{1}{2}}\|\mathbf{I}_{\mathrm{F}_b}^0\Phi\|_{0,\mathrm{F}}. \tag{6.42}$$

Below, we start to prove Lemmas 4.7, 4.8, and 4.9. Lemma 4.7 is a direct consequence of Lemmas 4.5, 6.5, and 6.10, and it indicates that the norm $\|\mathbf{I}_\mathrm{F}^0\Phi\|_{\mathcal{X}_{\Gamma_i}}$ cannot be bounded only by $\|\Phi\|_{\mathcal{X}_{\Gamma_i}}$ (compare to the estimate (6.2)).

Proof of Lemma 4.8. Using (6.40) and the relations

$$\mathbf{I}^0_{\Delta_i}\Phi=\sum_{\mathrm{F}\subset\Gamma_i}\mathbf{I}^t_{\mathrm{F}}(\mathbf{I}^0_{\mathrm{F}_b}\Phi)|_{\mathrm{F}},\quad \mathbf{I}^t_{\mathrm{F}}(\mathbf{I}^0_{\mathrm{F}_b}\Phi)|_{\mathrm{F}}=\mathbf{I}^t_{\mathrm{F}}\Phi-\mathbf{I}^0_{\mathrm{F}_\partial}\Phi$$

and the facts that $\mathbf{I}^t_{\mathrm{F}}\Phi)|_{\Gamma_i\backslash\bar{\mathrm{F}}}=\mathbf{0}$ but $(\mathbf{I}^0_{\mathrm{F}_b}\Phi)|_{\Gamma_i\backslash\bar{\mathrm{F}}}\neq\mathbf{0}$, we can write

$$\begin{aligned}\mathrm{div}_\tau(\mathbf{I}^0_{\Delta_i}\Phi)&=\mathrm{div}_\tau\left(\sum_{\mathrm{F}\subset\Gamma_i}\mathbf{I}^t_{\mathrm{F}}\Phi-\sum_{\mathrm{F}\subset\Gamma_i}\mathbf{I}^0_{\mathrm{F}}\Phi\right)=\mathrm{div}_\tau\left(\Phi-\sum_{\mathrm{F}\subset\Gamma_i}\mathbf{I}^0_{\mathrm{F}_\partial}\Phi\right)\\&=\mathrm{div}_\tau\Phi-\sum_{\mathrm{F}\subset\Gamma_i}\mathrm{div}_\tau(\mathbf{I}^0_{\mathrm{F}_\partial}\Phi)=\sum_{\mathrm{F}\subset\Gamma_i}(\mathbf{I}^t_{\mathrm{F}}\mathrm{div}_\tau(\Phi)|_{\mathrm{F}}-\mathrm{div}_\tau(\mathbf{I}^0_{\mathrm{F}_\partial}\Phi))\\&=\sum_{\mathrm{F}\subset\Gamma_i}\varphi_{\mathrm{F}_b}.\end{aligned}$$

This leads to

$$\|\mathbf{I}^0_{\Delta_i}\Phi\|_{-\frac12,\Gamma_i}\le\sum_{\mathrm{F}\subset\Gamma_i}\|\mathbf{I}^0_{\mathrm{F}_b}\Phi\|_{-\frac12,\mathrm{F}}\,,\quad \|\mathrm{div}_\tau(\mathbf{I}^0_{\Delta_i}\Phi)\|_{-\frac12,\Gamma_i}\le\sum_{\mathrm{F}\subset\Gamma_i}\|\varphi_{\mathrm{F}_b}\|_{-\frac12,\Gamma_i}.$$

Using these two estimates, together with Lemmas 6.11 and 6.8, we have

$$\|\mathbf{I}^0_{\Delta_i}\Phi\|_{\mathcal{X}_{\Gamma_i}}\le Ch^{\frac12}[1+\log(d/h)]^{\frac12}\sum_{\mathrm{F}\subset\Gamma_i}(d^{-1}\|\mathbf{I}^0_{\mathrm{F}_b}\Phi\|_{0,\mathrm{F}}+\|\varphi_{\mathrm{F}_b}\|_{0,\mathrm{F}}).\tag{6.43}$$

Substituting (6.17) and (6.37) into (6.43), we obtain the desired result. □

Proof of Lemma 4.9. By Lemma 6.10 we have

$$\begin{aligned}&\|\mathrm{div}_\tau[\mathbf{I}^0_{\mathrm{F}_\partial}(\Upsilon_{\partial\mathrm{F}}(\Pi_0\mathbf{v})\times\mathbf{n})]\|^2_{-\frac12,\Gamma_k}\\&\le C([1+\log(d/h)]^2\|\mathrm{div}_\tau[\Upsilon_{\partial\mathrm{F}}(\Pi_0\mathbf{v})\times\mathbf{n}|_{\Gamma_k}]\|^2_{-\frac12,\Gamma_k}\\&\quad+[1+\log(d/h)]\|\Upsilon_{\partial\mathrm{F}}(\Pi_0\mathbf{v})\times\mathbf{n}\|^2_{*,\mathrm{F}_b}).\end{aligned}\tag{6.44}$$

It is easy to see that

$$\|\Upsilon_{\partial\mathrm{F}}(\Pi_0\mathbf{v})\times\mathbf{n}_k\|^2_{*,\mathrm{F}_b}=\|\Upsilon_{\partial\mathrm{F}}(\Pi_0\mathbf{v})\times\mathbf{n}_k\|^2_{*,\mathrm{F}_b}\le C\|(\Pi_0\mathbf{v})\times\mathbf{n}\|^2_{*,\mathrm{F}_b}.\tag{6.45}$$

Since $\Upsilon_{\partial\mathrm{F}}(\Pi_0\mathbf{v})$ is a constant vector, we have

$$\mathrm{div}_\tau(\Upsilon_{\partial\mathrm{F}}(\Pi_0\mathbf{v})\times\mathbf{n}|_{\Gamma_k})=0\quad\text{in}\quad H^{-\frac12}(\Gamma_k).$$

Hence

$$\|\mathrm{div}_\tau(\Upsilon_{\partial\mathrm{F}}(\Pi_0\mathbf{v})\times\mathbf{n}|_{\Gamma_k})\|_{-\frac12,\Gamma_k}=0.$$

Substituting (6.45) and the above inequality into (6.44) yields

$$\|\mathrm{div}_\tau[\mathbf{I}^0_{\mathrm{F}_\partial}(\Upsilon_{\partial\mathrm{F}}(\Pi_0\mathbf{v})\times\mathbf{n})]\|^2_{-\frac12,\Gamma_k}\le C[1+\log(d/h)]\|(\Pi_0\mathbf{v})\times\mathbf{n}\|^2_{*,\mathrm{F}_b}.\tag{6.46}$$

On the other hand, it follows from Lemmas 6.11 and 6.4 that

$$\begin{aligned}&d^{-1}\|\mathbf{I}^0_{\mathrm{F}_\partial}(\Upsilon_{\partial\mathrm{F}}(\Pi_0\mathbf{v})\times\mathbf{n})\|_{-\frac12,\Gamma_k}=d^{-1}\|\mathbf{I}^0_{\mathrm{F}_\partial}(\Upsilon_{\partial\mathrm{F}}(\Pi_0\mathbf{v})\times\mathbf{n})\|_{-\frac12,\mathrm{F}}\\&\le C(d^{-1}\|\Upsilon_{\partial\mathrm{F}}(\Pi_0\mathbf{v})\times\mathbf{n}\|_{-\frac12,\mathrm{F}}+d^{-1}h[1+\log(d/h)]^{\frac12}\|\Upsilon_{\partial\mathrm{F}}(\Pi_0\mathbf{v})\times\mathbf{n}\|_{*,\mathrm{F}_b}).\end{aligned}\tag{6.47}$$

However, for any $\Psi \in (H^{\frac{1}{2}}(\mathrm{F}))^3$, we have

$$\begin{aligned} d^{-1}|\langle \Upsilon_{\partial \mathrm{F}}(\Pi_0 \mathbf{v}) \times \mathbf{n}, \Psi\rangle_{\mathrm{F}}| &\le d^{-1}\|\Upsilon_{\partial \mathrm{F}}(\Pi_0 \mathbf{v}) \times \mathbf{n}\|_{0,\mathrm{F}}\, \|\Psi\|_{0,\mathrm{F}} \\ &\le C d^{-\frac{1}{2}} \|\Upsilon_{\partial \mathrm{F}}(\Pi_0 \mathbf{v} \times \mathbf{n})\|_{0,\mathrm{F}}\, \|\Psi\|_{\frac{1}{2},\mathrm{F}} \\ &\le C d^{\frac{1}{2}} |\Upsilon_{\partial \mathrm{F}}(\Pi_0 \mathbf{v} \times \mathbf{n})|\, \|\Psi\|_{\frac{1}{2},\mathrm{F}} \\ &\le C \|(\Pi_0 \mathbf{v}) \times \mathbf{n}\|_{0,\partial \mathrm{F}}\, \|\Psi\|_{\frac{1}{2},\mathrm{F}} \\ &\le C \|(\Pi_0 \mathbf{v}) \times \mathbf{n}\|_{*,\mathrm{F}_b}\, \|\Psi\|_{\frac{1}{2},\mathrm{F}}, \end{aligned}$$

which implies

$$d^{-1}\|\Upsilon_{\partial \mathrm{F}}(\Pi_0 \mathbf{v}) \times \mathbf{n}\|_{-\frac{1}{2},\mathrm{F}} \le C \|(\Pi_0 \mathbf{v}) \times \mathbf{n}\|_{*,\mathrm{F}_b}.$$

Plugging this and (6.45) in (6.47) leads to

$$d^{-1}\|\mathbf{I}^0_{\mathrm{F}_\partial}(\Upsilon_{\partial \mathrm{F}}(\Pi_0 \mathbf{v}) \times \mathbf{n})\|_{-\frac{1}{2},\Gamma_k} \le C[1 + \log(d/h)]^{\frac{1}{2}} \|(\Pi_0 \mathbf{v}) \times \mathbf{n}\|_{*,\mathrm{F}_b},$$

which, together with Lemmas 6.4 and 6.9, gives the desired result. □

Acknowledgments. The authors wish to thank two anonymous referees for many constructive comments which led to a great improvement in the results and the presentation of the paper.

REFERENCES

[1] D. Arnold, R. Falk, and R. Winther, *Multigrid in H(div) and H(**curl**)*, Numer. Math., 85 (2000), pp. 175–195.

[2] A. Alonso and A. Valli, *Some remarks on the characterization of the space of tangential traces of $H(\mathbf{curl};\Omega)$ and the construction of an extension operator*, Manuscripta Math., 89 (1986), pp. 159–178.

[3] A. Alonso and A. Valli, *An optimal domain decomposition preconditioner for low-frequency time-harmonic Maxwell equations*, Math. Comp., 68 (1999), pp. 607–631.

[4] C. Amrouche, C. Bernardi, M. Dauge, and V. Girault, *Vector potentials in three-dimensional nonsmooth domains*, Math. Methods Appl. Sci., 21 (1998), pp. 823–864.

[5] F. Assous, P. Degond, E. Heintzé, P. Raviart, and J. Segre, *On a finite-element method for solving the three-dimensional Maxwell equations*, J. Comput. Phys., 109 (1993), pp. 222–237.

[6] M. Birman and M. Solomyak, *L_2-theory of the Maxwell operator in arbitrary domains*, Russian Math. Surveys, 42 (1987), pp. 75–96.

[7] J. Bramble, J. Pasciak, and A. Schatz, *The construction of preconditioners for elliptic problems by substructuring* IV, Math. Comp., 53 (1989), pp. 1–24.

[8] J. Bramble and J. Xu, *Some estimates for a weighted L^2 projection*, Math. Comp., 56 (1991), pp. 463–476.

[9] M. Cessenat, *Mathematical Methods in Electromagnetism*, World Scientific, River Edge, NJ, 1998.

[10] Z. Chen, Q. Du, and J. Zou, *Finite element methods with matching and nonmatching meshes for Maxwell equations with discontinuous coefficients*, SIAM J. Numer. Anal., 37 (2000), pp. 1542–1570.

[11] P. Ciarlet, Jr. and J. Zou, *Finite element convergence for the Darwin model to Maxwell's equations*, RAIRO Modél. Math. Anal. Numér., 31 (1997), pp. 213–249.

[12] P. Ciarlet, Jr. and J. Zou, *Fully discrete finite element approaches for time-dependent Maxwell's equations*, Numer. Math., 82 (1999), pp. 193–219.

[13] M. Dryja, B. F. Smith, and O. B. Widlund, *Schwarz analysis of iterative substructuring algorithms for elliptic problems in three dimensions*, SIAM J. Numer. Anal., 31 (1994), pp. 1662–1694.

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

[15] J. Gopalakrishnan and J. Pasciak, *Overlapping Schwarz preconditioners for indefinite time harmonic Maxwell's equations*, Math. Comp., 72 (2003), pp. 1–15.

[16] J. Pasciak and J. Zhao, *Overlapping Schwarz methods in H (curl) on nonconvex domains*, East-West J. Numer. Anal., 10 (2002), pp. 221–234.

[17] R. Hiptmair, *Multigrid method for Maxwell's equations*, SIAM J. Numer. Anal., 36 (1998), pp. 204–225.

[18] Q. Hu and G. Liang, *A general framework to construct interface preconditioners*, Chinese J. Numer. Math. Appl., 21 (1999), pp. 83–95.

[19] Q. Hu, G. Liang, and J. Lui, *Construction of a preconditioner for domain decomposition methods with polynomial multipliers*, J. Comput. Math., 19 (2001), pp. 213–224.

[20] Q. Hu and J. Zou, *A Non-overlapping Domain Decomposition Method for Maxwell's Equations in Three Dimensions*, Technical report CUHK 2001-13(232), Department of Mathematics, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong, 2002.

[21] R. Dautray and J.-L. Lions, *Mathematical Analysis and Numerical Methods for Science and Technology*, Springer-Verlag, New York, 1988.

[22] P. Monk, *Analysis of a finite element method for Maxwell's equations*, SIAM J. Numer. Anal., 29 (1992), pp. 714–729.

[23] J. Nédélec, *Mixed finite elements in R^3*, Numer. Math., 35 (1980), pp. 315–341.

[24] R. Nicolaides and D. Wang, *Convergence analysis of a covolume scheme for Maxwell's equations in three dimensions*, Math. Comp., 67 (1998), pp. 947–963.

[25] B. F. Smith, *A domain decomposition algorithm for elliptic problems in three dimensions*, Numer. Math., 60 (1991), pp. 219–234.

[26] B. F. Smith, P. Bjorstad, and W. Gropp, *Domain Decomposition: Parallel Multilevel Methods for Elliptic Partial Differential Equations*, Cambridge University Press, Cambridge, UK, 1996.

[27] P. Tallec, *Domain decomposition methods in computational mechanics*, Comput. Mech. Adv., 2 (1994), pp. 1321–220.

[28] A. Toselli, *Overlapping Schwarz methods for Maxwell's equations in three dimensions*, Numer. Math., 86 (2000), pp. 733–752.

[29] A. Toselli and A. Klawonn, *A FETI domain decomposition method for edge element approximations in two dimensions with discontinuous coefficients*, SIAM J. Numer. Anal., 39 (2001), pp. 932–956.

[30] A. Toselli, O. B. Widlund, and B. I. Wohlmuth, *An iterative substructuring method for Maxwell's equations in two dimensions*, Math. Comp., 70 (2001), pp. 935–949.

[31] B. I. Wohlmuth, A. Toselli, and O. B. Widlund, *An iterative substructuring method for Raviart–Thomas vector fields in three dimensions*, SIAM J. Numer. Anal., 37 (2000), pp. 1657–1676.

[32] J. Xu, *Iterative methods by space decomposition and subspace correction*, SIAM Rev., 34 (1992), pp. 581–613.

[33] J. Xu and J. Zou, *Some nonoverlapping domain decomposition methods*, SIAM Rev., 40 (1998), pp. 857–914.

SIAM J. Numer. Anal.
Vol. 41, No. 5, pp. 1709–1728

A NONOVERLAPPING DOMAIN DECOMPOSITION METHOD FOR ORTHOGONAL SPLINE COLLOCATION PROBLEMS[*]

BERNARD BIALECKI[†] AND MAKSYMILIAN DRYJA[‡]

Abstract. A nonoverlapping domain decomposition approach is used on uniform and matching grids to first define and then to compute the orthogonal spline collocation solution of the Dirichlet boundary value problem for Poisson's equation on an L-shaped region. We prove existence and uniqueness of the collocation solution and derive optimal order H^s-norm error bounds for $s = 0, 1, 2$. The collocation solution on two interfaces is computed using the preconditioned conjugate gradient method, and the collocation solution on three squares is computed by a matrix decomposition method that uses fast Fourier transforms. The total cost of the algorithm is $O(N^2 \log N)$, where the number of unknowns in the collocation solution is $O(N^2)$.

Key words. spline collocation, nonoverlapping domain decomposition, Sobolev norms, Steklov–Poincaré operator, preconditioned conjugate gradient method, separation of variables, fast Fourier transforms

AMS subject classifications. 65N35, 65N12, 65N15, 65N22, 65N55

DOI. 10.1137/S0036142901399793

1. Introduction.

A thorough presentation of overlapping and nonoverlapping domain decomposition methods for solving finite difference and finite element boundary value problems is given in [8, 15, 16] and references therein. Some overlapping methods are considered in [4, 5, 11, 12, 13, 17] for the solution of boundary value problems discretized by orthogonal spline collocation (OSC). In [1], a nonoverlapping method is developed for computing the OSC solution of Poisson's equation on a rectangle. However, to the best of our knowledge, no analysis of nonoverlapping OSC methods for nonrectangular regions is available. In the present paper, we use a nonoverlapping domain decomposition approach on uniform and matching grids to first define and then to compute the OSC solution of the model Dirichlet boundary value problem for Poisson's equation on an L-shaped region. In principle, the proposed approach to define and then to compute the OSC solution can also be used for quasi-uniform and matching grids, more general partial differential equations with variable coefficients, and more general regions with sides parallel to the coordinate axes. Of course, any such extensions may require generalizations of the present proofs and modifications of the employed computational techniques.

Let Ω be the L-shaped region given by $\Omega = \bigcup_{i=1}^{3} \Omega_i \cup \bigcup_{i=1}^{2} \Gamma_i$, where the squares

$$\Omega_1 = (0,1) \times (0,1), \qquad \Omega_2 = (1,2) \times (0,1), \qquad \Omega_3 = (0,1) \times (1,2)$$

and the two interfaces

$$\Gamma_1 = \{1\} \times (0,1), \qquad \Gamma_2 = (0,1) \times \{1\}.$$

[*] Received by the editors December 17, 2001; accepted for publication (in revised form) January 3, 2003; published electronically October 28, 2003.

http://www.siam.org/journals/sinum/41-5/39979.html

[†] Department of Mathematical and Computer Sciences, Colorado School of Mines, Golden, CO 80401-1887 (bbialeck@mines.edu). The research of this author was supported in part by National Science Foundation grant DMS-9805827.

[‡] Department of Mathematics, Warsaw University, Banacha 2, 02-097 Warsaw, Poland (dryja@mimuw.edu.pl).

We consider the model Dirichlet boundary value problem for Poisson's equation

$$\Delta u = f \quad \text{in} \quad \Omega, \qquad u = g \quad \text{on} \quad \partial\Omega, \tag{1.1}$$

which is approximated using a domain decomposition approach and OSC discretization. On each square Ω_i, $i = 1, 2, 3$, the collocation solution is a piecewise Hermite bicubic that satisfies Poisson's equation at the collocation points in the square. The collocation solution is continuous throughout the region Ω, and its normal derivative is continuous at the collocation points on the interfaces Γ_1 and Γ_2. However, continuity of the normal derivative across the interfaces is not guaranteed. We prove existence and uniqueness of the collocation solution and derive optimal order H^s-norm error bounds for $s = 0, 1, 2$. The solution of the collocation problem is reduced to finding the collocation solution on the interfaces. The collocation Steklov–Poincaré operator corresponding to the interfaces is self-adjoint and positive definite with respect to the discrete inner product associated with the collocation points on the interfaces. The right-hand side in the collocation Steklov–Poincaré equation is obtained by solving a collocation problem on each square. With the use of a fast Fourier transform (FFT) matrix decomposition method of [7], this is accomplished at a cost of $O(N^2 \log N)$, where the number of unknowns in the collocation solution is $O(N^2)$. The collocation solution on the interfaces is computed using the preconditioned conjugate gradient (PCG) method with a preconditioner obtained from two collocation Steklov–Poincaré operators corresponding to two pairs of the adjacent squares. It is shown that this preconditioner is spectrally equivalent to the interface operator with spectral constants 1/2 and 2. The cost of each PCG iteration is $O(N^2)$. Once the collocation solution is computed on the interfaces, the collocation solution on each square is obtained at a cost $O(N^2 \log N)$ using the FFT matrix decomposition method of [7]. With the number of PCG iterations proportional to $\log N$, the total cost of the algorithm is $O(N^2 \log N)$.

In our earlier study of OSC for (1.1), we used two additional basis functions associated with the re-entrant corner to formulate a piecewise Hermite bicubic OSC scheme in which the number of unknowns was equal to the number of the collocation points in Ω. With the vertical interface Γ_1 dividing Ω into two rectangles only, we used continuity of the normal derivative of the collocation solution across Γ_1 to obtain a linear system for the coefficients of the collocation solution restricted to Γ_1. Surprisingly the resulting interface matrix was nonsymmetric even though it was an algebraic counterpart of the Steklov–Poincaré operator. The interface system was formed explicitly and then solved by Gauss elimination, which led to an algorithm whose cost was $O(N^3)$. We believe that the nonsymmetry of the interface matrix was a consequence of using two additional basis functions associated with the re-entrant corner.

An outline of this paper is as follows. In section 2, we introduce OSC concepts and state and prove basic results. The OSC problem is defined and analyzed in section 3. In section 4, we formulate an algorithm for solving the OSC problem. The solution of the interface problem is discussed in section 5. The total cost of solving the OSC problem is given in section 6. Finally, in section 7, we present numerical results.

2. Preliminaries. Let N be a positive integer. We set $h = 1/N$ and introduce $t_k = kh$, $k = 0, \dots, N$. Let P_3 be the set of polynomials of degree ≤ 3, and let $\mathcal{M}(0,1)$, $\mathcal{M}^0(0,1)$ be the spaces of piecewise Hermite cubics on $[0,1]$ defined by

$$\begin{aligned} \mathcal{M}(0,1) &= \{v \in C^1[0,1] : v|_{[t_{k-1},t_k]} \in P_3, k = 1, \dots, N\}, \\ \mathcal{M}^0(0,1) &= \{v \in \mathcal{M}(0,1) : v(0) = v(1) = 0\}. \end{aligned} \tag{2.1}$$

Let $\mathcal{G} = \{\xi_l\}_{l=1}^{2N}$ be the set of collocation points in $[0,1]$ given by

$$\xi_{2k-1} = t_{k-1} + h\eta_1, \qquad \xi_{2k} = t_{k-1} + h\eta_2, \tag{2.2}$$

where $k = 1, \ldots, N$, and

$$\eta_1 = \frac{3-\sqrt{3}}{6}, \qquad \eta_2 = \frac{3+\sqrt{3}}{6}. \tag{2.3}$$

The following result is a consequence of Lemma 2.3 in [10].

LEMMA 2.1. *Any $V \in \mathcal{M}^0(0,1)$ is uniquely determined by its values on the set $\mathcal{G}$.*

For V and W defined on $\mathcal{G}$, we introduce

$$\langle V, W\rangle_{\mathcal{G}} = \frac{h}{2}\sum_{\xi\in\mathcal{G}}(VW)(\xi), \qquad \|V\|_{\mathcal{G}} = \sqrt{\langle V, V\rangle}. \tag{2.4}$$

Throughout the paper, C denotes a generic positive constant that is independent of h. The following result is a special case of Lemma 3.1 in [10].

LEMMA 2.2. *For any $V, W \in \mathcal{M}(0,1)$,*

$$\langle -V'', W\rangle_{\mathcal{G}} = \int_0^1 (V'W')(t)\,dt - V'W|_0^1 + \frac{2}{3}Ch^5\sum_{k=1}^{N} V_k^{(3)}W_k^{(3)},$$

where, for $k = 1, \ldots, N$, $V_k^{(3)} = V^{(3)}(t)$, $W_k^{(3)} = W^{(3)}(t)$, $t \in (t_{k-1}, t_k)$.

The next result is a special case of Lemma 3.2 in [10].

LEMMA 2.3. *For any $V \in \mathcal{M}(0,1)$,*

$$Ch^5\sum_{k=1}^{N}\left[V_k^{(3)}\right]^2 \le \|V'\|^2_{L^2(0,1)},$$

where C is the same as in Lemma 2.2.

The next result is inequality (3.4) in [10] (cf. the first inequality in (5.25) of [14]).

LEMMA 2.4. *For any $V \in \mathcal{M}(0,1)$,*

$$\langle V, V\rangle_{\mathcal{G}} \le C\|V\|^2_{L^2(0,1)}.$$

The next result is a generalization of the second inequality in (5.25) of [14].

LEMMA 2.5. *For any $V \in \mathcal{M}(0,1)$,*

$$\|V\|^2_{L^2(0,1)} \le C\left(\|V\|^2_{\mathcal{G}} + hV^2(0) + hV^2(1)\right).$$

Proof. For $V \in \mathcal{M}(0,1)$, we introduce

$$\tilde{V}(t) = \begin{cases} V(0)g(1-t/h), & t \in [t_0, t_1], \\ 0, & t \in [t_1, t_{N-1}, \\ V(1)g(1+(t-1)/h), & t \in [t_{N-1}, t_N], \end{cases} \tag{2.5}$$

where $g(x) = -2x^2 + 3x^2$. Then it is easy to verify that $\tilde{V} \in \mathcal{M}(0,1)$, and $\tilde{V}(t_k) = V(t_k)$, $k = 0, N$. Hence applying the second inequality in (5.25) of [14] to $V - \tilde{V} \in \mathcal{M}^0(0,1)$, and using the triangle inequality, we have

$$\|V - \tilde{V}\|_{L^2(0,1)} \le C\|V - \tilde{V}\|_{\mathcal{G}} \le C\left(\|V\|_{\mathcal{G}} + \|\tilde{V}\|_{\mathcal{G}}\right). \tag{2.6}$$

The triangle inequality and (2.6) yield

$$\|V\|_{L^2(0,1)} \le \|V-\tilde{V}\|_{L^2(0,1)} + \|\tilde{V}\|_{L^2(0,1)} \le C\left(\|V\|_{\mathcal{G}} + \|\tilde{V}\|_{\mathcal{G}} + \|\tilde{V}\|_{L^2(0,1)}\right). \tag{2.7}$$

Using (2.5), it is easy to show that

$$\|\tilde{V}\|^2_{\mathcal{G}}, \|\tilde{V}\|^2_{L^2(0,1)} \le C\left(hV^2(0) + hV^2(1)\right). \tag{2.8}$$

Hence the required result follows from (2.7), the Cauchy–Schwarz inequality, and (2.8). □

Let $\delta_{n,k}$ be the Kronecker delta. The following two lemmas are Lemmas 2.1 and 2.2 in [1].

LEMMA 2.6. *There exist eigenfunctions $\psi_n \in \mathcal{M}^0(0,1)$, $n = 1,\dots,2N$, such that*

$$-\psi_n''(\xi) = \lambda_n \psi_n(\xi), \qquad \xi \in \mathcal{G},$$
$$\langle \psi_n, \psi_k \rangle_{\mathcal{G}} = \frac{h}{2}\delta_{n,k}, \qquad n,k = 1,\dots,2N,$$

where the eigenvalues λ_n, $n = 1,\dots,2N$, all of which are positive, are given by the formulas in Lemma 2.1 *of* [1].

LEMMA 2.7. *For each λ_n, $n = 1,\dots,2N$, of Lemma* 2.6 *there exists a unique $v_n \in \mathcal{M}(0,1)$ such that*

$$v_n''(\xi) = \lambda_n v_n(\xi), \qquad \xi \in \mathcal{G}, \qquad v_n(0) = 0, \qquad v_n(1) = 1.$$

Moreover, a nonzero value of $v_n'(1)$ is given by (2.4)*–*(2.6) *in* [1].

Let B and Z be the $2N \times 2N$ matrices defined, respectively, in (2.9) and (2.10) of [1]. The following lemma follows easily from Lemma 2.3 in [1].

LEMMA 2.8. *Let ψ_n, $n = 1,\dots,2N$, be as in Lemma* 2.6. *Assume $v \in \mathcal{M}^0(0,1)$, and hence $v = \sum_{n=1}^{2N} \alpha_n \psi_n$. If $\vec{\alpha} = [\alpha_1,\dots,\alpha_{2N}]^T$, then*

$$[v(\xi_1),\dots,v(\xi_{2N})]^T = BZ\vec{\alpha}, \qquad \vec{\alpha} = Z^T B^T [v(\xi_1),\dots,v(\xi_{2N})]^T.$$

The following remark follows from (2.9) and (2.10) in [1].

Remark 2.1. The cost of multiplying a vector by B or B^T is $O(N)$, and with the use of FFTs the cost of multiplying a vector by Z or Z^T is $O(N \log N)$.

3. The OSC problem. Let $\mathcal{M}(1,2)$ be the space of piecewise Hermite cubics on $[1,2]$ defined by

$$\mathcal{M}(1,2) = \{v \in C^1[1,2] : v|_{[t_{k-1},t_k]} \in P_3, k = N+1,\dots,2N\},$$

where $t_k = kh$, $k = N,\dots,2N$. We introduce the following spaces of piecewise Hermite bicubics:

$$\mathcal{M}_1 = \mathcal{M}(0,1) \otimes \mathcal{M}(0,1), \qquad \mathcal{M}_2 = \mathcal{M}(1,2) \otimes \mathcal{M}(0,1),$$
$$\mathcal{M}_3 = \mathcal{M}(0,1) \otimes \mathcal{M}(1,2), \tag{3.1}$$

$$X_i = \{v \in \mathcal{M}_i : v = 0 \text{ on } \partial\Omega \cap \partial\Omega_i\}, \qquad i = 1,2,3, \tag{3.2}$$

where $\mathcal{M}(0,1)$ is given by (2.1). We note that $v(1,1) = 0$ for any $v \in X_1$.

Let $\tilde{\mathcal{G}} = \{\xi_l\}_{l=2N+1}^{4N}$ be the set of the collocation points in $[1, 2]$, where ξ_{2k-1}, ξ_{2k}, $k = N+1, \dots, 2N$, are given by (2.2)–(2.3). Then the sets of the collocation points in Ω_1, Ω_2, Ω_3 are defined, respectively, by

$$\mathcal{G}_1 = \mathcal{G} \times \mathcal{G}, \qquad \mathcal{G}_2 = \tilde{\mathcal{G}} \times \mathcal{G}, \qquad \mathcal{G}_3 = \mathcal{G} \times \tilde{\mathcal{G}},$$

where $\mathcal{G} = \{\xi_l\}_{l=1}^{2N}$ is given by (2.2)–(2.3).

Let $\tilde{g}$ be the piecewise Hermite cubic interpolant of g on $\partial\Omega$. The OSC problem for (1.1) consists in finding $U_i \in \mathcal{M}_i$, $i = 1, 2, 3$, such that

$$\Delta U_i(\xi) = f(\xi), \quad \xi \in \mathcal{G}_i, \qquad U_i = \tilde{g} \ \text{ on } \ \partial\Omega \cap \partial\Omega_i, \qquad i = 1, 2, 3, \tag{3.3}$$

and

$$\frac{\partial^j U_1}{\partial x^j}(1, \xi) = \frac{\partial^j U_2}{\partial x^j}(1, \xi), \qquad \frac{\partial^j U_1}{\partial y^j}(\xi, 1) = \frac{\partial^j U_3}{\partial y^j}(\xi, 1), \qquad \xi \in \mathcal{G}, \qquad j = 0, 1. \tag{3.4}$$

It follows from (3.3), (3.4) with $j = 0$ and Lemma 2.1 that

$$U_1|_{\overline{\Gamma}_1} = U_2|_{\overline{\Gamma}_1}, \qquad U_1|_{\overline{\Gamma}_2} = U_3|_{\overline{\Gamma}_2}. \tag{3.5}$$

However, in general

$$\frac{\partial U_1}{\partial x}(1, 1) \neq \frac{\partial U_2}{\partial x}(1, 1), \qquad \frac{\partial U_1}{\partial y}(1, 1) \neq \frac{\partial U_3}{\partial y}(1, 1).$$

To carry out the analysis of the OSC problem, we introduce, for $V_i \in X_i$, $i = 1, 2, 3$,

$$\|(V_1, V_2, V_3)\|_h^2 = \sum_{i=1}^{3} \|V_i\|_{i,h}^2, \tag{3.6}$$

where

$$\|V_1\|_{1,h}^2 = \frac{h}{2} \sum_{\eta \in \mathcal{G}} \left\| \frac{\partial V_1}{\partial x}(\cdot, \eta) \right\|_{L^2(0,1)}^2 + \frac{h}{2} \sum_{\xi \in \mathcal{G}} \left\| \frac{\partial V_1}{\partial y}(\xi, \cdot) \right\|_{L^2(0,1)}^2, \tag{3.7}$$

$$\|V_2\|_{2,h}^2 = \frac{h}{2} \sum_{\eta \in \mathcal{G}} \left\| \frac{\partial V_2}{\partial x}(\cdot, \eta) \right\|_{L^2(1,2)}^2 + \frac{h}{2} \sum_{\xi \in \tilde{\mathcal{G}}} \left\| \frac{\partial V_2}{\partial y}(\xi, \cdot) \right\|_{L^2(0,1)}^2, \tag{3.8}$$

$$\|V_3\|_{3,h}^2 = \frac{h}{2} \sum_{\eta \in \tilde{\mathcal{G}}} \left\| \frac{\partial V_3}{\partial x}(\cdot, \eta) \right\|_{L^2(0,1)}^2 + \frac{h}{2} \sum_{\xi \in \mathcal{G}} \left\| \frac{\partial V_3}{\partial y}(\xi, \cdot) \right\|_{L^2(1,2)}^2. \tag{3.9}$$

It is not difficult to show that, for any stepsize h, $\|\cdot\|_h$ defined by (3.6) is a norm on $X_1 \times X_2 \times X_3$.

For $i = 1, 2, 3$ and V_i, W_i defined on $\mathcal{G}_i$, we introduce

$$\langle V_i, W_i \rangle_{\mathcal{G}_i} = \frac{h^2}{4} \sum_{\xi \in \mathcal{G}_i} (V_i W_i)(\xi), \qquad \|V_i\|_{\mathcal{G}_i} = \sqrt{\langle V_i, V_i \rangle_{\mathcal{G}_i}}. \tag{3.10}$$

To prove existence and uniqueness of the OSC solution, we require the following lemma.

LEMMA 3.1. *If $W_i \in X_i$, $i = 1, 2, 3$, and*

$$\frac{\partial^j W_1}{\partial x^j}(1,\xi) = \frac{\partial^j W_2}{\partial x^j}(1,\xi), \quad \frac{\partial^j W_1}{\partial y^j}(\xi,1) = \frac{\partial^j W_3}{\partial y^j}(\xi,1), \quad j = 0,1, \quad \xi \in \mathcal{G}, \tag{3.11}$$

then

$$\|(W_1, W_2, W_3)\|_h^2 \le \sum_{i=1}^{3} \langle -\Delta W_i, W_i \rangle_{\mathcal{G}_i} \le \frac{5}{3} \|(W_1, W_2, W_3)\|_h^2.$$

Proof. Using (3.10), we obtain

$$\sum_{i=1}^{3} \langle -\Delta W_i, W_i \rangle_{\mathcal{G}_i} = I_x + I_y, \tag{3.12}$$

where

$$\begin{aligned} I_x = {} & \frac{h}{2} \sum_{\eta \in \mathcal{G}} \left\{ \frac{h}{2} \sum_{\xi \in \mathcal{G}} \left(-\frac{\partial^2 W_1}{\partial x^2} W_1 \right)(\xi,\eta) + \frac{h}{2} \sum_{\xi \in \tilde{\mathcal{G}}} \left(-\frac{\partial^2 W_2}{\partial x^2} W_2 \right)(\xi,\eta) \right\} \\ & + \frac{h}{2} \sum_{\eta \in \tilde{\mathcal{G}}} \frac{h}{2} \sum_{\xi \in \mathcal{G}} \left(-\frac{\partial^2 W_3}{\partial x^2} W_3 \right)(\xi,\eta), \end{aligned}$$

$$\begin{aligned} I_y = {} & \frac{h}{2} \sum_{\xi \in \mathcal{G}} \left\{ \frac{h}{2} \sum_{\eta \in \mathcal{G}} \left(-\frac{\partial^2 W_1}{\partial y^2} W_1 \right)(\xi,\eta) + \frac{h}{2} \sum_{\eta \in \tilde{\mathcal{G}}} \left(-\frac{\partial^2 W_3}{\partial y^2} W_3 \right)(\xi,\eta) \right\} \\ & + \frac{h}{2} \sum_{\xi \in \tilde{\mathcal{G}}} \frac{h}{2} \sum_{\eta \in \mathcal{G}} \left(-\frac{\partial^2 W_2}{\partial y^2} W_2 \right)(\xi,\eta). \end{aligned}$$

It follows from Lemma 2.2 and (3.11) that

$$\begin{aligned} I_x = {} & \frac{h}{2} \sum_{\eta \in \mathcal{G}} \left\{ \left\| \frac{\partial W_1}{\partial x}(\cdot,\eta) \right\|_{L^2(0,1)}^2 + \frac{2}{3} C h^5 \sum_{k=1}^{N} \left[W_{1,k}^{(3,0)}(\eta) \right]^2 \right\} \\ & + \frac{h}{2} \sum_{\eta \in \mathcal{G}} \left\{ \left\| \frac{\partial W_2}{\partial x}(\cdot,\eta) \right\|_{L^2(1,2)}^2 + \frac{2}{3} C h^5 \sum_{k=N+1}^{2N} \left[W_{2,k}^{(3,0)}(\eta) \right]^2 \right\} \\ & + \frac{h}{2} \sum_{\eta \in \tilde{\mathcal{G}}} \left\{ \left\| \frac{\partial W_3}{\partial x}(\cdot,\eta) \right\|_{L^2(0,1)}^2 + \frac{2}{3} C h^5 \sum_{k=1}^{N} \left[W_{3,k}^{(3,0)}(\eta) \right]^2 \right\}, \end{aligned} \tag{3.13}$$

where, for $x \in (t_{k-1}, t_k)$,

$$W_{1,k}^{(3,0)}(\eta) = \frac{\partial^3 W_1}{\partial x^3}(x,\eta), \quad W_{2,k}^{(3,0)}(\eta) = \frac{\partial^3 W_2}{\partial x^3}(x,\eta), \quad W_{3,k}^{(3,0)}(\eta) = \frac{\partial^3 W_3}{\partial x^3}(x,\eta).$$

Hence the required inequalities follow from (3.13), a similar expression for I_y, and Lemma 2.3. □

THEOREM 3.1. *The OSC problem* (3.3)–(3.4) *has a unique solution.*

Proof. We assume that $U_i^I, U_i^{II} \in \mathcal{M}_i$, $i = 1, 2, 3$, are two solutions of the OSC problem (3.3)–(3.4). Let $W_i = U_i^I - U_i^{II}$, $i = 1, 2, 3$. Since $\|\cdot\|_h$ is a norm on $X_1 \times X_2 \times X_3$, Lemma 3.1 gives $U_i^I = U_i^{II}$, $i = 1, 2, 3$. This in turn implies existence and uniqueness of the collocation solution since the number of degrees of freedom in the OSC problem is equal to the number of constraints. □

The L^2-norm convergence analysis of the OSC problem is based on the following lemma.

LEMMA 3.2. *If h is sufficiently small, then*

$$\sum_{i=1}^{3} \|V_i\|^2_{L^2(\Omega_i)} \le C\|(V_1, V_2, V_3)\|_h^2, \qquad V_i \in X_i, \qquad i = 1, 2, 3.$$

Proof. Using (3.6)–(3.9) and the Poincaré inequality and interchanging the order in which summation and integration are carried out, we have

$$\begin{aligned} C\|(V_1, V_2, V_3)\|_h^2 &\ge \frac{h}{2} \sum_{\xi \in \mathcal{G}} \left\| \frac{\partial V_1}{\partial y}(\xi, \cdot) \right\|^2_{L^2(0,1)} \\ &+ \int_0^1 \|V_1(x, \cdot)\|_{\mathcal{G}}^2 \, dx + \int_1^2 \|V_2(x, \cdot)\|_{\mathcal{G}}^2 \, dx + \int_1^2 \|V_3(\cdot, y)\|_{\mathcal{G}}^2 \, dy. \end{aligned}$$

Adding and subtracting $hV_1^2(x, 1)$ to and from $\|V_1(x, \cdot)\|_{\mathcal{G}}^2$, and using Lemma 2.5, we obtain

$$\begin{aligned} C\|(V_1, V_2, V_3)\|_h^2 &\ge \frac{h}{2} \sum_{\xi \in \mathcal{G}} \left\| \frac{\partial V_1}{\partial y}(\xi, \cdot) \right\|^2_{L^2(0,1)} - h \int_0^1 V_1^2(x, 1) \, dx \\ &+ C\|V_1\|^2_{L^2(\Omega_1)} + \|V_2\|^2_{L^2(\Omega_2)} + \|V_3\|^2_{L^2(\Omega_3)}. \end{aligned} \tag{3.14}$$

We note that $V_1(\cdot, 1) \in \mathcal{M}^0(0, 1)$ and $V_1(\xi, 1) = \int_0^1 \frac{\partial V_1}{\partial y}(\xi, y) \, dy$, $\xi \in \mathcal{G}$. Hence Lemma 2.5 and the Cauchy–Schwarz inequality give

$$\int_0^1 V_1^2(x, 1) \, dx \le C \frac{h}{2} \sum_{\xi \in \mathcal{G}} V_1^2(\xi, 1) \le C \frac{h}{2} \sum_{\xi \in \mathcal{G}} \left\| \frac{\partial V_1}{\partial y}(\xi, \cdot) \right\|^2_{L^2(0,1)}. \tag{3.15}$$

Thus the required result follows from (3.14) and (3.15). □

THEOREM 3.2. *Assume that u is the solution of* (1.1) *such that* u, $\partial u/\partial x$, $\partial u/\partial y$, *and* $\partial^2 u/\partial x \partial y$ *are continuous on* $\overline{\Omega}$ *and such that* $u|_{\Omega_i} \in H^6(\Omega_i)$, $i = 1, 2, 3$. *Let* $U_i \in \mathcal{M}_i$, $i = 1, 2, 3$, *satisfy* (3.3)–(3.4). *Then, for h sufficiently small,*

$$\sum_{i=1}^{3} \|u - U_i\|_{H^s(\Omega_i)} \le Ch^{4-s} \sum_{i=1}^{3} \|u\|_{H^6(\Omega_i)}, \qquad s = 0, 1, 2.$$

Proof. Let $\tilde{u}_1 \in \mathcal{M}_1$ be the piecewise Hermite bicubic interpolant of $u|_{\Omega_1}$, that is,

$$\frac{\partial^{i+j} \tilde{u}_1}{\partial x^i \partial y^j}(t_k, t_l) = \frac{\partial^{i+j} u}{\partial x^i \partial y^j}(t_k, t_l), \qquad k, l = 0, \dots, N, \qquad i, j = 0, 1,$$

and let, for $i = 2, 3$, the piecewise Hermite bicubic interpolant $\tilde{u}_i \in \mathcal{M}_i$ of $u|_{\Omega_i}$ be defined in a similar way. Let

$$v_i = u|_{\Omega_i} - \tilde{u}_i, \qquad W_i = \tilde{u}_i - U_i, \qquad i = 1, 2, 3. \tag{3.16}$$

Then it follows from [9] that

$$\|v_i\|_{H^s(\Omega_i)} \le Ch^{4-s}\|u\|_{H^4(\Omega_i)}, \qquad i = 1, 2, 3, \qquad s = 0, 1, 2, \tag{3.17}$$

and (3.3), (1.1), and (3.4) imply that $W_i \in X_i$, $i = 1, 2, 3$, and that (3.11) holds. Also, using (3.16), (3.3), and (1.1), we have

$$\sum_{i=1}^{3}\langle -\Delta W_i, W_i\rangle_{\mathcal{G}_i} = \sum_{i=1}^{3}\langle \Delta v_i, W_i\rangle_{\mathcal{G}_i}. \tag{3.18}$$

First we bound $\langle \Delta v_1, W_1\rangle_{\mathcal{G}_1}$ on the right-hand side of (3.18). It follows from (2.20) in [2] and the proof of (2.22) in [2] that

$$\begin{aligned} &\|\Delta v_1\|_{\mathcal{G}_1} \le Ch^3\|u\|_{H^5(\Omega_1)}, \\ &S_1 \equiv \left(\frac{h^2}{16}\sum_{k,l=1}^{N}\left[\sum_{i,j=0}^{1}\Delta v_1(\xi_{2k-i}, \xi_{2l-j})\right]^2\right)^{1/2} \le Ch^4\|u\|_{H^6(\Omega_1)}. \end{aligned} \tag{3.19}$$

Let

$$\sigma_{k,l} = \frac{1}{4}\sum_{i,j=0}^{1} W_1(\xi_{2k-i}, \xi_{2l-j}), \qquad k, l = 1, \dots, N. \tag{3.20}$$

By (3.20), the Cauchy–Schwarz inequality, Lemma 2.4, the Poincaré inequality, and (3.7), we obtain

$$h^2\sum_{k,l=1}^{N}\sigma_{k,l}^2 \le \frac{h}{2}\sum_{\eta\in\mathcal{G}}\|W_1(\cdot,\eta)\|_{\mathcal{G}}^2 \le C\frac{h}{2}\sum_{\eta\in\mathcal{G}}\left\|\frac{\partial W_1}{\partial x}(\cdot,\eta)\right\|_{L^2(0,1)}^2 \le C\|W_1\|_{1,h}^2. \tag{3.21}$$

Also, by the Cauchy–Schwarz inequality, for $k, l = 1, \dots, N$, $i, j = 0, 1$, we have

$$[W_1(\xi_{2k-1}, \xi_{2l-j}) - W_1(\xi_{2k}, \xi_{2l-j})]^2 \le h\left\|\frac{\partial W_1}{\partial x}(\cdot, \xi_{2l-j})\right\|_{L^2(t_{k-1},t_k)}^2,$$

$$[W_1(\xi_{2k-i}, \xi_{2l-1}) - W_1(\xi_{2k-i}, \xi_{2l})]^2 \le h\left\|\frac{\partial W_1}{\partial y}(\xi_{2k-i}, \cdot)\right\|_{L^2(t_{l-1},t_l)}^2,$$

which imply

$$\begin{aligned} &\sum_{i,j=0}^{1}[W_1(\xi_{2k-i}, \xi_{2l-j}) - \sigma_{k,l}]^2 \\ &\le Ch\left(\sum_{j=0}^{1}\left\|\frac{\partial W_1}{\partial x}(\cdot, \xi_{2l-j})\right\|_{L^2(t_{k-1},t_k)}^2 + \sum_{i=0}^{1}\left\|\frac{\partial W_1}{\partial y}(\xi_{2k-i}, \cdot)\right\|_{L^2(t_{l-1},t_l)}^2\right), \end{aligned} \tag{3.22}$$

since, for example, (3.20) gives

$$\begin{aligned} &W_1(\xi_{2k-1}, \xi_{2l-1}) - \sigma_{k,l} = \frac{1}{4}[W_1(\xi_{2k-1}, \xi_{2l-1}) - W_1(\xi_{2k}, \xi_{2l-1})] \\ &+\frac{1}{4}[W_1(\xi_{2k-1}, \xi_{2l-1}) - W_1(\xi_{2k-1}, \xi_{2l})] \\ &+\frac{1}{4}[W_1(\xi_{2k-1}, \xi_{2l-1}) - W_1(\xi_{2k}, \xi_{2l}) \pm W_1(\xi_{2k}, \xi_{2l-1})]. \end{aligned}$$

It follows from (3.22) and (3.7) that

$$S_2 \equiv \left(\frac{h^2}{4} \sum_{k,l=1}^{N} \sum_{i,j=0}^{1} [W_1(\xi_{2k-i}, \xi_{2l-j}) - \sigma_{k,l}]^2 \right)^{1/2} \leq Ch\|W_1\|_{1,h}. \tag{3.23}$$

Using the Cauchy–Schwarz inequality, (3.23), (3.19), and (3.21), we obtain

$$\begin{aligned} \langle \Delta v_1, W_1 \rangle_{\mathcal{G}_1} &= \frac{h^2}{4} \sum_{k,l=1}^{N} \sum_{i,j=0}^{1} \Delta v_1(\xi_{2k-i}, \xi_{2l-j})[W_1(\xi_{2k-i}, \xi_{2l-j}) - \sigma_{k,l}] \\ &+ \frac{h^2}{4} \sum_{k,l=1}^{N} \sigma_{k,l} \sum_{i,j=0}^{1} \Delta v_1(\xi_{2k-i}, \xi_{2l-j}) \leq \|\Delta v_1\|_{\mathcal{G}_1} S_2 + \left(h^2 \sum_{k,l=1}^{N} \sigma_{k,l}^2 \right)^{1/2} S_1 \\ &\leq Ch^4 \|W_1\|_{1,h} \|u\|_{H^6(\Omega_1)}. \end{aligned}$$

In a similar way, we can show that

$$\langle \Delta v_i, W_i \rangle_{\mathcal{G}_i} \leq Ch^4 \|W_i\|_{i,h} \|u\|_{H^6(\Omega_i)}, \qquad i = 2, 3.$$

Hence the Cauchy–Schwarz inequality and (3.6) give

$$\sum_{i=1}^{3} \langle \Delta v_i, W_i \rangle_{\mathcal{G}_i} \leq Ch^4 \|(W_1, W_2, W_3)\|_h \sum_{i=1}^{3} \|u\|_{H^6(\Omega_i)}.$$

Using this inequality, (3.18), Lemmas 3.1 and 3.2, and inverse inequalities [9], we obtain

$$\sum_{i=1}^{3} \|W_i\|_{H^s(\Omega_i)} \leq Ch^{4-s} \sum_{i=1}^{3} \|u\|_{H^6(\Omega_i)}, \qquad s = 0, 1, 2. \tag{3.24}$$

The required inequalities now follow from (3.16), the triangle inequality, (3.17), and (3.24). □

It should be noted that while the exponent on h in the error bounds of Theorem 3.2 is optimal, the smoothness assumption on u is not. The same proof shows that Theorem 3.2 holds for quasi-uniform and matching grids.

4. The algorithm for solving OSC problem.

Taking advantage of uniform partitions of Ω_i, $i = 1, 2, 3$, we develop an FFT algorithm for the solution of the OSC problem (3.3)–(3.4). In the case of quasi-uniform partitions, the corresponding problems in this algorithm can be solved very efficiently by multilevel methods of [6].

Assume that $U_i \in M_i$, $i = 1, 2, 3$, satisfy (3.3)–(3.4). Let $U_{\Gamma_i} \in \mathcal{M}^0(0, 1)$, $i = 1, 2$, be defined by (cf. (3.5))

$$U_{\Gamma_1}(t_k) = U_1(1, t_k) = U_2(1, t_k), \qquad k = 1, \ldots, N-1, \tag{4.1}$$

$$U'_{\Gamma_1}(t_k) = \frac{\partial U_1}{\partial y}(1, t_k) = \frac{\partial U_2}{\partial y}(1, t_k), \qquad k = 0, \ldots, N, \tag{4.2}$$

$$U_{\Gamma_2}(t_k) = U_1(t_k, 1) = U_3(t_k, 1), \qquad k = 1, \ldots, N-1, \tag{4.3}$$

$$U'_{\Gamma_2}(t_k) = \frac{\partial U_1}{\partial x}(t_k, 1) = \frac{\partial U_3}{\partial x}(t_k, 1), \qquad k = 0, \ldots, N. \tag{4.4}$$

For $i = 1, 2, 3$, let $\hat{U}_i \in \mathcal{M}_i$ be such that

$$\Delta \hat{U}_i(\xi) = f(\xi), \quad \xi \in \mathcal{G}_i, \qquad \hat{U}_i = \tilde{g} \quad \text{on} \quad \partial\Omega \cap \partial\Omega_i, \tag{4.5}$$

$$\hat{U}_1(1, t_k) = \hat{U}_1(t_k, 1) = \hat{U}_2(1, t_k) = \hat{U}_3(t_k, 1) = 0, \qquad k = 1, \dots, N-1, \tag{4.6}$$

$$\frac{\partial \hat{U}_1}{\partial y}(1, t_k) = \frac{\partial \hat{U}_1}{\partial x}(t_k, 1) = \frac{\partial \hat{U}_2}{\partial y}(1, t_k) = \frac{\partial \hat{U}_3}{\partial x}(t_k, 1) = 0, \qquad k = 0, \dots, N, \tag{4.7}$$

and let

$$\tilde{U}_i = U_i - \hat{U}_i. \tag{4.8}$$

Then it follows from (4.8), (3.3), and (4.1)–(4.7) that $\tilde{U}_i \in X_i$, $i = 1, 2, 3$, and that

$$\Delta \tilde{U}_1(\xi) = 0, \qquad \xi \in \mathcal{G}_1, \qquad \tilde{U}_1|_{\overline{\Gamma}_1} = U_{\Gamma_1}, \qquad \tilde{U}_1|_{\overline{\Gamma}_2} = U_{\Gamma_2}, \tag{4.9}$$

$$\Delta \tilde{U}_2(\xi) = 0, \qquad \xi \in \mathcal{G}_2, \qquad \tilde{U}_2|_{\overline{\Gamma}_1} = U_{\Gamma_1}, \tag{4.10}$$

$$\Delta \tilde{U}_3(\xi) = 0, \qquad \xi \in \mathcal{G}_3, \qquad \tilde{U}_3|_{\overline{\Gamma}_2} = U_{\Gamma_2}. \tag{4.11}$$

Using (4.8) and (3.4) with $j = 1$, we also have

$$\begin{aligned} \frac{\partial \tilde{U}_1}{\partial x}(1, \xi) - \frac{\partial \tilde{U}_2}{\partial x}(1, \xi) &= \frac{\partial \hat{U}_2}{\partial x}(1, \xi) - \frac{\partial \hat{U}_1}{\partial x}(1, \xi), \qquad \xi \in \mathcal{G}, \\ \frac{\partial \tilde{U}_1}{\partial y}(\xi, 1) - \frac{\partial \tilde{U}_3}{\partial y}(\xi, 1) &= \frac{\partial \hat{U}_3}{\partial y}(\xi, 1) - \frac{\partial \hat{U}_1}{\partial y}(\xi, 1), \qquad \xi \in \mathcal{G}. \end{aligned} \tag{4.12}$$

Based on these derivations, we arrive at the following algorithm to compute $U_i \in \mathcal{M}_i$, $i = 1, 2, 3$, satisfying (3.3)–(3.4):

(4.13)
Step 1. With $\hat{U}_i \in \mathcal{M}_i$, $i = 1, 2, 3$, defined by (4.5)–(4.7), compute the right-hand sides of (4.12).
Step 2. Compute $U_{\Gamma_i} \in \mathcal{M}^0(0, 1)$, $i = 1, 2$, such that $\tilde{U}_i \in X_i$, $i = 1, 2, 3$, satisfy (4.9)–(4.12).
Step 3. Compute $U_i \in \mathcal{M}_i$, $i = 1, 2, 3$, satisfying (3.3) and (4.1)–(4.4).

In the remainder of the paper, we explain how to carry out each step of this algorithm.

5. The interface problem.

This section is concerned with performing step 2 of algorithm (4.13), which is equivalent to solving the interface problem.

5.1. Collocation Steklov–Poincaré operator *K*.

Let $K : [\mathcal{M}^0(0, 1)]^2 \to [\mathcal{M}^0(0, 1)]^2$ be defined for $V_{\Gamma_i} \in \mathcal{M}^0(0, 1)$, $i = 1, 2$, by

$$K(V_{\Gamma_1}, V_{\Gamma_2}) = (W_{\Gamma_1}, W_{\Gamma_2}), \tag{5.1}$$

where $W_{\Gamma_i} \in \mathcal{M}^0(0, 1)$, $i = 1, 2$, are uniquely determined by (cf. Lemma 2.1)

$$W_{\Gamma_1}(\xi) = \frac{\partial V_1}{\partial x}(1, \xi) - \frac{\partial V_2}{\partial x}(1, \xi), \quad W_{\Gamma_2}(\xi) = \frac{\partial V_1}{\partial y}(\xi, 1) - \frac{\partial V_3}{\partial y}(\xi, 1), \quad \xi \in \mathcal{G}, \tag{5.2}$$

and where $V_i \in X_i$, $i = 1, 2, 3$, satisfy

$$\Delta V_1(\xi) = 0, \qquad \xi \in \mathcal{G}_1, \qquad V_1|_{\overline{\Gamma}_1} = V_{\Gamma_1}, \qquad V_1|_{\overline{\Gamma}_2} = V_{\Gamma_2}, \tag{5.3}$$

$$\Delta V_2(\xi) = 0, \qquad \xi \in \mathcal{G}_2, \qquad V_2|_{\overline{\Gamma}_1} = V_{\Gamma_1}, \tag{5.4}$$

$$\Delta V_3(\xi) = 0, \qquad \xi \in \mathcal{G}_3, \qquad V_3|_{\overline{\Gamma}_2} = V_{\Gamma_2}. \tag{5.5}$$

Then step 2 of algorithm (4.13) is equivalent to finding $U_{\Gamma_i} \in \mathcal{M}^0(0,1)$, $i = 1, 2$, such that

$$K(U_{\Gamma_1}, U_{\Gamma_2}) = (F_{\Gamma_1}, F_{\Gamma_2}), \tag{5.6}$$

where $F_{\Gamma_i} \in \mathcal{M}^0(0,1)$, $i = 1, 2$, are given by

$$F_{\Gamma_1}(\xi) = \frac{\partial \hat{U}_2}{\partial x}(1,\xi) - \frac{\partial \hat{U}_1}{\partial x}(1,\xi), \qquad F_{\Gamma_2}(\xi) = \frac{\partial \hat{U}_3}{\partial y}(\xi,1) - \frac{\partial \hat{U}_1}{\partial y}(\xi,1), \qquad \xi \in \mathcal{G},$$

and where $\hat{U}_i \in \mathcal{M}_i$, $i = 1, 2, 3$, satisfy (4.5)–(4.7).

We define the inner product in $[\mathcal{M}^0(0,1)]^2$ by

$$\langle (V_{\Gamma_1}, V_{\Gamma_2}), (W_{\Gamma_1}, W_{\Gamma_2}) \rangle = \sum_{i=1}^{2} \langle V_{\Gamma_i}, W_{\Gamma_i} \rangle_{\mathcal{G}}, \tag{5.7}$$

where $\langle \cdot, \cdot \rangle_{\mathcal{G}}$ is given in (2.4).

THEOREM 5.1. *The operator* $K : [\mathcal{M}^0(0,1)]^2 \to [\mathcal{M}^0(0,1)]^2$ *defined by* (5.1)–(5.2) *is self-adjoint and positive definite with respect to the inner product* (5.7).

Proof. To prove that K is self-adjoint, we have to show that, for $V_{\Gamma_i}, W_{\Gamma_i} \in \mathcal{M}^0(0,1)$, $i = 1, 2$,

$$\langle K(V_{\Gamma_1}, V_{\Gamma_2}), (W_{\Gamma_1}, W_{\Gamma_2}) \rangle = \langle (V_{\Gamma_1}, V_{\Gamma_2}), K(W_{\Gamma_1}, W_{\Gamma_2}) \rangle.$$

Using (5.1)–(5.2) and (5.7), we have

$$\begin{aligned} &\langle K(V_{\Gamma_1}, V_{\Gamma_2}), (W_{\Gamma_1}, W_{\Gamma_2}) \rangle \\ &= \left\langle \frac{\partial V_1}{\partial x}(1,\cdot) - \frac{\partial V_2}{\partial x}(1,\cdot), W_{\Gamma_1} \right\rangle_{\mathcal{G}} + \left\langle \frac{\partial V_1}{\partial y}(\cdot,1) - \frac{\partial V_3}{\partial y}(\cdot,1), W_{\Gamma_2} \right\rangle_{\mathcal{G}}, \end{aligned} \tag{5.8}$$

where $V_i \in X_i$, $i = 1, 2, 3$, satisfy (5.3)–(5.5). Let $W_i \in X_i$, $i = 1, 2, 3$, satisfy

$$\Delta W_1(\xi) = 0, \qquad \xi \in \mathcal{G}_1, \qquad W_1|_{\overline{\Gamma}_1} = W_{\Gamma_1}, \qquad W_1|_{\overline{\Gamma}_2} = W_{\Gamma_2}, \tag{5.9}$$

$$\Delta W_2(\xi) = 0, \qquad \xi \in \mathcal{G}_2, \qquad W_2|_{\overline{\Gamma}_1} = W_{\Gamma_1}, \tag{5.10}$$

$$\Delta W_3(\xi) = 0, \qquad \xi \in \mathcal{G}_3, \qquad W_3|_{\overline{\Gamma}_2} = W_{\Gamma_2}. \tag{5.11}$$

It follows from (5.3), Lemma 2.2, and (5.9) that

$$\begin{aligned} 0 &= \frac{h^2}{4} \sum_{(\xi,\eta)\in\mathcal{G}_1} (-\Delta V_1 W_1)(\xi,\eta) \\ &= \frac{h}{2} \sum_{\eta\in\mathcal{G}} \frac{h}{2} \sum_{\xi\in\mathcal{G}} \left(-\frac{\partial^2 V_1}{\partial x^2} W_1 \right)(\xi,\eta) + \frac{h}{2} \sum_{\xi\in\mathcal{G}} \frac{h}{2} \sum_{\eta\in\mathcal{G}} \left(-\frac{\partial^2 V_1}{\partial y^2} W_1 \right)(\xi,\eta) \\ &= \frac{h}{2} \sum_{\eta\in\mathcal{G}} \left[\int_0^1 \left(\frac{\partial V_1}{\partial x} \frac{\partial W_1}{\partial x} \right)(x,\eta)\, dx - \left(\frac{\partial V_1}{\partial x} W_{\Gamma_1} \right)(1,\eta) \right. \\ &\quad \left. + \frac{2}{3} C h^5 \sum_{k=1}^{N} (V_{1,k}^{(3,0)} W_{1,k}^{(3,0)})(\eta) \right] \\ &\quad + \frac{h}{2} \sum_{\xi\in\mathcal{G}} \left[\int_0^1 \left(\frac{\partial V_1}{\partial y} \frac{\partial W_1}{\partial y} \right)(\xi,y)\, dy - \left(\frac{\partial V_1}{\partial y} W_{\Gamma_2} \right)(\xi,1) \right. \\ &\quad \left. + \frac{2}{3} C h^5 \sum_{k=1}^{N} (V_{1,k}^{(0,3)} W_{1,k}^{(0,3)})(\xi) \right], \end{aligned} \tag{5.12}$$

where, for $t \in (t_{k-1}, t_k)$,

$$V_{1,k}^{(3,0)}(\eta) = \frac{\partial^3 V_1}{\partial x^3}(t, \eta), \qquad W_{1,k}^{(3,0)}(\eta) = \frac{\partial^3 W_1}{\partial x^3}(t, \eta),$$
$$V_{1,k}^{(0,3)}(\xi) = \frac{\partial^3 V_1}{\partial y^3}(\xi, t), \qquad W_{1,k}^{(0,3)}(\xi) = \frac{\partial^3 W_1}{\partial y^3}(\xi, t).$$

In a similar way, using (5.4), (5.5), Lemma 2.2, (5.10), and (5.11), we obtain

$$\begin{aligned} 0 &= \frac{h^2}{4} \sum_{(\xi,\eta)\in\mathcal{G}_2} (-\Delta V_2 W_2)(\xi, \eta) \\ &= \frac{h}{2} \sum_{\eta\in\mathcal{G}} \left[\int_1^2 \left(\frac{\partial V_2}{\partial x} \frac{\partial W_2}{\partial x} \right)(x, \eta)\, dx + \left(\frac{\partial V_2}{\partial x} W_{\Gamma_1} \right)(1, \eta) \right. \\ &\left. + \frac{2}{3} C h^5 \sum_{k=N+1}^{2N} (V_{2,k}^{(3,0)} W_{2,k}^{(3,0)})(\eta) \right] \\ &+ \frac{h}{2} \sum_{\xi\in\tilde{\mathcal{G}}} \left[\int_0^1 \left(\frac{\partial V_2}{\partial y} \frac{\partial W_2}{\partial y} \right)(\xi, y)\, dy + \frac{2}{3} C h^5 \sum_{k=1}^{N} (V_{2,k}^{(0,3)} W_{2,k}^{(0,3)})(\xi) \right] \end{aligned} \tag{5.13}$$

and

$$\begin{aligned} 0 &= \frac{h^2}{4} \sum_{(\xi,\eta)\in\mathcal{G}_3} (-\Delta V_3 W_3)(\xi, \eta) \\ &= \frac{h}{2} \sum_{\eta\in\tilde{\mathcal{G}}} \left[\int_0^1 \left(\frac{\partial V_3}{\partial x} \frac{\partial W_3}{\partial x} \right)(x, \eta)\, dx + \frac{2}{3} C h^5 \sum_{k=1}^{N} (V_{3,k}^{(3,0)} W_{3,k}^{(3,0)})(\eta) \right] \\ &+ \frac{h}{2} \sum_{\xi\in\mathcal{G}} \left[\int_1^2 \left(\frac{\partial V_3}{\partial y} \frac{\partial W_3}{\partial y} \right)(\xi, y)\, dy + \left(\frac{\partial V_3}{\partial y} W_{\Gamma_2} \right)(\xi, 1) \right. \\ &\left. + \frac{2}{3} C h^5 \sum_{k=N+1}^{2N} (V_{3,k}^{(0,3)} W_{3,k}^{(0,3)})(\xi) \right]. \end{aligned} \tag{5.14}$$

Hence (5.8) and (5.12)–(5.14) give

$$\langle K(V_{\Gamma_1}, V_{\Gamma_2}), (W_{\Gamma_1}, W_{\Gamma_2}) \rangle = \sum_{i=1}^{3} I_i(V_i, W_i), \tag{5.15}$$

where

$$\begin{aligned} I_1(V_1, W_1) &= \frac{h}{2} \sum_{\eta\in\mathcal{G}} \left[\int_0^1 \left(\frac{\partial V_1}{\partial x} \frac{\partial W_1}{\partial x} \right)(x, \eta)\, dx + \frac{2}{3} C h^5 \sum_{k=1}^{N} (V_{1,k}^{(3,0)} W_{1,k}^{(3,0)})(\eta) \right] \\ &+ \frac{h}{2} \sum_{\xi\in\mathcal{G}} \left[\int_0^1 \left(\frac{\partial V_1}{\partial y} \frac{\partial W_1}{\partial y} \right)(\xi, y)\, dy + \frac{2}{3} C h^5 \sum_{k=1}^{N} (V_{1,k}^{(0,3)} W_{1,k}^{(0,3)})(\xi) \right], \end{aligned} \tag{5.16}$$

$$\begin{aligned} I_2(V_2,W_2) &= \frac{h}{2}\sum_{\eta\in\mathcal{G}}\left[\int_1^2 \left(\frac{\partial V_2}{\partial x}\frac{\partial W_2}{\partial x}\right)(x,\eta)\,dx + \frac{2}{3}Ch^5\sum_{k=N+1}^{2N}(V_{2,k}^{(3,0)}W_{2,k}^{(3,0)})(\eta)\right] \\ &\quad + \frac{h}{2}\sum_{\xi\in\tilde{\mathcal{G}}}\left[\int_0^1 \left(\frac{\partial V_2}{\partial y}\frac{\partial W_2}{\partial y}\right)(\xi,y)\,dy + \frac{2}{3}Ch^5\sum_{k=1}^{N}(V_{2,k}^{(0,3)}W_{2,k}^{(0,3)})(\xi)\right], \end{aligned} \tag{5.17}$$

$$\begin{aligned} I_3(V_3,W_3) &= \frac{h}{2}\sum_{\eta\in\tilde{\mathcal{G}}}\left[\int_0^1 \left(\frac{\partial V_3}{\partial x}\frac{\partial W_3}{\partial x}\right)(x,\eta)\,dx + \frac{2}{3}Ch^5\sum_{k=1}^{N}(V_{3,k}^{(3,0)}W_{3,k}^{(3,0)})(\eta)\right] \\ &\quad + \frac{h}{2}\sum_{\xi\in\mathcal{G}}\left[\int_1^2 \left(\frac{\partial V_3}{\partial y}\frac{\partial W_3}{\partial y}\right)(\xi,y)\,dy + \frac{2}{3}Ch^5\sum_{k=N+1}^{2N}(V_{3,k}^{(0,3)}W_{3,k}^{(0,3)})(\xi)\right]. \end{aligned} \tag{5.18}$$

Equations (5.15)–(5.18) imply that K is self-adjoint.

Clearly, (5.15)–(5.18) show that $\langle K(V_{\Gamma_1},V_{\Gamma_2}),(V_{\Gamma_1},V_{\Gamma_2})\rangle \geq 0$. To prove that K is positive definite, we assume $\langle K(V_{\Gamma_1},V_{\Gamma_2}),(V_{\Gamma_1},V_{\Gamma_2})\rangle = 0$. Then (5.16) gives

$$\frac{\partial V_1}{\partial x}(t,\xi) = \frac{\partial V_1}{\partial y}(\xi,t) = 0, \qquad t\in[0,1], \qquad \xi\in\mathcal{G}.$$

Since $V_1\in X_1$,

$$V_1(0,\xi) = V_1(\xi,0) = 0, \qquad \xi\in\mathcal{G}.$$

Therefore,

$$V_1(t,\xi) = V_1(\xi,t) = 0, \qquad t\in[0,1], \qquad \xi\in\mathcal{G}.$$

Taking $t=1$ and using Lemma 2.1 and (5.3), we obtain $V_{\Gamma_1} = V_{\Gamma_2} = 0$. □

It follows from Theorem 5.1 that the PCG method is a natural choice for solving the interface problem (5.6).

5.2. Collocation preconditioner P.

In this section, we define a preconditioner P for the operator K of (5.1)–(5.2) and show that K and P are spectrally equivalent.

Let $P:[\mathcal{M}^0(0,1)]^2\to[\mathcal{M}^0(0,1)]^2$ be defined for $V_{\Gamma_i}\in\mathcal{M}^0(0,1)$, $i=1,2$, by

$$P(V_{\Gamma_1},V_{\Gamma_2}) = (W_{\Gamma_1},W_{\Gamma_2}), \tag{5.19}$$

where $W_{\Gamma_i}\in\mathcal{M}^0(0,1)$, $i=1,2$, are uniquely determined by (cf. (5.2))

$$W_{\Gamma_1}(\xi) = \frac{\partial V_1^v}{\partial x}(1,\xi) - \frac{\partial V_2}{\partial x}(1,\xi), \quad W_{\Gamma_2}(\xi) = \frac{\partial V_1^h}{\partial y}(\xi,1) - \frac{\partial V_3}{\partial y}(\xi,1), \quad \xi\in\mathcal{G}, \tag{5.20}$$

and where $V_i\in X_i$, $i=2,3$, satisfy (5.4), (5.5), while $V_1^v, V_1^h\in X_1$ satisfy

$$\Delta V_1^v(\xi) = 0, \qquad \xi\in\mathcal{G}_1, \qquad V_1^v|_{\overline{\Gamma}_1} = V_{\Gamma_1}, \qquad V_1^v|_{\overline{\Gamma}_2} = 0, \tag{5.21}$$

$$\Delta V_1^h(\xi) = 0, \qquad \xi\in\mathcal{G}_1, \qquad V_1^h|_{\overline{\Gamma}_1} = 0, \qquad V_1^h|_{\overline{\Gamma}_2} = V_{\Gamma_2}. \tag{5.22}$$

THEOREM 5.2. *The operator* $P:[\mathcal{M}^0(0,1)]^2\to[\mathcal{M}^0(0,1)]^2$ *defined by* (5.19)–(5.20) *is self-adjoint and positive definite with respect to the inner product* (5.7).

Proof. Following the proof of Theorem 5.1, and using (5.19)–(5.20), (5.7), we have

$$\begin{aligned}&\langle P(V_{\Gamma_1},V_{\Gamma_2}),(W_{\Gamma_1},W_{\Gamma_2})\rangle\\ &=\left\langle \frac{\partial V_1^v}{\partial x}(1,\cdot)-\frac{\partial V_2}{\partial x}(1,\cdot),W_{\Gamma_1}\right\rangle_{\mathcal{G}}+\left\langle \frac{\partial V_1^h}{\partial y}(\cdot,1)-\frac{\partial V_3}{\partial y}(\cdot,1),W_{\Gamma_2}\right\rangle_{\mathcal{G}},\end{aligned}\tag{5.23}$$

where $V_i \in X_i$, $i=2,3$, satisfy (5.4), (5.5) and $V_1^v, V_1^h \in X_1$ satisfy (5.21), (5.22). Let $W_i \in X_i$, $i=2,3$, satisfy (5.10), (5.11), and let $W_1^v, W_1^h \in X_1$ satisfy

$$\Delta W_1^v(\xi)=0,\quad \xi\in\mathcal{G}_1,\quad W_1^v|_{\overline{\Gamma}_1}=W_{\Gamma_1},\quad W_1^v|_{\overline{\Gamma}_2}=0,\tag{5.24}$$

$$\Delta W_1^h(\xi)=0,\quad \xi\in\mathcal{G}_1,\quad W_1^h|_{\overline{\Gamma}_1}=0,\quad W_1^h|_{\overline{\Gamma}_2}=W_{\Gamma_2}.\tag{5.25}$$

Then, using (5.21), (5.22), Lemma 2.2, (5.24), and (5.25), we obtain (cf. (5.12))

$$\begin{aligned}0&=\frac{h^2}{4}\sum_{(\xi,\eta)\in\mathcal{G}_1}(-\Delta V_1^v W_1^v)(\xi,\eta)\\ &=\frac{h}{2}\sum_{\eta\in\mathcal{G}}\left[\int_0^1\left(\frac{\partial V_1^v}{\partial x}\frac{\partial W_1^v}{\partial x}\right)(x,\eta)\,dx-\left(\frac{\partial V_1^v}{\partial x}W_{\Gamma_1}\right)(1,\eta)\right.\\ &\left.+\frac{2}{3}Ch^5\sum_{k=1}^N(V_{1,k}^{v(3,0)}W_{1,k}^{v(3,0)})(\eta)\right]\\ &+\frac{h}{2}\sum_{\xi\in\mathcal{G}}\left[\int_0^1\left(\frac{\partial V_1^v}{\partial y}\frac{\partial W_1^v}{\partial y}\right)(\xi,y)\,dy+\frac{2}{3}Ch^5\sum_{k=1}^N(V_{1,k}^{v(0,3)}W_{1,k}^{v(0,3)})(\xi)\right]\end{aligned}\tag{5.26}$$

and

$$\begin{aligned}0&=\frac{h^2}{4}\sum_{(\xi,\eta)\in\mathcal{G}_1}(-\Delta V_1^h W_1^h)(\xi,\eta)\\ &=\frac{h}{2}\sum_{\eta\in\mathcal{G}}\left[\int_0^1\left(\frac{\partial V_1^h}{\partial x}\frac{\partial W_1^h}{\partial x}\right)(x,\eta)\,dx+\frac{2}{3}Ch^5\sum_{k=1}^N(V_{1,k}^{h(3,0)}W_{1,k}^{h(3,0)})(\eta)\right]\\ &+\frac{h}{2}\sum_{\xi\in\mathcal{G}}\left[\int_0^1\left(\frac{\partial V_1^h}{\partial y}\frac{\partial W_1^h}{\partial y}\right)(\xi,y)\,dy-\left(\frac{\partial V_1^h}{\partial y}W_{\Gamma_2}\right)(\xi,1)\right.\\ &\left.+\frac{2}{3}Ch^5\sum_{k=1}^N(V_{1,k}^{h(0,3)}W_{1,k}^{h(0,3)})(\xi)\right].\end{aligned}\tag{5.27}$$

Since V_i, W_i, $i=2,3$, are the same as in the proof of Theorem 5.1, (5.13) and (5.14) hold true. Hence (5.23)–(5.27), (5.13), and (5.14) give (cf. (5.15))

$$\langle P(V_{\Gamma_1},V_{\Gamma_2}),(W_{\Gamma_1},W_{\Gamma_2})\rangle=I_1(V_1^v,W_1^v)+I_1(V_1^h,W_1^h)+\sum_{i=2}^3 I_i(V_i,W_i),\tag{5.28}$$

where I_i, $i=1,2,3$, are defined in (5.16)–(5.18), respectively. Equations (5.28) and (5.16)–(5.18) imply that P is self-adjoint and positive definite. □

THEOREM 5.3. *The operators K and P are spectrally equivalent with respect to the inner product* (5.7). *In fact, for $V_{\Gamma_1},V_{\Gamma_2}\in\mathcal{M}^0(0,1)$,*

$$\frac{1}{2}\langle P(V_{\Gamma_1},V_{\Gamma_2}),(V_{\Gamma_1},V_{\Gamma_2})\rangle\le\langle K(V_{\Gamma_1},V_{\Gamma_2}),(V_{\Gamma_1},V_{\Gamma_2})\rangle\le 2\langle P(V_{\Gamma_1},V_{\Gamma_2}),(V_{\Gamma_1},V_{\Gamma_2})\rangle.$$

Table 5.1
$\kappa(P^{-1/2}KP^{-1/2})$ *and* $\kappa(K)$.

N	4	8	16	32	64
$\kappa(P^{-1/2}KP^{-1/2})$	2.51	2.62	2.69	2.75	2.79
$\kappa(K)$	8.96	17.79	35.50	70.93	141.81

Proof. Assume $V_{\Gamma_i} \in \mathcal{M}^0(0,1)$, $i = 1, 2$. Then it follows from (5.15) that

$$\langle K(V_{\Gamma_1}, V_{\Gamma_2}), (V_{\Gamma_1}, V_{\Gamma_2})\rangle = \sum_{i=1}^{3} I_i(V_i, V_i), \tag{5.29}$$

where $V_i \in X_i$, $i = 1, 2, 3$, satisfy (5.3)–(5.4), and I_i, $i = 1, 2, 3$, are defined in (5.16)–(5.18). In a similar way, it follows from (5.28) that

$$\langle P(V_{\Gamma_1}, V_{\Gamma_2}), (V_{\Gamma_1}, V_{\Gamma_2})\rangle = I_1(V_1^v, V_1^v) + I_1(V_1^h, V_1^h) + \sum_{i=2}^{3} I_i(V_i, V_i), \tag{5.30}$$

where $V_i \in X_i$, $i = 2, 3$, satisfy (5.4), (5.5) and $V_1^v, V_1^h \in X_1$ satisfy (5.21), (5.22).

Since V_1, V_1^v, and $V_1^h \in X_1$ satisfy (5.3), (5.21), and (5.22), respectively, we have $V_1 = V_1^v + V_1^h$, which, by (5.16) and the inequality $(\alpha + \beta)^2 \leq 2(\alpha^2 + \beta^2), \alpha, \beta \in R$, gives

$$I_1(V_1, V_1) \leq 2[I_1(V_1^v, V_1^v) + I_1(V_1^h, V_1^h)]. \tag{5.31}$$

Hence (5.29), (5.30), and (5.31) imply that $K \leq 2P$.

Since $V_2 \in X_2$ and $V_1^v \in X_1$ satisfy (5.4) and (5.21), respectively, we have $V_1^v(x, y) = V_2(2 - x, y)$, $x, y \in [0, 1]$, which, by (5.16), (5.17), and the symmetry of $\mathcal{G}$ and $\tilde{\mathcal{G}}$ about 1, gives

$$I_1(V_1^v, V_1^v) = I_2(V_2, V_2). \tag{5.32}$$

In a similar way, for $V_3 \in X_3$ and $V_1^h \in X_1$ satisfying, respectively, (5.5) and (5.22), we have $V_1^h(x, y) = V_3(x, 2 - y)$, $x, y \in [0, 1]$, which, by (5.16), (5.18), and the symmetry of $\mathcal{G}$ and $\tilde{\mathcal{G}}$ about 1, gives

$$I_1(V_1^h, V_1^h) = I_3(V_3, V_3). \tag{5.33}$$

Hence (5.30), (5.32), (5.33), and (5.29) imply that $P \leq 2K$. □

With the preconditioner P, the convergence rate of the PCG method applied to (5.6) depends on $\kappa(P^{-1/2}KP^{-1/2})$, where, for a self-adjoint and positive definite operator A, $\kappa(A) = \lambda_{max}(A)/\lambda_{min}(A)$. Since $P^{-1/2}KP^{-1/2}$ and $P^{-1}K$ have the same eigenvalues,

$$\kappa(P^{-1/2}KP^{-1/2}) = \lambda_{max}(P^{-1}K)/\lambda_{min}(P^{-1}K).$$

This formula was used to compute $\kappa(P^{-1/2}KP^{-1/2})$ numerically for several values of N. The results in Table 5.1 confirm the theoretical result $\kappa(P^{-1/2}KP^{-1/2}) \leq 4$ of Theorem 5.3 and also indicate that $\kappa(K) = O(N)$.

5.3. Solving with P and multiplying by K. In this section, we discuss the cost of solving an operator equation with the preconditioner P and the cost of multiplying by the operator K. Throughout this section, ψ_n, γ_n, $n = 1, \dots, 2N$, are as in Lemma 2.6, and v_n, $n = 1, \dots, 2N$, are as in Lemma 2.7.

It follows from the definition (5.19)–(5.20) of the operator P that, given $W_{\Gamma_i} \in \mathcal{M}^0(0,1)$, $i = 1, 2$, solving

$$P(V_{\Gamma_1}, V_{\Gamma_2}) = (W_{\Gamma_1}, W_{\Gamma_2})$$

for $V_{\Gamma_i} \in \mathcal{M}^0(0,1)$, $i = 1, 2$, is equivalent to solving two independent problems. The first problem consists in finding $V_{\Gamma_1} \in \mathcal{M}^0(0,1)$ such that

$$\frac{\partial V_1^v}{\partial x}(1,\xi) - \frac{\partial V_2}{\partial x}(1,\xi) = W_{\Gamma_1}(\xi), \qquad \xi \in \mathcal{G}, \tag{5.34}$$

where $V_1^v \in X_1$ satisfies (5.21) and $V_2 \in X_2$ satisfies (5.4). The second problem consists in finding $V_{\Gamma_2} \in \mathcal{M}^0(0,1)$ such that

$$\frac{\partial V_1^h}{\partial y}(\xi,1) - \frac{\partial V_3}{\partial y}(\xi,1) = W_{\Gamma_2}(\xi), \qquad \xi \in \mathcal{G},$$

where $V_1^h \in X_1$ satisfies (5.22) and $V_3 \in X_3$ satisfies (5.5). We explain how to solve (5.34) using separation of variables; the second problem can be solved in a similar way.

Using Lemmas 2.6 and 2.7, it is easy to show that V_1^v defined by

$$V_1^v(x,y) = \sum_{n=1}^{2N} \alpha_n v_n(x) \psi_n(y), \qquad x, y \in [0,1],$$

where the α_n are arbitrary constants, belongs to X_1 and satisfies $\Delta V_1^v(\xi) = 0$, $\xi \in \mathcal{G}_1$, and $V_1^v|_{\overline{\Gamma}_2} = 0$. Also, it is easy to verify that V_2 defined by

$$V_2(x,y) = V_1^v(2-x,y), \qquad x \in [1,2], \qquad y \in [0,1],$$

belongs to X_2 and satisfies $\Delta V_2(\xi) = 0$, $\xi \in \mathcal{G}_2$. Moreover, V_{Γ_1} defined by

$$V_{\Gamma_1}(y) = V_1^v(1,y) = V_2(1,y) = \sum_{n=1}^{2N} \alpha_n \psi_n(y), \qquad y \in [0,1],$$

belongs to $\mathcal{M}^0(0,1)$. Since $W_{\Gamma_1} \in \mathcal{M}^0(0,1)$, we have

$$W_{\Gamma_1}(y) = \sum_{n=1}^{2N} \beta_n \psi_n(y), \qquad y \in [0,1].$$

Therefore, (5.34) becomes

$$\sum_{n=1}^{2N} 2\alpha_n v_n'(1) \psi_n(\xi) = \sum_{n=1}^{2N} \beta_n \psi_n(\xi), \qquad \xi \in \mathcal{G},$$

which gives

$$\alpha_n = \frac{\beta_n}{2v_n'(1)}, \qquad n = 1, \dots, 2N,$$

where $v_n'(1)$ is as referred to in Lemma 2.7. If we introduce

$$\vec{\alpha} = [\alpha_1, \dots, \alpha_{2N}]^T, \qquad \vec{\beta} = [\beta_1, \dots, \beta_{2N}]^T,$$

then Lemma 2.8 yields

$$\vec{\beta} = Z^T B^T [W_{\Gamma_1}(\xi_1), \dots, W_{\Gamma_1}(\xi_{2N})]^T, \qquad [V_{\Gamma_1}(\xi_1), \dots, V_{\Gamma_1}(\xi_{2N})]^T = BZ\vec{\alpha}.$$

Thus, given $W_{\Gamma_1}(\xi)$, $\xi \in \mathcal{G}$, Remark 2.1 implies that $V_{\Gamma_1}(\xi)$, $\xi \in \mathcal{G}$, can be computed at a cost $O(N \log N)$.

If follows from the definition (5.1)–(5.2) of the operator K that, given $V_{\Gamma_1}, V_{\Gamma_2} \in \mathcal{M}^0(0,1)$, the computation of

$$(W_{\Gamma_1}, W_{\Gamma_2}) = K(V_{\Gamma_1}, V_{\Gamma_2}), \qquad W_{\Gamma_1}, W_{\Gamma_2} \in \mathcal{M}^0(0,1),$$

involves solving collocation problems (5.3)–(5.5). Let $V_1, V_1^v, V_1^h \in X_1$ be, respectively, solutions of (5.3), (5.21), (5.22). Then $V_1 = V_1^v + V_1^h$. Moreover, if $V_2 \in X_2$ and $V_3 \in X_3$ are solutions of (5.4) and (5.5), respectively, then, by the symmetry of $\mathcal{G}$ and $\tilde{\mathcal{G}}$ about 1, we have $V_2(x,y) = V_1^v(2-x, y)$, $x \in [1,2]$, $y \in [0,1]$, and $V_3(x,y) = V_1^h(x, 2-y)$, $x \in [0,1]$, $y \in [1,2]$. Hence, it follows from (5.2) that

$$W_{\Gamma_1}(\xi) = 2\frac{\partial V_1^v}{\partial x}(1,\xi) + \frac{\partial V_1^h}{\partial x}(1,\xi), \qquad W_{\Gamma_2}(\xi) = \frac{\partial V_1^v}{\partial y}(\xi,1) + 2\frac{\partial V_1^h}{\partial y}(\xi,1), \qquad \xi \in \mathcal{G}.$$

We explain how to compute $\frac{\partial V_1^v}{\partial x}(1,\xi)$, $\frac{\partial V_1^v}{\partial y}(\xi,1), \xi \in \mathcal{G}$; $\frac{\partial V_1^h}{\partial x}(1,\xi)$, $\frac{\partial V_1^h}{\partial y}(\xi,1)$, $\xi \in \mathcal{G}$, can be computed in a similar way. Since $V_{\Gamma_1} \in \mathcal{M}^0(0,1)$, we have

$$V_{\Gamma_1}(y) = \sum_{n=1}^{2N} \alpha_n \psi_n(y).$$

Lemmas 2.6 and 2.7 imply that

$$V_1^v(x,y) = \sum_{n=1}^{2N} \alpha_n v_n(x) \psi_n(y) \tag{5.35}$$

is a solution of (5.21). Hence

$$\frac{\partial V_1^v}{\partial x}(1,y) = \sum_{n=1}^{2N} \alpha_n v_n'(1) \psi_n(y),$$

where $v_n'(1)$ is referred to in Lemma 2.7. If we introduce

$$\vec{\alpha} = [\alpha_1, \dots, \alpha_{2N}]^T, \qquad \vec{\beta} = [\alpha_1 v_1'(1), \dots, \alpha_{2N} v_{2N}'(1)]^T,$$

then Lemma 2.8 yields

$$\vec{\alpha} = Z^T B^T [V_{\Gamma_1}(\xi_1), \dots, V_{\Gamma_1}(\xi_{2N})]^T, \qquad \left[\frac{\partial V_1^v}{\partial x}(1,\xi_1), \dots, \frac{\partial V_1^v}{\partial x}(1,\xi_{2N})\right]^T = BZ\vec{\beta}.$$

Thus, given $V_{\Gamma_1}(\xi)$, $\xi \in \mathcal{G}$, Remark 2.1 implies that $\frac{\partial V_1^v}{\partial x}(1,\xi)$, $\xi \in \mathcal{G}$, can be computed at a cost $O(N \log N)$. On the other hand, (5.35) gives

$$\frac{\partial V_1^v}{\partial y}(\xi,1) = \sum_{n=1}^{2N} \alpha_n v_n(\xi) \psi_n'(1), \qquad \xi \in \mathcal{G}.$$

Since explicit formulas for $v_n(\xi)$, $\xi \in \mathcal{G}$, and $\psi_n'(1)$, $n = 1, \ldots, 2N$, are known, $\frac{\partial V_1^v}{\partial y}(\xi, 1)$, $\xi \in \mathcal{G}$, can be computed at a cost $O(N^2)$ by multiplying the vector $[\alpha_1 \psi_1'(1), \ldots, \alpha_{2N} \psi_{2N}'(1)]^T$ by the matrix $C = (c_{k,n})_{k,n=1}^{2N}$, where $c_{k,n} = v_n(\xi_k)$.

6. Total cost of solving OSC problem. In this section, we give the total cost of solving OSC problem (3.3)–(3.4) using algorithm (4.13).

Step 1 of algorithm (4.13) involves computing $\frac{\partial \hat{U}_1}{\partial x}(1, \xi)$, $\frac{\partial \hat{U}_1}{\partial y}(\xi, 1)$, $\frac{\partial \hat{U}_2}{\partial x}(1, \xi)$, $\frac{\partial \hat{U}_3}{\partial y}(\xi, 1)$, $\xi \in \mathcal{G}$, where $\hat{U}_i \in \mathcal{M}_i$, $i = 1, 2, 3$, satisfies (4.5)–(4.7). Since only $\frac{\partial \hat{U}_1}{\partial x}(1, \xi)$, $\frac{\partial \hat{U}_1}{\partial y}(\xi, 1)$, $\xi \in \mathcal{G}$, are required when solving (4.5) with $i = 1$ and (4.6)–(4.7) for $\hat{U}_1$, these values can be obtained by applying steps 1 and 2 and only a part of step 3 of Algorithm II of section 3.3 in [7]. The costs of steps 1 and 2 of this algorithm are $O(N^2 \log N)$ and $O(N^2)$, respectively, and the cost of the part of step 3 is $O(N^2)$. When performing these calculations, we save the results obtained in step 2 of Algorithm II of [7] when solving partially for $\hat{U}_1$. In a similar way, we compute $\frac{\partial \hat{U}_2}{\partial x}(1, \xi)$, $\frac{\partial \hat{U}_3}{\partial y}(\xi, 1)$, $\xi \in \mathcal{G}$.

Step 2 of algorithm (4.13) is carried out using the PCG method with P as a preconditioner for K. It follows from section 5.3 that the cost of this step is $O(mN^2)$, where m is the number of PCG iterations.

In step 3 of algorithm (4.13), we have to compute $U_i \in \mathcal{M}_i$, $i = 1, 2, 3$, satisfying (3.3) and (4.1)–(4.4). It follows from (4.8) that $U_1 = \hat{U}_1 + \tilde{U}_1$, where $\hat{U}_1 \in \mathcal{M}_1$ is a solution of (4.5) with $i = 1$ and (4.6)–(4.7), and $\tilde{U}_1 \in X_1$ is a solution of (4.9). First, we apply steps 1 and 2 of Algorithm II of [7] when solving (4.9) for $\tilde{U}_1$. Because of the zero right-hand side in the first equation of (4.9), the cost of step 1 in this algorithm becomes $O(N^2)$, while the cost of step 2 remains $O(N^2)$. Next, we add the results obtained in step 2 of Algorithm II of [7] when solving for $\hat{U}_1$ and $\tilde{U}_1$. Finally, we perform, at a cost $O(N^2 \log N)$, step 3 of Algorithm II of [7] to obtain U_1. In a similar way, we compute U_2 and U_3.

It follows from the discussion in this section that the total cost of algorithm (4.13) for solving the OSC problem is $O(N^2 \log N) + O(mN^2)$, where m is the number of PCG iterations in the solution of the interface problem. With m proportional to $\log N$, the total cost becomes $O(N^2 \log N)$.

7. Numerical results. We used the method of this paper to solve (1.1) with f corresponding to the exact solution

$$u(x, y) = e^{x+y}.$$

The algorithm was run in double precision on a Gateway PC E-2000 400. The initial guess in the PCG method (for the solution of the interface problem) was 0, and the number of PCG iterations was set to $\log_2 N + 4$. Convergence rates in various norms were determined using the formula

$$\text{rate } = \frac{\log(e_{N/2}/e_N)}{\log 2},$$

where e_N is the error corresponding to the $N \times N$ partition of Ω_1.

In Table 7.1, we present errors and the corresponding convergence rates for U_1 using Sobolev norms. As expected, the convergence rates for the L^2-, H^1-, and H^2-norms are 4, 3, and 2, respectively.

Table 7.1
Sobolev norm errors and convergence rates.

	$\|u-U_1\|_{L^2(\Omega_1)}$		$\|u-U_1\|_{H^1(\Omega_1)}$		$\|u-U_1\|_{H^2(\Omega_1)}$	
N	Error	Rate	Error	Rate	Error	Rate
4	4.33–05		4.08–04		1.05–02	
8	2.72–06	3.992	5.08–05	3.005	2.63–03	1.998
16	1.70–07	3.998	6.34–06	3.001	6.58–04	1.999
32	1.07–08	4.000	7.93–07	3.000	1.64–04	2.000
64	6.66–10	4.000	9.91–08	3.000	4.11–05	2.000

Table 7.2
Maximum nodal errors and convergence rates.

	$\|u-U_1\|_{C_h}$		$\|(u-U_1)_x\|_{C_h}$		$\|(u-U_1)_y\|_{C_h}$		$\|(u-U_1)_{xy}\|_{C_h}$	
N	Error	Rate	Error	Rate	Error	Rate	Error	Rate
4	9.71–06		6.48–05		6.48–05		1.61–03	
8	6.16–07	3.979	5.65–06	3.519	5.65–06	3.519	2.57–04	2.649
16	3.85–08	4.000	4.82–07	3.551	4.82–07	3.551	4.13–05	2.635
32	2.41–09	3.996	4.04–08	3.578	4.04–08	3.578	6.66–06	2.634
64	1.51–10	4.001	3.33–09	3.601	3.34–09	3.596	1.07–06	2.638

Table 7.3
CPU times for the OSC method.

N	64	128	256	512
CPU time	0.49	1.99	8.29	33.91

In Table 7.2, we give errors and the corresponding convergence rates for U_1 using the maximum nodal norm defined by

$$\|w\|_{C_h} = \max_{0\le k,l\le N} |w(t_k,t_l)|.$$

The convergence rate for $\|u-U_1\|_{C_h}$ is 4, while the convergence rates for $\|(u-U_1)_x\|_{C_h}$ and $\|(u-U_1)_y\|_{C_h}$ appear to be between 3.5 and 4. It was shown in [3] that the piecewise Hermite bicubic OSC solution U of Poisson's equation on Ω_1, with the nonzero Dirichlet boundary conditions approximated by the piecewise Hermite cubic interpolant, possesses superconvergence phenomena; that is, the convergence rates for $\|(u-U)_x\|_{C_h}$ and $\|(u-U)_y\|_{C_h}$ are 4. Numerical results, identical to those in Tables 7.1 and 7.2, were also obtained by first forming the symmetric positive definite matrix corresponding to the operator K and then solving the resulting linear system using Cholesky's method. Hence the presented convergence rates for $\|(u-U_1)_x\|_{C_h}$ and $\|(u-U_1)_y\|_{C_h}$ are due to the proposed OSC scheme for an L-shaped region and not an insufficient number of PCG iterations.

Finally, in Table 7.3, we give the CPU times for our OSC domain decomposition method. Clearly, as N increases by a factor of 2, the CPU time increases approximately by a factor of 4. This observation supports the theoretical result that the total cost of our algorithm is $O(N^2 \log N)$.

8. Concluding remarks.

We used a nonoverlapping domain decomposition approach to define the OSC solution of the Dirichlet boundary value problem for Poisson's equation on an L-shaped region. We proved existence and uniqueness of the collocation solution and derived optimal order H^s-norm error bounds for $s = 0, 1, 2$. The collocation solution on the interfaces is computed using the PCG method with a preconditioner obtained from two collocation Steklov–Poincaré operators corresponding to two pairs of the adjacent squares. The collocation solution on each square is obtained using the FFT matrix decomposition method. The total cost of computing the collocation solution is $O(N^2 \log N)$, where the number of unknowns in the collocation solution is $O(N^2)$.

REFERENCES

[1] B. Bialecki, *A fast domain decomposition Poisson solver on a rectangle for Hermite bicubic orthogonal spline collocation*, SIAM J. Numer. Anal., 30 (1993), pp. 425–434.

[2] B. Bialecki, *Convergence analysis of orthogonal spline collocation for elliptic boundary value problems*, SIAM J. Numer. Anal., 35 (1998), pp. 617–631.

[3] B. Bialecki, *Superconvergence of the orthogonal spline collocation solution of Poisson's equation*, Numer. Methods Partial Differential Equations, 15 (1999), pp. 285–303.

[4] B. Bialecki, X.-C. Cai, M. Dryja, and G. Fairweather, *An additive Schwarz algorithm for piecewise Hermite bicubic orthogonal spline collocation*, in Proceedings of the Sixth International Conference on Domain Decomposition Methods in Science and Engineering, Contemp. Math. 57, AMS, Providence, RI, 1994, pp. 237–244.

[5] B. Bialecki and S. D. Dillery, *Fourier analysis of Schwarz alternating methods for piecewise Hermite bicubic orthogonal spline collocation*, BIT, 33 (1993), pp. 634–646.

[6] B. Bialecki and M. Dryja, *Multilevel additive and multiplicative methods for orthogonal spline collocation problems*, Numer. Math., 77 (1997), pp. 35–58.

[7] B. Bialecki, G. Fairweather, and K. R. Bennett, *Fast direct solvers for piecewise Hermite bicubic orthogonal spline collocation equations*, SIAM J. Numer. Anal., 29 (1992), pp. 156–173.

[8] T. F. Chan and T. P. Mathew, *Domain decomposition algorithms*, in Acta Numerica, 1994, Cambridge University Press, Cambridge, UK, 1994, pp. 61–143.

[9] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North–Holland, Amsterdam, 1978.

[10] J. Douglas, Jr. and T. Dupont, *Collocation Methods for Parabolic Equations in a Single Space Variable*, Lecture Notes in Math. 385, Springer-Verlag, New York, 1974.

[11] S. Kim and S. Kim, *Estimating convergence factors of Schwarz algorithms for orthogonal spline collocation method*, Bull. Korean Math. Soc., 36 (1999), pp. 363–370.

[12] Y.-L. Lai, A. Hadjidimos, and E. N. Houstis, *A generalized Schwarz splitting method based on Hermite collocation for elliptic boundary value problems*, Appl. Numer. Math., 21 (1996), pp. 265–290.

[13] G. Mateescu, C. J. Ribbens, and L. T. Watson, *A domain decomposition preconditioner for Hermite collocation problems*, Numer. Methods Partial Differential Equations, 19 (2003), pp. 135–151.

[14] P. Percell and M. F. Wheeler, *A C^1 finite element collocation method for elliptic equations*, SIAM J. Numer. Anal., 17 (1980), pp. 605–622.

[15] A. Quarteroni and A. Valli, *Domain Decomposition Methods for Partial Differential Equations*, Oxford University Press, New York, 1999.

[16] B. F. Smith, P. E. Bjørstad, and W. D. Gropp, *Domain Decomposition. Parallel Multilevel Methods for Elliptic Partial Differential Equations*, Cambridge University Press, Cambridge, UK, 1996.

[17] E. G. Yanik, *A Schwarz alternating procedure using spline collocation methods*, Internat. J. Numer. Methods Engrg., 28 (1989), pp. 621–627.

SIAM J. NUMER. ANAL.
Vol. 41, No. 5, pp. 1729–1750

COMPUTING ACOUSTIC WAVES IN AN INHOMOGENEOUS MEDIUM OF THE PLANE BY A COUPLING OF SPECTRAL AND FINITE ELEMENTS[*]

SALIM MEDDAHI[†], ANTONIO MÁRQUEZ[‡], AND VIRGINIA SELGAS[†]

Abstract. In this paper we analyze a Galerkin procedure, based on a combination of finite and spectral elements, for approximating a time-harmonic acoustic wave scattered by a bounded inhomogeneity. The finite element method used to approximate the near field in the region of inhomogeneity is coupled with a nonlocal boundary condition, which consists in a linear integral equation. This integral equation is discretized by a spectral Galerkin approximation method.

We provide error estimates for the Galerkin method, propose fully discrete schemes based on elementary quadrature formulas, and show that the perturbation due to this numerical integration gives rise to a quasi-optimal rate of convergence. We also suggest a method for implementing the algorithm using the preconditioned GMRES method and provide some numerical results.

Key words. exterior boundary value problem, Helmholtz equation, variational formulation, integral equation, finite element, spectral method

AMS subject classifications. 65N30, 65F10

DOI. 10.1137/S0036142902406624

1. Introduction.

The purpose of this paper is to introduce a new fully discrete method for a boundary element and finite element coupling strategy applied to an acoustic scattering problem in the plane. The difficulty related to the fact that the acoustic field extends over the whole space has been tackled in the literature by different strategies. For example, the problem may be posed in a bounded domain by reducing it to the Lippman–Schwinger integral equation (cf. [11]). However, the computational cost associated to the discretization of such an equation can be excessive: it requires numerical integration for singular volume integrals, and it leads to linear systems of equations with nonsparse matrices (cf. [10]).

The approaches based on a finite element approximation method require *absorbing boundary conditions* prescribed on an artificial boundary Γ enclosing the region of inhomogeneity. These boundary conditions that incorporate (approximately) the far-field effects into the finite element model may be of local (differential) or global type.

Most of the differential absorbing boundary conditions use a circle (or a sphere) as an artificial boundary, and they are more exact the larger the radius of the circle is (cf. [7, 2]). This can lead to a large nondimensional wave number in a scaled model and renders the numerical solution more difficult to compute. It is also worth mentioning here the *perfectly matched layer method* introduced recently by Bérenger [3].

Choosing a circle as an artificial boundary, one may also compute a series representation of the exterior solution by separation of variables and obtain on the way global absorbing boundary conditions (cf. [6, 13, 14]). However, when the region of

[*] Received by the editors April 29, 2002; accepted for publication (in revised form) March 21, 2003; published electronically October 28, 2003.
http://www.siam.org/journals/sinum/41-5/40662.html

[†] Departamento de Matemáticas, Universidad de Oviedo, Calvo Sotelo s/n, 33007 Oviedo, Spain (salim@orion.ciencias.uniovi.es, selgas@orion.ciencias.uniovi.es). The research of Salim Meddahi was supported by *MCYT* through the project BFM2000-1324. The research of Virginia Selgas was supported by Ministerio de Educación, Cultura y Deporte through grant AP2001-2318.

[‡] Departamento de Construcción, Universidad de Oviedo, Campus de Viesques, 33203 Gijón, Spain (marquez@itma.edv.uniovi.es).

inhomogeneity is anisotropic, one may again be obliged to compute the numerical solution in a large domain. For more information about the previous methods we refer to the survey of Ihlenburg [8] and the references given therein.

We consider in this paper a nonlocal absorbing boundary condition based on boundary integral operators defined on Γ. It is a discretization procedure that combines finite elements (FEM) and boundary elements (BEM). In fact, we use the well-known Johnson–Nedelec BEM–FEM formulation introduced in [9] for an exterior Poisson problem and also used successfully for the Stokes system (cf. [18, 19]). We point out here that, in the Johnson–Nedelec method, the boundary integral equation coupled with the finite element problem must be posed on a smooth artificial interface in order to ensure the compactness of the double-layer potential. This is essential in the analysis of the discrete problem (see [9, 15] and the analysis given in section 5 of this paper).

Usually, the discrete problem is posed on a domain with a piecewise polynomial boundary that approximates the smooth artificial boundary in order to use isoparametric finite elements (cf. [9]). This strategy has a serious drawback since it renders difficult the approximation of the nearly singular boundary integrals by simple quadratures. Recently, a more efficient method that discretizes the integral operators on their natural boundary has been proposed. It permits one to design fully discrete schemes that require few kernel evaluations while preserving the stability and convergence properties obtained when the integrals are computed exactly (cf. [15, 16, 18]). This discretization method relies on exact triangulations of the domain. Hence curved triangles are needed all along the auxiliary interface.

Furthermore, these new BEM–FEM formulations permit one to approximate the periodic representation of the unknown defined on the boundary by trigonometric polynomials (cf. [17]). We apply here this mixed scheme that combines a finite element method and a spectral method to solve an acoustic scattering problem. Our analysis shows that stability and convergence are obtained for the Johnson–Nedelec method without any constraint involving the mesh size h for the interior finite elements and the dimension $2n$ of the space of trigonometric polynomials used for the approximation on the boundary. We also introduce a fully discrete scheme that requires elementary quadrature rules and converges at a quasi-optimal rate.

We notice that, as we have a spectral convergence for the variable defined on the boundary, few degrees of freedom are needed on the interface boundary in order to attain the order of convergence imposed by the finite element method. This permits us to eliminate the periodic unknown at matricial level by a static condensation process and reduce the complexity of the linear systems. The preconditioned GMRES method is used to solve the reduced linear systems of equations. The iterative method requires solving a short sequence of standard interior elliptic finite element problems. Furthermore, we do not need to store the unstructured and nonsymmetric global matrix, and the problems we have to solve during each iteration process are standard. We see from the numerical experiments that the method seems to be stable in the sense that the number of iterations does not increase with the finite element degrees of freedom.

We point out that our discretization method combines standard techniques of approximation since the schemes obtained here for the integral equation are directly derived from those described in [12, 5]; see also the references given therein. Nevertheless, to our knowledge, the only work where a coupling of finite elements and spectral methods is exploited to solve exterior Helmholtz problems is given by Kirsch and Monk in [20]. Their method relies on a Lagrange multiplier approach that is conceptually more complicated than the one we propose here, and they do not include the effects of numerical integration in their analysis. We also point out that Kirsch and Monk use a Nyström method for the unknown boundary, while we use a spectral Galerkin method. Our choice simplifies the analysis but, from the point of view of implementation, the computational work associated to a Nyström scheme is less than the computational work corresponding to a Galerkin method. We overcome this disadvantage by providing a fully discrete Galerkin method that may be interpreted in practice as a collocation method.

The paper is organized as follows. In the first part, which consists of sections 2, 3, and 4, we introduce the model problem, derive the Galerkin discretization of the Johnson–Nedelec method, and give its convergence analysis. In section 5, we describe the quadrature rules that we use to obtain the fully discrete schemes, and section 6 is devoted to the convergence analysis of these completely discrete problems. Finally, we present our numerical results in section 7.

1.1. Notation and Sobolev spaces.

In what follows, we deal with complex valued functions, and the symbol $\imath$ is used for $\sqrt{-1}$. We denote by $\overline{\alpha}$ the conjugate of a complex number $\alpha \in \mathbb{C}$ and by $|\alpha|$ its modulus. Let Ω be a bounded open set of $\mathbb{R}^2$. We denote by $\|\cdot\|_{0,\Omega}$ the $L^2(\Omega)$-norm corresponding to the inner product $\int_\Omega f\overline{g}\, d\boldsymbol{x}$. More generally, for any $m \in \mathbb{N}$, $\|\cdot\|_{m,\Omega}$ stands for the norm of the Sobolev space $H^m(\Omega)$; see [1].

On the other hand, we will also consider periodic Sobolev spaces. Let $\mathcal{C}^\infty_{2\pi}$ be the space of 2π-periodic and infinitely differentiable complex valued functions of a single variable. Given $g \in \mathcal{C}^\infty_{2\pi}$, we define its Fourier coefficients as

$$\widehat{g}(k) := \frac{1}{2\pi}\int_0^{2\pi} g(s)e^{-\imath ks}\, ds.$$

Then, for $p \in \mathbb{R}$, we define the Sobolev space $H^p(0, 2\pi)$ to be the completion of $\mathcal{C}^\infty_{2\pi}$ with the norm

$$\|g\|_p := \left(\sum_{k\in\mathbb{Z}}(1+|k|^2)^p|\widehat{g}(k)|^2\right)^{1/2}.$$

It is well known that $H^p(0, 2\pi)$ are Hilbert spaces and $H^p(0, 2\pi) \subset H^q(0, 2\pi)$ for every $p > q$, the inclusion being dense and compact; see [12]. Moreover, the $L^2(0,\ 2\pi)$-inner product $\int_0^{2\pi} \lambda(s)\overline{\eta}(s)\, ds$ can be extended to represent the duality of $H^{-p}(0, 2\pi)$ and $H^p(0, 2\pi)$ for all p.

Throughout this paper, C will denote positive constants, not necessarily the same at different occurrences, which are independent of the parameters h and n.

2. The model problem.

Let $\theta \in \mathcal{C}^2(\mathbb{R}^2)$ be a given function that satisfies $\operatorname{Re}\theta(\boldsymbol{x}) > 0$ and $\operatorname{Im}\theta(\boldsymbol{x}) \geq 0$ for all $\boldsymbol{x} \in \mathbb{R}^2$. We also assume that the function $1-\theta(\boldsymbol{x})$ has a compact support in $\mathbb{R}^2$. Let $k > 0$ be given together with a function w that satisfies the Helmholtz equation $\Delta w + k^2 w = 0$ in all $\mathbb{R}^2$. We seek the $u : \mathbb{R}^2 \to \mathbb{C}$ solution of

$$\begin{aligned} \Delta u + k^2\theta(\boldsymbol{x})\, u &= 0 && \text{in } \mathbb{R}^2, \\ u &= w + u^s && \text{in } \mathbb{R}^2 \end{aligned} \tag{2.1}$$

that satisfies the outgoing Sommerfeld radiation condition

$$\frac{\partial u^s}{\partial r} - \imath k u^s = o\left(\frac{1}{\sqrt{r}}\right), \tag{2.2}$$

when $r := |\boldsymbol{x}| \to \infty$ uniformly for all directions $\boldsymbol{x}/|\boldsymbol{x}|$.

The system (2.1)–(2.2) governs the propagation of time-harmonic acoustic waves of small amplitude in a slowly varying inhomogeneous and absorbing medium. The wave motion is caused by a time-harmonic incident field of amplitude w. A common choice for w is the plane wave $w(\boldsymbol{x}) := \exp(\imath k \boldsymbol{d} \cdot \boldsymbol{x})$, where $\boldsymbol{d}$ is a fixed unit vector. The solution u of our problem is determined by the scattered field u^s that satisfies the Sommerfeld radiation condition (2.2). We refer to [5] and [11] for more information about the physical background of the problem.

Let us introduce an artificial boundary Γ such that the support of $1-\theta$ lays in its interior. Then Γ separates $\mathbb{R}^2$ into a bounded domain Ω and an unbounded region Ω_e exterior to Γ.

We introduce the bilinear form

$$a^k(z,v) := \int_\Omega \nabla z \cdot \nabla v \, d\boldsymbol{x} - k^2 \int_\Omega \theta(\boldsymbol{x}) z v \, d\boldsymbol{x} \quad \forall z, v \in H^1(\Omega).$$

It is straightforward that u satisfies in Ω the following variational formulation:

$$\begin{aligned} &\text{find } u \in H^1(\Omega) \text{ such that} \\ &a^k(u,v) - \textstyle\int_\Gamma \frac{\partial u}{\partial \boldsymbol{\nu}} v \, d\sigma = 0 \quad \forall v \in H^1(\Omega), \end{aligned} \tag{2.3}$$

where the unit normal $\boldsymbol{\nu}$ on Γ is directed into Ω_e.

On the other hand, using a Green formula, the radiation condition (2.2), and the fact that $\Delta u^s + k^2 u^s = 0$ in Ω_e, one arrives at the integral representation

$$u^s(\boldsymbol{x}) = \int_\Gamma \frac{\partial E(\boldsymbol{x},\boldsymbol{y})}{\partial \boldsymbol{\nu}_y} u^s(\boldsymbol{y}) \, d\sigma_y - \int_\Gamma E(\boldsymbol{x},\boldsymbol{y}) \frac{\partial u^s}{\partial \boldsymbol{\nu}} \, d\sigma_y \quad \forall \boldsymbol{x} \in \Omega_e, \tag{2.4}$$

where

$$E(\boldsymbol{x},\boldsymbol{y}) := \frac{\imath}{4} H_0^{(1)}(k|\boldsymbol{x}-\boldsymbol{y}|)$$

is the radial outgoing fundamental solution of the Helmholtz equation and $H_0^{(1)}$ stands for the Hankel function of order 0 and first type. The Johnson–Nedelec BEM–FEM method introduced in [9] uses a boundary integral identity relating on Γ the trace of u^s and its normal derivative $\frac{\partial u^s}{\partial \boldsymbol{\nu}}$. This boundary integral equation arises from the integral representation formula (2.4) and the classical jump conditions for the double-layer potential. Our purpose is to perform the coupling of this boundary equation with (2.3), but let us first introduce some notation and basic properties.

In what follows, we choose Γ to be an infinitely differentiable boundary, and we denote by $\boldsymbol{x} : \mathbb{R} \to \mathbb{R}^2$ a regular 2π-periodic parametric representation of this curve:

$$|\boldsymbol{x}'(s)| > 0 \quad \forall s \in \mathbb{R} \quad \text{and} \quad \boldsymbol{x}(s) = \boldsymbol{x}(t) \quad \text{iff} \quad t - s \in 2\pi\mathbb{Z}.$$

Therefore, we can identify any function defined on Γ with a 2π-periodic function. We can also define the parameterized trace on Γ as the linear continuous extension of

$$\begin{aligned} \gamma : \mathcal{C}^\infty(\overline{\Omega}) &\to L^2(0,\, 2\pi), \\ u &\mapsto \gamma u(\cdot) := u|_\Gamma(\boldsymbol{x}(\cdot)) \end{aligned}$$

to $H^1(\Omega)$. The resulting linear application $\gamma : H^1(\Omega) \to H^{1/2}(0, 2\pi)$ is bounded and onto (cf. Theorem 8.15 of [12]).

We introduce the parameterized versions of the simple and double-layer acoustic potentials

$$\mathcal{S}g(s) := \int_0^{2\pi} V(s,t)g(t)dt \qquad \text{and} \qquad \mathcal{D}g(s) := \int_0^{2\pi} K(s,t)g(t)dt,$$

where

$$V(s,t) := \frac{\imath}{4} H_0^{(1)}(k|\boldsymbol{x}(s) - \boldsymbol{x}(t)|)$$

and

$$K(s,t) := -\frac{k\imath}{4} H_1^{(1)}(k|\boldsymbol{x}(t) - \boldsymbol{x}(s)|) \frac{x_2'(t)(x_1(t) - x_1(s)) - x_1'(t)(x_2(t) - x_2(s))}{|\boldsymbol{x}(t) - \boldsymbol{x}(s)|}$$

with $H_1^{(1)}$ being the Hankel function of first type and order one.

Parameterizing the integral identity relating u^s and $\frac{\partial u^s}{\partial \boldsymbol{\nu}}$ on Γ yields (a similar strategy is used in [15, 18])

$$\gamma u^s = \left(\frac{1}{2}\mathcal{I} + \mathcal{D}\right) \gamma u^s - \mathcal{S}\xi, \tag{2.5}$$

where $\mathcal{I}$ is the identity operator and the auxiliary unknown ξ is given in terms of the normal derivative of u^s on Γ by

$$\xi := |\boldsymbol{x}'| \, \frac{\partial u^s}{\partial \boldsymbol{\nu}} \circ \boldsymbol{x}.$$

Combining (2.3) with a variational version of (2.5), we arrive at the following global weak formulation of (2.1)–(2.2):

$$\begin{aligned} &\text{find } u \in H^1(\Omega) \text{ and } \xi \in H^{-1/2}(0, 2\pi) \text{ such that} \\ &a^k(u, v) - \int_0^{2\pi} \xi(t)\gamma v(t)\, dt = \int_0^{2\pi} \lambda(t)\gamma v(t)\, dt \quad \forall v \in H^1(\Omega), \\ &\qquad b(u, \mu) + c(\xi, \mu) = b(w, \mu) \qquad\qquad \forall \mu \in H^{-1/2}(0, 2\pi), \end{aligned} \tag{2.6}$$

where $\lambda := |\boldsymbol{x}'| \, \frac{\partial w}{\partial \boldsymbol{\nu}} \circ \boldsymbol{x}$,

$$c(\xi, \mu) := \int_0^{2\pi} \mu(t)(\mathcal{S}\xi)(t)\, dt, \qquad \text{and} \qquad b(v, \mu) := \int_0^{2\pi} \mu(t) \left(\frac{1}{2}\mathcal{I} - \mathcal{D}\right) (\gamma v)(t)\, dt.$$

Moreover, defining the Hilbert space $\mathbf{M} := H^1(\Omega) \times H^{-1/2}(0, 2\pi)$ and denoting $\mathbf{u} := (u, \xi)$, $\mathbf{v} := (v, \mu) \in \mathbf{M}$, we may formulate (2.6) equivalently by

$$\begin{aligned} &\text{find } \mathbf{u} \in \mathbf{M} \text{ such that} \\ &A(\mathbf{u}, \mathbf{v}) = L(\mathbf{v}) \quad \forall \mathbf{v} \in \mathbf{M}, \end{aligned} \tag{2.7}$$

where $L(\mathbf{v}) = \int_0^{2\pi} \lambda(t)\gamma v(t)\, dt + 2b(w, \mu)$ and

$$A(\mathbf{u}, \mathbf{v}) = a^k(u, v) + 2c(\xi, \mu) + 2b(u, \mu) - \int_0^{2\pi} \xi(t)\gamma v(t)\, dt.$$

3. Existence and uniqueness. We will first give a brief account of some fundamental tools that concern the properties of $\mathcal{S}$ and $\mathcal{D}$ when mapping between Sobolev spaces. For $n = 0, 1, 2$ we introduce the auxiliary integral operators

$$\Lambda_n(\xi)(t) := \frac{-1}{2\pi}\int_0^{2\pi}\left(\sin\frac{t-s}{2}\right)^n \log\left(\frac{4}{e}\sin^2\frac{t-s}{2}\right)\xi(s)\,ds.$$

Lemma 3.1. *For any $p \in \mathbb{R}$ the mappings $\Lambda_0 : H^p(0,2\pi) \to H^{p+1}(0,2\pi)$, $\mathcal{S} : H^p(0,2\pi) \to H^{p+1}(0,2\pi)$, $\mathcal{D} : H^p(0,2\pi) \to H^{p+2}(0,2\pi)$, and $\mathcal{S}-\Lambda_0 : H^p(0,2\pi) \to H^{p+3}(0,2\pi)$ are bounded. Furthermore, Λ_0 is $H^{-1/2}(0,2\pi)$-elliptic; i.e., there exists $\alpha > 0$ such that*

$$\int_0^{2\pi}\overline{\xi}(t)(\Lambda_0\xi)(t)\,dt \geq \alpha\|\xi\|^2_{-1/2} \quad \forall\xi \in H^{-1/2}(0,2\pi). \tag{3.1}$$

Proof. Let $f_m(t) := \exp(\imath mt)$. One may deduce easily from the property (see [12])

$$\Lambda_0 f_m = \frac{1}{\max(1,|m|)} f_m \qquad \forall m \in \mathbb{Z} \tag{3.2}$$

that Λ_0 is a pseudodifferential operator of order -1 and that it satisfies (3.1).

In general, given a function $D(t,s)$ which is in $\mathcal{C}^\infty_{2\pi}$ with respect to each variable, it is shown in [21] that $\mu \mapsto \Lambda_n(D(t,\cdot)\mu(\cdot))$ is a pseudodifferential operator of order $-n-1$. Therefore, the results for $\mathcal{S}$, $\mathcal{D}$, and $\mathcal{S}-\Lambda_0$ are obtained by noticing that there exist two functions D_1 and D_2 that belong to $\mathcal{C}^\infty_{2\pi}$ in each of their two variables such that

$$(\mathcal{S}\mu)(t) = (\Lambda_0\mu)(t) + \Lambda_2(D_1(t,\cdot)\mu(\cdot)) + (\mathcal{F}\mu)(t)$$

and

$$(\mathcal{D}\mu)(t) = \Lambda_1(D_2(t,\cdot)\mu(\cdot))(t) + (\mathcal{G}\mu)(t),$$

where $\mathcal{F}$ and $\mathcal{G}$ are integral operators with 2π-periodic and infinitely differentiable kernels. □

Theorem 3.2. *Assume that k^2 is not an eigenvalue of the Laplacian in Ω with a Dirichlet boundary condition on Γ. Then problem* (2.7) *has a unique solution.*

Proof. We introduce the bilinear form

$$A_0(\mathbf{u},\mathbf{v}) := a(u,v) + 2\int_0^{2\pi}\mu(t)(\Lambda_0\xi)(t)\,dt - \int_0^{2\pi}\gamma v(t)\,\xi(t)\,dt + \int_0^{2\pi}\gamma u(t)\,\mu(t)\,dt,$$

where $a(u,v) := \int_\Omega \nabla u \cdot \nabla v\,d\boldsymbol{x} + \int_\Omega uv\,d\boldsymbol{x}$. We deduce from Lemma 3.1 that $A_0(\cdot,\cdot)$ is bounded on $\mathbf{M}\times\mathbf{M}$ and $\mathbf{M}$-elliptic:

$$\mathrm{Re}[A_0(\mathbf{v},\overline{\mathbf{v}})] \geq \min(1,\alpha)\|\mathbf{v}\|^2_{\mathbf{M}} \quad \forall\mathbf{v} \in \mathbf{M}.$$

Let $\mathbf{M}'$ be the dual of $\mathbf{M}$ pivotal to $L^2(\Omega) \times L^2(0,\,2\pi)$. Then $\mathbf{M} \subset L^2(\Omega) \times L^2(0,\,2\pi) \subset \mathbf{M}'$ with dense inclusions. We denote by $[\cdot,\,\cdot]$ the duality bracket between $\mathbf{M}$ and $\mathbf{M}'$. We consider the continuous linear mappings $\mathcal{A} : \mathbf{M} \to \mathbf{M}'$ and $\mathcal{A}_0 : \mathbf{M} \to \mathbf{M}'$ defined by

$$[\mathcal{A}\mathbf{u},\,\mathbf{v}] := A(\mathbf{u},\,\mathbf{v}) \quad \text{and} \quad [\mathcal{A}_0\mathbf{u},\,\mathbf{v}] := A_0(\mathbf{u},\,\mathbf{v})$$

for all $\mathbf{u}$, $\mathbf{v}$ in $\mathbf{M}$.

Now it is clear that $\mathcal{A}_0$ is an isomorphism, and we deduce easily from Lemma 3.1 and the compactness of the canonical injection from $H^1(\Omega)$ into $L^2(\Omega)$ that $\mathcal{A} - \mathcal{A}_0 : \mathbf{M} \mapsto \mathbf{M}'$ is compact. Hence $\mathcal{A}$ is a Fredholm operator of index zero. Thus the theorem reduces to prove uniqueness of the solution for (2.7).

Let $(u_0, \xi_0) \in H^1(\Omega) \times H^{-1/2}(0, 2\pi)$ be a solution of (2.6) with $w = 0$. We introduce the function

$$\widetilde{u}(\boldsymbol{x}) := \begin{cases} u_0(\boldsymbol{x}), & \boldsymbol{x} \in \Omega, \\ z(\boldsymbol{x}) := \displaystyle\int_0^{2\pi} \frac{\partial E}{\partial \boldsymbol{\nu}_y}(\boldsymbol{x}, \boldsymbol{x}(t)) u_0(\boldsymbol{x}(t)) |\boldsymbol{x}'(t)| dt - \int_0^{2\pi} E(\boldsymbol{x}, \boldsymbol{x}(t)) \xi_0(t) dt, & \boldsymbol{x} \in \Omega_e. \end{cases}$$

It is easy to verify that u_0 solves the equation

$$\Delta u_0 + k^2 \theta(\boldsymbol{x})\, u_0 = 0 \qquad \text{in } \Omega. \tag{3.3}$$

On the other hand, z solves the Helmholtz equation

$$\Delta z + k^2 z = 0 \qquad \text{in } \Omega_e \tag{3.4}$$

and satisfies the Sommerfeld radiation condition (2.2). Furthermore, using the jump relations of the acoustic potential layers (see section 2 of [21]), we obtain the following identities on Γ:

$$\gamma z = \left(\frac{1}{2}\mathcal{I} + \mathcal{D}\right) \gamma u_0 - \mathcal{S}\xi_0, \tag{3.5}$$

$$|\boldsymbol{x}'(t)| \frac{\partial z}{\partial \boldsymbol{\nu}}(\boldsymbol{x}(t)) = -\mathcal{H}\gamma u_0 + \left(\frac{1}{2}\mathcal{I} - \mathcal{D}^*\right) \xi_0, \tag{3.6}$$

where $\mathcal{D}^*$ is the adjoint of operator $\mathcal{D}$, i.e.,

$$\mathcal{D}^* g(t) := \int_0^{2\pi} K(s,t) g(s) ds.$$

We refer to [21] for the definition of the hypersingular operator $\mathcal{H}$. We point out that we are using a parameterized version of this operator.

By virtue of (2.5), (3.5) directly yields the identity

$$\gamma z = \gamma u_0. \tag{3.7}$$

Now we also deduce from the integral representation of z in Ω_e and the jump conditions that

$$|\boldsymbol{x}'(t)| \frac{\partial z}{\partial \boldsymbol{\nu}}(\boldsymbol{x}(t)) = -\mathcal{H}\gamma z + \left(\frac{1}{2}\mathcal{I} - \mathcal{D}^*\right) |\boldsymbol{x}'(t)| \frac{\partial z}{\partial \boldsymbol{\nu}}(\boldsymbol{x}(t)). \tag{3.8}$$

Subtracting (3.8) from (3.6) and using (3.7), we obtain that

$$\left(\frac{1}{2}\mathcal{I} - \mathcal{D}^*\right) \left(|\boldsymbol{x}'(t)| \frac{\partial z}{\partial \boldsymbol{\nu}}(\boldsymbol{x}(t)) - \xi_0 \right) = 0. \tag{3.9}$$

Theorem 3.3.4. of [21] proves that, under our hypothesis on k, operator $\frac{1}{2}\mathcal{I} - \mathcal{D}^*$ is one-to-one and thus (3.9) provides the identity

$$\frac{\partial z}{\partial \boldsymbol{\nu}} = \frac{\partial u_0}{\partial \boldsymbol{\nu}} \quad \text{on } \Gamma. \tag{3.10}$$

Now, (3.3), (3.4), (3.7), and (3.10) show that $\widetilde{u}$ is a solution of (2.1)–(2.2) when $w = 0$, and Theorem 8.7 of [5] ensures that such a function $\widetilde{u}$ should vanish identically in all $\mathbb{R}^2$, and the result follows. □

4. Finite elements with curved triangles.

Let N be a given integer. We consider the equidistant subdivision $\{\frac{i\pi}{N}; \quad i = 0, \dots, 2N-1\}$ of the interval $[0, 2\pi]$ with $2N$ grid points. We denote by Ω_h the polygonal domain whose vertices lying on Γ are $\{\boldsymbol{x}(\frac{i\pi}{N}); i = 0, \dots, 2N-1\}$. Let τ_h be a regular triangulation of $\overline{\Omega}_h$ by triangles T of diameter h_T not greater than $\max|\boldsymbol{x}'(s)|h$ with $h := \frac{\pi}{N}$. We assume that any vertex of a triangle lying on the boundary of Ω_h belongs to $\{\boldsymbol{x}(\frac{i\pi}{N}); i = 0, \dots, 2N-1\}$.

We obtain from τ_h a triangulation $\widetilde{\tau}_h$ of $\overline{\Omega}$ by replacing each triangle of τ_h with one side along $\partial\Omega_h$ by the corresponding curved triangle.

Let T be a curved triangle of $\widetilde{\tau}_h$. We denote its vertices by $\mathbf{a}_1^T$, $\mathbf{a}_2^T$, and $\mathbf{a}_3^T$, numbered in such a way that $\mathbf{a}_2^T$ and $\mathbf{a}_3^T$ are the endpoints of the curved side of T. Let $t_i, t_{i+1} \in [0, 1]$ be such that $\boldsymbol{x}(t_i) = \mathbf{a}_2^T$ and $\boldsymbol{x}(t_{i+1}) = \mathbf{a}_3^T$. Then $\boldsymbol{\varphi}(t) := \boldsymbol{x}\,(t_i + t\,h)$ $(t \in [0, 1])$ is a parameterization of the curved side of T. Let $\widehat{T}$ be the reference triangle with vertices $\widehat{\mathbf{a}}_1 := (0, 0)$, $\widehat{\mathbf{a}}_2 := (1, 0)$, and $\widehat{\mathbf{a}}_3 := (0, 1)$. Consider the affine map $\mathbf{G}_T$ defined by $\mathbf{G}_T(\widehat{\mathbf{a}}_i) = \mathbf{a}_i^T$ for $i \in \{1, 2, 3\}$. Consider also the function $\boldsymbol{\Theta}_T : \widehat{T} \to \mathbb{R}^2$,

$$\boldsymbol{\Theta}_T(\widehat{\boldsymbol{x}}) := \frac{\hat{x}_1}{1 - \hat{x}_2}\left(\boldsymbol{\varphi}(\hat{x}_2) - (1 - \hat{x}_2)\mathbf{a}_2^T - \hat{x}_2\mathbf{a}_3^T\right),$$

where the limiting value has to be taken as $\hat{x}_2$ goes to 1. Then there exists $h_0 > 0$ such that if $h \in (0, h_0)$, T is the range of $\widehat{T}$ by the $\mathcal{C}^\infty$ and the one-to-one mapping $\mathbf{F}_T : \widehat{T} \to \mathbb{R}^2$ given by

$$\mathbf{F}_T := \mathbf{G}_T + \boldsymbol{\Theta}_T.$$

Moreover, each side of $\widehat{T}$ is mapped onto the corresponding side of T; i.e., $\boldsymbol{\Theta}_T(0, t) = \boldsymbol{\Theta}_T(t, 0) = (0, 0)$ and $\mathbf{F}_T(t, 1 - t) = \boldsymbol{\varphi}(t)$ for all $t \in [0, 1]$. This type of diffeomorphism was first proposed by Zlámal [25] and studied by Scott [22]. If T is a straight (interior) triangle, we take the curving perturbation $\boldsymbol{\Theta}_T \equiv \mathbf{0}$, and thus $\mathbf{F}_T$ is the usual affine map from the reference triangle. This hypothesis will be implicit in the following. In the finite element analysis we need estimates on the derivatives of $\mathbf{F}_T$ and $\mathbf{F}_T^{-1}$. These estimates are classical in the affine case, and they are proven in Theorem 22.4 of [24] (cf. also [22]) when T is a curved triangle. We collect the properties used in the forthcoming analysis in the following lemma.

Lemma 4.1. *For all $h \in (0, h_0)$, the Jacobian J_T of $\mathbf{F}_T$ does not vanish on a neighborhood of $\widehat{T}$, and the following estimates hold:*

$$C_1 h_T^2 \leq |J_T(\widehat{\boldsymbol{x}})| \leq C_2 h_T^2 \quad \forall \widehat{\boldsymbol{x}} \in \widehat{T}, \tag{4.1}$$

$$\max_{\widehat{\boldsymbol{x}} \in \widehat{T}} |\partial^\alpha (\mathbf{F}_T)_i(\widehat{\boldsymbol{x}})| \leq C h_T^{|\alpha|}, \quad \max_{\boldsymbol{x} \in T} |\partial^\alpha (\mathbf{F}_T^{-1})_i(\boldsymbol{x})| \leq C h_T^{-1} \tag{4.2}$$

for all $i = 1, 2$ and for all multi-index α such that $|\alpha| = 1, 2$.

A finite element is defined on T by a triplet $(T, P_1(T), \Sigma_T)$, where $P_1(T)$ is the image under $\mathbf{F}_T$ of the space $P_1(\widehat{T})$ of polynomials of degree no greater than 1 on $\widehat{T}$,

$$P_1(T) := \{p : T \to \mathbb{C};\ p = \hat{p} \circ \mathbf{F}_T^{-1},\ \hat{p} \in P_1(\widehat{T})\},$$

and $\Sigma_T = \{N_i; i = 1, 2, 3\}$ is a set of linear functionals defined by $N_i(\phi) = \phi(\mathbf{a}_{i,T})$ for all $\phi \in P_1(T)$ and for $i = 1, 2, 3$.

Interpolation error bounds on curved triangles are obtained by the technique used generally in the affine case (cf. [4] or [24]). Namely, if h is sufficiently small, one may readily prove with the aid of Lemma 4.1 that there exists a constant C independent of T such that

$$\|v - \pi_T v\|_{1,T} \leq\ C h_T \|v\|_{2,T} \quad \forall v \in H^2(T), \tag{4.3}$$

where $\pi_T v \in P_1(T)$ is uniquely determined by $\pi_T v(\mathbf{a}_i^T) = v(\mathbf{a}_i^T)$ for all $i = 1, 2, 3$. Notice that the norm $\|\cdot\|_{2,T}$ in (4.3) may be substituted by the seminorm $|\cdot|_{2,T}$ when T is a straight triangle.

We define the finite-dimensional subspace $V_h \subset H^1(\Omega)$ by

$$V_h = \{v \in H^1(\Omega); \quad v|_T \in P_1(T) \quad \forall T \in \widetilde{\tau}_h\}.$$

We deduce from (4.3) that

$$\inf_{v \in V_h} \|u - v\|_{1,\Omega} \leq\ C h\, \|u\|_{2,\Omega} \qquad \forall u \in H^2(\Omega). \tag{4.4}$$

Let n be a given integer. We consider the $2n$-dimensional space

$$T_n := \left\{ \sum_{j=0}^{n} a_j \cos jt + \sum_{j=1}^{n-1} b_j \sin jt; \quad a_j, b_j \in \mathbb{C} \right\}.$$

We have the following approximation property (cf. [21]):

$$\inf_{\mu \in T_n} \|\lambda - \mu\|_s \leq 2^{t-s} n^{s-t} \|\lambda\|_t \quad \forall \lambda \in H^t(0, 2\pi), \forall t \geq s. \tag{4.5}$$

We will also need the two inverse inequalities (cf. [21])

$$\|\mu\|_q \leq \left(\frac{1}{2n}\right)^{p-q} \|\mu\|_p \quad \forall \mu \in T_n, \quad \forall p \leq q \tag{4.6}$$

and

$$\|\mu\|_\infty \leq C\, n \|\mu\|_{-1/2} \quad \forall \mu \in T_n, \tag{4.7}$$

where we denoted $\|\mu\|_\infty := \max_{t \in [0,\, 2\pi]} |\mu(t)|$.

5. The discrete problem. Let $\mathbf{M}_\delta = V_h \times T_n$, where $\delta := (h, 1/n)$ is the discretization parameter. In terms of this notation, the discrete version of (2.7) is given by

$$\begin{aligned} &\text{find } \mathbf{u}_\delta \in \mathbf{M}_\delta \text{ such that} \\ &A(\mathbf{u}_\delta, \mathbf{v}) = L(\mathbf{v}) \quad \forall \mathbf{v} \in \mathbf{M}_\delta. \end{aligned} \tag{5.1}$$

THEOREM 5.1. *Assume that k^2 is not an eigenvalue of the Laplacian in Ω with a Dirichlet boundary condition on Γ. For all δ small enough, problem* (5.1) *has a unique solution. Moreover, the Galerkin method is stable, and we have Céa's estimate*

$$\|\mathbf{u}-\mathbf{u}_\delta\|_{\mathbf{M}} \leq C_1 \inf_{\mathbf{v}\in\mathbf{M}_\delta} \|\mathbf{u}-\mathbf{v}\|_{\mathbf{M}}.$$

In case the exact solution belongs to $H^2(\Omega)\times H^{\sigma-1/2}(0,2\pi)$ for some $\sigma>0$, we have

$$\|u-u_h\|_{1,\Omega}+\|\xi-\xi_n\|_{-1/2} \leq C_2\left(h\|u\|_{2,\Omega}+n^{-\sigma}\|\xi\|_{\sigma-1/2}\right).$$

Proof. The theorem is a consequence of a classical result for compact perturbations of operator equations. Indeed, let us consider the auxiliary problem

$$\begin{aligned}&\text{find } \mathbf{z}\in\mathbf{M} \text{ such that}\\ &A_0(\mathbf{z},\mathbf{v}) = L(\mathbf{v}) \quad \forall\mathbf{v}\in\mathbf{M}\end{aligned} \tag{5.2}$$

and its discrete counterpart

$$\begin{aligned}&\text{find } \mathbf{z}_\delta\in\mathbf{M}_\delta \text{ such that}\\ &A_0(\mathbf{z}_\delta,\mathbf{v}) = L(\mathbf{v}) \quad \forall\mathbf{v}\in\mathbf{M}_\delta.\end{aligned} \tag{5.3}$$

Both (5.2) and (5.3) are well posed by virtue of the $\mathbf{M}$-ellipticity of $A_0(\cdot,\cdot)$. Furthermore, Céa's lemma, the approximation properties (4.4) and (4.5), and the density of smooth functions in $\mathbf{M}$ yield

$$\lim_{\delta\to 0}\inf_{\mathbf{v}\in\mathbf{M}_\delta}\|\mathbf{z}-\mathbf{v}\|_{\mathbf{M}} = 0. \tag{5.4}$$

Now we proved that problem (2.7) is also well posed and that it is a compact perturbation of (5.2) (see Theorem 3.2). Under these hypotheses, Theorem 13.7 of [12] shows that, if δ is sufficiently small, (5.1) is also well posed and convergent. Finally, the convergence implies Céa's estimate, and the last assertion of the theorem follows from the approximation properties (4.4) and (4.5). □

6. Description of the fully discrete method.

6.1. Approximation of $a^k(\cdot,\cdot)$ on $V_h\times V_h$.

Consider first a quadrature formula on the reference triangle

$$\widehat{Q}(\widehat{\phi}) := \sum_{l=1}^{L}\widehat{\omega}_l\widehat{\phi}(\widehat{\boldsymbol{z}}_l) \simeq \int_{\widehat{T}}\widehat{\phi}\,d\widehat{x}$$

with weights $\widehat{\omega}_l>0$ such that $\sum_{l=1}^{L}\widehat{\omega}_l=\frac{1}{2}$. On each $T\in\widetilde{\tau}_h$ we define

$$Q_T(\phi) := \widehat{Q}(|J(F_T)|\phi\circ F_T) = \sum_{l=1}^{L}\widehat{\omega}_l|J(F_T)|(\widehat{\boldsymbol{z}}_l)\phi(F_T(\widehat{\boldsymbol{z}}_l)) \simeq \int_T \phi(\boldsymbol{x})\,d\boldsymbol{x}.$$

This induces us to define an approximation $a_h^k(\cdot,\cdot)$ of $a^k(\cdot,\cdot)$ by

$$a_h^k(u,v) = \sum_{K\in\widetilde{\tau}_h} Q_K(\nabla u\cdot\nabla v - k^2\theta uv) \qquad \forall u,v\in V_h.$$

6.2. Approximation of $c(\cdot,\cdot)$ on $T_n \times T_n$. For all continuous and 2π-periodic functions g, we consider the composite trapezoidal rule

$$\mathcal{Q}_n(g) := \frac{\pi}{n} \sum_{i=0}^{2n-1} g\left(\frac{i\pi}{n}\right)$$

associated to the partition of $[0,\ 2\pi]$ into $2n$ grid points.

We also need to construct approximations for the improper integral

$$(\Lambda_0 g)(t) := -\frac{1}{2\pi}\int_0^{2\pi} \log\left(\frac{4}{e}\sin^2\frac{t-s}{2}\right) g(s)\, ds. \tag{6.1}$$

We can proceed as in [12] and obtain a quadrature formula replacing $g(s)$ in (6.1) by its trigonometric interpolation polynomial

$$(\mathcal{P}_n g)(s) := \sum_{j=0}^{2n-1} g\left(\frac{j\pi}{n}\right) L_j(s),$$

where the Lagrange basis is given by

$$L_j(s) := \frac{1}{2n}\left\{1 + 2\sum_{k=1}^{n-1}\cos k\left(s - \frac{j\pi}{n}\right) + \cos n\left(s - \frac{j\pi}{n}\right)\right\} \quad \forall j = 0, \ldots, 2n-1.$$

Therefore, we obtain the formula

$$\widetilde{\mathcal{Q}}_n g(t) := \sum_{j=0}^{2n-1} R_j^{(n)}(t) g\left(\frac{j\pi}{n}\right),$$

where, for $j = 0, \ldots, 2n-1$, the weights

$$R_j^{(n)}(t) = \frac{1}{2n} + \frac{1}{n}\sum_{m=1}^{n-1}\frac{1}{m}\cos m\left(t - \frac{j\pi}{n}\right) + \frac{1}{2n^2}\cos n\left(t - \frac{j\pi}{n}\right)$$

are given explicitly by evaluating the integrals $(\Lambda_0 L_j)(t)$ with the aid of (3.2) (cf. [12]).

Using the splitting (cf. [5])

$$V(t,s) = -\frac{1}{2\pi}V_1(t,s)\log\left(\frac{4}{e}\sin^2\frac{t-s}{2}\right) + V_2(t,s) \tag{6.2}$$

of the single-layer acoustic potential, where $V_1(t,s) := \frac{1}{2}J_0(k|\boldsymbol{x}(t) - \boldsymbol{x}(s)|)$ and J_0 is the Bessel function of order zero, we obtain

$$c(\xi,\mu) = \int_0^{2\pi} \Lambda_0(V_1(t,\cdot)\xi(\cdot))(t)\mu(t)\, dt + \int_0^{2\pi}\left(\int_0^{2\pi} V_2(t,s)\xi(s)\, ds\right)\mu(t)\, dt. \tag{6.3}$$

Hereafter, taking into account that $V_1(\cdot,\cdot)$ and $V_2(\cdot,\cdot)$ are in $\mathcal{C}^\infty_{2\pi}$ with respect to each variable, the first term of the right-hand side in (6.3) may be approximated by using the quadrature rule $\widetilde{\mathcal{Q}}_n$ for the internal integral and $\mathcal{Q}_n$ for the external one. The two-dimensional quadrature rule derived from $\mathcal{Q}_n$ is applied to the second term. In other words, we are introducing an approximation of the bilinear form $c(\cdot,\cdot)$ on $T_n \times T_n$ given by

$$c_n(\xi,\mu) := \mathcal{Q}_n[\widetilde{\mathcal{Q}}_n[V_1(t,\cdot)\xi(\cdot)]\mu(t)] + \mathcal{Q}_n[\mathcal{Q}_n[V_2(t,\cdot)\xi(\cdot)]\mu(t)],$$

which may also be written in matricial form as follows:

$$c_n(\xi,\mu) = \sum_{i=0}^{2n-1}\left(\sum_{j=0}^{2n-1}\mathbf{C}_{i,j}\xi\left(\frac{j\pi}{n}\right)\right)\mu\left(\frac{i\pi}{n}\right).$$

The entries of the symmetric $2n \times 2n$ matrix $\mathbf{C}$ are

$$\mathbf{C}_{i,j} := \frac{\pi}{n}R_j^{(n)}\left(\frac{i\pi}{n}\right)V_1\left(\frac{i\pi}{n},\frac{j\pi}{n}\right) + \frac{\pi^2}{n^2}V_2\left(\frac{i\pi}{n},\frac{j\pi}{n}\right).$$

6.3. Approximation of $\boldsymbol{b(\cdot,\cdot)}$ on $\boldsymbol{V_h \times T_n}$. We point out that the kernel $K(\cdot,\cdot)$ associated to the bilinear form $b(\cdot,\cdot)$ is continuous but not derivable; therefore, it is necessary to split it, as we did for $V(\cdot,\cdot)$ in (6.2), before using any quadrature rule. We follow [5] and write

$$K(t,s) = -\frac{1}{2\pi}K_1(t,s)\log\left(\frac{4}{e}\sin^2\frac{t-s}{2}\right) + K_2(t,s), \tag{6.4}$$

with

$$K_1(s,t) := -\frac{k}{2}J_1(k|\boldsymbol{x}(t)-\boldsymbol{x}(s)|)\frac{x_2'(s)(x_1(t)-x_1(s)) - x_1'(s)(x_2(t)-x_2(s))}{|\boldsymbol{x}(t)-\boldsymbol{x}(s)|},$$

and J_1 being the Bessel function of order one. Here again, it turns out that K_1 and K_2 belong to $\mathcal{C}^\infty_{2\pi}$ in each variable.

We introduce the composite trapezoidal rule

$$\mathcal{Q}_N(g) := \frac{\pi}{N}\sum_{i=0}^{2N-1} g\left(\frac{i\pi}{N}\right)$$

associated to the uniform partition of $[0,\, 2\pi]$ into $2N$ grid points. Given $v \in V_h$ and $\mu \in T_n$, our strategy consists in approximating

$$b(v,\mu) = \frac{1}{2}\int_0^{2\pi}\gamma v(t)\mu(t)\,dt - \int_0^{2\pi}\Lambda_0(K_1(\cdot,s)\mu(\cdot))(s)\,\gamma v(s)\,ds \\ - \int_0^{2\pi}\left(\int_0^{2\pi}K_2(t,s)\gamma v(s)\,ds\right)\mu(t)\,dt$$

by employing $\mathcal{Q}_n$, $\widetilde{\mathcal{Q}}_n$, and $\mathcal{Q}_N$ as follows:

$$b_\delta(v,\mu) := \frac{1}{2}\int_0^{2\pi}\gamma v(t)\mu(t)\,dt - \mathcal{Q}_N[\widetilde{\mathcal{Q}}_n[K_1(\cdot,s)\mu(\cdot)]\gamma v(s)] \\ - \mathcal{Q}_N[\mathcal{Q}_n[K_2(\cdot,s)\mu(\cdot)]\gamma v(s)].$$

In other words,

$$b_\delta(v,\mu) = \frac{1}{2}\int_0^{2\pi}\gamma v(t)\mu(t)\,dt - \sum_{j=0}^{2N-1}\left(\sum_{i=0}^{2n-1}\mathbf{B}_{i,j}\mu\left(\frac{i\pi}{n}\right)\right)\gamma v\left(\frac{j\pi}{N}\right),$$

where $\mathbf{B}$ is the $2n \times 2N$ matrix with entries

$$\mathbf{B}_{i,j} := \frac{\pi}{N} R_i^{(n)}\left(\frac{j\pi}{N}\right) K_1\left(\frac{i\pi}{n}, \frac{j\pi}{N}\right) + \frac{\pi^2}{nN} K_2\left(\frac{i\pi}{n}, \frac{j\pi}{N}\right).$$

We note here that we did not propose a quadrature rule for $\int_0^{2\pi} \gamma v(t)\mu(t)\,dt$ since this integral may be easily evaluated analytically.

We are now in a position to propose a completely discrete version of the Galerkin scheme (5.1):

$$\text{(6.5)} \qquad \begin{aligned} &\text{find } \mathbf{u}_\delta^* \in \mathbf{M}_\delta \text{ such that} \\ &A_\delta(\mathbf{u}_\delta^*, \mathbf{v}) = L_\delta(\mathbf{v}) \quad \forall \mathbf{v} \in \mathbf{M}_\delta, \end{aligned}$$

where

$$A_\delta(\mathbf{u}, \mathbf{v}) = a_h^k(u, v) + 2c_n(\xi, \mu) + 2b_\delta(u, \mu) - \int_0^{2\pi} \xi(t)\gamma v(t)\,dt$$

and

$$L_\delta(\mathbf{v}) = \int_0^{2\pi} (\mathcal{P}_n \lambda)\,\gamma v\,dt + 2b_\delta(\Pi_h w, \mu),$$

with $\Pi_h : \mathcal{C}^0(\overline{\Omega}) \to V_h$ being the global Lagrange interpolation operator.

6.4. Matrix form of the fully discrete problem. Let us denote by $\mathbf{A}^k$ the symmetric $M_h \times M_h$ matrix whose entries are given by the complex numbers $\mathbf{A}_{i,j}^k := a_h^k(\varphi_i, \varphi_j)$, where $\{\varphi_i\,;\ i = 1, \ldots, M_h\}$ is the usual nodal basis of V_h. If we set

$$u_h(\boldsymbol{x}) = \sum_{i=1}^{M_h} \mathbf{u}_i \varphi_i(\boldsymbol{x}), \qquad \xi_h(t) = \sum_{i=0}^{2n-1} \boldsymbol{\xi}_i L_i(t),$$

then the matricial interpretation of (6.5) takes the form

$$\text{(6.6)} \qquad \begin{pmatrix} \mathbf{A}^k & -\mathbf{R}^t \\ \mathbf{K} & \mathbf{C} \end{pmatrix} \begin{pmatrix} \mathbf{u} \\ \boldsymbol{\xi} \end{pmatrix} = \begin{pmatrix} \mathbf{R}^t \boldsymbol{\lambda} \\ \mathbf{K}\mathbf{w} \end{pmatrix},$$

where

$$\mathbf{K}_{i,j} = b_\delta(\varphi_j, L_i) \quad \text{and} \quad \mathbf{R}_{i,j} = \int_0^{2\pi} \gamma \varphi_j(t) L_i(t)\,dt$$

for all $i = 0, \ldots, 2n-1$ and $j = 1, \ldots, M_h$. We also denoted by $\mathbf{w}$ the M_h-vector whose components are given by the values of the incident wave w at the nodes of the triangulation $\widetilde{\tau}_h$ and $\boldsymbol{\lambda}_k := \lambda(\frac{k\pi}{n})$ for $k = 0, \ldots, 2n-1$.

7. Analysis of the fully discrete method.

7. Analysis of the fully discrete method. We begin our analysis with the following classical result (see [4] or [24]).

LEMMA 7.1. *There exists $h_0 \in (0, 1]$ such that*

$$|a^k(u, v) - a_h^k(u, v)| \le Ch\|u\|_{1,\Omega}\|v\|_{1,\Omega} \quad \forall u, v \in V_h,$$

for some constant $C > 0$ independent of h for all $h \le h_0$.

We recall that $f_k(t) := \exp(\imath m t)$ and that $\mathcal{P}_n$ is the trigonometric interpolation operator. It is easy to check that $\mathcal{P}_n f_{2kn+m} = \mathcal{P}_n f_m$ for all $m, k \in \mathbb{Z}$. Furthermore, $\mathcal{P}_n f_m = f_m$ if $-n < m < n$ and $\mathcal{P}_n f_n = \frac{1}{2}(f_n + f_{-n})$. The following estimate is essential in the subsequent analysis, and we deduce it by just adapting the proof of Theorem 11.8 of [12].

Lemma 7.2. *Assume that* $p > 1/2$*; then there exists a constant* $C > 0$ *such that*

$$\|af_m - \mathcal{P}_n(af_m)\|_\infty \le C n^{-p+1/2} \|a\|_p \qquad \forall a \in H^p(0, 2\pi), \quad \forall m \in \mathbb{Z}.$$

Proof. Using the Fourier expansion

$$a(t) = \sum_{\rho=-n+1}^{n} \sum_{k\in\mathbb{Z}} \widehat{a}_{2kn+\rho} f_{2kn+\rho}(t)$$

of a, where $\widehat{a}_k := 1/2\pi \int_0^{2\pi} a(t) f_{-k}(t)\, dt$, we deduce that

$$(\mathcal{I} - \mathcal{P}_n)(af_m) = \sum_{\rho=-n+1}^{n} \sum_{k\in\mathbb{Z}} \widehat{a}_{2kn+\rho} (f_{2kn+\rho+m} - \mathcal{P}_n f_{\rho+m}).$$

Notice that we may always write $m = 2k_0 n + \overline{m}$ with $k_0 \in \mathbb{Z}$ and $-n < \overline{m} \le n$. We denote by $\varepsilon(\overline{m}) := \frac{\overline{m}}{|\overline{m}|}$ the sign of $\overline{m}$ with the convention $\varepsilon(0) = 1$. We remark that $-n < \overline{\rho} := \varepsilon(\overline{m})n - \overline{m} \le n$ and write $(\mathcal{I} - \mathcal{P}_n)(af_m) = S_1 + S_2$ with

$$S_1 = \sum_{\substack{\rho=-n+1\\ \rho\neq\overline{\rho}}}^{n} \sum_{k\in\mathbb{Z}^*} \widehat{a}_{2kn+\rho} (f_{2kn+\rho+m} - f_{\rho+m})$$

and

$$S_2 = \sum_{k\in\mathbb{Z}} \widehat{a}_{2kn+\overline{\rho}} \left(f_{2kn+\overline{\rho}+m} - \frac{1}{2} f_n - \frac{1}{2} f_{-n} \right),$$

where $\mathbb{Z}^* = \mathbb{Z} - \{0\}$.

A first bound of S_1 is obtained by using the Cauchy–Schwarz inequality

$$|S_1(t)|^2 \le \left(2 \sum_{\substack{\rho=-n+1\\ \rho\neq\overline{\rho}}}^{n} \sum_{k\in\mathbb{Z}^*} |\widehat{a}_{2kn+\rho}| \right)^2 \le 8n \sum_{\substack{\rho=-n+1\\ \rho\neq\overline{\rho}}}^{n} \left(\sum_{k\in\mathbb{Z}^*} |\widehat{a}_{2kn+\rho}| \right)^2.$$

The Cauchy–Schwarz inequality gives again

$$\begin{aligned}
\left(\sum_{k\in\mathbb{Z}^*} |\widehat{a}_{2kn+\rho}| \right)^2 &\le \sum_{k\in\mathbb{Z}^*} \frac{1}{(2kn+\rho)^{2p}} \sum_{k\in\mathbb{Z}^*} (2kn+\rho)^{2p} |\widehat{a}_{2kn+\rho}|^2 \\
&\le n^{-2p} \sum_{k\in\mathbb{Z}^*} \frac{1}{(2k+\frac{\rho}{n})^{2p}} \sum_{k\in\mathbb{Z}^*} (2kn+\rho)^{2p} |\widehat{a}_{2kn+\rho}|^2 \\
&\le C_1\, n^{-2p} \sum_{k\in\mathbb{Z}^*} (2kn+\rho)^{2p} |\widehat{a}_{2kn+\rho}|^2,
\end{aligned}$$

where $C_1 = \max_{t\in[-1,\,1]} \sum_{k\in\mathbb{Z}^*} \frac{1}{(2k+t)^{2p}} < \infty$ since $p > 1/2$. We deduce that

$$|S_1(t)|^2 \leq 8C_1\, n^{-2p+1} \sum_{\substack{\rho=-n+1\\ \rho\neq\overline{\rho}}}^{n} \sum_{k\in\mathbb{Z}^*} (2kn+\rho)^{2p}|\widehat{a}_{2kn+\rho}|^2 \quad \forall t\in[0,\ 2\pi]. \tag{7.1}$$

On the other hand,

$$|S_2(t)|^2 \leq \left(2\sum_{k\in\mathbb{Z}}|\widehat{a}_{2kn+\overline{\rho}}|\right)^2 \leq 4n^{-2p}\sum_{k\in\mathbb{Z}}\frac{1}{(2k+\varepsilon(\overline{m})-\frac{\overline{m}}{n})^{2p}}\sum_{k\in\mathbb{Z}}(2kn+\overline{\rho})^{2p}|\widehat{a}_{2kn+\overline{\rho}}|^2,$$

and then

$$|S_2(t)|^2 \leq 4\left(\max_{\rho\in[-1,\,1]}\sum_{k\in\mathbb{Z}}\frac{1}{(2k+\varepsilon(\overline{m})+\rho)^{2p}}\right)n^{-2p}\sum_{k\in\mathbb{Z}}(2kn+\overline{\rho})^{2p}|\widehat{a}_{2kn+\overline{\rho}}|^2 \tag{7.2}$$

for all $t\in[0,\ 2\pi]$.

Summing (7.1) and (7.2), we obtain that

$$|(\mathcal{I}-\mathcal{P}_n)(af_m)|^2 \leq C_2\, n^{-2p+1}\sum_{k\in\mathbb{Z}} k^{2p}|\widehat{a}_k|^2 \quad \forall t\in[0,\ 2\pi],$$

and the result follows. □

LEMMA 7.3. *There exists a constant C independent of n such that*

$$|c(\xi,\mu)-c_n(\xi,\mu)| \leq Cn^{-\sigma}\|\xi\|_{-1/2}\|\mu\|_{-1/2} \quad \forall\sigma>0,$$

for all ξ and μ in T_n.

Proof. We begin with the decomposition

$$\begin{aligned} c(\xi,\mu)-c_n(\xi,\mu) = \int_0^{2\pi}\Lambda_0(V_1(t,\cdot)\xi(\cdot))\mu(t)\,dt - \mathcal{Q}_n[\widetilde{\mathcal{Q}}_n[V_1(t,\cdot)\xi(\cdot)]\mu(t)] \\ + \int_0^{2\pi}\int_0^{2\pi}V_2(t,s)\xi(s)\mu(t)\,dsdt - \mathcal{Q}_n[\mathcal{Q}_n[V_2(t,\cdot)\xi(\cdot)]\mu(t)]. \end{aligned}$$

It is proved in Lemma 7 of [17] that there exists a constant C such that

$$\left|\int_0^{2\pi}\int_0^{2\pi}V_2(t,s)\xi(s)\mu(t)\,dsdt - \mathcal{Q}_n[\mathcal{Q}_n[V_2(t,\cdot)\xi(\cdot)]\mu(t)]\right| \leq Cn^{-\sigma}\|\xi\|_{-1/2}\|\mu\|_{-1/2}$$

for all $\sigma>0$.

Now, we have to prove the same estimate for the remaining term, which may be written

$$\int_0^{2\pi}\Lambda_0(V_1(t,\cdot)\xi(\cdot))\mu(t)\,dt - \mathcal{Q}_n[\widetilde{\mathcal{Q}}_n[V_1(t,\cdot)\xi(\cdot)]\mu(t)] = \int_0^{2\pi}E_1(t)\,dt + \mathcal{Q}_n[E_2(t)\mu(t)],$$

where

$$E_1(t) = (\mathcal{I}-\mathcal{P}_n)[\Lambda_0(V_1(t,\cdot)\xi(\cdot))\,\mu(t)] \quad\text{and}\quad E_2(t) = \Lambda_0(\mathcal{I}-\mathcal{P}_n)\,(V_1(t,\cdot)\xi(\cdot))\,.$$

We develop V_1 as

$$V_1(t,s) = \sum_{m\in\mathbb{Z}} \widehat{a}_m(t) f_m(s),$$

with $\widehat{a}_m(t) := 1/2\pi \int_0^{2\pi} V_1(t,s) f_{-m}(s)\,ds$, and we deduce that

$$\begin{aligned}\Lambda_0(V_1(\cdot,s)\xi(s))(t)\mu(t) &= \sum_{i,j=-n}^{n} \left(\sum_{m\in\mathbb{Z}} \widehat{a}_m(t)\Lambda_0(f_{m+j})(t) \right) f_i(t)\xi_j\mu_i \\ &= \sum_{i,j=-n}^{n} \left(\sum_{m\in\mathbb{Z}} \widehat{a}_m(t) \frac{f_{m+j+i}(t)}{\max(1,|m+j|)} \right) \xi_j \mu_i,\end{aligned}$$

where ξ_i and μ_i are the coefficients of $\xi,\mu \in T_n$ in the basis f_i. Therefore,

$$E_1(t) = \sum_{i,j=-n}^{n} \left(\sum_{m\in\mathbb{Z}} \frac{1}{\max(1,|m+j|)} (\mathcal{I}-\mathcal{P}_n)(\widehat{a}_m(t) f_{m+j+i}(t)) \right) \xi_j\mu_i,$$

and applying Lemma 7.2 with $p = \sigma + \frac{5}{2}$, we deduce that

$$|E_1(t)| \leq C_1\, n^{-\sigma-2} \left(\sum_{m\in\mathbb{Z}} \|\widehat{a}_m\|_{\sigma+5/2} \right) \sum_{i=-n}^{n} |\mu_i| \sum_{j=-n}^{n} |\xi_j| \quad \forall t \in [0,\, 2\pi].$$

Now, on the one hand, using (4.6) with $p = -1/2$ and $q = 0$, we have

$$\sum_{i=-n}^{n} |\mu_i| \leq \sqrt{2n}\|\mu\|_0 \leq 2n\|\mu\|_{-1/2} \quad \forall \mu \in T_n, \tag{7.3}$$

and on the other hand, the regularity of V_1 implies that

$$\sum_{m\in\mathbb{Z}} \|\widehat{a}_m\|_{\sigma+5/2} \leq \left(\sum_{m\in\mathbb{Z}} \frac{1}{1+m^2} \right)^{1/2} \left(\sum_{m,k\in\mathbb{Z}} (1+m^2)(1+k^2)^{\sigma+5/2}\, |\widehat{a}_{m,k}|^2 \right)^{1/2} < \infty,$$

where $\widehat{a}_{m,k} := 1/2\pi \int_0^{2\pi} \widehat{a}_m(t) f_{-k}(t)\,dt$. Consequently, we have the estimate

$$\left| \int_0^{2\pi} E_1(t)\,dt \right| \leq C_2\, n^{-\sigma} \|\mu\|_{-1/2} \|\xi\|_{-1/2}. \tag{7.4}$$

It remains to bound E_2. In this case we develop V_1 with respect to the variable t, considering s as a parameter:

$$V_1(t,s) = \sum_{m\in\mathbb{Z}} \widehat{b}_m(s) f_m(t), \qquad \left(\widehat{b}_m(s) := \frac{1}{2\pi} \int_0^{2\pi} V_1(t,s) f_{-m}(t)\,dt \right).$$

It follows that

$$E_2(t) = \sum_{j=-n}^{n} \xi_j \sum_{m\in\mathbb{Z}} \Lambda_0(\mathcal{I}-\mathcal{P}_n)(\widehat{b}_m f_j) f_m(t).$$

Let us denote by $\mathcal{C}^0_{2\pi}$ the space of 2π-periodic and continuous functions. By virtue of the Sobolev imbedding $H^1(0,2\pi) \hookrightarrow \mathcal{C}^0_{2\pi}$, we have that $\Lambda_0 :\ L^2(0,2\pi) \to \mathcal{C}^0_{2\pi}$ is bounded and

$$\|\Lambda_0(\mathcal{I}-\mathcal{P}_n)(\widehat{b}_m f_j)\|_\infty \leq C_3\|(\mathcal{I}-\mathcal{P}_n)(\widehat{b}_m f_j)\|_0 \leq \sqrt{2\pi}C_3\|(\mathcal{I}-\mathcal{P}_n)(\widehat{b}_m f_j)\|_\infty.$$

Thus applying Lemma 7.2 with $p=\sigma+5/2$ yields

$$|E_2(t)| \leq Cn^{-\sigma-2}\left(\sum_{m\in\mathbb{Z}}\|\widehat{b}_m\|_{\sigma+5/2}\right)\sum_{j=-n}^{n}|\xi_j| \quad \forall t\in[0,\ 2\pi],$$

and, here again, (7.3) and the regularity of V_1 permit us to obtain the bound

$$\|E_2(t)\|_\infty = \|(\mathcal{I}-\mathcal{P}_n)\,(V_1(t,\cdot)\xi(\cdot))\|_\infty \leq C_5\, n^{-\sigma-1}\|\xi\|_{-1/2}. \tag{7.5}$$

Now, by definition of the trapezoidal rule

$$\mathcal{Q}_n[E_2(t)\mu(t)] = \frac{\pi}{n}\sum_{i=0}^{2n-1}E_2(t_i)\mu(t_i)$$

and by virtue of (4.7), we obtain the inequality

$$|\mathcal{Q}_n[E_2(t)\mu(t)]| \leq 2\pi\|E_2(t)\|_\infty\|\mu\|_\infty \leq C_6\, n^{-\sigma}\|\xi\|_{-1/2}\|\mu\|_{-1/2}$$

that, joined to (7.4), gives the result. □

LEMMA 7.4. *There exists a constant C independent of $\delta=(h,1/n)$ such that*

$$|b(v,\mu)-b_\delta(v,\mu)| \leq C\left(h\frac{n\sqrt{\log n}}{N}+n^{-\sigma}\right)\|\mathbf{v}\|_{\mathbf{M}} \quad \forall\sigma>0,$$

for all $\mathbf{v} := (v,\mu)\in\mathbf{M}_\delta$.

Proof. We begin by noting that the parameterized trace $t\to\gamma v(\mathbf{x}(t))$ of a function $v\in V_h$ belongs to the space T_h of 2π-periodic, continuous, and piecewise linear functions on the uniform partition of $[0,\ 2\pi]$ into $2N$ subintervals.

We have the decomposition

$$\begin{aligned}b(v,\mu)-b_\delta(v,\mu) &= \int_0^{2\pi}\Lambda_0(K_1(\cdot,s)\mu(\cdot))\,\gamma v(s)\,ds - \mathcal{Q}_N[\widetilde{\mathcal{Q}}_n[K_1(\cdot,s)\mu(\cdot)]\gamma v(s)]\\ &\quad+\int_0^{2\pi}\int_0^{2\pi}K_2(t,s)\gamma v(s)\mu(t)\,dsdt - \mathcal{Q}_N[\mathcal{Q}_n[K_2(\cdot,s)\mu(\cdot)]\gamma v(s)].\end{aligned}$$

One can proceed as in Lemma 8 of [17] and prove that there exists a constant C_1 such that

$$\begin{aligned}&\left|\int_0^{2\pi}\int_0^{2\pi}K_2(t,s)\gamma v(s)\mu(t)\,dsdt - \mathcal{Q}_N[\mathcal{Q}_n[K_2(\cdot,s)\mu(\cdot)]\gamma v(s)]\right|\\ &\qquad\qquad\leq C_1\,(n^{-\sigma}+h)\|\mu\|_{-1/2}\|v\|_{1,\Omega}.\end{aligned}$$

We introduce the error operator $\mathcal{E}_N(g) := \int_0^{2\pi}g-\mathcal{Q}_N(g)$ corresponding to the quadrature formula $\mathcal{Q}_N$ and write

$$\begin{aligned}&\int_0^{2\pi}\Lambda_0(K_1(\cdot,s)\mu(\cdot))\gamma v(s)\,ds - \mathcal{Q}_N[\widetilde{\mathcal{Q}}_n[K_1(\cdot,s)\mu(\cdot)]\gamma v(s)]\\ &\qquad\qquad= \mathcal{E}_N[B_1(s)\gamma v] + \mathcal{Q}_N[B_2(s)\gamma v],\end{aligned}$$

where

$$B_1(s) = \Lambda_0(K_1(\cdot, s)\mu(\cdot)) \qquad \text{and} \qquad B_2(s) = \Lambda_0(\mathcal{I} - \mathcal{P}_n)\left(K_1(\cdot, s)\mu(\cdot)\right).$$

A simple change of variable yields

$$\mathcal{E}_N[B_1(s)\gamma v(s)] = h \sum_{i=0}^{2N-1} \mathcal{E}(B_1^i)\gamma v\left(\frac{i\pi}{N}\right),$$

where $\mathcal{E}(g) := \int_0^1 g(t)\,dt - \frac{g(1)+g(0)}{2}$ is the error operator of the basic trapezoidal rule on the reference interval [0, 1] and

$$B_1^i(s) := sB_1\left(\frac{(i-1)\pi}{N} + sh\right) + (1-s)B_1\left(\frac{i\pi}{N} + sh\right).$$

It follows readily from the Bramble–Hilbert lemma that

$$|\mathcal{E}(B_1^i)| \le C_2 \left\| \frac{d^2}{ds^2}(sB_1(t_{i-1} + sh) + (1-s)B_1(t_i + sh)) \right\|_\infty \le C_3 h^2 \|B_1''\|_\infty.$$

Using that for any $\varepsilon > 0$, $H^{\varepsilon+1/2}(0, 2\pi)$ is imbedded continuously in the space of 2π-periodic and continuous functions, we deduce that

$$\|B_1''\|_\infty \le C(\varepsilon)\|B_1''\|_{\varepsilon+1/2} \le C(\varepsilon)\|B_1\|_{\varepsilon+5/2}.$$

Moreover, it is easy to show that $C(\varepsilon) = \left(\sum_{k\in\mathbb{Z}} 1/\max(1, |k|)^{1+2\varepsilon}\right)^{1/2}$ behaves like $1/\sqrt{\varepsilon}$. Now, as we have already noticed in Lemma 3.1, $D_2(t, s) = K_1(t, s)/(\sin\frac{t-s}{2})$ belongs to $\mathcal{C}^\infty_{2\pi}$ in each of its two variables. It follows that we may write $B_1(s) = \Lambda_1(D_2(\cdot, s)\mu)$ and

$$\|\Lambda_1(D_2(\cdot, s)\mu)\|_{5/2+\varepsilon} \le C(\varepsilon)\|\mu\|_{1/2+\varepsilon} \le C(\varepsilon)n^{1+\varepsilon}\|\mu\|_{-1/2}$$

since $\mu \mapsto \Lambda_1(D_2(\cdot, s)\mu)$ is a pseudodifferential operator of index -2. We also used the inverse inequality (4.6).

Putting together the last estimates, we obtain, after using the Cauchy–Schwarz inequality and the fact that the norms

$$v(\boldsymbol{x}(t)) \to \|v(\boldsymbol{x}(t))\|_0 \qquad \text{and} \qquad v(\boldsymbol{x}(t)) \to \left(h \sum_{i=0}^{2N-1} \left|\gamma v\left(\frac{i\pi}{N}\right)\right|^2\right)^{1/2}$$

are uniformly equivalent on T_h, that

$$|\mathcal{E}_N[B_1(s)\gamma v(s)]| \le C_4\, h^2 \frac{n^{1+\varepsilon}}{\sqrt{\varepsilon}} \|\mu\|_{-1/2} \sum_{i=0}^{2N-1} h \left|\gamma v\left(\frac{i\pi}{N}\right)\right| \le C_5\, h^2 \frac{n^{1+\varepsilon}}{\sqrt{\varepsilon}} \|\mu\|_{-1/2}\|\gamma v\|_0$$

for all $\varepsilon > 0$. Furthermore, we notice that the function $\varepsilon \mapsto n^{1+\varepsilon}/\sqrt{\varepsilon}$ attains its minimum at a value proportional to $n\sqrt{\log n}$ when $\varepsilon = \frac{1}{2\log n}$, and hence

$$|\mathcal{E}_N[B_1(s)\gamma v(s)]| \le C_6\, h \frac{n\sqrt{\log n}}{N} \|\mu\|_{-1/2}\|v\|_{1,\Omega} \quad \forall \mu \in T_n, \quad \forall v \in V_h. \tag{7.6}$$

On the other hand,

$$|\mathcal{Q}_N[B_2(s)\gamma v(s)]| = h\left|\sum_{i=0}^{2N-1} B_2\left(\frac{i\pi}{N}\right)\gamma v\left(\frac{i\pi}{N}\right)\right|$$
$$\leq \|B_2\|_\infty \sum_{i=0}^{2N-1} h\left|\gamma v\left(\frac{i\pi}{N}\right)\right| \leq C_7\, \|B_2\|_\infty \|\gamma v\|_0,$$

and estimate (7.5) yields (after substituting V_1 by K_1)

$$\|B_2\|_\infty = \|\Lambda_0(\mathcal{I}-\mathcal{P}_n)\,(K_1(\cdot,s)\mu(\cdot))\|_\infty \leq C_8\, n^{-\sigma}\|\mu\|_{-1/2} \quad \forall \mu \in T_n.$$

Finally,

$$|\mathcal{Q}_N[B_2(s)\gamma v(s)]| \leq C_9\, n^{-\sigma}\|\mu\|_{-1/2}\|v\|_{1,\Omega} \quad \forall \mu \in T_n, \quad \forall v \in V_h,$$

and the result follows from the last inequality and (7.6). □

We conclude our analysis with the following result.

THEOREM 7.5. *Assume that k^2 is not an eigenvalue of the Laplacian in Ω with a Dirichlet boundary condition on Γ. We also assume that there exists $\kappa > 0$ such that $\frac{n}{N} \leq \kappa$. Then, for $\delta = (h, 1/n)$ small enough, the fully discrete scheme* (6.5) *is well posed and convergent. Moreover, we have*

$$\|u - u_h^*\|_{1,\Omega} + \|\xi - \xi_n^*\|_{-1/2} \leq\; C\,(h\sqrt{\log n} + n^{-\sigma})\big(\|u\|_{2,\Omega} + \|\xi\|_{\sigma-1/2}\big)$$

in case the exact solution belongs to $H^2(\Omega) \times H^{\sigma-1/2}(0,2\pi)$ for some $\sigma > 0$.

Proof. On the one hand, the convergence and stability of the Galerkin method (5.1) (see Theorem 5.1) is equivalent to the uniform inf-sup condition

$$\sup_{\mathbf{v}\in\mathbf{M}_\delta} \frac{A(\mathbf{u},\mathbf{v})}{\|\mathbf{v}\|_{\mathbf{M}}} \geq C\,\|\mathbf{u}\|_{\mathbf{M}} \qquad \forall \mathbf{u} \in \mathbf{M}_\delta.$$

On the other hand, Lemmas 7.3, 7.1, and 7.4 yield

$$(7.7) \qquad |A(\mathbf{u},\mathbf{v}) - A_\delta(\mathbf{u},\mathbf{v})| \leq C_1\,(h\sqrt{\log n} + n^{-\sigma})\|\mathbf{u}\|_{\mathbf{M}}\|\mathbf{v}\|_{\mathbf{M}} \quad \forall \mathbf{u},\mathbf{v} \in \mathbf{M}_\delta,$$

which permits us to deduce by standard arguments that, for δ small enough, $A_\delta(\cdot,\cdot)$ also satisfies a uniform inf-sup condition, and hence problem (6.5) has a unique solution.

Now, the second Strang lemma [23] gives the abstract estimate

$$\|\mathbf{u}_h^* - \mathbf{v}\|_{\mathbf{M}} \leq\; C_2\,\left(\sup_{\mathbf{z}\in\mathbf{M}_\delta}\frac{|A(\mathbf{u}-\mathbf{v},\mathbf{z})|}{\|\mathbf{z}\|_{\mathbf{M}}} + \sup_{\mathbf{z}\in\mathbf{M}_\delta}\frac{|A(\mathbf{v},\mathbf{z}) - A_\delta(\mathbf{v},\mathbf{z})|}{\|\mathbf{z}\|_{\mathbf{M}}}\right.$$
$$\left. + \sup_{\mathbf{z}\in\mathbf{M}_\delta}\frac{|L(\mathbf{z}) - L_\delta(\mathbf{z})|}{\|\mathbf{z}\|_{\mathbf{M}}}\right)$$

for all $\mathbf{v} \in \mathbf{M}_\delta$.

We can see from the right-hand side of the last inequality that it remains only to estimate the difference $L(\cdot) - L_\delta(\cdot)$ on $\mathbf{M}_\delta$ to obtain the asymptotic convergence annunciated in the theorem. We have that

$$|L(\mathbf{v}) - L_\delta(\mathbf{v})| \leq \left|\int_0^{2\pi}(\lambda - \mathcal{P}_n\lambda)\gamma v\, dt\right| + 2|b(w,\mu) - b_\delta(\Pi_h w,\mu)|.$$

On the one hand,

$$\left| \int_0^{2\pi} (\lambda - \mathcal{P}_n \lambda) \gamma v \, dt \right| \leq \|(\lambda - \mathcal{P}_n \lambda)\|_0 \|\gamma v\|_0 \leq C_3 \, n^{-\sigma} \|\lambda\|_\sigma \|v\|_{1,\Omega}$$

by virtue of the trace theorem and the well-known trigonometric interpolation error estimate in Sobolev spaces (cf. Theorem 11.8 of [12]).

On the other hand, using Lemma 7.4,

$$\begin{aligned} |b(w,\mu) - b_\delta(\Pi_h w, \mu)| &\leq |b(w - \Pi_h w, \mu)| + |b(\Pi_h w, \mu) - b_\delta(\Pi_h w, \mu)| \\ &\leq C_4 \left(\|w - \Pi_h w\|_{1,\Omega} \|\mu\|_{-1/2} + (h\sqrt{\log n} + n^{-\sigma}) \|\Pi_h w\|_{1,\Omega} \|\mu\|_{-1/2} \right), \end{aligned}$$

and by virtue of the Lagrange interpolation error estimate derived from (4.3) we deduce that

$$|b(w,\mu) - b_\delta(\Pi_h w, \mu)| \leq C_5 (h\sqrt{\log n} + n^{-\sigma}) \|w\|_{2,\Omega} \|\mu\|_{-1/2},$$

and consequently

$$|L(\mathbf{v}) - L_\delta(\mathbf{v})| \leq C_6 (h\sqrt{\log n} + n^{-\sigma}) (\|w\|_{2,\Omega} + \|\lambda\|_\sigma) \|\mathbf{v}\|_{\mathbf{M}}. \tag{7.8}$$

We deduce from the triangle inequality

$$\|\mathbf{u} - \mathbf{u}_h^*\|_{\mathbf{M}} \leq \|\mathbf{u} - \mathbf{u}_h\|_{\mathbf{M}} + \|\mathbf{u}_h^* - \mathbf{u}_h\|_{\mathbf{M}}$$

and the second Strang inequality that the asymptotic behavior of the fully discrete method is a direct consequence of (7.7), (7.8), the boundness of $A(\cdot,\cdot)$, and Theorem 5.1. □

We point out that the asymptotic behavior predicted by the last theorem is in fact too pessimistic. In practice, as we will show in the next section, n can be blocked at a very low value (which is independent of h) without influencing the accuracy of the method. Consequently, the convergence may still be considered as linear in h.

8. Numerical results.

We test our numerical method by using the same example given in section 5 of [20]. In all of what follows, the artificial boundary Γ is the ellipse of minor and major semiaxes $a = 1.1$ and $b = 1.5$. The incident wave is given by $w(\boldsymbol{x}) := \exp(\imath k x_1)$ and $\theta(\boldsymbol{x}) := 1 + \psi(|\boldsymbol{x}|)$ with

$$\psi(t) := \begin{cases} (1 - t^4)^2 & \text{for } t \in [0,\ 1], \\ 0 & \text{elsewhere.} \end{cases}$$

As $\theta(\boldsymbol{x})$ is radially symmetric, it is possible to use a separation of variable method and compute u explicitly in terms of the series

$$u(r\cos\phi, r\sin\phi) = \sum_{m=-\infty}^{\infty} y_m(r) \exp(\imath m \phi),$$

where y_m satisfies an integral equation described in section 5 of [20]. In the following, we compare our solution with

$$\overline{u}(r\cos\phi, r\sin\phi) = \sum_{m=-20}^{20} \overline{y}_m(r) \exp(\imath m \phi),$$

Table 8.1

Convergence history and number of iterations of the method for different values of the parameter n when $k = 1$ and $h = 2\pi/128$.

$2n$	$\text{error}_{\text{max}}$	Iterations
128	8.30×10^{-4}	10
64	8.62×10^{-4}	7
32	8.62×10^{-4}	6
16	8.62×10^{-4}	6
8	2.0×10^{-3}	6

Table 8.2

Convergence history and number of iterations of the method for different values of the parameter n when $k = 5$ and $h = 2\pi/256$.

$2n$	$\text{error}_{\text{max}}$	Iterations
256	3.22×10^{-2}	51
128	3.22×10^{-2}	42
64	3.22×10^{-2}	41
32	3.21×10^{-2}	40
16	3.75×10^{-1}	36

Table 8.3

Convergence history and number of iterations of the method for different values of the parameter h when $k = 2$ and $2n = 16$.

h	$\text{error}_{\text{max}}$	iterations
$2\pi/32$	7.57×10^{-2}	16
$2\pi/64$	2.29×10^{-2}	15
$2\pi/128$	5.8×10^{-3}	15
$2\pi/256$	1.5×10^{-3}	14

where $\overline{y}_m$ is an approximation of y_m obtained by solving the corresponding integral equation by a Nyström method as suggested in [20]. In all the tables, $\text{error}_{\text{max}} := \max |u_h(\mathbf{a}) - \overline{u}(\mathbf{a})|$, where the maximum is taken over the vertices $\mathbf{a}$ of $\widetilde{\mathcal{T}}_h$.

We take in Table 8.1 $k = 1$ and $h = 2\pi/128$, and we take in Table 8.2 $k = 5$ and $h = 2\pi/256$. In both cases we decrease the spectral parameter n until we obtain the smallest value that preserves the order of accuracy. We can see that the number of degrees of freedom is drastically reduced. This justifies the following strategy used to solve the linear systems: We eliminate the boundary variable from the system of linear equations (6.6) to obtain the reduced system

$$\left(\mathbf{A}^k + \mathbf{R}^t\mathbf{C}^{-1}\mathbf{K}\right)\mathbf{u} = \mathbf{f}, \tag{8.1}$$

where $\mathbf{f} := \mathbf{R}^t\boldsymbol{\lambda} + (\mathbf{R}^t\mathbf{C}^{-1}\mathbf{K})\mathbf{w}$. The system of equations (8.1) is solved by a GMRES method using $\mathbf{A} := (a(\varphi_i, \varphi_j))_{i,j}$ as a preconditioner. We use a version of GMRES without restarts. We take as an initial guess the solution of the Helmholtz equation in the bounded domain Ω, and iterations are continued until $\|r_{k+1}\|_2/\|r_k\|_2 < 10^{-6}$, where r_k is the kth residual.

Each iteration of the GMRES method entails the solution of a linear system with a full but small matrix $\mathbf{C}$ and another linear system with the sparse stiffness matrix

A. Furthermore, as **A** is symmetric and positive definite, this can be performed by a direct method through a Cholesky decomposition for matrix **A**. Table 8.3 shows the number of iterations against h with $n = 8$ and $k = 2$. The numerical results suggest that the method has a number of iterations bounded independently of the critical parameter h.

REFERENCES

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

[2] A. Bayliss, M. Gunzburger, and E. Tukel, *Boundary conditions for the numerical solution of elliptic equations in exterior regions*, SIAM J. Appl. Math., 42 (1982), pp. 430–451.

[3] J. Bérenger, *A perfectly matched layer for the absorption of electromagnetic waves*, J. Comput. Phys., 114 (1994), pp. 185–200.

[4] Ph. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North–Holland, Amsterdam, 1978.

[5] D. Colton and R. Kress, *Inverse Acoustic and Electromagnetic Scattering Theory*, 2nd ed., Springer-Verlag, Berlin, 1998.

[6] D. Givoli, *Numerical Methods for Problems in Infinite Domains*, Elsevier, Amsterdam, 1992.

[7] C. Goldstein, *The finite element method with non-uniform mesh sizes applied to the exterior Helmholtz problem*, Numer. Math., 38 (1981), pp. 61–82.

[8] F. Ihlenburg, *Finite Element Analysis of Acoustic Scattering*, Springer-Verlag, New York, 1998.

[9] C. Johnson and J. C. Nédélec, *On the coupling of boundary integral and finite element methods*, Math. Comp., 35 (1980), pp. 557–566.

[10] S. Johnson and M. Tracey, *Inverse scattering solutions by a sinc basis, multiple source, moment method* I: *Theory*, Ultrason. Imaging, 5 (1983), pp. 361–375.

[11] A. Kirsch, *An Introduction to the Mathematical Theory of Inverse Problems*, Springer-Verlag, New York, 1996.

[12] R. Kress, *Linear Integral Equations*, 2nd ed., Springer-Verlag, New York, 1999.

[13] R. Li, *On the coupling of BEM and FEM for exterior problems for the Helmholtz equation*, Math. Comp., 68 (1999), pp. 945–953.

[14] M. Masmoudi, *Numerical solution for exterior problems*, Numer. Math., 51 (1987), pp. 87–101.

[15] S. Meddahi, *An optimal iterative process for the Johnson–Nedelec method of coupling boundary and finite elements*, SIAM J. Numer. Anal., 35 (1998), pp. 1393–1415.

[16] S. Meddahi, M. González, and P. Pérez, *On a FEM–BEM formulation for an exterior quasilinear problem in the plane*, SIAM J. Numer. Anal., 37 (2000), pp. 1820–1837.

[17] S. Meddahi and A. Márquez, *A combination of spectral and finite elements methods for an exterior problem in the plane*, Appl. Numer. Math., 43 (2002), pp. 275–295.

[18] S. Meddahi and F.-J. Sayas, *A fully discrete BEM-FEM for the exterior Stokes problem in the plane*, SIAM J. Numer. Anal., 37 (2000), pp. 2082–2102.

[19] S. Meddahi and A. Márquez, *New implementation techniques for the exterior Stokes problem in the plane*, J. Comput. Phys., 172 (2001), pp. 685–703.

[20] A. Kirsch and P. Monk, *An analysis of the coupling of finite–element and Nyström methods in acoustic scattering*, IMA J. Numer. Anal., 14 (1994), pp. 523–544.

[21] J. Sarannen and G. Vainikko, *Periodic Integral and Pseudodifferential Equations with Numerical Approximation*, Springer-Verlag, New York, 2002.

[22] R. Scott, *Interpolated boundary conditions in the finite element method*, SIAM J. Numer. Anal., 12 (1975), pp. 404–427.

[23] G. Strang, *Variational crimes in the finite element method*, in The Mathematical Foundations of the Finite Element Method with Applications to Partial Differential Equations, A. Aziz, ed., Academic Press, New York, 1972, pp. 689–710.

[24] A. Ženíšek, *Nonlinear Elliptic and Evolution Problems and Their Finite Element Approximations*, Academic Press, London, 1990.

[25] M. Zlámal, *Curved elements in the finite element method* I, SIAM J. Numer. Anal., 10 (1973), pp. 229–240.

SIAM J. NUMER. ANAL.
Vol. 41, No. 5, pp. 1751–1772

ERROR ESTIMATES FOR LOW-ORDER ISOPARAMETRIC QUADRILATERAL FINITE ELEMENTS FOR PLATES[*]

RICARDO G. DURÁN[†], ERWIN HERNÁNDEZ[‡], LUIS HERVELLA-NIETO[§], ELSA LIBERMAN[¶], AND RODOLFO RODRÍGUEZ[‖]

Abstract. This paper deals with the numerical approximation of the bending of a plate modeled by Reissner–Mindlin equations. It is well known that, in order to avoid locking, some kind of reduced integration or mixed interpolation has to be used when solving these equations by finite element methods. In particular, one of the most widely used procedures is based on the family of elements called *MITC* (mixed interpolation of tensorial components). We consider two lowest-order methods of this family on quadrilateral meshes.

Under mild assumptions we obtain optimal H^1 and L^2 error estimates for both methods. These estimates are valid with constants independent of the plate thickness. We also obtain error estimates for the approximation of the plate vibration problem. Finally, we report some numerical experiments showing the very good behavior of the methods, even in some cases not covered by our theory.

Key words. Reissner–Mindlin, *MITC* methods, isoparametric quadrilaterals

AMS subject classifications. 65N30, 74S05, 74K20

DOI. 10.1137/S0036142902409410

1. Introduction.

The Reissner–Mindlin model is the most widely used for the analysis of thin or moderately thick elastic plates. It is now very well understood that standard finite element methods applied to this model produce very unsatisfactory results due to the so-called *locking* phenomenon. Therefore, some special method based on reduced integration or mixed interpolation has to be used. Among them, the mixed interpolation of tensorial components (*MITC*) methods introduced by Bathe and Dvorkin in [7] or variants of them are very likely the most used in practice.

A great number of papers dealing with the mathematical analysis of this kind of method have been published (see, for example, [2, 6, 10, 12, 13, 18, 20, 23]). In those papers, optimal order error estimates, valid uniformly on the plate thickness, have been obtained for several methods. However, although some of the most commonly used elements in engineering applications are the isoparametric quadrilaterals (indeed, the original Bathe and Dvorkin paper deals with these elements), no available result seems to exist for this case.

[*] Received by the editors June 9, 2002; accepted for publication (in revised form) April 30, 2003; published electronically October 28, 2003.

http://www.siam.org/journals/sinum/41-5/40941.html

[†] Departamento de Matemática, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, 1428 – Buenos Aires, Argentina (rduran@dm.uba.ar). The research of this author was supported by ANPCyT under grant PICT 03-05009 and by CONICET under grant PIP 0660/98.

[‡] Departamento de Matemática, Universidad Técnica Federico Santa María, Casilla 110-V, Valparaíso, Chile (erwin.hernandez@mat.utfsm.cl). The research of this author was supported by FONDECYT 2.000.114 (Chile).

[§] Departamento de Matemáticas, Universidade da Coruña, 15071 – A Coruña, Spain (luisher@udc.es). The research of this author was partially supported by FONDAP in Applied Mathematics.

[¶] Comisión de Investigaciones Científicas de la Provincia de Buenos Aires and Departamento de Matemática, Facultad de Ciencias Exactas, Universidad Nacional de La Plata, C.C. 172, 1900 – La Plata, Argentina (elsali@mate.undl.edu.ar).

[‖] GI^{2}MA, Departamento de Ingeniería Matemática, Universidad de Concepción, Casilla 160-C, Concepción, Chile (rodolfo@ing-mat.udec.cl). The research of this author was partially supported by FONDAP in Applied Mathematics (Chile).

On the other hand, it has been recently noted that the extension to general quadrilaterals of convergence results valid for rectangular elements is not straightforward, and, furthermore, the order of convergence can deteriorate when nonstandard finite elements are used in distorted quadrilaterals, even if they satisfy the usual shape regularity assumption (see [3, 4]).

The aim of this paper is to analyze two low-order methods based on quadrilateral meshes. One is the original $MITC4$ introduced in [7], while the other one is an extension to the quadrilateral case of a method introduced in [12] for triangular elements. (From now on the latter will be called $DL4$.) We are interested not only in load problems but also in the determination of the free vibration modes of the plate.

For nested uniform meshes of rectangles, an optimal order error estimate in H^1 norm has been proved in [6] for $MITC4$. However, the regularity assumptions on the exact solution required in that paper are not optimal. These assumptions have been weakened in [12], but they are still not optimal. Let us remark that to obtain approximation results for the plate vibration spectral problem, it is important to remove this extra regularity assumption.

On the other hand, for low-order elements as those considered here, an optimal error estimate in L^2 norm is difficult to obtain because of the consistency term arising in the error equation. For triangular elements, such an estimate has been only recently proved in [13]. However, the proof given in that paper cannot be extended straightforwardly, even for the case of rectangular elements.

In this paper we prove optimal in order and regularity H^1 and L^2 error estimates for both methods, $MITC4$ and $DL4$, under appropriate assumptions on the family of meshes. As a consequence, following the arguments in [13], we also obtain optimal error estimates for the approximation of the corresponding plate vibration spectral problem.

In order to prove the H^1 error estimate for $MITC4$, we require an additional assumption on the meshes (which is satisfied, for instance, by uniform refinements of any starting mesh). Instead, no assumption other than the usual shape regularity is needed for $DL4$.

On the other hand, a further assumption on the meshes is made to prove the L^2 error estimates: the meshes must be formed by higher-order perturbations of parallelograms. This restriction is related to approximation properties of the Raviart–Thomas elements which are used in our arguments and do not hold for general quadrilateral elements. However, this assumption is only needed for extremely refined meshes. Indeed, the L^2 estimate holds for any regular mesh as long as the mesh-size is comparable with the plate thickness. Moreover, we believe that this quasi-parallelogram assumption is of a technical character. In fact, the numerical experiments reported here seem to show that it is not necessary.

The rest of the paper is organized as follows. In section 2, we recall Reissner–Mindlin equations and introduce the two discrete methods. We prove optimal-order error estimates for both methods in H^1 and L^2 norms in sections 3 and 4, respectively. In section 5, we prove error estimates for the spectral plate vibration problem. Finally, in section 6, we report some numerical experiments.

Throughout the paper we denote by C a positive constant not necessarily the same at each occurrence but always independent of the mesh-size and the plate thickness.

2. Statement of the problem.

2.1. Reissner–Mindlin model. Let $\Omega \times (-\frac{t}{2}, \frac{t}{2})$ be the region occupied by an undeformed elastic plate of thickness t, where Ω is a convex polygonal domain of $\mathbb{R}^2$.

In order to describe the deformation of the plate, we consider the Reissner–Mindlin model, which is written in terms of the rotations $\beta = (\beta^1, \beta^2)$ of the fibers initially normal to the plate's midsurface and the transverse displacement w. The following equations describe the plate's response to conveniently scaled transversal and shear loads $f \in \mathrm{L}^2(\Omega)$ and $\theta \in \mathrm{L}^2(\Omega)^2$, respectively (see, for instance, [9, 13]).

Problem 2.1. *Find* $(\beta, w) \in \mathrm{H}^1_0(\Omega)^2 \times \mathrm{H}^1_0(\Omega)$ *such that*

$$\left\{\begin{array}{ll} a(\beta,\eta) + (\gamma, \nabla v - \eta) = (f,v) + \dfrac{t^2}{12}(\theta,\eta) & \forall (\eta, v) \in \mathrm{H}^1_0(\Omega)^2 \times \mathrm{H}^1_0(\Omega), \\ \gamma = \dfrac{\kappa}{t^2}(\nabla w - \beta). & \end{array}\right. \tag{2.1}$$

In this expression, $\kappa := Ek/2(1+\nu)$ is the shear modulus, with E being the Young modulus, ν the Poisson ratio, and k a correction factor. We have also introduced the shear stress γ and denoted by $(\cdot,\cdot)$ the standard L^2 inner product. Finally, a is the $\mathrm{H}^1_0(\Omega)^2$ elliptic bilinear form defined by

$$a(\beta,\eta) := \frac{E}{12(1-\nu^2)} \int_\Omega \left[\sum_{i,j=1}^{2} (1-\nu)\varepsilon_{ij}(\beta)\varepsilon_{ij}(\eta) + \nu \operatorname{div}\beta \operatorname{div}\eta \right],$$

with $\varepsilon_{ij}(\beta) = \frac{1}{2}(\partial\beta_i/\partial x_j + \partial\beta_j/\partial x_i)$ being the components of the linear strain tensor.

Let us remark that we have included in our formulation the shear load term $\frac{t^2}{12}(\theta,\eta)$ since it arises naturally when considering the free vibration plate problem. In fact, it is simple to see that the free vibration modes of the plate are determined by

$$t^3 a(\beta,\eta) + \kappa t \int_\Omega (\nabla w - \beta)\cdot(\nabla v - \eta) = \omega^2 \left(t \int_\Omega \rho\, wv + \frac{t^3}{12}\int_\Omega \rho\, \beta\cdot\eta \right) \qquad \forall (\eta, v) \in \mathrm{H}^1_0(\Omega)^2 \times \mathrm{H}^1_0(\Omega),$$

where ω denotes the angular vibration frequency, β and w the rotation and transversal displacement amplitudes, respectively, and ρ the plate density (see [13] for further details). Thus, rescaling the problem with $\lambda := \rho\,\omega^2/t^2$, we obtain the following, which is the spectral problem associated to Problem 2.1.

Problem 2.2. *Find* $\lambda \in \mathbb{R}$ *and* $0 \neq (\beta, w) \in \mathrm{H}^1_0(\Omega)^2 \times \mathrm{H}^1_0(\Omega)$ *such that*

$$\left\{\begin{array}{ll} a(\beta,\eta) + (\gamma, \nabla v - \eta) = \lambda \left[(w,v) + \dfrac{t^2}{12}(\beta,\eta) \right] & \forall (\eta, v) \in \mathrm{H}^1_0(\Omega)^2 \times \mathrm{H}^1_0(\Omega), \\ \gamma = \dfrac{\kappa}{t^2}(\nabla w - \beta). & \end{array}\right.$$

This paper deals with the finite element approximation of Problems 2.1 and 2.2. It is well known that both are well-posed (see [9] and [13]). Furthermore, we will use the following regularity result for the solution of (2.1) (see [2]):

$$\|\beta\|_{2,\Omega} + \|w\|_{2,\Omega} + \|\gamma\|_{0,\Omega} + t\,\|\gamma\|_{1,\Omega} \leq C\left(t^2\|\theta\|_{0,\Omega} + \|f\|_{0,\Omega}\right) \leq C\,|(\theta,f)|_t\,, \tag{2.2}$$

where, for any open subset O of Ω and any integer k, $\|\cdot\|_{k,O}$ denotes the standard norm of $\mathrm{H}^k(O)$ or $\mathrm{H}^k(O)^2$, as corresponds, and $|(\cdot,\cdot)|_t$ is the norm in $\mathrm{L}^2(\Omega)^2 \times \mathrm{L}^2(\Omega)$ induced by the weighted inner product on the right-hand side of the first equation in (2.1) (see [13]).

2.2. Discrete problems. In what follows, we consider two lowest-degree methods on isoparametric quadrilateral meshes for the approximation of Problem 2.1: the so-called *MITC*4 (see [7]) and an extension to quadrilaterals of a method introduced in [12] that we call *DL*4. Both methods are based on relaxing the shear terms in (2.1) by introducing an interpolation operator called a *reduction operator*.

Let $\{\mathcal{T}_h\}$ be a family of decompositions of Ω into convex quadrilaterals, satisfying the usual condition of regularity (see, for instance, [19]); i.e., there exist constants $\sigma > 1$ and $0 < \varrho < 1$ independent of h such that

$$h_K \le \sigma \rho_K, \qquad |\cos \vartheta_{iK}| \le \varrho, \quad i = 1, 2, 3, 4, \qquad \forall K \in \mathcal{T}_h,$$

where h_K is the diameter of K, ρ_K the diameter of the largest circle contained in K, and ϑ_{iK}, $i = 1, 2, 3, 4$, the four angles of K.

Let $\widehat{K} := [0, 1]^2$ be the reference element. We denote by $Q_{i,j}(\widehat{K})$ the space of polynomials of degree less than or equal to i in the first variable and to j in the second one. Also, we set $Q_k(\widehat{K}) = Q_{k,k}(\widehat{K})$.

Let $K \in \mathcal{T}_h$. We denote by F_K a bilinear mapping of $\widehat{K}$ onto K, with Jacobian matrix and determinant denoted by DF_K and J_{F_K}, respectively. The regularity assumptions above lead to

$$ch_K^2 \le J_{F_K} \le Ch_K^2,$$

with c and C depending only on σ and ϱ (see [19]). In particular, $J_{F_K} > 0$, and hence F_K is a one-to-one map. Let ℓ_i, $i = 1, 2, 3, 4$, be the edges of K; then $\ell_i = F_K(\widehat{\ell}_i)$, with $\widehat{\ell}_i$ being the edges of $\widehat{K}$. Let $\widehat{\tau}_i$ be a unit vector tangent to $\widehat{\ell}_i$ on the reference element; then $\tau_i := DF_K\widehat{\tau}_i / \|DF_K\widehat{\tau}_i\|$ is a unit vector tangent to ℓ_i on K (see Figure 2.1).

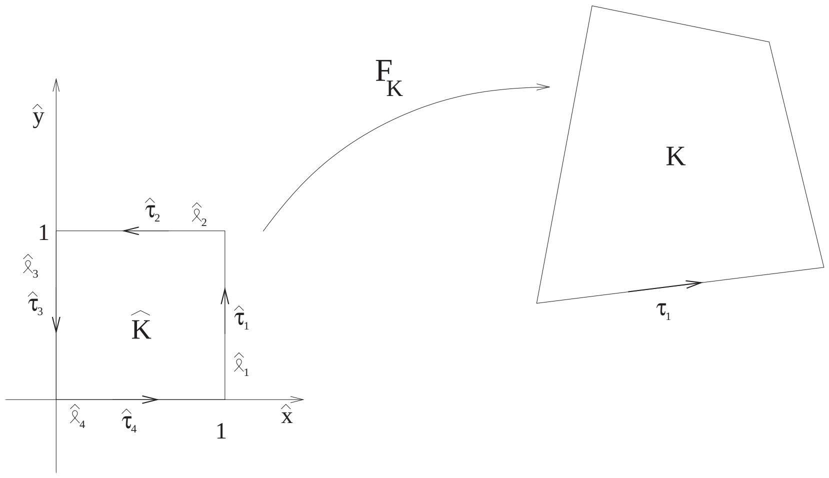


FIG. 2.1. *Bilinear mapping onto an element* $K \in \mathcal{T}_h$.

Let

$$\mathcal{N}(\widehat{K}) := \left\{ \widehat{p} : \ \widehat{p} \in Q_{0,1}(\widehat{K}) \times Q_{1,0}(\widehat{K}) \right\},$$

and, from this space, we define through covariant transformation

$$\mathcal{N}(K) := \left\{ p : \ p \circ F_K = DF_K^{-t} \widehat{p}, \ \widehat{p} \in \mathcal{N}(\widehat{K}) \right\}.$$

Let us remark that the mapping between $\mathcal{N}(K)$ and $\mathcal{N}(\widehat{K})$ is a kind of "Piola transformation" for the "rot" operator, $\operatorname{rot} p := \partial p_1/\partial x_2 - \partial p_2/\partial x_1$. (The Piola transformation is defined for the "div" operator in, for example, [9].) Then we have the following results which are easily established (see [23, 24]):

$$\int_{\ell_i} p \cdot \tau_i = \int_{\widehat{\ell}_i} \widehat{p} \cdot \widehat{\tau}_i, \qquad i = 1, 2, 3, 4, \tag{2.3}$$

and

$$(\operatorname{rot} p) \circ F_K = J_{F_K}^{-1} \, \widehat{\operatorname{rot}} \, \widehat{p} \qquad \text{in } \widehat{K}. \tag{2.4}$$

We define the lowest-order rotated Raviart–Thomas space (see [21, 24])

$$\Gamma_h := \left\{ \psi \in \mathrm{H}_0(\operatorname{rot}, \Omega) : \ \psi|_K \in \mathcal{N}(K) \ \forall K \in \mathcal{T}_h \right\},$$

which will be used to approximate the shear stress γ. We remark that, since $\Gamma_h \subset \mathrm{H}_0(\operatorname{rot}, \Omega)$, the tangential component of a function in Γ_h must be continuous along interelement boundaries and vanish on $\partial\Omega$. In fact, the integrals (2.3) of these tangential components are the degrees of freedom defining an element of Γ_h.

We consider the "interpolation" operator

$$R : \mathrm{H}^1(\Omega)^2 \cap \mathrm{H}_0(\operatorname{rot}, \Omega) \longrightarrow \Gamma_h, \tag{2.5}$$

defined by (see [21])

$$\int_\ell R\psi \cdot \tau_\ell = \int_\ell \psi \cdot \tau_\ell \qquad \forall \text{ edge } \ell \text{ of } \mathcal{T}_h, \tag{2.6}$$

where, from now on, τ_ℓ denotes a unit vector tangent to ℓ. Clearly, the operator R satisfies $\forall \psi \in \mathrm{H}^1(\Omega)^2$

$$\int_K \operatorname{rot}(\psi - R\psi) = 0 \qquad \forall K \in \mathcal{T}_h. \tag{2.7}$$

Taking into account the rotation mentioned above, it is proved in Theorem III.4.4 of [14] that

$$\| \operatorname{rot} R\psi \|_{0,\Omega} \leq C \|\psi\|_{1,\Omega} \tag{2.8}$$

and

$$\|\psi - R\psi\|_{0,\Omega} \leq Ch \|\psi\|_{1,\Omega}. \tag{2.9}$$

To approximate the transverse displacements, we will use the space of standard bilinear isoparametric elements

$$W_h := \left\{ v \in \mathrm{H}_0^1(\Omega) : \ v|_K \in Q(K) \ \forall K \in \mathcal{T}_h \right\},$$

where, $\forall K \in \mathcal{T}_h$, $Q(K) := \left\{ p \in \mathrm{L}^2(K) : \ p \circ F_K \in Q_1(\widehat{K}) \right\}$.

The following lemma establishes some relations between the spaces Γ_h and W_h.

LEMMA 2.1. *The following properties hold:*

$$\nabla W_h = \{\mu \in \Gamma_h : \ \operatorname{rot} \mu = 0\}$$

and

$$R(\nabla w) = \nabla(w^{\mathrm{I}}) \qquad \forall w \in \mathrm{H}^2(\Omega),$$

where w^{I} is the Lagrange interpolant of w on W_h.

Proof. For $\mu \in \Gamma_h$ and $K \in \mathcal{T}_h$, let $\widehat{\mu} \in \mathcal{N}(\widehat{K})$ be such that $\mu|_K \circ F_K = DF_K^{-\mathrm{t}}\widehat{\mu}$. Then, according to (2.4), we have $\operatorname{rot} \mu|_K \circ F_K = J_{F_K}^{-1} \widehat{\operatorname{rot}}\, \widehat{\mu}$. Hence, since $J_{F_K} > 0$, $\operatorname{rot} \mu = 0$ if and only if $\widehat{\operatorname{rot}}\, \widehat{\mu} = 0$.

On the other hand, note that if $\widehat{\mu} \in \mathcal{N}(\widehat{K})$, then $\widehat{\mu} = (a+b\widehat{y}, c+d\widehat{x})$, with $a, b, c, d \in \mathbb{R}$, and $\widehat{\operatorname{rot}}\, \widehat{\mu} = d - b$. Therefore, $\widehat{\operatorname{rot}}\, \widehat{\mu} = 0$ if and only if $\widehat{\mu} = (a + d\widehat{y}, c + d\widehat{x}) = \widehat{\nabla}\widehat{v}$ for $\widehat{v} = a\widehat{x} + c\widehat{y} + d\widehat{x}\widehat{y} \in Q_1(\widehat{K})$.

Thus $\operatorname{rot} \mu|_K = 0$ if and only if $\mu|_K = (DF_K^{-\mathrm{t}}\widehat{\mu}) \circ F_K^{-1} = \nabla v$, with $v = \widehat{v} \circ F_K^{-1} \in Q(K)$.

To prove the second property, since we have already proved that $\nabla w^{\mathrm{I}} \in \Gamma_h$, it is enough to show that the degrees of freedom defining $R(\nabla w)$ and ∇w^{I} coincide. Indeed, consider an edge ℓ with end points A and B as in Figure 2.2. Then,

$$\int_\ell R(\nabla w) \cdot \tau_\ell = \int_\ell \nabla w \cdot \tau_\ell = w(B) - w(A) = w^{\mathrm{I}}(B) - w^{\mathrm{I}}(A) = \int_\ell \nabla w^{\mathrm{I}} \cdot \tau_\ell,$$

and we conclude the proof. $\square$

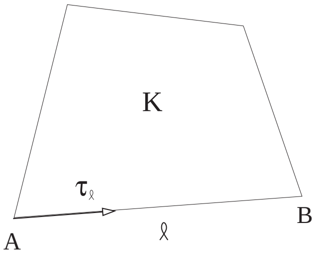


FIG. 2.2. *Geometry of K.*

The two methods that we analyze in this paper differ only in the space used to approximate the rotations. Let us now specify them:

*MITC*4: The spaces W_h and Γ_h are the ones defined above, whereas the space of standard isoparametric bilinear functions is used for the rotations; namely,

$$H_h^1 := \left\{\eta \in \mathrm{H}_0^1(\Omega)^2 : \ \eta|_K \in Q(K)^2 \ \forall K \in \mathcal{T}_h\right\}.$$

*DL*4: While for this method W_h and Γ_h are the same as for *MITC*4, the space for the rotations is enriched by using a rotation of a space used for the approximation of the Stokes problem in [14].

In fact, for each edge $\widehat{\ell}_i$ of $\widehat{K}$, $i = 1, 2, 3, 4$, let $\widehat{p}_i$ be cubic functions vanishing on $\widehat{\ell}_j$ for $j \neq i$. Namely, $\widehat{p}_1 = \widehat{x}\widehat{y}(1-\widehat{y})$, $\widehat{p}_2 = \widehat{x}\widehat{y}(1-\widehat{x})$, $\widehat{p}_3 = \widehat{y}(1-\widehat{x})(1-\widehat{y})$, and $\widehat{p}_4 = \widehat{x}(1-\widehat{x})(1-\widehat{y})$ (see Figure 2.1). Then we define $p_i := (\widehat{p}_i \circ F_K^{-1})\tau_i$, and we set

$$H_h^2 := \left\{\eta \in \mathrm{H}_0^1(\Omega)^2 : \ \eta|_K \in Q(K)^2 \oplus \langle p_1, p_2, p_3, p_4\rangle \ \forall K \in \mathcal{T}_h\right\}.$$

From now on we use H_h to denote any of the two spaces H_h^1 or H_h^2. In both methods we use R defined by (2.5)–(2.6) as *reduction operator*. Then, the discretization of Problem 2.1 can be written in both cases as follows.

Problem 2.3. *Find* $(\beta_h, w_h) \in H_h \times W_h$ *such that*

$$\left\{\begin{array}{ll} a(\beta_h,\eta) + (\gamma_h, \nabla v - R\eta) = (f, v) + \dfrac{t^2}{12}(\theta,\eta) & \forall(\eta, v) \in H_h \times W_h, \\ \gamma_h = \dfrac{\kappa}{t^2}(\nabla w_h - R\beta_h). & \end{array}\right. \tag{2.10}$$

On the other hand, the discretization of Problem 2.2 is as follows.

Problem 2.4. *Find* $\lambda_h \in \mathbb{R}$ *and* $0 \neq (\beta_h, w_h) \in H_h \times W_h$ *such that*

$$\left\{\begin{array}{ll} a(\beta_h,\eta) + (\gamma_h, \nabla v - R\eta) = \lambda_h \left[(w_h, v) + \dfrac{t^2}{12}(\beta_h,\eta)\right] & \forall(\eta, v) \in H_h \times W_h, \\ \gamma_h = \dfrac{\kappa}{t^2}(\nabla w_h - R\beta_h). & \end{array}\right.$$

Existence and uniqueness of solution for Problem 2.3 follow easily (see [12]). Regarding Problem 2.4, it leads to a well-posed generalized matrix eigenvalue problem, since the bilinear form in the right-hand side of the first equation is an inner product.

3. $\mathbf{H}^1$ error estimates.

To prove optimal error estimates in H^1 norm, we will use the abstract theory developed in [12]. In particular, sufficient conditions to obtain such estimates have been settled in Theorem 3.1 of this reference. By virtue of Lemma 2.1, this theorem reads as follows in our case.

Theorem 3.1. *Let* H_h, W_h, Γ_h, *and the operator* R *be defined as above. Let* (β, w, γ) *and* (β_h, w_h, γ_h) *be the solutions of* (2.1) *and* (2.10), *respectively. If there exist* $\widetilde{\beta} \in H_h$ *and an operator* $\Pi : \mathrm{H}_0(\mathrm{rot},\Omega) \cap \mathrm{H}^1(\Omega)^2 \longrightarrow \Gamma_h$ *satisfying*

$$\|\beta - \widetilde{\beta}\|_{1,\Omega} \leq Ch\|\beta\|_{2,\Omega}, \tag{3.1}$$

$$\|\eta - \Pi\eta\|_{0,\Omega} \leq Ch\|\eta\|_{1,\Omega} \qquad \forall \eta \in \mathrm{H}^1(\Omega)^2 \cap \mathrm{H}_0(\mathrm{rot},\Omega), \tag{3.2}$$

and

$$\mathrm{rot}\left(\frac{t^2}{\kappa}\Pi\gamma + R\widetilde{\beta}\right) = 0, \tag{3.3}$$

then the following error estimate holds true:

$$\|\beta - \beta_h\|_{1,\Omega} + t\|\gamma - \gamma_h\|_{0,\Omega} + \|w - w_h\|_{1,\Omega} \leq Ch\left(\|\beta\|_{2,\Omega} + t\|\gamma\|_{1,\Omega} + \|\gamma\|_{0,\Omega}\right).$$

Then, our next step is to construct an approximation $\widetilde{\beta}$ of β and an operator Π satisfying the hypotheses of the previous theorem for each one of the methods $MITC4$ and $DL4$.

3.1. *MITC4.* Several studies have been carried out for this method in, for example, [6], [12], and [17]. Since the variational equations for plates have a certain similitude with those of the Stokes problem, the main results are based on properties already known for the latter. An order h of convergence is obtained in those references only for uniform meshes of square elements. Moreover, more regularity of the solutions is also required. Although these results can be adapted for parallelogram meshes, they cannot be extended to general quadrilateral ones.

In what follows we obtain error estimates optimal in order and regularity for this method on somewhat more general meshes. We assume specifically the following condition.

ASSUMPTION 3.1. *The mesh $\mathcal{T}_h$ is a refinement of a coarser partition $\mathcal{T}_{2h}$, obtained by joining the midpoints of each opposite edge in each $M \in \mathcal{T}_{2h}$ (called macroelement). In addition, $\mathcal{T}_{2h}$ is a similar refinement of a still coarser regular partition $\mathcal{T}_{4h}$.*

Let

$$Q_h := \left\{q_h \in \mathrm{L}^2_0(\Omega) : \ q_h|_K = c_K, \ c_K \in \mathbb{R}, \ \forall K \in \mathcal{T}_h\right\},$$

where $\mathrm{L}^2_0(\Omega) := \left\{q \in \mathrm{L}^2(\Omega) : \ \int_\Omega q = 0\right\}$. Note that, for parallelogram meshes, we have $Q_h = \operatorname{rot}\Gamma_h$, but this does not hold for general quadrilateral meshes.

For each macroelement $M \in \mathcal{T}_{2h}$, we introduce four functions q_i, $i = 1, 2, 3, 4$, taking the values 1 and -1 according to the pattern of Figure 3.1.

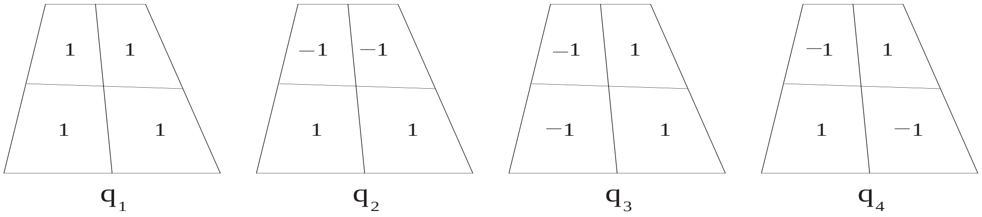


FIG. 3.1. *Bases for the macroelements.*

Let

$$Q_{h4} := \{q_h \in Q_h : \ q_h|_M = c_M q_4, \ c_M \in \mathbb{R} \ \forall M \in \mathcal{T}_{2h}\},$$

and let $\widetilde{Q}_h$ be its $\mathrm{L}^2(\Omega)$ orthogonal complement on Q_h; then

$$\widetilde{Q}_h := \{q_h \in Q_h : \ q_h|_M \in \langle q_1, q_2, q_3\rangle \ \forall M \in \mathcal{T}_{2h}\}.$$

We associate to these spaces the subspace of H^1_h defined by

$$\widetilde{H}^1_h := \left\{\eta_h \in H^1_h : \ \int_\Omega \operatorname{rot}\eta_h \, q_h = 0 \quad \forall q_h \in Q_{h4}\right\}.$$

The following lemma provides the approximation $\widetilde{\beta}$ required by Theorem 3.1. Moreover, this $\widetilde{\beta} \in \widetilde{H}^1_h$, and this fact will be used below to define the operator Π required by the same theorem.

Lemma 3.2. *Let $\beta \in \mathrm{H}_0^1(\Omega)$. Then there exists $\widetilde{\beta} \in \widetilde{H}_h^1$ such that*

$$\int_\Omega \operatorname{rot}(\widetilde{\beta} - \beta) q_h = 0 \qquad \forall q_h \in \widetilde{Q}_h,$$

and the estimate (3.1) *holds true.*

Proof. The proof follows from the results in section VI.5.4 of [9] by changing "div" to "rot" and rotating the fields 90°, which in their turn are based on the results for isoparametric elements in [22] (see also [19]). □

Our next step is to define the operator Π satisfying the requirements of Theorem 3.1. To do this, we will use a particular projector $\widetilde{P}$ onto $\operatorname{rot}\Gamma_h$.

We have already mentioned that, in general, $Q_h \neq \operatorname{rot}\Gamma_h$. In fact, it is simple to show that

$$\operatorname{rot}\Gamma_h = \left\{ \sum_{K \in \mathcal{T}_h} \frac{c_K}{J_{F_K}} \chi_K : \; c_K \in \mathbb{R} \; \forall K \in \mathcal{T}_h \right\} \cap \mathrm{L}_0^2(\Omega), \tag{3.4}$$

where χ_K denotes the characteristic function of K.

For each macroelement $M \in \mathcal{T}_{2h}$, we consider the bilinear mapping F_M as shown in Figure 3.2. Therefore, for any $\eta_h \in \Gamma_h$ we have

$$\operatorname{rot}\eta_h|_M = \frac{1}{J_{F_M}} \sum_{i=1}^4 c_i \chi_{K_i},$$

where K_i are the four elements in M (see Figure 3.2).

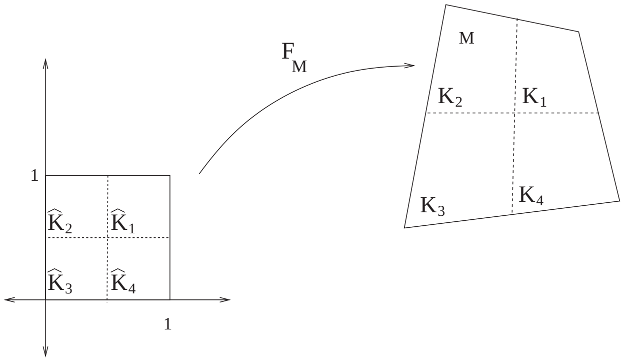


Fig. 3.2. *Bilinear mapping on macroelements.*

We define $\widetilde{P} : \mathrm{L}_0^2(\Omega) \longrightarrow \operatorname{rot}\Gamma_h$ as follows: given $p \in \mathrm{L}_0^2(\Omega)$,

$$\forall M = \bigcup_{i=1}^4 K_i \in \mathcal{T}_{2h}, \qquad \widetilde{P}p|_M = \sum_{i=1}^4 \frac{c_i}{J_{F_M}} \chi_{K_i},$$

with c_i chosen such that

$$\int_M \widetilde{P}p\, q_i = \int_M p q_i, \quad i = 1, 2, 3, \qquad \text{and} \qquad \int_M \widetilde{P}p\, q_4 = 0.$$

Straightforward computations show that $\widetilde{P}$ is well defined by the equations above and that they can be equivalently written

$$\int_\Omega \widetilde{P}p\, q_h = \int_\Omega pq_h \quad \forall q_h \in \widetilde{Q}_h \qquad \text{and} \qquad \int_\Omega \widetilde{P}p\, q_h = 0 \quad \forall q_h \in Q_{h4}. \tag{3.5}$$

The following properties of this operator will be used in what follows.

Lemma 3.3. *The following estimates hold* $\forall p \in \mathrm{L}^2(\Omega)$:

$$\|p - \widetilde{P}p\|_{0,\Omega} \le C\|p\|_{0,\Omega}, \tag{3.6}$$

$$\|p - \widetilde{P}p\|_{-1,\Omega} \le Ch\|p\|_{0,\Omega}. \tag{3.7}$$

Proof. To verify (3.6) it is enough to prove that $\|\widetilde{P}p\|_{0,\Omega} \le C\|p\|_{0,\Omega}$. From the definition of $\widetilde{P}$ we have

$$\int_M (\widetilde{P}p)^2 = \int_M \widetilde{P}p \left(\sum_{i=1}^4 \frac{c_i}{J_{F_M}} \chi_{K_i}\right) \le \frac{1}{\min\limits_M J_{F_M}} \int_M \widetilde{P}p \left(\sum_{i=1}^4 c_i \chi_{K_i}\right).$$

On the other hand, if we write $\sum_{i=1}^4 c_i\chi_{K_i}$ in terms of the basis functions q_i, we obtain $\sum_{i=1}^4 c_i\chi_{K_i} = \sum_{i=1}^4 d_i q_i$, with d_i related to c_i by

$$\begin{pmatrix} d_1 \\ d_2 \\ d_3 \\ d_4 \end{pmatrix} = \frac{1}{2} \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \\ 1 & -1 & 1 & -1 \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \\ c_3 \\ c_4 \end{pmatrix}.$$

Hence

$$|d_i| \le 2 \max_{1\le j\le 4} |c_j|, \qquad i = 1, 2, 3, 4.$$

Therefore, from the definition of $\widetilde{P}$ we have

$$\begin{aligned} \int_M \widetilde{P}p \left(\sum_{i=1}^4 c_i\chi_{K_i}\right) &= \int_M \widetilde{P}p \left(\sum_{i=1}^4 d_i q_i\right) = \sum_{i=1}^3 d_i \left(\int_M pq_i\right) \\ &\le \|p\|_{0,M} \left(\sum_{i=1}^3 |d_i|\, \|q_i\|_{0,M}\right) \le C|M|^{1/2}\|p\|_{0,M} \left(\max_{1\le j\le 4} |c_j|\right) \\ &\le C \max_M J_{F_M} \|p\|_{0,M} \|\widetilde{P}p\|_{0,M}, \end{aligned}$$

where we have used that

$$\int_{K_j} (\widetilde{P}p)^2 = c_j^2 \int_{K_j} \frac{1}{J_{F_M}^2} \ge \frac{|K_j|}{\max\limits_M J_{F_M}^2} c_j^2$$

and that $|M| \le C|K_j|$, $j = 1, 2, 3, 4$, with C depending only on σ and ϱ. Now, using the inequalities above and noting that, for a quadrilateral regular mesh, $\max_M J_{F_M} \le C \min_M J_{F_M}$ with a constant C independent of h, we obtain (3.6).

To verify (3.7), let $P : \mathrm{L}^2(\Omega) \longrightarrow \widetilde{Q}_h$ be the orthogonal projection onto $\widetilde{Q}_h$. Let $v \in \mathrm{H}_0^1(\Omega)$ be such that $\|v\|_{1,\Omega} = 1$. By the definition of P, (3.6), and the fact that $\widetilde{Q}_h$ contains the piecewise constants over $\mathcal{T}_{2h}$, we have

$$\begin{aligned} (p - \widetilde{P}p, v) &= (p - \widetilde{P}p, v - Pv) \le \|p - \widetilde{P}p\|_{0,\Omega} \|v - Pv\|_{0,\Omega} \\ &\le C\|p\|_{0,\Omega}\|v - Pv\|_{0,\Omega} \le Ch\|p\|_{0,\Omega}\|v\|_{1,\Omega}. \end{aligned}$$

Thus we conclude (3.7). □

Now we are in order to define an operator Π as required in Theorem 3.1.

LEMMA 3.4. *Let (β, w, γ) be the solution of (2.1) and $\widetilde{\beta} \in \widetilde{H}_h^1$ be as in Lemma 3.2. Then there exists an operator $\Pi : \mathrm{H}_0(\mathrm{rot}, \Omega) \cap \mathrm{H}^1(\Omega)^2 \longrightarrow \Gamma_h$ such that (3.2) and (3.3) hold true.*

Proof. For $\eta \in \mathrm{H}_0(\mathrm{rot}, \Omega) \cap \mathrm{H}^1(\Omega)^2$, let $\Pi\eta := R(\eta - L\eta)$, where $L\eta := \mathbf{curl}\, \phi := (-\partial\phi/\partial x_2, \partial\phi/\partial x_1)$, with $\phi \in \mathrm{H}^1(\Omega)$ being a solution of

$$-\Delta\phi = \mathrm{rot}\, R\eta - \widetilde{P}(\mathrm{rot}\, R\eta) \quad \text{in } \Omega,$$

with homogeneous Neumann boundary conditions. Note that this problem is compatible since its right-hand side belongs to $\mathrm{rot}\,\Gamma_h \subset \mathrm{L}_0^2(\Omega)$. Then the standard estimates for the Neumann problem yield

$$\|L\eta\|_{m+1,\Omega} \leq \|\mathrm{rot}\, R\eta - \widetilde{P}(\mathrm{rot}\, R\eta)\|_{m,\Omega}, \qquad m = -1, 0. \tag{3.8}$$

Also note that

$$\mathrm{rot}\, L\eta = -\Delta\phi = \mathrm{rot}\, R\eta - \widetilde{P}(\mathrm{rot}\, R\eta). \tag{3.9}$$

From the definition of the operator Π, we have

$$\|\eta - \Pi\eta\|_{0,\Omega} \leq \|\eta - R\eta\|_{0,\Omega} + \|RL\eta\|_{0,\Omega}.$$

The first term on the right-hand side is bounded by (2.9), while for the second term we use again (2.9), (3.8), Lemma 3.3, and (2.8) to obtain

$$\begin{aligned}\|RL\eta\|_{0,\Omega} &\leq \|L\eta - RL\eta\|_{0,\Omega} + \|L\eta\|_{0,\Omega} \leq Ch\|L\eta\|_{1,\Omega} + \|L\eta\|_{0,\Omega} \\ &\leq Ch\|\mathrm{rot}\, R\eta - \widetilde{P}(\mathrm{rot}\, R\eta)\|_{0,\Omega} + C\|\mathrm{rot}\, R\eta - \widetilde{P}(\mathrm{rot}\, R\eta)\|_{-1,\Omega} \\ &\leq Ch\|\mathrm{rot}\, R\eta\|_{0,\Omega} \leq Ch\|\eta\|_{1,\Omega}.\end{aligned}$$

Thus we conclude (3.2).

To prove (3.3), note that (2.7) and Lemma 3.2 yield

$$\int_\Omega \mathrm{rot}\left[R(\widetilde{\beta} - \beta)\right] q_h = 0 \qquad \forall q_h \in \widetilde{Q}_h,$$

whereas, since $\widetilde{\beta} \in \widetilde{H}_h^1$, from (2.7) and the definition of $\widetilde{H}_h^1$ we have

$$\int_\Omega \mathrm{rot}\, R\widetilde{\beta}\, q_h = \int_\Omega \mathrm{rot}\, \widetilde{\beta}\, q_h = 0 \qquad \forall q_h \in Q_{h4}.$$

Hence, since $\mathrm{rot}\, R\widetilde{\beta} \in \mathrm{rot}\,\Gamma_h$, from (3.5) we conclude that $\widetilde{P}(\mathrm{rot}\, R\beta) = \mathrm{rot}\, R\widetilde{\beta}$. Therefore,

$$\mathrm{rot}\, R\widetilde{\beta} = \widetilde{P}(\mathrm{rot}\, R\beta) = -\frac{t^2}{\kappa}\widetilde{P}(\mathrm{rot}\, R\gamma), \tag{3.10}$$

because of the definition of γ in (2.1) and the fact that $\mathrm{rot}\, R(\nabla w)$ vanishes, as a consequence of Lemma 2.1.

On the other hand, note that

$$\mathrm{rot}\, RL\gamma = \mathrm{rot}\, L\gamma. \tag{3.11}$$

Indeed, $\operatorname{rot} RL\gamma$ and $\operatorname{rot} L\gamma$ both belong to $\operatorname{rot}\Gamma_h$ (the latter because of (3.9)). Then, from the characterization (3.4) of this space, it is enough to verify that $\int_K \operatorname{rot} RL\gamma = \int_K \operatorname{rot} L\gamma \ \forall K \in \mathcal{T}_h$, which in its turn is a consequence of (2.7). Therefore, from the definition of Π, (3.11), and (3.9), we obtain

$$\operatorname{rot}\Pi\gamma = \operatorname{rot} R(\gamma - L\gamma) = \operatorname{rot} R\gamma - \operatorname{rot} L\gamma = \widetilde{P}(\operatorname{rot} R\gamma),$$

which together with (3.10) allows us to conclude (3.2). □

3.2. ***DL*4.** The convergence of this method follows immediately from that of *MITC*4. However, we have an alternative proof valid for any regular mesh without the need of Assumption 3.1.

In this case, to define the approximation $\widetilde{\beta}$ of β and the operator Π satisfying the hypotheses of Theorem 3.1, we use some known results for the Stokes problem (see Girault and Raviart [14]).

LEMMA 3.5. *There exists $\widetilde{\beta} \in H_h^2$ such that* (3.1) *holds true. Furthermore, $R\widetilde{\beta} = R\beta$.*

Proof. By using results from [14] (section 3.1, chapter II) and taking into account a rotation of the space $\mathrm{H}(\operatorname{div},\Omega)$, it follows that for $\beta \in \mathrm{H}_0^1(\Omega)^2$ there exists $\widetilde{\beta} \in H_h^2$ such that

$$\int_\ell (\widetilde{\beta} - \beta)\cdot\tau_\ell = 0 \qquad \forall \ell \in \mathcal{T}_h,$$

and

$$|\widetilde{\beta} - \beta|_{m,\Omega} \le Ch^{k-m}|\beta|_{k,\Omega}, \qquad m = 0, 1, \quad k = 1, 2.$$

Then $R(\widetilde{\beta} - \beta) = 0$ because of the definition of R, whereas (3.1) corresponds to the inequality above for $k = 2$ and $m = 1$. □

LEMMA 3.6. *There exists an operator $\Pi : \mathrm{H}_0(\operatorname{rot},\Omega) \cap \mathrm{H}^1(\Omega)^2 \longrightarrow \Gamma_h$ such that* (3.3) *and* (3.2) *hold.*

Proof. Because of the previous lemma we have $R(\widetilde{\beta} - \beta) = 0$. On the other hand, $\operatorname{rot} R(\nabla w) = 0$ because of Lemma 2.1. Then it is enough to take $\Pi = R$ to obtain (3.3), whereas (3.2) follows from (2.9). □

3.3. Main result in $\mathbf{H}^1$ norm. Now we are in position to establish the error estimates. As above, in the case of *MITC*4, we consider meshes satisfying Assumption 3.1.

THEOREM 3.7. *Given $(\theta, f) \in \mathrm{L}^2(\Omega)^2 \times \mathrm{L}^2(\Omega)$, let (β, w) and (β_h, w_h) be the solutions of Problems* 2.1 *and* 2.3, *respectively. Then there exists a constant C, independent of t and h, such that*

$$\|\beta - \beta_h\|_{1,\Omega} + \|w - w_h\|_{1,\Omega} \le Ch|(\theta, f)|_t.$$

Proof. The proof is a direct consequence of Lemmas 3.2, 3.4, 3.5 and 3.6, Theorem 3.1, and the a priori estimate (2.2). □

4. $\mathbf{L}^2$ error estimates. Our next goal is to prove L^2 error estimates optimal in order and regularity. To do this, we follow the techniques in [13], where a triangular element similar to *DL*4 is analyzed, although the arguments therein cannot be directly applied to our case. Let us remark that, in the case of *MITC*4, this result completes the analysis carried out in [10, 18] for higher-order methods.

Our proofs are based on a standard Nitsche duality argument. However, since the methods are nonconforming, additional consistency terms also arise. Then higher-order estimates must be proved for these terms too, which is the most delicate part of the paper.

First, we introduce the dual problem corresponding to (2.1). Let $(\varphi, u) \in \mathrm{H}_0^1(\Omega)^2 \times \mathrm{H}_0^1(\Omega)$ be the solution of

$$\left\{ \begin{array}{l} a(\eta, \varphi) + (\nabla v - \eta, \delta) = (v, w - w_h) + (\eta, \beta - \beta_h) \\ \qquad\qquad\qquad\qquad\qquad\qquad \forall (\eta, v) \in \mathrm{H}_0^1(\Omega)^2 \times \mathrm{H}_0^1(\Omega), \\ \delta = \dfrac{\kappa}{t^2}(\nabla u - \varphi). \end{array} \right. \tag{4.1}$$

By taking $\eta = 0$ in (4.1), we have

$$\operatorname{div} \delta = w_h - w. \tag{4.2}$$

An a priori estimate analogous to (2.2) yields for this problem

$$\|\varphi\|_{2,\Omega} + \|u\|_{2,\Omega} + \|\delta\|_{0,\Omega} + t\,\|\delta\|_{1,\Omega} \leq C\left(\|\beta - \beta_h\|_{0,\Omega} + \|w - w_h\|_{0,\Omega}\right). \tag{4.3}$$

The arguments in the proof of Lemma 3.4 in [13] can be used in our case leading to the following result.

Lemma 4.1. *Given $(\theta, f) \in \mathrm{L}^2(\Omega)^2 \times \mathrm{L}^2(\Omega)$, let (β, w, γ) and (β_h, w_h, γ_h) be the solutions of* (2.1) *and* (2.10), *respectively. Let (φ, u, δ) be the solution of* (4.1). *Let $\widetilde{\varphi} \in H_h$ be the vector field associated to φ by Lemma* 3.2 *or* 3.5 *for MITC*4 *or DL*4, *respectively. Then there exists a constant C, independent of t and h, such that*

$$\|\beta - \beta_h\|_{0,\Omega} + \|w - w_h\|_{0,\Omega} \leq Ch^2|(\theta, f)|_t + \frac{|(\beta_h - R\beta_h, \delta)| + |(\gamma, \widetilde{\varphi} - R\widetilde{\varphi})|}{\|\beta - \beta_h\|_{0,\Omega} + \|w - w_h\|_{0,\Omega}}.$$

Our next step is to prove that the last term in the inequality above is $\mathcal{O}(h^2)$ too. A similar result has been proved in [13] in the case of triangular meshes. That proof relies on a technical result for the rotated Raviart–Thomas interpolant R (Lemma 3.3 of that reference). It is easy to check that the arguments given there do not apply for quadrilateral elements. Therefore, we need to introduce new arguments, and this is the aim of the following lemma.

Lemma 4.2. *Given $\zeta \in H(\operatorname{div}, \Omega)$ and $\psi \in \mathrm{H}_0^1(\Omega)^2$, there holds*

$$|(\zeta, \psi - R\psi)| \leq Ch^2 \left(\sum_K |R\psi - \psi|_{1,K}^2 \right)^{1/2} \|\operatorname{div} \zeta\|_{0,\Omega} + Ch\|\operatorname{rot}(R\psi - \psi)\|_{0,\Omega}\,\|\zeta\|_{0,\Omega}.$$

Proof. For $K \in \mathcal{T}_h$, let $s_K \in \mathrm{H}^1(K)$ be a solution of

$$-\Delta s_K = \operatorname{rot}(R\psi - \psi) \quad \text{in } K,$$

with homogeneous Neumann boundary conditions. By virtue of (2.7) we know that the above problem is compatible. Hence s_K satisfies

$$\|\operatorname{curl} s_K\|_{m+1,K} \leq C\|\operatorname{rot}(R\psi - \psi)\|_{m,K}, \qquad m = -1, 0. \tag{4.4}$$

The Laplace equation above can be equivalently written

$$\operatorname{rot}\left[\operatorname{curl} s_K - (R\psi - \psi)\right] = 0,$$

and, hence, there exists $r_K \in \mathrm{H}^1(K)$ (unique up to an additive constant) such that

$$\nabla r_K = \operatorname{curl} s_K - (R\psi - \psi). \tag{4.5}$$

Moreover, from the homogeneous Neumann boundary condition satisfied by s_K, we have $\nabla r_K \cdot \tau_\ell = -R\psi \cdot \tau_\ell + \psi \cdot \tau_\ell$ for each edge ℓ of K. Thus, if we define $G \in \mathrm{L}^2(\Omega)^2$ such that $G|_K = \nabla r_K$, then $G \in \mathrm{H}_0(\mathrm{rot}, \Omega)$ and $\operatorname{rot} G = 0$.

Hence there exists $r \in \mathrm{H}^1(\Omega)/\mathbb{R}$ such that $G = \nabla r$ in Ω. Furthermore, since $G \in \mathrm{H}_0(\mathrm{rot}, \Omega)$, r can be chosen in $\mathrm{H}_0^1(\Omega)$ and the additive constants defining r_K on each $K \in \mathcal{T}_h$ can be fixed as to satisfy $r|_K = r_K$.

Let A and B be as in Figure 2.2. Then, because of (2.6), we have

$$r(B) = r(A) + \int_\ell \nabla r_K \cdot \tau_\ell = r(A) + \int_\ell (-R\psi + \psi) \cdot \tau_\ell = r(A).$$

Thus r vanishes at all nodes of $\mathcal{T}_h$, since $r|_{\partial\Omega} = 0$. Hence a standard scaling argument on each element K yields $\|r\|_{0,K} \leq Ch^2 |r_K|_{2,K}$ (see, for instance, [11]), and then, by using (4.5) and (4.4), we have

$$\begin{aligned} \|r\|_{0,K} &\leq Ch^2 \left|\nabla r_K\right|_{1,K} \leq Ch^2 \left(\left|\operatorname{curl} s_K\right|_{1,K} + \left|R\psi - \psi\right|_{1,K} \right) \\ &\leq Ch^2 \left[\left\|\operatorname{rot}(R\psi - \psi)\right\|_{0,K} + \left|R\psi - \psi\right|_{1,K} \right] \leq Ch^2 \left|R\psi - \psi\right|_{1,K}. \end{aligned} \tag{4.6}$$

On the other hand, let $(\cdot,\cdot)_K$ be the usual inner product in $\mathrm{L}^2(K)$ and P the orthogonal projection onto the constant functions. Because of (2.7), we have, $\forall \eta \in \mathrm{H}_0^1(\Omega)$,

$$\frac{\big(\operatorname{rot}(R\psi - \psi), \eta\big)_K}{\|\eta\|_{1,K}} = \frac{\big(\operatorname{rot}(R\psi - \psi), \eta - P\eta\big)_K}{\|\eta\|_{1,K}} \leq \frac{\left\|\operatorname{rot}(R\psi - \psi)\right\|_{0,K} \left\|\eta - P\eta\right\|_{0,K}}{\|\eta\|_{1,K}}.$$

Hence

$$\left\|\operatorname{rot}(R\psi - \psi)\right\|_{-1,K} \leq Ch \left\|\operatorname{rot}(R\psi - \psi)\right\|_{0,K}.$$

Now, let $S \in \mathrm{L}^2(\Omega)^2$ be such that $S|_K = \operatorname{curl} s_K$. Therefore, because of (4.4) we have

$$\begin{aligned} \|S\|_{0,\Omega}^2 &= \sum_{K \in \mathcal{T}_h} \left\|\operatorname{curl} s_K\right\|_{0,K}^2 \leq \sum_{K \in \mathcal{T}_h} \left\|\operatorname{rot}(R\psi - \psi)\right\|_{-1,K}^2 \\ &\leq Ch^2 \sum_{K \in \mathcal{T}_h} \left\|\operatorname{rot}(R\psi - \psi)\right\|_{0,K}^2 \leq Ch^2 \left\|\operatorname{rot}(R\psi - \psi)\right\|_{0,\Omega}^2. \end{aligned} \tag{4.7}$$

Finally, from (4.5) we obtain

$$|(\zeta, \psi - R\psi)| = \left| \int_\Omega \zeta \cdot \nabla r + \int_\Omega \zeta \cdot S \right| \leq \|\operatorname{div} \zeta\|_{0,\Omega} \|r\|_{0,\Omega} + \|\zeta\|_{0,\Omega} \|S\|_{0,\Omega},$$

and the lemma follows by using (4.6) and (4.7). □

To obtain a bound of the consistency term in Lemma 4.1, there remains only to estimate the terms involving $(R\psi - \psi)$ of the previous lemma. To this aim, we use the analogue of Theorem 4.3 in [24] applied to our situation in the space $\mathrm{H}(\mathrm{rot}, \Omega)$, which reads

$$|\psi - R\psi|_{1,K} \leq C \left(|\psi|_{1,K} + h_K \left|\operatorname{rot} \psi\right|_{1,K} \right) \leq \|\psi\|_{2,K} \tag{4.8}$$

and

$$\|\operatorname{rot}(\psi - R\psi)\|_{0,K} \le C\left(\frac{\delta_K}{h_K}|\operatorname{rot}\psi|_{0,K} + h_K|\operatorname{rot}\psi|_{1,K}\right), \tag{4.9}$$

where δ_K is a measure of the deviation of the quadrilateral K from a parallelogram, as defined in Figure 4.1.

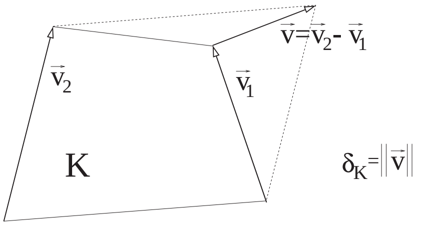


FIG. 4.1. *Geometrical definition of δ_K.*

Note that for shape-regular meshes clearly $\delta_K/h_K \le C\ \forall K \in \mathcal{T}_h$. On the other hand, $\{\mathcal{T}_h\}$ is said to be a family of *asymptotically parallelogram* meshes when there exists a constant C such that $\max_{K\in\mathcal{T}_h}(\delta_K/h_K) \le Ch$ for all the meshes.

Now we are in position to estimate the consistency term in Lemma 4.1.

LEMMA 4.3. *Let β_h, δ, γ, and $\widetilde{\varphi}$ be as in Lemma* 4.1. *Then there holds*

$$\frac{|(\beta_h - R\beta_h, \delta)| + |(\gamma, \widetilde{\varphi} - R\widetilde{\varphi})|}{\|\beta - \beta_h\|_{0,\Omega} + \|w - w_h\|_{0,\Omega}} \le Ch\left(h + t\max_{K\in\mathcal{T}_h}\frac{\delta_K}{h_K}\right)|(\theta, f)|_t.$$

Proof. First, we have

$$|(\beta_h - R\beta_h, \delta)| \le \left|\left((\beta_h - \beta) - R(\beta_h - \beta), \delta\right)\right| + |(\beta - R\beta, \delta)|.$$

By using (2.9), Theorem 3.7, and (4.3), we obtain

$$\begin{aligned}\left|\left((\beta_h - \beta) - R(\beta_h - \beta), \delta\right)\right| &\le Ch\|\beta_h - \beta\|_{1,\Omega}\|\delta\|_{0,\Omega} \le Ch^2|(\theta, f)|_t\|\delta\|_{0,\Omega}\\ &\le Ch^2|(\theta, f)|_t\left(\|\beta - \beta_h\|_{0,\Omega} + \|w - w_h\|_{0,\Omega}\right).\end{aligned}$$

On the other hand, by the definition of γ in (2.1) and the estimate (2.2), we have

$$|\operatorname{rot}\beta|_{0,K} = \frac{t^2}{\kappa}|\operatorname{rot}\gamma|_{0,K} \le Ct|(\theta, f)|_t.$$

Then, by using Lemma 4.2, (4.8), (4.9), the estimate above, (2.2), (4.2), and (4.3), we have

$$\begin{aligned}|(\beta - R\beta, \delta)| &\le Ch^2\left(\sum_K|R\beta - \beta|_{1,K}^2\right)^{1/2}\|\operatorname{div}\delta\|_{0,\Omega} + Ch\|\operatorname{rot}(R\beta - \beta)\|_{0,\Omega}\|\delta\|_{0,\Omega}\\ &\le Ch^2\|\beta\|_{2,\Omega}\|\operatorname{div}\delta\|_{0,\Omega} + Ch\left(\max_{K\in\mathcal{T}_h}\frac{\delta_K}{h_K}|\operatorname{rot}\beta|_{0,\Omega} + h|\operatorname{rot}\beta|_{1,\Omega}\right)\|\delta\|_{0,\Omega}\\ &\le Ch^2|(\theta, f)|_t\|\operatorname{div}\delta\|_{0,\Omega} + Ch\left(h + t\max_{K\in\mathcal{T}_h}\frac{\delta_K}{h_K}\right)|(\theta, f)|_t\|\delta\|_{0,\Omega}\\ &\le Ch\left(h + t\max_{K\in\mathcal{T}_h}\frac{\delta_K}{h_K}\right)|(\theta, f)|_t\left(\|\beta - \beta_h\|_{0,\Omega} + \|w - w_h\|_{0,\Omega}\right).\end{aligned}$$

The term $|(\gamma, \widetilde{\varphi} - R\widetilde{\varphi})|$ can be bounded almost identically by using Lemma 3.2 for *MITC*4 or Lemma 3.5 for *DL*4 to estimate $\|\widetilde{\varphi} - \varphi\|_{1,\Omega}$ and the fact that

$$-\operatorname{div}\gamma = f \qquad \text{in } \Omega,$$

which follows by taking $\eta = 0$ in the first equation of (2.1). Therefore, we obtain

$$|(\gamma, \widetilde{\varphi} - R\widetilde{\varphi})| \le Ch\left(h + t\max_{K\in\mathcal{T}_h}\frac{\delta_K}{h_K}\right)|(\theta, f)|_t\left(\|\beta - \beta_h\|_{0,\Omega} + \|w - w_h\|_{0,\Omega}\right),$$

which allows us to conclude the proof. □

Finally, we can establish an $\mathrm{L}^2(\Omega)$ error estimate. As above, in the case of *MITC*4 elements, we consider meshes satisfying Assumption 3.1.

THEOREM 4.4. *Given $(\theta, f) \in \mathrm{L}^2(\Omega)^2 \times \mathrm{L}^2(\Omega)$, let (β, w) and (β_h, w_h) be the solutions of Problems* 2.1 *and* 2.3*, respectively. Then there exists a constant C, independent of t and h, such that*

$$\|\beta - \beta_h\|_{0,\Omega} + \|w - w_h\|_{0,\Omega} \le Ch\left(h + t\max_{K\in\mathcal{T}_h}\frac{\delta_K}{h_K}\right)|(\theta, f)|_t\,.$$

Proof. The proof is a direct consequence of Lemmas 4.1 and 4.3. □

COROLLARY 4.5. *The following error estimate holds for any family of asymptotically parallelogram meshes:*

$$\|\beta - \beta_h\|_{0,\Omega} + \|w - w_h\|_{0,\Omega} \le Ch^2\,|(\theta, f)|_t\,.$$

Remark 4.1. The asymptotically parallelogram assumption on the meshes is not necessary as long as $h > \alpha t$ for α fixed. Indeed, according to Theorem 4.4, for general regular meshes with $h > \alpha t$, we have

$$\|\beta - \beta_h\|_{0,\Omega} + \|w - w_h\|_{0,\Omega} \le C_\alpha h^2|(\theta, f)|_t.$$

Note that the condition $h > \alpha t$ is fulfilled in practice for reasonably large values of α.

5. The spectral problem.

The aim of this section is to study how the eigenvalues and eigenfunctions of Problem 2.4 approximate those of Problem 2.2. We do this in the framework of the abstract spectral approximation theory as stated, for instance, in the monograph by Babuška and Osborn [5]. In order to use this theory, we define operators T and T_h associated to the continuous and discrete spectral problems, respectively.

We consider the operator

$$T : \mathrm{L}^2(\Omega)^2 \times \mathrm{L}^2(\Omega) \longrightarrow \mathrm{L}^2(\Omega)^2 \times \mathrm{L}^2(\Omega),$$

defined by $T(\theta, f) := (\beta, w)$, where $(\beta, w) \in \mathrm{H}^1_0(\Omega)^2 \times \mathrm{H}^1_0(\Omega)$ is the solution of Problem 2.1. Note that T is compact as a consequence of estimate (2.2). Since the operator is clearly self-adjoint with respect to $(\cdot,\cdot)_t$, then, apart from $\mu = 0$, its spectrum consists of a sequence of finite multiplicity real eigenvalues converging to zero. Note that λ is an eigenvalue of Problem 2.2 if and only if $\mu := 1/\lambda$ is an eigenvalue of T, with the same multiplicity and corresponding eigenfunctions.

As shown in [13], each eigenvalue μ of Problem 2.1 converges to some limit μ_0 when the thickness $t \to 0$. Indeed, μ_0 is an eigenvalue of the operator associated with the Kirchhoff model of the same plate (see Lemma 2.1 in [13]). From now on, for simplicity, we assume that $\mu = 1/\lambda$ is an eigenvalue of T which converges to a simple eigenvalue μ_0, as t goes to zero (see section 2 in [13] for further discussions).

Now, analogously to the continuous case, we introduce the operator

$$T_h : \mathrm{L}^2(\Omega)^2 \times \mathrm{L}^2(\Omega) \longrightarrow \mathrm{L}^2(\Omega)^2 \times \mathrm{L}^2(\Omega),$$

defined by $T_h(\theta, f) := (\beta_h, w_h)$, where $(\beta_h, w_h) \in H_h \times W_h$ is the solution of Problem 2.3. The operator T_h is also self-adjoint with respect to $(\cdot,\cdot)_t$. Clearly, the eigenvalues of T_h are given by $\mu_h := 1/\lambda_h$, with λ_h being the strictly positive eigenvalues of Problem 2.4, and the corresponding eigenfunctions coincide.

As a consequence of Theorem 3.7, for each simple eigenvalue μ of T, there is exactly one eigenvalue μ_h of T_h converging to μ as h goes to zero (see, for instance, [16]). The following theorem shows optimal t-independent error estimates. Let us remark that the results of this theorem are valid for both methods $MITC4$ and $DL4$, although, for the former, under Assumption 3.1 on the meshes as in the previous section.

THEOREM 5.1. *Let λ and λ_h be simple eigenvalues of Problems* 2.2 *and* 2.4*, respectively, such that $\lambda_h \to \lambda$ as $h \to 0$. Let (β, w) and (β_h, w_h) be corresponding eigenfunctions normalized in the same manner. Then, under the assumptions stated above, there exists $C > 0$ such that, for t and h small enough, there holds*

$$\|\beta - \beta_h\|_{1,\Omega} + \|w - w_h\|_{1,\Omega} \leq Ch.$$

Furthermore, for any family of asymptotically parallelogram meshes, there hold

$$\|\beta - \beta_h\|_{0,\Omega} + \|w - w_h\|_{0,\Omega} \leq Ch^2$$

and

$$|\lambda - \lambda_h| \leq Ch^2.$$

Proof. The proof, which relies on Theorem 3.7 and Corollary 4.5, is essentially the same as those of Theorems 2.1, 2.2, and 2.3 in [13]. □

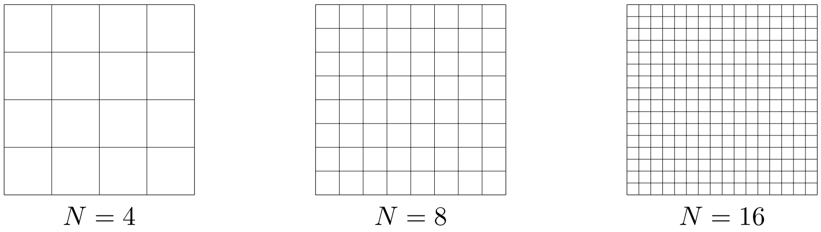
6. Numerical experiments. In this section, we report some numerical experiments carried out with both methods applied to the spectral problem, Problem 2.2.

First, we have tested the two methods by using different meshes, not necessarily satisfying the assumptions in the theorems above. We have considered a square clamped moderately thick plate of side-length L and thickness-to-span ratio $t/L = 0.1$. We report the results obtained with both types of elements using the following three families of meshes:

$\mathcal{T}_h^{\mathrm{U}}$ consists of uniform subdivisions of the domain into $N \times N$ subsquares for $N = 4, 8, 16, \ldots$ (see Figure 6.1). Clearly, these are parallelogram meshes satisfying Assumption 3.1.

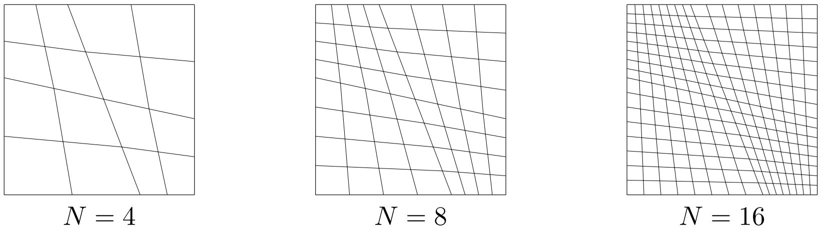
$\mathcal{T}_h^{\mathrm{A}}$ consists of "uniform" refinements of a nonuniform mesh obtained by splitting the square into four quadrilaterals. Each refinement step is obtained by subdividing each quadrilateral into the other four, by connecting the midpoints of the opposite edges. Thus we obtain a family of $N \times N$ asymptotically parallelogram-shape-regular meshes as shown in Figure 6.2. Furthermore, for $N = 4, 8, 16, \ldots$, these meshes satisfy Assumption 3.1.

$\mathcal{T}_h^{\mathrm{T}}$ consists of partitions of the domain into $N \times N$ congruent trapezoids, all similar to the trapezoid with vertices $(0,0)$, $(1/2,0)$, $(1/2,2/3)$, and $(0,1/3)$, as shown in Figure 6.3. Clearly, these are not asymptotically parallelogram meshes and they do not satisfy Assumption 3.1.



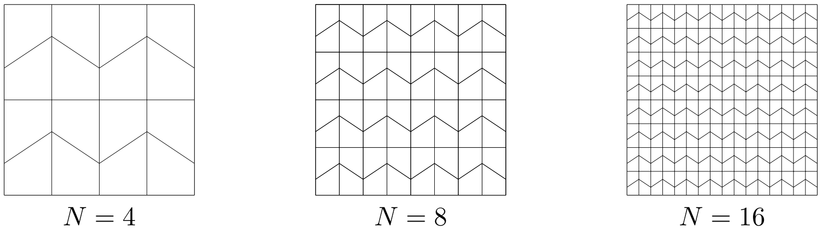
$N = 4$ $N = 8$ $N = 16$

FIG. 6.1. *Uniform square meshes* $\mathcal{T}_h^{\mathrm{U}}$.



$N = 4$ $N = 8$ $N = 16$

FIG. 6.2. *Asymptotically parallelogram meshes* $\mathcal{T}_h^{\mathrm{A}}$.



$N = 4$ $N = 8$ $N = 16$

FIG. 6.3. *Trapezoidal meshes* $\mathcal{T}_h^{\mathrm{T}}$.

Let us remark that the third family was used in [3, 4] to show that the order of convergence of some finite elements deteriorate on these meshes in spite of the fact that they are shape-regular.

We have computed approximations of the free-vibration angular frequencies $\omega = t\sqrt{\lambda/\rho}$ corresponding to the lowest-frequency vibration modes of the plate. In order to compare the obtained results with those in [1], we present the computed frequencies ω_{mn}^h in the nondimensional form

$$\hat{\omega}_{mn} := \omega_{mn}^h L \left[\frac{2(1+\nu)\rho}{E}\right]^{1/2},$$

m and n being the numbers of half-waves occurring in the mode shapes in the x and y directions, respectively.

Tables 6.1 and 6.2 show the four lowest-vibration frequencies computed by our method with successively refined meshes of each type, $\mathcal{T}_h^{\mathrm{U}}$, $\mathcal{T}_h^{\mathrm{A}}$, and $\mathcal{T}_h^{\mathrm{T}}$. Each table also includes the values of the vibration frequencies obtained by extrapolating the computed ones as well as the estimated order of convergence. Finally, each table also includes in its last column the results reported in [1]. In every case, we have used a Poisson ratio $\nu = 0.3$ and a correction factor $k = 0.8601$. The reported nondimensional frequencies are independent of the remaining geometrical and physical parameters, except for the thickness-to-span ratio.

Table 6.1
Scaled vibration frequencies $\hat{\omega}_{mn}$ computed with MITC4.

Mesh	Mode	$N = 16$	$N = 32$	$N = 64$	Extrap.	Order	[1]
$\mathcal{T}_h^{\mathrm{U}}$	$\hat{\omega}_{11}$	1.6055	1.5946	1.5919	1.5910	2.01	1.591
	$\hat{\omega}_{21}$	3.1042	3.0550	3.0429	3.0389	2.03	3.039
	$\hat{\omega}_{12}$	3.1042	3.0550	3.0429	3.0389	2.03	3.039
	$\hat{\omega}_{22}$	4.3534	4.2850	4.2681	4.2625	2.02	4.263
$\mathcal{T}_h^{\mathrm{A}}$	$\hat{\omega}_{11}$	1.6073	1.5951	1.5921	1.5911	2.01	1.591
	$\hat{\omega}_{21}$	3.1094	3.0563	3.0433	3.0390	2.02	3.039
	$\hat{\omega}_{12}$	3.1190	3.0586	3.0438	3.0390	2.03	3.039
	$\hat{\omega}_{22}$	4.3711	4.2894	4.2692	4.2626	2.02	4.263
$\mathcal{T}_h^{\mathrm{T}}$	$\hat{\omega}_{11}$	1.6112	1.5961	1.5923	1.5910	1.99	1.591
	$\hat{\omega}_{21}$	3.1129	3.0575	3.0436	3.0388	1.99	3.039
	$\hat{\omega}_{12}$	3.1306	3.0618	3.0446	3.0388	2.00	3.039
	$\hat{\omega}_{22}$	4.3916	4.2955	4.2708	4.2622	1.96	4.263

Table 6.2
Scaled vibration frequencies $\hat{\omega}_{mn}$ computed with DL4.

Mesh	Mode	$N = 16$	$N = 32$	$N = 64$	Extrap.	Order	[1]
$\mathcal{T}_h^{\mathrm{U}}$	$\hat{\omega}_{11}$	1.5956	1.5922	1.5913	1.5910	1.98	1.591
	$\hat{\omega}_{21}$	3.0711	3.0470	3.0409	3.0388	1.99	3.039
	$\hat{\omega}_{12}$	3.0711	3.0470	3.0409	3.0388	1.99	3.039
	$\hat{\omega}_{22}$	4.3136	4.2754	4.2657	4.2624	1.98	4.263
$\mathcal{T}_h^{\mathrm{A}}$	$\hat{\omega}_{11}$	1.5929	1.5915	1.5912	1.5910	1.94	1.591
	$\hat{\omega}_{21}$	3.0592	3.0441	3.0402	3.0388	1.96	3.039
	$\hat{\omega}_{12}$	3.0732	3.0476	3.0411	3.0389	1.98	3.039
	$\hat{\omega}_{22}$	4.3136	4.2756	4.2658	4.2624	1.96	4.263
$\mathcal{T}_h^{\mathrm{T}}$	$\hat{\omega}_{11}$	1.5927	1.5914	1.5911	1.5910	2.21	1.591
	$\hat{\omega}_{21}$	3.0606	3.0445	3.0403	3.0388	1.94	3.039
	$\hat{\omega}_{12}$	3.0654	3.0453	3.0405	3.0390	2.05	3.039
	$\hat{\omega}_{22}$	4.3131	4.2754	4.2657	4.2623	1.96	4.263

It can be clearly seen that both methods converge for the three types of meshes with an optimal $\mathcal{O}(h^2)$ order. Hence none of the two particular assumptions on the meshes (Assumption 3.1 and the assumption of being asymptotically parallelogram) seem to be actually necessary.

As a second test, we have made a numerical experiment to assess the stability of the methods as the thickness t goes to zero. We have used a sequence of clamped plates with decreasing values of the thickness-to-span ratios: $t/L = 0.1$, 0.01, 0.001, 0.0001. All the other geometrical and physical parameters have been taken as in the previous test.

We have computed again approximations of the free-vibration angular frequencies $\omega = t\sqrt{\lambda/\rho}$. The quotients ω/t are known to converge to the corresponding vibration frequencies of an identical Kirchhoff plate (i.e., to the frequencies obtained from the Kirchhoff model for the deflection of a similar zero-thickness ideal plate; see Lemma 2.1 from [13]). Because of this, we present now the computed frequencies ω_{mn}^h scaled in the following manner:

$$\tilde{\omega}_{mn} := \omega_{mn}^h \frac{L}{t} \left[\frac{2(1+\nu)\rho}{E} \right]^{1/2}.$$

Table 6.3
Scaled vibration frequency $\tilde{\omega}_{11}$ computed with $DL4$ for different thickness-to-span ratios t/L.

Mesh	t/L	$N = 16$	$N = 32$	$N = 64$	Extrap.	Order
$\mathcal{T}_h^{\mathrm{U}}$	0.1	15.9561	15.9220	15.9133	15.9104	1.98
	0.01	17.5778	17.5485	17.5412	17.5387	1.99
	0.001	17.5975	17.5685	17.5612	17.5588	2.00
	0.0001	17.5976	17.5687	17.5614	17.5590	2.00
	0 (extrap.)	17.5977	17.5687	17.5614	17.5590	2.00
$\mathcal{T}_h^{\mathrm{A}}$	0.1	15.9286	15.9151	15.9116	15.9104	1.94
	0.01	17.5368	17.5382	17.5385	17.5387	1.87
	0.001	17.5563	17.5580	17.5586	17.5588	1.74
	0.0001	17.5565	17.5582	17.5588	17.5590	1.74
	0 (extrap.)	17.5565	17.5582	17.5588	17.5590	1.76
$\mathcal{T}_h^{\mathrm{T}}$	0.1	15.9272	15.9141	15.9113	15.9105	2.21
	0.01	17.5681	17.5450	17.5395	17.5377	2.05
	0.001	17.5901	17.5671	17.5608	17.5585	1.89
	0.0001	17.5903	17.5673	17.5611	17.5588	1.89
	0 (extrap.)	17.5903	17.5674	17.5611	17.5588	1.89

The obtained results have been qualitatively similar for both methods. We report only those obtained with $DL4$, since the performance of $MITC4$ has been assessed in many other papers (see, for instance, [8], as well as [15] for the vibration problem).

We present in Table 6.3 the results for the lowest-frequency vibration mode, with the same meshes as in the previous test. In each case, for each thickness-to-span ratio t/L, we have computed again an extrapolated, more accurate value of the scaled vibration frequency and the estimated order of convergence. Finally we have also estimated by extrapolation the limit values of the scaled frequencies $\tilde{\omega}_{mn}$ as t goes to zero.

Table 6.4
Extrapolated values as $(t/L) \to 0$ of the scaled vibration frequencies $\tilde{\omega}_{mn}$ computed with $DL4$.

Mesh	Mode	$N = 16$	$N = 32$	$N = 64$	Extrap.	Order
$\mathcal{T}_h^{\mathrm{U}}$	$\hat{\omega}_{11}$	17.5977	17.5687	17.5614	17.5590	2.00
	$\hat{\omega}_{21}$	36.2064	35.9115	35.8374	35.8125	1.99
	$\hat{\omega}_{12}$	36.2064	35.9115	35.8374	35.8126	1.99
	$\hat{\omega}_{22}$	53.4123	52.9570	52.8428	52.8045	1.99
$\mathcal{T}_h^{\mathrm{A}}$	$\hat{\omega}_{11}$	17.5565	17.5583	17.5588	17.5590	1.76
	$\hat{\omega}_{21}$	35.9947	35.8590	35.8243	35.8123	1.97
	$\hat{\omega}_{12}$	36.2003	35.9102	35.8371	35.8124	1.99
	$\hat{\omega}_{22}$	53.3174	52.9353	52.8374	52.8037	1.97
$\mathcal{T}_h^{\mathrm{T}}$	$\hat{\omega}_{11}$	17.5904	17.5673	17.5611	17.5588	1.89
	$\hat{\omega}_{21}$	36.0770	35.8823	35.8303	35.8113	1.90
	$\hat{\omega}_{12}$	36.2500	35.9259	35.8412	35.8112	1.94
	$\hat{\omega}_{22}$	53.5074	52.9936	52.8526	52.7993	1.87

Note that the extrapolated values for each thickness-to-span ratio are almost identical for the three meshes. Moreover, although the estimated orders of convergence seem to deteriorate a bit as t/L goes to zero for the nonuniform meshes, the values obtained with these meshes are better than those computed with the uniform mesh (i.e., closer to the extrapolated ones), even for the coarser meshes. Therefore, this test suggests that the method is locking-free for any kind of regular mesh.

Finally, we report in Table 6.4 the corresponding extrapolated values as t/L goes to zero for the four lowest-scaled vibration frequencies. It can be seen from this table that the results are essentially the same as for $\tilde{\omega}_{11}$. Furthermore, the computed orders of convergence are even closer to 2.

Further experiments with *MITC*4 have been reported in [15], including other boundary conditions and the extension of this method to compute the vibration modes of Naghdi shells.

Acknowledgment. This work was partially done by three of the authors (Ricardo G. Durán, Elsa Liberman, and Rodolfo Rodríguez) at the Mathematical Science Research Institute, Berkeley, CA, in April 2000, during the Workshop on *Elastic Shells: Modeling, Analysis and Numerics*, chaired by Douglas N. Arnold in the program in *Numerical and Applied Mathematics* organized by Ivo M. Babuška. They are grateful for the support and stimulating atomosphere at the Institute.

REFERENCES

[1] B. S. Al Janabi and E. Hinton, *A study of the free vibrations of square plates with various edge conditions*, in Numerical Methods and Software for Dynamic Analysis of Plates and Shells, E. Hinton, ed., Pineridge Press, Swansea, UK, 1988, pp. 167–204.

[2] D. N. Arnold and R. S. Falk, *A uniformly accurate finite element method for the Reissner–Mindlin plate*, SIAM J. Numer. Anal., 26 (1989) pp. 1276–1290.

[3] D. N. Arnold, D. Boffi, and R. S. Falk, *Approximation by quadrilateral finite elements*, Math. Comp., 71 (2002), pp. 909–922.

[4] D. N. Arnold, D. Boffi, R. S. Falk, and L. Gastaldi, *Finite element approximation on quadrilateral meshes*, Comm. Numer. Methods Engrg., 17 (2001), pp. 805–812.

[5] I. Babuška and J. Osborn, *Eigenvalue problems*, in Handbook of Numerical Analysis, Vol. II, P. G. Ciarlet and J. L. Lions, eds., North–Holland, Amsterdam, 1991, pp. 641–787.

[6] K.-J. Bathe and F. Brezzi, *On the convergence of a four-node plate bending element based on Mindlin/Reissner plate theory and a mixed interpolation*, in The Mathematics of Finite Elements and Applications V, J. R. Whiteman, ed., Academic Press, London, 1985, pp. 491–503.

[7] K.-J. Bathe and E. N. Dvorkin, *A four-node plate bending element based on Mindlin/Reissner plate theory and a mixed interpolation*, Internat. J. Numer. Methods Engrg., 21 (1985), pp. 367–383.

[8] K.-J. Bathe, A. Iosilevich, and D. Chapelle, *An evaluation of the MITC shell elements*, Comput. & Structures, 75 (2000), pp. 1–30.

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

[10] F. Brezzi, M. Fortin, and R. Stenberg, *Quasi-optimal error bounds for approximation of shear-stresses in Mindlin-Reissner plate models*, Math. Models Methods Appl. Sci., 1 (1991), pp. 125–151.

[11] P. G. Ciarlet, *Basic error estimates for elliptic problems*, in Handbook of Numerical Analysis, Vol. II, P. G. Ciarlet and J. L. Lions, eds., North–Holland, Amsterdam, 1991, pp. 17–351.

[12] R. Durán and E. Liberman, *On mixed finite element methods for the Reissner-Mindlin plate model*, Math. Comp., 58 (1992), pp. 561–573.

[13] R. Durán, L. Hervella-Nieto, E. Liberman, R. Rodríguez, and J. Solomin, *Approximation of the vibration modes of a plate by Reissner-Mindlin equations*, Math. Comp., 68 (1999), pp. 1447–1463.

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

[15] E. Hernández, L. Hervella-Nieto, and R. Rodríguez, *Computation of the vibration modes of plates and shells by low-order MITC quadrilateral finite elements*, Comput. & Structures, 81 (2003), pp. 615–628.

[16] T. Kato, *Perturbation Theory for Linear Operators*, Springer-Verlag, Berlin, 1995.

[17] E. Liberman, *Sobre la convergencia de métodos de elementos finitos para el modelo de placas de Reissner-Mindlin*, Ph.D. thesis, Universidad Nacional de La Plata, Argentina, 1995.

[18] P. Peisker and D. Braess, *Uniform convergence of mixed interpolated elements for Reissner-Mindlin plates*, M2AN Math. Model. Numer. Anal., 26 (1992), pp. 557–574.

[19] J. Pitkäranta and R. Stenberg, *Error bounds for the approximation of the Stokes problem using bilinear/constant elements on irregular quadrilateral meshes*, in The Mathematics of Finite Elements and Applications V, J. R. Whiteman, ed., Academic Press, London, 1985, pp. 325–334.

[20] J. Pitkäranta and M. Suri, *Design principles and error analysis for reduced-shear plate-bending finite elements*, Numer. Math., 75 (1996), pp. 223–266.

[21] P. A. Raviart and J. M. Thomas, *A mixed finite element method for second order elliptic problems*, in Mathematical Aspects of Finite Element Methods, Lecture Notes in Mathematics 606, I. Galligani and E. Magenes, eds., Springer-Verlag, Berlin, 1977, pp. 292–315.

[22] R. Stenberg, *Analysis of mixed finite element methods for the Stokes problem: A unified approach*, Math. Comp., 42 (1984), pp. 9–23.

[23] R. Stenberg and M. Suri, *An hp error analysis of MITC plate elements*, SIAM J. Numer. Anal., 34 (1997), pp. 544–568.

[24] J. M. Thomas, *Sur l'analyse numérique des méthodes d'éléments finis hybrides et mixtes*, Thèse de Doctorat d'Etat, Université Pierre et Marie Curie, Paris 6, France, 1977.

SIAM J. NUMER. ANAL.
Vol. 41, No. 5, pp. 1773–1784

$O(h^2)$ GLOBAL POINTWISE ALGORITHMS FOR DELAY QUADRATIC PROBLEMS IN THE CALCULUS OF VARIATIONS[*]

OM PRAKASH AGRAWAL[†] AND JOHN GREGORY[‡]

Abstract. The main purpose of this paper is to give numerical algorithms and the error analysis for delay quadratic problems in the calculus of variations. These methods are new, efficient, and accurate and have a global a priori error of $O(h^2)$, where h is the distance between any two successive node points.

We also derive the results for the general, numerical, delay problem, but we focus on proving our results in the simpler quadratic case since the extra technical numerical details have been given previously by the second author. In addition, the authors have previously shown how to reformulate general delay constrained problems in optimal control theory/constrained calculus of variations as unconstrained delay problems. Thus our numerical results and methods will hold for these general constrained problems also.

Finally, we note that our algorithm, which solves the stationary condition(s) numerically, avoids the more difficult problems of piecing the solution of second order equations together and requires less smoothness in the solution. Thus we replace difficult second order boundary value problems with the easier task of approximating definite integrals involving first order derivatives.

Key words. numerical delay extremal solutions, $O(h^2)$ global error

AMS subject classifications. 49K25, 65D99, 65K10

DOI. 10.1137/S0036142902410623

1. Introduction.

The specific purpose of this paper is to give numerical methods and efficient algorithms for extremal values of functionals in the form

$$J(y) = \frac{1}{2}\int_a^b [r(t)y'^2(t) + \bar{r}_\tau(t)y'^2_\tau(t) + p(t)y^2(t) + \bar{p}_\tau(t)y^2_\tau(t)]dt, \tag{1.1}$$

where $y_\tau(t) \equiv y(t-\tau)$, etc. and $0 < \tau < b-a$. This presentation represents the quadratic form theory for delay problems associated with the more general integral

$$I(y) = \int_a^b f(t, y, y', y_\tau, y'_\tau)dt. \tag{1.2}$$

For either (1.1) or (1.2), we assume

$$y(t) = \alpha(t), \quad t \in [a-\tau, \alpha], \tag{1.3}$$

and the usual smoothness conditions as in [3] to pose a well-defined problem.

The problems of the above type have been considered by several investigators. Eller and Aggarwal [5] presented an optimal control solution for a linear time varying system with delay. The formulation is complex because it requires solution of a system of partial differential equations. Several investigators have used Walsh, block pulse,

[*] Received by the editors July 2, 2002; accepted for publication (in revised form) April 10, 2003; published electronically October 28, 2003.
http://www.siam.org/journals/sinum/41-5/41062.html

[†] Department of Mechanical Engineering, Southern Illinois University Carbondale, Carbondale, IL 62901 (om@engr.siu.edu).

[‡] Department of Mathematics, Southern Illinois University Carbondale, Carbondale, IL 62901 (jgregory@math.siu.edu).

shifted Legendre, and other polynomial approaches to solve linear and nonlinear optimal control problems with delay [4, 5, 9, 11, 14]. Although these papers develop algorithms to solve the problems, they do not provide convergence proof for the algorithms. References [12] and [10, 13, 15] presented formulations and algorithms for optimal control computations for linear and nonlinear time delay systems. Although these papers provide some error bound and convergence criteria, they do not develop expressions for the order of convergence.

A complete variational theory for (1.2) has been given by the authors in [3, 1, 2], including constrained delay problems in optimal control theory/calculus of variations. Thus our methods in this paper will allow us to obtain numerical results for delay variational problems with linear constraints and, with a little more work, a *complete* numerical theory for general, constrained, delay problems. We have chosen to concentrate on (1.1) since the results are so important in applied problems, and by doing so we will avoid difficult technical problems and problems of exposition for the reader.

Our paper is divided as follows. In section 2 we give the basic variational theory for (1.1). This motivates section 3, where we derive our basic algorithm for (1.1). In section 4 we derive the local error for our algorithm and then our main result, that the algorithm has a global a priori error of $O(h^2)$, where h is the distance between node points.

In section 5 we derive our basic results a second way, using the more general results for (1.2) applied to (1.1). In section 6 we consider an example to demonstrate our numerical results.

2. The basic variational theory. The purpose of this section is to derive the necessary analytic conditions for the stationary conditions for our delay problem in (1.1), which, in turn, leads to our numerical algorithm in section 4. The three conditions we obtain, which are necessary (and sufficient) for the stationary condition $J'(y,z) = 0$ for all admissible arcs $z(t)$, exactly yield the tools to get an analytic/closed form solution if one can be easily calculated.

We will see in section 4 that our numerical methods follow from solving $J'(y_h, z_k) = 0$, where z_k is the kth linear spline element and $y_h(t)$ is the numerical solution for step size h.

Thus we have the following analytic results for (1.1).

Theorem 1. *If $y(t)$ yields an extremal value for $J(y)$ in* (1.1)*, then*

(i) $r(t)y'(t) = \int_{b-\tau}^{t} p(s)y(s)ds + c_1$ *or* $\frac{d}{dt}(ry') = py$ *for* $b-\tau < t < b$,

(i)$_T$ $y'(b) = 0$ *if* $y(b)$ *is not given,*

(ii) $R(t)y'(t) = \int_{a}^{t} P(s)y(s)ds + c_2$ *or* $\frac{d}{dt}(Ry') = Py$ *for* $a < t < b-\tau$, *and*

(ii)$_T$ $R(t)y'(t)|_{(b-\tau)^-} = r(t)y'(t)|_{(b-\tau)^+}$,

where c_1 and c_2 are constants and

$$R(t) = [r(t) + \bar{r}(t+\tau)], \qquad P(t) = [p(t) + \bar{p}(t)]. \tag{2.1}$$

We note that (i) is the well-known Euler–Lagrange equation for nondelay problems which, in this case, holds on $b-\tau < t < b$ while (ii) is the same with the delay. Condition (i)$_T$ is the expected transversality condition at $t = b$. There is no transversality condition at $t = a$ since $y(a) = \alpha(a)$ is specified in (1.3). Condition (ii)$_T$ is a piecing or consistency condition for the two second order equations. Thus we have two equations and three boundary conditions as required. Surprisingly, this delay problem, which initially looks quite formidable, can be profitably viewed as a translation problem on $a < t < b-\tau$, a classical problem on $b-\tau < t < b$, and a consistency condition at $t = b-\tau$.

To obtain these results we begin with $F'(0) = 0$, where $F(\epsilon) = J(y + \epsilon z)$, ϵ in $\mathbb{R}$. Thus

$$
\begin{aligned}
0 &= \frac{d}{d\epsilon} J(y+\epsilon z)\bigg|_{\epsilon=0} = J'(y,z) = \int_a^b [r(t)y'(t)z'(t) + \bar{r}_\tau(t)y'_\tau(t)z'_\tau(t) + p(t)y(t)z(t) \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad + \bar{p}_\tau(t)y_\tau(t)z_\tau(t)]dt \\
&= \int_a^{b-\tau} \{[r(t) + \bar{r}(t+\tau)]y'(t)z'(t) + [p(t) + \bar{p}(t+\tau)]y(t)z(t)\}dt \\
&\qquad + \int_{b-\tau}^{b} (r(t)y'(t)z'(t) + p(t)y(t)z(t))dt \\
&= \int_a^{b-\tau} [R(t)y'(t)z'(t) + P(t)y(t)z(t)]dt + \int_{b-\tau}^{b} (r(t)y'(t)z'(t) + p(t)y(t)z(t))dt \\
&= \int_a^{b-\tau} \left[R(t)y'(t) - \int_a^t P(s)y(s)ds\right] z'(t)dt \\
&\qquad + \int_{b-\tau}^{b} \left[r(t)y'(t) - \int_{b-\tau}^{t} p(s)y(s)ds\right] z'(t)dt \\
&\qquad + z(t)\int_a^t P(s)y(s)ds\bigg|_a^{b-\tau} + z(t)\int_{b-\tau}^{t} p(s)y(s)ds\bigg|_{b-\tau}^{b},
\end{aligned}
$$

where we have used integration by parts in the last equality and the change of variables $\bar{t} = t + \tau$ and then $t = \bar{t}$ for the previous equality.

We obtain (i) by choosing $z(t) \equiv 0$ on $[a, b-\tau]$ and $z(b) = 0$. Specifically, a fundamental lemma in the classical calculus of variations has the result that $\int_{t_1}^{t_2} q(t)z'(t)dt = 0$ for all $z(t)$ such that $z(t_1) = z(t_2) = 0$ implies $q(t) \equiv c_1$ on $[t_1, t_2]$.

Choosing $z(t) \equiv 0$ on $[a, b-\tau]$ and using (i) imply (i)$_T$ since

$$
\begin{aligned}
0 &= \int_{b-\tau}^{b} c_1 z'(t)dt + z(t)\int_{b-\tau}^{t} P(s)y(s)ds\bigg|_{b-\tau}^{b} \\
&= c_1[z(b) - z(b-\tau)] + z(b)\int_{b-\tau}^{b} P(s)y(s)ds \\
&= z(b)\left[c_1 + \int_{b-\tau}^{b} P(s)y(s)ds\right] = r(b)[R(b)y'(b)]
\end{aligned}
$$

so that if $y(b)$ is not specified, $z(b)$ is not zero, and hence $y'(b) = 0$.

To obtain (ii) we use (i) and (i)$_T$, and hence if $z(a) = z(b-\tau) = 0$, the fundamental lemma implies the result.

Finally, the earlier results imply

$$
\begin{aligned}
0 &= c_2[z(b-\tau) - z(a)] + c_1[z(b) - z(b-\tau)] + z(b-\tau)\int_a^{b-\tau} P(t)y(t)dt \\
&\qquad + z(b)\int_{b-\tau}^{b} p(t)y(t)dt \\
&= z(b-\tau)\left[-c_1 + c_2 + \int_a^{b-\tau} P(t)y(t)dt\right] + z(b)\left[c_1 + \int_{b-\tau}^{b} p(t)z(t)dt\right] \\
&= t(b-\tau)[-c_1 + R(b-\tau)y'(b-\tau)^-] + z(b)[r(b)y'(b)].
\end{aligned}
$$

Now $z(b)y'(b) = 0$, as before, so

$$c_1 = R(b-\tau)y'(b-\tau)^-,$$

and the result follows by taking $t \to (b-\tau)^+$ in (i).

3. Numerical algorithm.

In this section we develop our numerical algorithm for the quadratic problem defined by (1.1) and (1.3).

Taking the variation of $J(y)$ in (1.1) along the z direction, we obtain the condition

$$J'(y,z) = \int_a^b [ry'z' + pyz + \bar{r}y'_\tau z'_\tau + \bar{p}y_\tau z_\tau]dt = 0, \tag{3.1}$$

where z is an arbitrary function except that it vanishes at $t = b$ and for $t \in [a-\tau,\, a]$.

Intuitively, with $z^{(k)}(t)$ defined below, as the linear spline element of size h, we replace the exact condition (3.1) for arbitrary admissible $z(t)$ with $J'_h(y_h, z_h) = 0$, where $y_h = \sum A_k z^{(k)}$ is the approximate solution for each h.

For simplicity in the discussion to follow, we define

$$I'_1(y,z) = \int_a^b [r(t)y'(t)z'(t) + p(t)y(t)z(t)]dt, \tag{3.2}$$

$$I'_2(y,z) = \int_a^b [\bar{r}(t)y'_\tau(t)z'_\tau(t) + \bar{p}(t)y_\tau(t)z_\tau(t)]dt. \tag{3.3}$$

Hence

$$J'(y,z) = I'_1(y,z) + I'_2(y,z). \tag{3.4}$$

We begin our construction by a partition of $[a, b]$. Thus let N be a positive integer. Divide the interval $[a,\, b]$ into N equal parts; then the length of each part is $h = (b-a)/N$. Label the node points as $t_k = a + kh$ for $k = 0, 1, \ldots, N$. For $k = 1, 2, \ldots, N-1$, define the one dimensional spline hat functions as

$$z^{(k)}(t) = \begin{cases} 0, & t \le t_{k-1}, \\ \dfrac{t - t_{k-1}}{h}, & t_{k-1} \le t \le t_k, \\ \dfrac{t_{k+1} - t}{h}, & t_k \le t \le t_{k+1}, \\ 0, & t \ge t_{k+1}. \end{cases} \tag{3.5}$$

Using (3.5), the variation of $I_1(y)$ along $z^{(k)}$ can be written as

$$\begin{aligned} I_1(y, z^{(k)}) = \frac{1}{h}\int_{t_{k-1}}^{t_k} & [r(t)y'(t) + p(t)y(t)(t - t_{k-1})]dt \\ & + \frac{1}{h}\int_{t_k}^{t_{k+1}} [-r(t)y'(t) + p(t)y(t)(t_{k+1} - t)]dt. \end{aligned} \tag{3.6}$$

For computational purposes, (3.6) can be approximated as

$$\begin{aligned} I'_1(y, z^{(k)}) \doteq r(t^*_{k-1}) & \frac{y_k - y_{k-1}}{h} + p(t^*_{k-1})\frac{y_{k-1} + y_k}{2}\frac{h}{2} \\ & - r(t^*_k)\frac{y_{k+1} - y_k}{h} + p(t^*_k)\frac{y_{k+1} + y_k}{2}\frac{h}{2}, \end{aligned} \tag{3.7}$$

where

$$t_{k-1}^* = \frac{t_k + t_{k-1}}{2} \quad \text{and} \quad t_k^* = \frac{t_k + t_{k+1}}{2} \tag{3.8}$$

and y_k, $k = 0, 1, \ldots, N$, are the numerical values of $y(t_k)$. To compute $I_2'(y, z^{(k)})$, note that $z_\tau^{(k)}$ will be zero until $t = t_{k-1} + \tau$. Hence $I_2'(y, z^{(k)})$ can be written as

$$\begin{aligned} I_2'(y, z^{(k)}) = &\int_{t_{k-1}+\tau}^{t_k+\tau} [y_\tau'(t) z_\tau'(t) \bar{r}(t) + \bar{p}(t) y_\tau(t) z_\tau(t)] dt \\ &+ \int_{t_k+\tau}^{t_{k+1}+\tau} [\bar{r}(t) y_\tau'(t) z_\tau'(t) + \bar{p}(t) y_\tau(t) z_\tau(t)] dt. \end{aligned} \tag{3.9}$$

By a direct shift of time axis, (3.9) reduces to

$$\begin{aligned} I_2'(y, z^{(k)}) = &\int_{t_{k-1}}^{t_k} [\bar{r}(t+\tau) y'(t) z'(t) + \bar{p}(t+\tau) y(t) z(t)] dt \\ &+ \int_{t_k}^{t_{k+1}} [\bar{r}(t+\tau) y'(t) z'(t) + \bar{p}(t+\tau) y(t) z(t)] dt. \end{aligned} \tag{3.10}$$

Following the approach presented above, (3.10) can be approximated as

$$\begin{aligned} I_2'(y, z^{(k)}) = &\bar{r}(t_{k-1}^* + \tau) \frac{y_k - y_{k-1}}{h} + \bar{p}(t_{k-1}^* + \tau) \frac{y_{k-1} + y_k}{2} \frac{h}{2} \\ &- \bar{r}(t_k^* + \tau) \frac{y_{k+1} - y_k}{h} + \bar{p}(t_k^* + \tau) \frac{y_{k+1} + y_k}{2} \frac{h}{2}. \end{aligned} \tag{3.11}$$

Combining (3.7) and (3.11), we get

$$\begin{aligned} I'(y, z^{(k)}) \quad = \quad &\{r(t_{k-1}^*) + \bar{r}(t_{k-1}^* + \tau)\} \frac{y_k - y_{k-1}}{h} \\ &+ \{p(t_{k-1}^*) + \bar{p}(t_{k-1}^* + \tau)\} (y_{k-1} + y_k) \frac{h}{4} \\ &- \{r(t_k^*) + \bar{r}(t_k^* + \tau)\} \frac{y_{k+1} - y_k}{h} \\ &+ \{p(t_k^*) + \bar{p}(t_k^* + \tau)\} (y_{k+1} + y_k) \frac{h}{4}. \end{aligned} \tag{3.12}$$

Notice that for $b - \tau \le t_k \le b$, the $\bar{r}(t+\tau)$ and $\bar{p}(t+\tau)$ terms drop out.

Therefore, using (2.1), (3.12) can be written as follows.

For $a \le t_k < b - \tau$, we have

$$\begin{aligned} &R(t_{k-1}^*) \frac{y_k - y_{k-1}}{h} + P(t_{k-1}^*) (y_{k-1} + y_k) \frac{h}{4} \\ &\qquad - R(t_k^*) \frac{y_{k+1} - y_k}{h} + P(t_k^*) (y_k + y_{k+1}) \frac{h}{4} = 0. \end{aligned} \tag{3.13}$$

For $b - \tau < t_k \le b$, we have

$$\begin{aligned} &r(t_{k-1}^*) \frac{y_k - y_{k-1}}{h} + p(t_{k-1}^*) (y_{k-1} + y_k) \frac{h}{4} \\ &\qquad - r(t_k^*) \frac{y_{k+1} - y_k}{h} + p(t_k^*) (y_k + y_{k+1}) \frac{h}{4} = 0. \end{aligned} \tag{3.14}$$

For $t_k = b - \tau$, we have

$$R(t^*_{k-1})\frac{y_k - y_{k-1}}{h} + P(t^*_{k-1})(y_{k-1} + y_k)\frac{h}{4} - r(t^*_k)\frac{y_{k+1} - y_k}{2h} + p(t^*_k)(y_k + y_{k+1})\frac{h}{8} = 0. \tag{3.15}$$

Equations (3.13)–(3.15) provide the numerical algorithm to solve the quadratic optimal control problem defined in section 5. If $y(b)$ is given, we have exactly the number of equations we need. Otherwise, the condition $(\mathrm{i})_T$ in Theorem 1 is approximated by $y_N - y_{N-1} = 0$.

Two final comments are important:

1. For convenience, we have assumed in (3.15) that $b - \tau$ is a node point of our partition. If it is not, the numerical process ignores this continuity condition, and we obtain the same $O(h^2)$ result.
2. Anyone can present a heuristic algorithm as we have done in (3.13)–(3.15). The justification of the local error being $O(h^q)$, with $q \geq 2$, shows that our algorithm makes sense. The global error establishes the main claim for this algorithm.

4. Derivation of local and global errors. The purpose of this section is to derive the local and global errors for algorithm (3.13)–(3.15). In the previous section, different expressions were given: $a \leq t_k < b - \tau$, $b - \tau < t_k \leq b$, and $t_k = b - \tau$. Accordingly, the expressions for the local error for the three cases must be derived separately.

For $a \leq t_k < b - \tau$, we define $L(t, h)$ as

$$\begin{aligned} L(t,h) = R(t^*_{k-1})&\frac{y(t_k) - y(t_{k-1})}{h} + P(t^*_{k-1})(y(t_k) + y(t_{k-1}))\frac{h}{4} \\ &-R(t^*_k)\frac{y(t_{k+1}) - y(t_k)}{h} + P(t^*_k)(y(t_k) + y(t_{k+1}))\frac{h}{4} \end{aligned} \tag{4.1}$$

as the local error in $a \leq t_k < b - \tau$. In the discussion to follow, unless the time parameter is specified, it will be assumed that the functions are computed at t_k. The Taylor series for $y(t_{k+1})$, $y(t_{k-1})$, $R(t^*_k)$, $R(t^*_{k-1})$, $P(t^*_k)$, and $P(t^*_{k-1})$ can be written as

$$\begin{bmatrix} y_{t_{k+1}} \\ y_{t_{k-1}} \end{bmatrix} = y \pm hy' + \frac{1}{2}h^2y'' \pm \frac{1}{6}h^3y''' + \cdots, \tag{4.2}$$

$$\begin{bmatrix} R(t^*_k) \\ R(t^*_{k-1}) \end{bmatrix} = R \pm \frac{1}{2}hR' + \frac{1}{8}h^2R'' \pm \frac{1}{48}h^3R''' + \cdots, \tag{4.3}$$

$$\begin{bmatrix} P(t^*_k) \\ P(t^*_{k-1}) \end{bmatrix} = P \pm \frac{1}{2}hP' + \frac{1}{8}h^2P'' \pm \frac{1}{48}h^3P''' + \cdots. \tag{4.4}$$

Substituting (4.2)–(4.4) into (4.1) and simplifying, we obtain

$$\begin{aligned} L(t,h) = (-Ry'' - R'y' + Py)h + \Big(&-\frac{1}{6}R'y''' - \frac{1}{8}R''y'' - \frac{1}{24}R'''y' \\ &+ \frac{1}{4}Py'' + \frac{1}{4}P'y' + \frac{1}{8}P''y\Big)h^3 + O(h^5). \end{aligned} \tag{4.5}$$

Using the Euler–Lagrange equation in Theorem 1, it follows that

$$L(t,h) = O(h^3) \qquad \text{in} \qquad a \le t_k < b-\tau. \tag{4.6}$$

For $b-\tau \le t_k \le b$ we define the local error $L(t,h)$ as

$$\begin{aligned} L(t,h) = \frac{1}{2}\Bigg[& r(t^*_{k-1})\frac{y(t_k)-y(t_{k-1})}{h} + p(t^*_{k-1})(y(t_k)+y(t_{k-1}))\frac{h}{4} \\ & - r(t^*_k)\frac{y(t_{k+1})-y(t_k)}{h} + p(t^*_k)(y(t_k)+y(t_{k+1}))\frac{h}{4}\Bigg]. \end{aligned} \tag{4.7}$$

The Taylor series for $r(t^*_k)$, $r(t^*_{k-1})$, $p(t^*_k)$, and $p(t^*_{k-1})$ can be written as

$$\begin{bmatrix} r(t^*_k) \\ r(t^*_{k-1}) \end{bmatrix} = r \pm \frac{1}{2}hr' + \frac{1}{8}h^2r'' \pm \frac{1}{48}h^3r''' + \cdots, \tag{4.8}$$

$$\begin{bmatrix} p(t^*_k) \\ p(t^*_{k-1}) \end{bmatrix} = p \pm \frac{1}{2}hp' + \frac{1}{8}h^2p'' \pm \frac{1}{48}h^3p''' + \cdots. \tag{4.9}$$

Substituting (4.2), (4.8), and (4.9) into (4.7) and simplifying, we obtain

$$\begin{aligned} L(t,h) = (-ry'' - r'y' + py)h + \Bigg(& -\frac{1}{6}r'y''' - \frac{1}{8}r''y'' - \frac{1}{24}r'''y' \\ & + \frac{1}{4}py'' + \frac{1}{4}p'y' + \frac{1}{8}p''y\Bigg) h^3 + O(h^5). \end{aligned} \tag{4.10}$$

Using (5.9), it follows, once again, that

$$L(t,h) = O(h^3) \qquad \text{in} \qquad b-\tau < t_k \le b. \tag{4.11}$$

For $t_k = b-\tau$ we define the local error $L(t,h)$ as

$$\begin{aligned} L(t,h) = {} & R(t^*_{k-1})\frac{y(t_k)-y(t_{k-1})}{h} + P(t^*_{k-1})(y(t_k)+y(t_{k-1}))\frac{h}{4} \\ & - r(t^*_k)\frac{y(t_{k+1})-y(t_k)}{2h} + p(t^*_k)(y(t_k)+y(t_{k+1}))\frac{h}{8}. \end{aligned} \tag{4.12}$$

For simplicity in the derivation to follow, define $t_k = b-\tau = c$. Since $t_k = b-\tau$ is a corner point, $y(t_{k-1})$ and $y(t_{k+1})$ must be expanded at c^- and c^+, respectively. Keeping this in mind, the Taylor series for $y(t_{k+1})$ and $y(t_{k-1})$ can be written as

$$y(t_{k-1}) = y - hy'(c^-) + \frac{1}{2}h^2y''(c^-) - \frac{1}{6}h^3y'''(c^-) + \cdots, \tag{4.13}$$

$$y(t_{k+1}) = y + hy'(c^+) + \frac{1}{2}h^2y''(c^+) - \frac{1}{6}h^3y'''(c^+) + \cdots. \tag{4.14}$$

It is assumed that $R(t)$, $P(t)$, $r(t)$, and $p(t)$ are sufficiently smooth at $t = b-\tau$. Therefore, the Taylor expansions of $R(t^*_{k-1})$, $P(t^*_{k-1})$, $r(t^*_k)$, and $p(t^*_k)$ given by (4.3), (4.4), (4.8), and (4.9), respectively, are still valid for $t_k = b - \tau$. Substituting (4.3), (4.4), (4.8), (4.9), (4.13), and (4.14) into (4.12) and simplifying, we obtain

$$\begin{aligned} L(t,h) = {}& Ry'(c^-) - \frac{1}{2}ry'(c^+) \\ & + \left[-\frac{1}{2}Ry''(c^-) - \frac{1}{2}R'y'(c^-) + \frac{1}{2}Py - \frac{1}{4}ry''(c^+) - \frac{1}{4}r'y'(c^+) + \frac{1}{4}py \right] h \\ & + \left[\frac{1}{6}Ry'''(c^-) + \frac{1}{4}R'y''(c^-) + \frac{1}{8}R''y'(c^-) - \frac{1}{4}Py'(c^-) - \frac{1}{4}P'y \right. \\ & \left. - \frac{1}{12}ry'''(c^+) - \frac{1}{8}r'y''(c^+) - \frac{1}{16}r''y'(c^+) + \frac{1}{8}p'y + \frac{1}{8}py'(c^+) \right] h^2 \\ & + \left[-\frac{1}{12}R'y'''(c^-) - \frac{1}{16}R''y''(c^-) - \frac{1}{48}R'''y'(c^-) + \frac{1}{8}Py''(c^-) \right. \\ & + \frac{1}{8}P'y'(c^-) + \frac{1}{16}P''y - \frac{1}{24}r'y'''(c^+) - \frac{1}{32}r''y''(c^+) \\ & \left. - \frac{1}{96}r'''y'(c^+) + \frac{1}{16}py''(c^+) + \frac{1}{16}p'y'(c^+) + \frac{1}{32}p''y \right] h^3 \\ & + O(h^4). \end{aligned} \tag{4.15}$$

Using (4.15), the differential (5.4) and (5.5), and the corner condition (5.6), it follows that

$$L(t,h) = O(h^2) \qquad \text{at} \qquad t_k = b - \tau. \tag{4.16}$$

We note, once again, that we have established consistency for our algorithm by showing that if $y(t)$ satisfies the conditions in Theorem 1 or (5.8)–(5.10), then $L(t,h) = O(h^3)$ for $t \neq b - \tau$ and $L(t,h) = O(h^2)$ for $t = b - \tau$.

Thus we have the following theorem.

THEOREM 2. *The algorithm* (3.13)–(3.15) *is consistent as described directly above.*

We note that the algorithm along with the values for $y(a)$ and $y(b)$, or $y(b)$ replaced by the numerical transversality condition for $y'(b) = 0$, $y_N - y_{N-1} = 0$, leads immediately to a well-defined set of numerical values y_k, $k = 0, 1, \ldots, N$, which are approximation values of $y(a_k)$, where $y(t)$ is a unique solution to our problem. The purpose of Theorem 3 is to see how good the approximation is.

THEOREM 3. *If* $e_h = y(a_k) - y_h(a_k)$*, where* $y(t)$ *is the unique solution described above and* $y_h(a_k) = y_k$ *for fixed* $h > 0$*, given by our algorithm as described above in Theorem* 2*, then for* h *sufficiently small there exists* $C > 0$ *independent of* h *such that*

$$\|e_h\|_\infty \leq Ch^2 \quad \mathit{or} \quad \|e_h\|_2 \leq Ch^2. \tag{4.17}$$

In fact, this theorem will hold in a variety of situations: (a) when y is an m vector, (b) when the algorithm is generated in the obvious way with nonquadratic f as in section 3, and (c) when a constraint delay problem is reformulated as an unconstrained problem.

For example, in (a) above, for $y(t)$ an m vector, we use $\|x\|_2$ to denote the 2-norm and $\|x\|_\infty$ to denote the max-norm; thus, if $x = (x^1, x^2, \ldots, x^m)$, then

$$\begin{aligned} \|x\|_2 &= \sqrt{(x^1)^2 + (x^2)^2 + \cdots + (x^m)^2}, \\ \|x\|_\infty &= \max_{1 \leq k \leq m} |x^k|, \end{aligned}$$

and the equality $\|x\|_\infty \leq \|x\|_2 \leq m\|x\|_\infty$ follows immediately.

The proof of Theorem 3 is very long and complicated. It involves proving a long string of inequalities by use of Rayleigh–Ritz methods and Gershgorin's theorem. Details are found in [7] when the boundary conditions are given and in [6] when they are arbitrary.

For completeness, the inequalities are

$$\begin{aligned} C_1 h\|E_h\|_2^2 &\overset{\textcircled{1}}{=} C_1 \frac{h^2}{h} E_h^T E_h \overset{\textcircled{2}}{\le} C_2 E_h^T \frac{1}{h} J_h^m E_h \\ &\overset{\textcircled{3}}{=} C_2 \int_a^b e'^T_h e'_h dt \overset{\textcircled{4}}{=} C_2(e_h, e_h) \overset{\textcircled{5}}{\le} C_3 H_0(e_h) \overset{\textcircled{6}}{\le} \bar{\bar{H}}_h(e_h) \\ &\overset{\textcircled{7}}{=} E_h^T M^h E_h \overset{\textcircled{8}}{=} h^3(E_h^T Q_h + Q_h^T E_h) + O(h^5) \\ &\overset{\textcircled{9}}{=} C_4 h^3 \|E_h\|_2 \|Q_h\|_2 \overset{\textcircled{10}}{\le} C_5 h^{5/2} \|E_h\|_2. \end{aligned}$$

We will not define the symbols, to protect the innocent, but "T" denotes transpose. In the quadratic case, $\textcircled{8}$ can be omitted. The result $M^h E_h$ is obtained by using Theorem 2 and the definition of $y_h(a_k)$.

Thus $\|E_h\|_2 \le Ch^{3/2}$, and the final result can be obtained as described in [7] or [6] or by the results for sparse symmetric matrix.

5. The general problem. The purpose of this section is to briefly describe the results for the more general, delay calculus of variations problem, as given in [1]. We will show that these results lead to those in section 2 and motivate how the numerical algorithm in the more general case is obtained.

Thus our basic problem is

$$\min I(y) = \int_a^b f(t, y(t), y'(t), y_\tau(t), y'_\tau(t))dt, \tag{5.1}$$

subject to the terminal conditions given by

$$y(t) = \alpha(t), \quad t \in [a-\tau, a], \tag{5.2}$$

$$y(b) = y_b, \tag{5.3}$$

where $y_\tau(t)$, etc., is given above. It is shown in [1] that the above problem leads to the differential equations

$$\frac{d}{dt} f_{y'}(t) + \frac{d}{dt} f_{y'_\tau}(t+\tau) = f_y(t) + f_{y_\tau}(t+\tau), \quad a \le t \le b-\tau, \tag{5.4}$$

$$\frac{d}{dt} f_{y'}(t) = f_y(t), \quad b-\tau \le t \le b, \tag{5.5}$$

and the corner condition

$$f_{y'}((b-\tau)^-) - f_{y'}((b-\tau)^+) + f_{y'_\tau}(t+\tau)\big|_{t=(b-\tau)^-} = 0. \tag{5.6}$$

Recalling that our specialized quadratic problem was

$$J(y) = \frac{1}{2}\int_a^b [r(t)y'^2(t) + \bar{r}_\tau(t)y'^2_\tau(t) + p(t)y^2(t) + \bar{p}_\tau(t)y_\tau^2(t)]dt, \tag{1.1}$$

comparing (5.1) and (1.1), we observe that for the quadratic problem

$$f(t,y(t),y'(t),y_\tau(t),y'_\tau(t))=\frac{1}{2}[r(t)y'^2(t)+p(t)y^2(t)+\bar{r}(t)y'^2_\tau(t)+\bar{p}(t)y^2_\tau(t)]. \tag{5.7}$$

Substituting (5.7) into (5.5)–(1.1), we obtain the differential equations

$$\frac{d}{dt}[R(t)y'(t)]=P(t)y(t),\qquad a\le t\le b-\tau, \tag{5.8}$$

$$\frac{d}{dt}\left[r(t)y'(t)\right]=p(t)y(t),\qquad b-\tau\le t\le b, \tag{5.9}$$

and the corner conditions

$$R(b^--\tau)y'(b^--\tau)-r(b^+-\tau)y'(b^+-\tau)=0. \tag{5.10}$$

These results are those of Theorem 1 in slightly different notation.

A numerical algorithm for (5.1)–(5.3) which also results in an error bound of order $O(h^2)$ will be presented in a later paper.

6. Example. In this section, we consider a numerical example to describe the algorithm developed in the paper, and we show that the numerical results agree with the error bound given in (4.17).

The problem is described as follows: Find the minimum of the functional

$$I(y)=\frac{1}{2}\int_0^\pi((y')^2+(y'_\tau)^2-y^2-y_\tau^2)dt \tag{6.1}$$

such that

$$y(t)=0,\qquad t\in[-\pi/4,0], \tag{6.2}$$

and

$$y(\pi)=1, \tag{6.3}$$

where $\tau=\pi/4$. In this example, we have $r(t)=\bar{r}(t)=-p(t)=-\bar{p}(t)=1$. Following the approach presented in Theorem 1, it can be demonstrated that the solution of above problem is

$$y(t)=\begin{cases}-2\sin(t), & 0\le t<3\pi/4,\\ -3\sin(t)-\cos(t), & 3\pi/4<t\le\pi.\end{cases} \tag{6.4}$$

For computational purposes, the time domain is divided into n equal divisions. Substituting the values of $r(t)$, $\bar{r}(t)$, $p(t)$, and $\bar{p}(t)$ into (3.13)–(3.15), we arrive at the following algorithm for this example.

For $0\le t_k<3\pi/4$,

$$\frac{1}{h}(-y_{k-1}+2y_k-y_{k+1})+\frac{h}{4}(y_{k-1}+2y_k+y_{k+1})=0. \tag{6.5}$$

For $3\pi/4<t_k\le\pi$,

$$\frac{1}{2h}(-y_{k-1}+2y_k-y_{k+1})+\frac{h}{8}(y_{k-1}+2y_k+y_{k+1})=0. \tag{6.6}$$

For $t_k = 3\pi/4$,

$$\frac{1}{2h}(-2y_{k-1} + 3y_k - y_{k+1}) + \frac{h}{8}(2y_{k-1} + 3y_k + y_{k+1}) = 0. \tag{6.7}$$

Here $h = \pi/n$, and y_k, $k = 1, \ldots, n-1$, are the computed values of $y(t)$ at $t = t_k = kh$. Note that $y_0 = 0$ and $y_n = 1$.

For numerical simulation, we take $n = 8$, 16, and 32. The table below shows the analytical results and the numerical errors $e_{h1} = e_h(h = 1/8)$, $e_{h2} = e_h(h = 1/16)$, and $e_{h3} = e_h(h = 1/32)$ for $n = 8$, 16, and 32 and the error ratios $r_1 = e_{h1}/e_{h2}$ and $r_2 = e_{h2}/e_{h3}$. From the table it is clear that as the time interval is reduced from h to $h/2$, the error reduces by a factor of 4. Thus the numerical results also justify our claim that our algorithm leads to an $O(h^2)$ global pointwise error.

t	$y(t)$	$e_{h1}(h = 1/8)$	$e_{h2}(h = 1/16)$	$e_{h3}(h = 1/32)$	$r_1 = e_{h1}/e_{h2}$	$r_2 = e_{h2}/e_{h3}$
0	0	0	0	0		
0.0982	-0.196			0.000665719		
0.1963	-0.3902		0.005347356	0.001325356		4.035
0.2945	-0.5806			0.001976645		
0.3927	-0.7654	0.043607135	0.010538135	0.002613135	4.138	4.033
0.4909	-0.9428			0.003229526		
0.589	-1.1111		0.015419534	0.003819534		4.037
0.6872	-1.2688			0.004383432		
0.7854	-1.4142	0.082106438	0.019846438	0.004916438	4.137	4.037
0.8836	-1.546			0.005419093		
0.9817	-1.6629		0.023680775	0.005870775		4.034
1.0799	-1.7638			0.006287471		
1.1781	-1.8478	0.110940935	0.026820935	0.006650935	4.136	4.033
1.2763	-1.9139			0.006969329		
1.3744	-1.9616		0.029149439	0.007229439		4.032
1.4726	-1.9904			0.007440547		
1.5708	-2	0.12659	0.03061	0.00759	4.136	4.033
1.669	-1.9904			0.007680547		
1.7671	-1.9616		0.031139439	0.007719439		4.034
1.8653	-1.9139			0.007699329		
1.9635	-1.8478	0.126990935	0.030720935	0.007620935	4.134	4.031
2.0617	-1.7638			0.007477471		
2.1598	-1.6629		0.029340775	0.007280775		4.03
2.2580	-1.5460			0.007019093		
2.3562	-1.4142	0.111796438	0.027046438	0.006706438	4.133	4.033
2.4544	-1.1302			0.006000601		
2.5525	-0.8352		0.021167913	0.005249913		4.032
2.6507	-0.5323			0.004452054		
2.7489	-0.2242	0.060212235	0.014577235	0.003616235	4.131	4.031
2.8471	0.0861			0.002746204		
2.9452	0.3955		0.007455314	0.001849314		4.031
3.0434	0.7011			0.000932306		
3.1416	1	0	0	0		

7. Conclusions. A numerical scheme has been presented for time delay quadratic problems in the calculus of variations. It has been demonstrated that the algorithm has a global a priori error of $O(h^2)$, where h is the distance between any two successive node points. As opposed to solving time delay differential equations, our methods require weaker smoothness conditions, and they avoid the need for satisfying the transversality and corner conditions explicitly.

A numerical example has been presented to demonstrate the numerical scheme and to verify the a priori error conditions. Numerical results agree with the proposed global error bound. Although numerical results have been presented in only one example, the algorithm can be applied to a large class of time delay quadratic problems.

REFERENCES

[1] O. P. Agrawal and J. Gregory, *The complete solution for constrained delay problems in the calculus of variations by unconstrained methods*, J. Math. Anal. Appl., 218 (1998), pp. 127–135.

[2] O. P. Agrawal and J. Gregory, *A complete solution for optimal control delay problems*, Util. Math., to appear.

[3] O. P. Agrawal, J. Gregory, and K. Pericak-Spector, *A Bliss-type multiplier rule for constrained variational problems with time delay*, J. Math. Anal. Appl., 210 (1997), pp. 702–711.

[4] S. A. Bianco, I. S. Sadek, and M. T. Kambule, *Optimal control of a class of time-delayed distributed systems by orthogonal functions*, J. Franklin Inst. B, 335 (1998), pp. 1477–1492.

[5] D. H. Eller and J. K. Aggarwal, *Optimal control of linear time-delay systems*, IEEE Trans. Automat. Control, 14 (1969), pp. 678–687.

[6] J. Gregory and C. Lin, *Discrete variable methods for the m-dependent variable nonlinear, extremal problem in the calculus of variations* II, SIAM J. Numer. Anal., 30 (1993), pp. 871–881.

[7] J. Gregory and R. S. Wang, *Discrete variable methods for the m-dependent variable, nonlinear extremal problem in the calculus of variations*, SIAM J. Numer. Anal., 27 (1990), pp. 470–487.

[8] C. Hwang and M. Chen, *A direct approach using the shifted Legendre series expansion for near optimal control of linear time-varying systems with multiple state and control delays*, Internat. J. Control, 43 (1986), pp. 1673–1692.

[9] C. Hwang and Y. P. Shih, *Optimal control of delay systems via block pulse functions*, J. Optim. Theory Appl., 45 (1985), pp. 101–112.

[10] K. Kaji and K. H. Wong, *Nonlinearly constrained time-delay optimal control problems*, J. Optim. Theory Appl., 82 (1994), pp. 295–313.

[11] T. T. Lee and S. C. Tsay, *Approximate solutions for linear time-delay systems via the Padé approximation and orthogonal polynomials expansions*, Control Theory Adv. Tech., 3 (1987), pp. 111–128.

[12] K. L. Teo, K. H. Wong, and D. J. Clements, *Optimal control computation for linear time-lag systems with linear terminal constraints*, J. Optim. Theory Appl., 44 (1984), pp. 509–526.

[13] K. L. Teo, K. H. Wong, and D. J. Clements, *A feasible directions algorithm for time-lag optimal control problems with control and terminal inequality constraints*, J. Optim. Theory Appl., 46 (1985), pp. 295–317.

[14] S. C. Tsay, I. L. Wu, and T. T. Lee, *Optimal control of linear time-delay systems via general orthogonal polynomials*, Internat. J. Systems Sci., 19 (1988), pp. 365–376.

[15] K. H. Wong, D. J. Clements, and K. L. Teo, *Optimal control computation for nonlinear time-lag systems*, J. Optim. Theory Appl., 47 (1985), pp. 91–107.

SIAM J. NUMER. ANAL.
Vol. 41, No. 5, pp. 1785–1823

ADAPTIVE WAVELET SCHEMES FOR NONLINEAR VARIATIONAL PROBLEMS[*]

ALBERT COHEN[†], WOLFGANG DAHMEN[‡], AND RONALD DEVORE[§]

Abstract. We develop and analyze wavelet based adaptive schemes for nonlinear variational problems. We derive estimates for convergence rates and corresponding work counts that turn out to be asymptotically optimal. Our approach is based on a new paradigm that has been put forward recently for a class of linear problems. The original problem is transformed first into an equivalent one which is well posed in the Euclidean metric ℓ_2. Then conceptually one seeks iteration schemes for the infinite dimensional problem that exhibits at least a fixed error reduction per step. This iteration is then realized approximately through an adaptive application of the involved operators with suitable dynamically updated accuracy tolerances. The main conceptual ingredients center around nonlinear tree approximation and the sparse evaluation of nonlinear mappings of wavelet expansions. We prove asymptotically optimal complexity for adaptive realizations of first order iterations and of Newton's method.

Key words. variational problems, wavelet representations, semilinear equations, mapping properties, gradient iteration, convergence rates, adaptive application of operators, sparse evaluation of nonlinear mappings of wavelet expansions, tree approximation, Newton's scheme

AMS subject classifications. 65J15, 65N12, 65N15, 35A15, 35A35, 35J60, 41A60, 46A45, 47H17

DOI. 10.1137/S0036142902412269

1. Introduction.

1.1. Background and objectives.

Adaptive wavelet schemes for numerically solving a wide class of variational problems have been recently studied in [8, 9] from the perspective of asymptotic estimates for convergence rates and corresponding work counts. The problems covered by that analysis include elliptic boundary integral equations and elliptic boundary value problems but also indefinite problems of elliptic type such as the Stokes problem. Two requirements were essential in this context: (i) the variational problem induces an operator $\mathcal{L}$ that is an isomorphism from some Hilbert space $\mathcal{H}$ onto its dual; (ii) this Hilbert space permits a wavelet characterization; i.e., the $\mathcal{H}$-norm of an element is equivalent to a weighted ℓ_2-norm of its wavelet coefficients. It could then be shown that certain adaptive schemes exhibit an *asymptotically optimal accuracy/work balance* within a certain range of convergence rates depending on the choice of wavelet bases. The precise meaning of this statement is explained in the Meta-Theorem below. To our knowledge for the above range of linear problems such complexity estimates have been established so far only for wavelet methods. Just recently, a similar result was proved for adaptive finite element methods for Laplace's

[*] Received by the editors July 26, 2002; accepted for publication (in revised form) May 1, 2003; published electronically October 28, 2003. This work was supported in part by the Office of Naval Research contract N0014-91-J1343, the Army Research Office contract DAAD 19-02-1-0028, the IHP-network (N.HPRN-CT-2002-00286) "Breaking Complexity," by the SFB 401, funded by the German Research Foundation, and by the Alexander von Humboldt Foundation.

http://www.siam.org/journals/sinum/41-5/41226.html

[†] Laboratoire d'Analyse Numerique, Universite Pierre et Marie Curie, 175 Rue du Chevaleret, 75013 Paris, France (cohen@ann.jussieu.fr, http://www.ann.jussieu.fr/∼cohen/).

[‡] Institut für Geometrie und Praktische Mathematik, RWTH Aachen, 52056 Aachen, Germany (dahmen@igpm.rwth-aachen.de, http://www.igpm.rwth-aachen.de/∼dahmen/).

[§] Department of Mathematics, University of South Carolina, Columbia, SC 29208 (devore@math.sc.edu, http://www.math.sc.edu/∼devore/).

equation in two space dimensions [4].

In this paper we wish to explore the convergence rates and the computational complexity of certain new adaptive wavelet schemes for *nonlinear problems* for which no results of the above type seem to be known so far.

Our primary concern here is *not* to develop a specific algorithm for a concrete application. We are rather interested in developing a numerically realizable *new algorithmic paradigm* in a fairly general context of nonlinear problems and in analyzing its principal complexity features. Therefore, the various algorithmic ingredients will at times not be discussed in full detail but only to an extent that clarifies their principal asymptotic complexity.

The new paradigm is based upon the adaptive evaluation of (linear and nonlinear) operators in the course of an ideal iteration for the *infinite dimensional* problem formulated in the wavelet coordinate domain. Such perturbed iterations will lead to an algorithm **SOLVE** that (with a proper initialization) produces for any target accuracy ϵ a finitely supported vector of coefficients $\bar{\mathbf{u}}(\epsilon)$ that approximates the array of wavelet coefficients of the exact solution (of the underlying variational problem) in ℓ_2 with accuracy ϵ. The choice of wavelet basis will then imply that the corresponding finite expansion approximates the exact solution with accuracy $C\epsilon$ in the energy norm, where C depends only on the wavelet basis. In order to identify the essential mechanisms governing such schemes, we will consider nonlinear variational problems on various levels of generality. The results will be purely asymptotic in nature. They reveal asymptotically optimal work/accuracy balances interrelating the achieved target accuracy with the required computational work and associated adaptively generated number of degrees of freedom. More precisely, we shall prove results of the following type.

META-THEOREM. *If the exact solution can be approximated as a linear combination of N wavelets (subject only to certain tree restrictions on the distribution of active coefficients) to accuracy of order N^{-s} (for a certain range of s), then the support of the output $\bar{\mathbf{u}}(\epsilon)$ of* **SOLVE** *for target accuracy ϵ grows at most as $\epsilon^{-1/s}$, uniformly in ϵ, and the computational complexity also stays proportional to the support size. In this sense, the scheme tracks the exact solution at asymptotically minimal cost.*

Note that the above-mentioned *tree restriction* on the permitted distribution of active coefficients is the analogue of locally refined meshes in the finite element context.

We shall outline now how we approach results of the above type.

1.2. The basic paradigm. The *classical* approach to numerically solving (linear and nonlinear) variational problems is concerned with the following issues:

(c1) well-posedness of the given variational problem;
(c2) discretization of the infinite dimensional problem so as to obtain a finite system of algebraic equations;
(c3) well-posedness of the finite system of equations and error analysis;
(c4) numerical solution of the finite system of equations.

It is important to note that (c1) is often hidden in the analysis and that (c3) is, in general, *not* a direct consequence of (c1). Typical examples even in the linear case are *saddle point problems.* It is well known that, for Galerkin discretizations to be stable, the trial spaces for the different solution components have to satisfy a certain compatibility condition (Ladyšhenskaya–Babŭska–Brezzi (LBB)-condition). For nonlinear problems one can often establish only *local* uniqueness of solutions so that some care is required to ensure that the discrete problems approximate the correct solution branch. Thus the discrete problems do not necessarily inherit the “nice properties” of the original infinite dimensional problem. Depending on the choice of the discretization, one might introduce "new difficulties." The typical obstructions encountered in (c4) are the *large size* of the discrete systems and possible *ill-conditioning.* The latter issue interferes with the need to resort to iterative solvers, due to the size and sparsity of the systems. Attempts to reduce computational complexity are often based on adaptive and hence possibly economic discretizations. A reliable control of adaptive refinements, however, depends usually in a sensitive way on the particular type of the problem, and rigorous complexity estimates are generally not available yet.

A *new paradigm* has been explored in [9] for *linear variational problems.* It aims at closely intertwining the analysis–discretization–solution process. The basic steps there read as follows:

(n1) well-posedness of the given variational problem;

(n2) transformation of the infinite dimensional problem into an *equivalent* problem in ℓ_2 which is *well posed* in the Euclidean metric;

(n3) the derivation of an iterative scheme for the infinite dimensional ℓ_2-problem that exhibits a fixed error reduction per iteration step;

(n4) numerical realization of the iterative scheme by an *adaptive application* of the involved infinite dimensional operators within some finite dynamically updated accuracy tolerances.

Thus the starting point (n1) is the same as (c1), although it takes a somewhat more exposed and explicit role in the new setting, as will be explained later. The main difference is that one aims at staying as long as possible with the infinite dimensional problem, which one hopes is given in a favorable format. Of course, it remains to see in each concrete case how to exploit (n2) in order to guarantee a fixed error reduction in (n3). We shall present several strategies regarding this task. Only at the very end, when it comes to applying the operators in the ideal iteration scheme (n4), does one enter the finite dimensional realm. However, the finite number of degrees of freedom is determined at each stage by the adaptive application of the operator so that at *no* stage is any specific trial space fixed. Roughly speaking, the "nice properties" of the infinite dimensional problem are preserved through adaptive evaluations. In fact, one can show that thereby compatibility conditions like the LBB-condition indeed become void [9, 13].

The main goal of the present paper is to show how to carry over this paradigm, already existing for linear problems, to the nonlinear setting. On a theoretical level, one then encounters three major issues, namely,

(a) the choice of tolerances in (n4) to ensure that the perturbed iteration converges to the correct solution;

(b) the design of economic approximate application schemes for the possibly nonlinear infinite dimensional operators;

(c) estimating the complexity of the scheme.

Here (a) means that any given *target accuracy* ϵ is achieved after finitely many steps. (b) is the most crucial part and will be discussed in detail in the course of the paper. Clearly (b) is closely related to (c). As in [8, 9, 13], we will measure complexity by the *number of adaptively generated degrees of freedom* $N = N(\epsilon)$ required by the adaptive scheme to achieve the target accuracy ϵ and the corresponding number of floating point operations (which, of course, is aimed at staying proportional to $N(\epsilon)$). Estimating the asymptotic *work/accuracy balance* $N(\epsilon) \leftrightarrow \epsilon$ will be a central theme in the subsequent developments. This part differs significantly from the classical error analysis for finite element methods and relies on concepts from *harmonic analysis* and *nonlinear approximation*.

Of course, on a practical level one will encounter in each concrete case further obstacles concerning quantitative information about constants and initial guesses. We shall discuss variational problems on a different level of generality in order to indicate possible strategies of acquiring such information or to identify those issues that require additional work.

Finally, a comment on (n3) is in order. Aiming at a fixed error reduction per iteration step means that one is content with a *first order* scheme. So why not go for faster iteration schemes? The answer to this question is not completely clear. Indeed, a higher order method may not automatically win for the following reason. Usually a higher order method is more costly in function evaluations. In the present context this means, according to (n4), it is more costly in the adaptive application of the full infinite dimensional operators within some dynamically updated accuracy tolerance. Preserving the higher order of the ideal iteration also in its perturbed form in connection with the higher demands of function evaluations may very well increase the cost of each iteration step so as to offset the potential gain of a better error reduction. So with regard to the objective of reaching a target accuracy at possibly low overall computational cost, the role of higher order schemes remains unclear. In fact, it will be seen that *asymptotic optimality* can indeed be achieved already with simple *first order outer iterations*. Nevertheless, we shall show that it is also possible to retain second order convergence of the adaptive version of Newton's scheme so as to arrive at an overall scheme with asymptotically optimal solution complexity, which may offer quantitative advantages over the first order versions.

1.3. Organization of material. The paper is organized as follows. In section 2 we describe (n1), (n2), and (n3) for a general setting that will host all subsequent specifications. In section 3 we distinguish several classes of variational problems to which the subsequent developments will refer frequently, namely, (L) *linear* problems, (SL) semilinear elliptic problems, and (GNL) more general nonlinear problems where we have to assume the existence of locally unique solutions. In section 4 we formulate the prototype of an adaptive perturbed first order iteration which is based on two main ingredients, namely, approximate *residual evaluations* and a certain *coarsening scheme*. In particular, the residual approximations involve the *adaptive* application of linear or nonlinear (infinite dimensional) operators. Assuming at this stage that these ingredients are indeed available, we address for the most general setting first only issue (a) to clarify for which choice of dynamically updated accuracy tolerances is convergence guaranteed. The remaining sections will be devoted to issues (b) and (c) for the problem types (L), (SL), and (GNL).

In section 5 we review briefly concrete realizations of these ingredients for the linear case (L) and indicate the concepts needed for their complexity analysis. This serves two purposes. First, these results will be used in the last section in connection with Newton iterations. Second, they motivate our treatment of the nonlinear case. In section 6 we introduce some new concepts needed to deal with nonlinear problems. They center upon *tree approximation* and related coarsening techniques. This enables us to formulate the notion of s^*-sparsity as the key criterion for controlling the complexity of the adaptive schemes in the nonlinear case. Drawing on several results from [10], we develop in section 7 adaptive evaluation schemes that are proven to be s^*-sparse and thus lead to asymptotically optimal results in the sense of the above Meta-Theorem. To our knowledge these are the first convergence and complexity estimates for adaptive solvers for nonlinear problems. Finally, in section

8 we develop an adaptive Newton scheme and analyze its complexity. It differs in essential ways from the schemes discussed in the previous sections which are based on first order iterations. In particular, we show that the quadratic convergence of the outer iteration can in some sense be preserved in the adaptive context.

2. The setting.

We describe now the setting for which the above paradigm will be discussed.

2.1. The general problem format.

The variational problems mentioned in step (n1) above will always have the following format. Let $\mathcal{H}$ be a Hilbert space with norm $\|\cdot\|_{\mathcal{H}}$, and let $\mathcal{H}'$ denote its dual endowed with the norm

$$\|v\|_{\mathcal{H}'} := \sup_{w\in\mathcal{H}} \frac{\langle w, v\rangle}{\|w\|_{\mathcal{H}}},$$

where $\langle\cdot,\cdot\rangle$ is the dual pairing between $\mathcal{H}$ and $\mathcal{H}'$ (with respect to L_2 as the pivot space). The Hilbert space $\mathcal{H}$ will always refer to a bounded domain Ω with spatial dimension d. Suppose that $f \in \mathcal{H}'$ and

$$F : \mathcal{H} \to \mathcal{H}' \tag{2.1}$$

is a (possibly nonlinear) mapping. We consider the numerical solution of the problem: Find $u \in \mathcal{H}$ such that

$$\langle v, F(u) - f\rangle =: \langle v, R(u)\rangle = 0 \quad \forall v \in \mathcal{H}. \tag{2.2}$$

The objective in (n1) is the identification of a suitable space $\mathcal{H}$ so that (2.2) is well posed in the following sense. Recall that the Frechét derivative $DR(z) = DF(z)$ is a mapping from $\mathcal{H}$ to $\mathcal{H}'$, defined by the duality

$$\langle v, DR(z)w\rangle = \lim_{h\to 0} \frac{1}{h}\langle v, R(z+hw) - R(z)\rangle. \tag{2.3}$$

The problem (2.2) is called *well posed* if F has the following properties:

A1. F possesses a continuous Frechét derivative; i.e., $R \in C^1(\mathcal{H}, \mathcal{H}')$ as a mapping $v \mapsto R(v)$.

A2. There exists a solution $u \in \mathcal{H}$ to (2.2), and in addition to (2.1) the Frechét derivative DF of F at v in some neighborhood $\mathcal{U}$ of u is an isomorphism from $\mathcal{H}$ onto $\mathcal{H}'$; i.e., for $v \in \mathcal{U}$ there exist positive finite constants $c_{v,F}, C_{v,F}$ such that

$$c_{v,F}\|w\|_{\mathcal{H}} \le \|DF(v)w\|_{\mathcal{H}'} \le C_{v,F}\|w\|_{\mathcal{H}} \quad \forall\, w \in \mathcal{H},\ v \in \mathcal{U}. \tag{2.4}$$

Clearly A2 ensures that the solution u is locally unique.

2.2. Wavelet coordinates and an equivalent ℓ_2-problem.

The transformations for (n2) will be based on suitable wavelet bases. For a detailed discussion of such bases, we refer the reader to the literature (see, e.g., [5, 6, 14, 11]) and collect here only the relevant facts. A *wavelet basis* $\Psi = \{\psi_\lambda : \lambda \in \mathcal{J}\} \subset \mathcal{H}$ has the following properties: The indices $\lambda \in \mathcal{J}$ encode typical information about the wavelet ψ_λ, namely, its type, its location $k(\lambda)$, and its scale $|\lambda|$.

We shall now explain the meaning of "suitable" in the present context. We will always assume that the wavelets have compact support $S_\lambda := \operatorname{supp}\psi_\lambda$, $\lambda \in \mathcal{J}$, which scales as $\operatorname{diam}(S_\lambda) \sim 2^{-|\lambda|}$.

Furthermore, aside from finitely many functions $\psi_\lambda, \lambda \in \mathcal{J}_\phi \subset \mathcal{J}$, $|\lambda| = j_0$, representing the coarsest scale j_0, the wavelets ψ_λ, $\lambda \in \mathcal{J}\backslash\mathcal{J}_\phi$, have *vanishing moments* of some order $m \in \mathbb{N}$; i.e., these wavelets are orthogonal to all polynomials of order m.

Finally, each $v \in \mathcal{H}$ has a unique expansion $\sum_{\lambda\in\mathcal{J}} v_\lambda\psi_\lambda$ such that

$$c_1\|\mathbf{v}\|_{\ell_2(\mathcal{J})} \le \left\|\sum_{\lambda\in\mathcal{J}} v_\lambda\psi_\lambda\right\|_{\mathcal{H}} \le C_1\|\mathbf{v}\|_{\ell_2(\mathcal{J})} \tag{2.5}$$

holds for some positive constants c_1, C_1; i.e., Ψ forms a *Riesz basis* for $\mathcal{H}$. Note that (unlike the quoted references) we have normalized the wavelets here in the energy space $\mathcal{H}$ associated with the variational problem (2.1), (2.2); i.e., $\|\psi_\lambda\|_{\mathcal{H}} = 1$, $\lambda \in \mathcal{J}$. Again such bases are known whenever $\mathcal{H}$ is a product of Sobolev spaces (or closed subspaces of Sobolev spaces, determined, e.g., by homogeneous boundary conditions or vanishing integral means).

In the following, we will always use boldface notation $\mathbf{v}$ to denote the wavelet coefficients of a given function $v \in \mathcal{H}$ with respect to the basis Ψ (and analogously for $u, w \in \mathcal{H}$).

Next note that by duality (2.5) implies

$$C_1^{-1}\|(\langle w, \psi_\lambda\rangle)_{\lambda\in\mathcal{J}}\|_{\ell_2(\mathcal{J})} \le \|w\|_{\mathcal{H}'} \le c_1^{-1}\|(\langle w, \psi_\lambda\rangle)_{\lambda\in\mathcal{J}}\|_{\ell_2(\mathcal{J})}. \tag{2.6}$$

We can now transform (2.2) into wavelet coordinates. Defining

$$\mathbf{R}(\mathbf{v}) := (\langle\psi_\lambda, R(v)\rangle : \lambda \in \mathcal{J}) \quad \text{whenever} \quad v = \sum_{\lambda\in\mathcal{J}} v_\lambda\psi_\lambda, \tag{2.7}$$

the original problem (2.2) is obviously equivalent to finding $\mathbf{u} \in \ell_2(\mathcal{J})$ so that

$$\mathbf{R}(\mathbf{u}) = \mathbf{0}. \tag{2.8}$$

Now note that the Jacobian $D\mathbf{R}(\mathbf{v}) = D\mathbf{F}(v)$ is given by

$$D\mathbf{R}(\mathbf{v}) = (\langle\psi_\lambda, DR(v)\psi_\nu\rangle)_{\lambda,\nu\in\mathcal{J}}\,, \tag{2.9}$$

where again $DR = DF$ is the Frechét derivative of the mapping R. Combining the norm equivalences (2.5), (2.6) with the mapping property (2.4) shows in what sense (2.8) is now well posed in ℓ_2; see, e.g., [9] for a proof.

Remark 2.1. Under the above assumptions A1 and A2, one has for any $v = \sum_{\lambda\in\mathcal{J}} v_\lambda\psi_\lambda \in \mathcal{U}$

$$c_1^2 c_{v,F}\|\mathbf{w}\|_{\ell_2(\mathcal{J})} \le \|D\mathbf{F}(\mathbf{v})\mathbf{w}\|_{\ell_2(\mathcal{J})} \le C_1^2 C_{v,F}\|\mathbf{w}\|_{\ell_2(\mathcal{J})}, \quad \mathbf{w} \in \ell_2(\mathcal{J}). \tag{2.10}$$

2.3. The basic iteration. According to (n3), we wish to devise an iterative scheme for the problem (2.8) such that each step reduces the current error at least by a fixed rate $\rho < 1$. The schemes we shall consider will have the form

$$\mathbf{u}^{n+1} = \mathbf{u}^n - \mathbf{B}_n\mathbf{R}(\mathbf{u}^n), \tag{2.11}$$

where the (infinite, possibly stage dependent) matrix $\mathbf{B}_n$ is yet to be chosen. For instance, $\mathbf{B}_n = \alpha\mathbf{I}$ corresponds to a fixed point or Richardson iteration, while for $\mathbf{B}_n := D\mathbf{R}(\mathbf{u}^n)^{-1}$ (2.11) becomes Newton's method.

We proceed now to discuss several instances of this setting.

3. The scope of reference problems. We shall address the variational problem (2.2) for the following different levels of generality:

(SL) semilinear elliptic boundary value problems, covering the case (L) of *linear* problems as a special case;

(GNL) general nonlinear problems.

The explicit discussion of (SL) will serve several purposes. First, this class is specific enough to permit a complete complexity analysis for a *globally convergent* adaptive scheme. In particular, we shall be able to obtain in this case concrete bounds on initial guesses or how to find a suitable damping parameter in $\mathbf{B}_n = \alpha\mathbf{I}$. Second, this class is a model representative for the interplay between a linear (diffusion) operator and a nonlinear part. On one hand, it covers linear problems (L) as special cases. Briefly reviewing the essential features of linear problems in this context comes in handy for three reasons. It provides a guideline for the treatment of nonlinear problems. It is a necessary prerequisite for the later discussion of Newton's method. Most importantly, in the presence of a nonlinearity, the complexity analysis for the linear case has to be modified in order to treat problems where both linear and nonlinear operators are involved. It will be instructive to see the conceptual distinctions and how the treatment of nonlinear problems builds on ingredients from the linear case.

Finally, in (GNL) we relax our assumptions on the structure of R to a great extent. We pay for this by making stronger assumptions on initial guesses and being content with *locally convergent first order* iterations on the infinite dimensional level.

We shall exemplify step (n3) for all three cases (L), (SL), and (GNL) in this order. Except for the last section, this will be based on first order iteration schemes for the underlying infinite dimensional problem. It will be seen along the way that it then suffices to employ stationary "preconditioners" $\mathbf{B}_n = \mathbf{B}$ to obtain asymptotically optimal complexity estimates (although more flexible nonstationary choices may well result in quantitative improvements in practical realizations). The use of truly nonstationary $\mathbf{B}_n$ will be necessary only in connection with Newton's method in section 8.

3.1. Semilinear (SL) and linear (L) elliptic problems. Suppose that $a(\cdot,\cdot)$ is a continuous bilinear form on a Hilbert space $\mathcal{H}$ endowed with the norm $\|\cdot\|_{\mathcal{H}}$, which is $\mathcal{H}$-elliptic; i.e., there exist positive constants c, C such that

$$c\|v\|_{\mathcal{H}}^2 \leq a(v,v), \quad a(v,w) \leq C\|v\|_{\mathcal{H}}\|w\|_{\mathcal{H}} \quad \forall\ v,w \in \mathcal{H}. \tag{3.1}$$

The simplest example is

$$a(v,u) := \langle \nabla v, \nabla u\rangle + \kappa\langle v,u\rangle, \quad \kappa \geq 0, \ \ \langle v,w\rangle = \int_\Omega vw, \tag{3.2}$$

and $\mathcal{H} = H_0^1(\Omega)$ (the space of functions with first order weak derivatives in L_2 whose traces vanish on the boundary $\Gamma = \partial\Omega$) endowed with the norm $\|v\|_{\mathcal{H}}^2 := \|\nabla v\|_{L_2(\Omega)}^2 + \kappa\|v\|_{L_2(\Omega)}^2$.

In principle, the subsequent analysis will also cover elliptic integral operators with positive order such as the hypersingular operator.

To introduce a nonlinearity, we suppose that $G : \mathbb{R} \to \mathbb{R}$ is a function with the following property:

P1. The mapping $v \mapsto G(v)$ takes $\mathcal{H}$ into its dual $\mathcal{H}'$ and is *stable* in the sense that

$$\|G(u) - G(v)\|_{\mathcal{H}'} \le C_G(\max\{\|u\|_{\mathcal{H}}, \|v\|_{\mathcal{H}}\})\|u - v\|_{\mathcal{H}}, \quad u, v \in \mathcal{H}, \tag{3.3}$$

where $t \to C_G(t)$ is a nondecreasing function of t.

The problem: Given $f \in \mathcal{H}'$, find $u \in \mathcal{H}$ such that

$$\langle v, F(u)\rangle := a(v, u) + \langle v, G(u)\rangle = \langle v, f\rangle \quad \forall \ \ v \in \mathcal{H} \tag{3.4}$$

is of the form (2.2) with $R(v) = F(v) - f$.

Remark 3.1. If we assume in addition to P1 that G is monotone (as in (3.6)), i.e., $(u - v)(G(u) - G(v)) \ge 0$ for $u, v \in \mathbb{R}$, then (3.4) has for every $f \in \mathcal{H}'$ a unique solution $u \in \mathcal{H}$. Moreover, the problem is well posed in the sense of (2.4) with constants $c_{v,F} := c$, $C_{v,F} := C + C_G(\|v\|_{\mathcal{H}})$, where c, C are the constants from (3.1) and $C_G(s)$ is the constant from (3.3) in P1.

Proof. The argument follows standard lines. Under the above assumptions it is easy to show that the operator F, defined by (3.4), is also monotone and coercive. One can then invoke the Browder–Minty theorem (see, e.g., [22, Theorem 9.45]) to conclude existence, while the strict monotonicity guaranteed by the quadratic part also ensures uniqueness. To confirm the validity of (2.4) with the above constants, let $\mathcal{A}$ be the linear operator defined by $\langle w, \mathcal{A}v\rangle = a(w, v)$, for all $w, v \in \mathcal{H}$, and note that, in view of (3.1),

$$c\|v\|_{\mathcal{H}} \le \|\mathcal{A}v\|_{\mathcal{H}'} \le C\|v\|_{\mathcal{H}}, \quad v \in \mathcal{H}, \tag{3.5}$$

with c, C from (3.1). Since we have $DF(v)w = \mathcal{A}w + G'(v)w$, the assertion follows easily from (3.5), P1, and the monotonicity of G. □

Remark 3.2. Alternatively one can argue that, under the above assumptions, G is of potential type so that (3.4) is the Euler equation of a convex minimization problem with a strictly convex functional; see, e.g., [24, Proposition 42.6].

As an example, it is not hard to verify that the weak formulation of the boundary value problem

$$-\Delta u + u^3 = f \quad \text{in} \ \ \Omega, \quad u = 0 \ \ \text{on} \ \ \partial\Omega, \tag{3.6}$$

is of the form (3.4), where for $\mathcal{H} = H_0^1(\Omega)$ the above assumptions hold for $d \le 3$, and that it satisfies the monotonicity assumption of Remark 3.1.

An equivalent ℓ_2-formulation (n2). We turn now to step (n2) in the present setting. In order to rewrite (3.4) in wavelet coordinates, let $\mathbf{A} = (a(\psi_\lambda, \psi_\nu))_{\lambda,\nu\in\mathcal{J}}$ denote the wavelet representation of the operator $\mathcal{A}$, and set $\mathbf{f} = (\langle\psi_\lambda, f\rangle : \lambda \in \mathcal{J})^T$. In addition, define in analogy to (2.7) $\mathbf{G}(\mathbf{v}) := (\langle\psi_\lambda, G(v)\rangle)_{\lambda\in\mathcal{J}}$. Then $u = \sum_{\lambda\in\mathcal{J}} u_\lambda\psi_\lambda$ is the unique solution of (3.4) if and only if $\mathbf{u}$ solves

$$\mathbf{R}(\mathbf{u}) := \mathbf{A}\mathbf{u} + \mathbf{G}(\mathbf{u}) - \mathbf{f} = \mathbf{0}. \tag{3.7}$$

Note that, in view of (2.6), f belongs to $\mathcal{H}'$ if and only if $\mathbf{f} \in \ell_2(\mathcal{J})$. Clearly, by our assumptions on $a(\cdot,\cdot)$, $\mathbf{A}$ is symmetric positive definite. Moreover, it follows from Remarks 2.1 and 3.1 that (2.10) holds for $\mathbf{R}$ in (3.7). In particular, this covers the case $G \equiv 0$, where (2.10) takes the form

$$c_A\|\mathbf{v}\|_{\ell_2(\mathcal{J})} \le \|\mathbf{A}\mathbf{v}\|_{\ell_2(\mathcal{J})} \le C_A\|\mathbf{v}\|_{\ell_2(\mathcal{J})}, \quad \mathbf{v} \in \ell_2(\mathcal{J}), \tag{3.8}$$

with $c_A = c_1^2c$, $C_A = C_1^2C$, and c_1, C_1, c, C from (2.5) and (3.1).

We end this section with the simple observation that monotonicity of G carries over into the discrete setting, namely,

$$(\mathbf{u}-\mathbf{v})^T(\mathbf{G}(\mathbf{u})-\mathbf{G}(\mathbf{v})) \geq 0, \quad \mathbf{u},\mathbf{v} \in \ell_2(\mathcal{J}). \tag{3.9}$$

In fact, denoting by $\tilde{\Psi}$ the dual basis to Ψ, we have by definition of $\mathbf{G}(\mathbf{u})$

$$u - v = \sum_{\lambda\in\mathcal{J}}(u_\lambda - v_\lambda)\psi_\lambda, \quad G(u)-G(v) = \sum_{\lambda\in\mathcal{J}}\langle G(u)-G(v),\psi_\lambda\rangle\tilde{\psi}_\lambda.$$

Thus (3.9) follows from monotonicity of G and biorthogonality. Due to (3.1), F is also monotone so that (3.9) holds also for $\mathbf{F}$.

Gradient iterations (n3). We now address (n3) for the above class of semilinear elliptic problems. The simplest option is to take $\mathbf{B}_n = \alpha\mathbf{I}$, which gives the iteration

$$\mathbf{u}^{n+1} = \mathbf{u}^n - \alpha\mathbf{R}(\mathbf{u}^n), \quad n \in \mathbb{N}_0. \tag{3.10}$$

We have to find some $\alpha > 0$ for which this iteration converges in $\ell_2(\mathcal{J})$ with a guaranteed error reduction $\rho < 1$. To this end, note that, by (2.8), $\mathbf{u}^{n+1}-\mathbf{u} = \mathbf{u}^n - \mathbf{u} - \alpha(\mathbf{R}(\mathbf{u}^n)-\mathbf{R}(\mathbf{u}))$, so that

$$\begin{aligned}\mathbf{u}^{n+1}-\mathbf{u} &= \left(\mathbf{I} - \alpha\int_0^1(\mathbf{A}+D\mathbf{G}(\mathbf{u}+s(\mathbf{u}^n-\mathbf{u})))ds\right)(\mathbf{u}^n-\mathbf{u})\\ &=: (\mathbf{I}-\alpha\mathbf{M}(\mathbf{u}^n,\mathbf{u}))\,(\mathbf{u}^n-\mathbf{u}).\end{aligned} \tag{3.11}$$

By (3.9) and (3.8), the smallest eigenvalue of the matrix $\mathbf{M}(\mathbf{u}^n,\mathbf{u})$ is bounded from below by c_A. To bound the spectral radius of $\mathbf{M}(\mathbf{u}^n,\mathbf{u})$, note that, by (2.5) and (2.6), one has $\|\mathbf{G}(\mathbf{v})-\mathbf{G}(\mathbf{w})\|_{\ell_2(\mathcal{J})} \leq \hat{C}(\max\{\|\mathbf{v}\|_{\ell_2(\mathcal{J})},\|\mathbf{w}\|_{\ell_2(\mathcal{J})}\})\|\mathbf{v}-\mathbf{w}\|_{\ell_2(\mathcal{J})}$, where $\hat{C}(s) := C_1^2C_G(C_1s)$ and $C_1, C_G(s)$ are the constants from (2.5) and (3.3) for G, respectively. It follows from (3.8) and (3.3) in P1 that $\|\mathbf{M}(\mathbf{u}^n,\mathbf{u})\|_{\ell_2(\mathcal{J})\to\ell_2(\mathcal{J})} \leq C_A + \hat{C}(\|\mathbf{u}-\mathbf{u}^n\|_{\ell_2(\mathcal{J})})$. Given some knowledge about the behavior of the stability constant $C_G(s)$ when s increases, we can estimate $\hat{C}(\|\mathbf{u}-\mathbf{u}^n\|_{\ell_2(\mathcal{J})})$ with the aid of an a priori estimate for $\|\mathbf{u}^n-\mathbf{u}\|_{\ell_2(\mathcal{J})}$. To this end, note that again by (2.8) and (3.7), one has for any $\mathbf{v}\in\ell_2(\mathcal{J})$

$$\begin{aligned}\|\mathbf{u}-\mathbf{v}\|_{\ell_2(\mathcal{J})}\|\mathbf{R}(\mathbf{v})\|_{\ell_2(\mathcal{J})} \geq (\mathbf{u}-\mathbf{v})^T(\mathbf{R}(\mathbf{v})-\mathbf{R}(\mathbf{u})) = (\mathbf{u}-\mathbf{v})^T\big(\mathbf{A}(\mathbf{u}-\mathbf{v})\\ + \mathbf{G}(\mathbf{u})-\mathbf{G}(\mathbf{v})\big) \geq c_A\|\mathbf{u}-\mathbf{v}\|^2_{\ell_2(\mathcal{J})},\end{aligned}$$

where we have used (3.8) and (3.9) in the last step. Hence we obtain

$$\|\mathbf{u}-\mathbf{v}\|_{\ell_2(\mathcal{J})} \leq c_A^{-1}\|\mathbf{R}(\mathbf{v})\|_{\ell_2(\mathcal{J})}, \quad \|\mathbf{u}\|_{\ell_2(\mathcal{J})} \leq c_A^{-1}\big(\|\mathbf{G}(\mathbf{0})\|_{\ell_2(\mathcal{J})} + \|\mathbf{f}\|_{\ell_2(\mathcal{J})}\big). \tag{3.12}$$

Thus, for any fixed initial guess $\mathbf{v} = \mathbf{u}^0$, we have a computable bound

$$\|\mathbf{u}^0-\mathbf{u}\|_{\ell_2(\mathcal{J})} \leq c_A^{-1}\big(\|\mathbf{R}(\mathbf{u}^0)\|_{\ell_2(\mathcal{J})}\big) =: \delta_0 \tag{3.13}$$

and therefore a bound for $\hat{C}(\delta_0)$.

Remark 3.3. Given $\mathbf{u}^0$ and δ_0 from (3.13), suppose that $\alpha > 0$ satisfies

$$0 < \alpha < 2/(C_A + \hat{C}(\delta_0)) \tag{3.14}$$

for $\hat{C}(\delta_0)$ defined above so that

$$\rho = \rho(\alpha) := \max\{|1-c_A\alpha|, |1-\alpha(C_A+\hat{C}(\delta_0))|\} < 1. \tag{3.15}$$

Then, denoting by $B_\delta(\mathbf{u})$ the ball of radius δ with center $\mathbf{u}$, we have

$$\sup_{\mathbf{v}\in B_{\delta_0}(\mathbf{u})} \|\mathbf{I}-\alpha\big(\mathbf{A}+D\mathbf{G}(\mathbf{v})\big)\|_{\ell_2(\mathcal{J})\to\ell_2(\mathcal{J})} =: \rho < 1. \tag{3.16}$$

Hence we obtain $\|\mathbf{I}-\alpha\mathbf{M}(\mathbf{u}^n,\mathbf{u})\|_{\ell_2(\mathcal{J})\to\ell_2(\mathcal{J})} \le \rho$, which, in view of (3.11), yields

$$\|\mathbf{u}-\mathbf{u}^n\|_{\ell_2(\mathcal{J})} \le \rho\|\mathbf{u}-\mathbf{u}^{n-1}\|_{\ell_2(\mathcal{J})}, \quad n\in\mathbb{N}. \tag{3.17}$$

Of course, a better error reduction would result from an optimal stage dependent step size α_n. Keeping Remark 3.2 in mind, one can show that (3.7) are the Euler equations of a strictly convex minimization problem on $\ell_2(\mathcal{J})$. For a given $\mathbf{u}^n$ the residual $\mathbf{r} := \mathbf{f}-\mathbf{A}\mathbf{u}^n-\mathbf{G}(\mathbf{u}^n)$ is the direction of the corresponding steepest descent starting from $\mathbf{u}^n$. The minimum along this direction is given by the zero $\alpha=\alpha_n$ of the function $g(\alpha) = \big(\mathbf{f}-\mathbf{A}(\mathbf{u}^n+\alpha\mathbf{r})-\mathbf{G}(\mathbf{u}^n+\alpha\mathbf{r})\big)^T\mathbf{r} = \mathbf{0}$. We shall later discuss ways of approximately evaluating the terms in $g(\alpha)$. Noting that $g'(\alpha) = -\mathbf{r}^T(\mathbf{A}+D\mathbf{G}(\mathbf{u}^n+\alpha\mathbf{r}))\mathbf{r}$, one could think of using such routines for performing a Newton step with the above initial guess for α to solve approximately $g(\alpha)=0$.

We conclude this section with a remark on the linear case (L), which is to find $u\in\mathcal{H}$ such that

$$a(v,u) = \langle v,f\rangle \quad \forall\, v\in\mathcal{H} \quad\Longleftrightarrow\quad \mathbf{A}\mathbf{u}=\mathbf{f}. \tag{3.18}$$

Note that (3.5) follows from (3.1) but may still hold for indefinite problems, which still implies, in view of Remark 2.1, the validity of (3.8). In this case, when $G\equiv 0$, the matrix $\mathbf{A}^T\mathbf{A}$ is symmetric positive definite, and the iteration

$$\mathbf{u}^{n+1} = \mathbf{u}^n - \alpha\mathbf{A}^T(\mathbf{A}\mathbf{u}^n-\mathbf{f}), \quad n=0,1,2,\dots, \tag{3.19}$$

converges with a fixed error reduction $\rho<1$, provided that $0<\alpha<2/C_A^2$; i.e., (3.19) has the form (2.11) with $\mathbf{B}_n := \alpha\mathbf{A}^T$. An analogue for the general nonlinear case (GNL) will be given below.

For saddle point problems there are actually alternatives that avoid squaring the problem (in wavelet coordinates). One option is to employ an Uzawa iteration for applying the Schur complement operator, which conceptually also leads to an iteration of the form (3.10) for the Schur complement [13, 17].

Of course, in either case, step (n4) requires eventually approximating the *weighted residual* $\mathbf{B}_n\mathbf{R}(\mathbf{u}^n)$, which in the above linear case amounts to approximating $\mathbf{f}$ and approximately evaluating the infinite matrix $\mathbf{A}$ (respectively, $\mathbf{A}^T$). We shall address this issue later in some detail.

3.2. The general nonlinear case—locally convergent schemes (GNL).

While the assumptions in the previous setting allow us to conclude convergence of the ideal infinite dimensional scheme for *any* initial guess $\mathbf{u}^0$, one often has to be content with weaker assumptions (and correspondingly weaker conclusions). In the literature, variational problems of the type (2.2) are frequently studied under general assumptions on R, such as A1 and A2, that typically guarantee local convergence of an iterative scheme to a locally unique solution provided that a sufficiently good initial guess is known; see, e.g., [21, 23].

Our plan here is to exemplify the above paradigm under assumptions A1 and A2, provided that a sufficiently good initial approximation is known. According to (n2), we consider again the equivalent formulation (2.8) in wavelet coordinates and turn to devising a suitable iteration of the form (2.11) that converges for a sufficiently good initial guess. To this end, we assume that

$$\mathbf{u}^0 \in B_\delta(\mathbf{u}) := \{\mathbf{v} : \|\mathbf{v} - \mathbf{u}\|_{\ell_2(\mathcal{J})} < \delta\}, \tag{3.20}$$

where δ will be specified below.

As mentioned before, a possible choice for $\mathbf{B}_n$ could involve the Jacobian, which leads to Newton's method. However, under the above weak assumptions on R, we wish to avoid at this point requiring higher order smoothness conditions and consider first the following much simpler option. An analogue to the least squares iteration (3.19) would be $\mathbf{B}_n := D\mathbf{R}(\mathbf{u}^n)^T$. An even simpler alternative, which is presumably less computationally demanding, is to take the *stationary* matrix

$$\mathbf{B} = D\mathbf{R}(\mathbf{u}^0)^T, \tag{3.21}$$

provided that δ is sufficiently small. Let us point out next that for a sufficiently good initial guess $\mathbf{u}^0$

$$\mathbf{W}(\mathbf{v}) := \mathbf{v} - \alpha D\mathbf{R}(\mathbf{u}^0)^T\mathbf{R}(\mathbf{v})$$

is a contraction on $B_\delta(\mathbf{u})$. In fact,

$$\begin{aligned}\mathbf{W}(\mathbf{z}) - \mathbf{W}(\mathbf{v}) &= (\mathbf{z} - \mathbf{v}) - \alpha D\mathbf{R}(\mathbf{u}^0)^T(\mathbf{R}(\mathbf{z}) - \mathbf{R}(\mathbf{v}))\\ &= \left(\mathbf{I} - \alpha D\mathbf{R}(\mathbf{u}^0)^T D\mathbf{R}(\mathbf{v})\right)(\mathbf{z} - \mathbf{v}) + \mathrm{o}(\|\mathbf{z} - \mathbf{v}\|_{\ell_2(\mathcal{J})})\\ &= \left(\mathbf{I} - \alpha D\mathbf{R}(\mathbf{u}^0)^T D\mathbf{R}(\mathbf{u}^0)\right)(\mathbf{z} - \mathbf{v}) + \mathrm{o}(\|\mathbf{z} - \mathbf{v}\|_{\ell_2(\mathcal{J})})\\ &\qquad + \mathrm{O}(\epsilon(\delta)\|\mathbf{z} - \mathbf{v}\|_{\ell_2(\mathcal{J})}),\end{aligned} \tag{3.22}$$

where we have used assumption A1 and where $\epsilon(\delta)$ tends to zero as $\delta \to 0$. By A1 and A2, $DR(u^0)$ is still an isomorphism from $\mathcal{H}$ onto $\mathcal{H}'$ when δ is sufficiently small. Thus, by Remark 2.1, the positive definite matrix $D\mathbf{R}(\mathbf{u}^0)^T D\mathbf{R}(\mathbf{u}^0)$ is an automorphism on $\ell_2(\mathcal{J})$. Therefore, for $\alpha > 0$ satisfying

$$\alpha\|D\mathbf{R}(\mathbf{u}^0)^T D\mathbf{R}(\mathbf{u}^0)\|_{\ell_2(\mathcal{J})\to\ell_2(\mathcal{J})} < 2, \tag{3.23}$$

$\mathbf{W}$ is a contraction on $B_\delta(\mathbf{u})$. Furthermore, the iterates

$$\mathbf{u}^{n+1} = \mathbf{u}^n - \alpha D\mathbf{R}(\mathbf{u}^0)^T\mathbf{R}(\mathbf{u}^n), \quad n = 0, 1, \ldots, \tag{3.24}$$

stay in $B_\delta(\mathbf{u})$. In fact, as above,

$$\begin{aligned}\mathbf{u}^{n+1} - \mathbf{u} &= \mathbf{u}^n - \mathbf{u} - \alpha D\mathbf{R}(\mathbf{u}^0)^T(\mathbf{R}(\mathbf{u}^n) - \mathbf{R}(\mathbf{u}))\\ &= (\mathbf{I} - \alpha D\mathbf{R}(\mathbf{u}^0)^T D\mathbf{R}(\mathbf{u}^0))(\mathbf{u}^n - \mathbf{u}) + \mathrm{o}(\|\mathbf{u}^n - \mathbf{u}\|_{\ell_2(\mathcal{J})})\\ &\qquad + \mathrm{O}(\epsilon(\delta)\|\mathbf{u}^n - \mathbf{u}\|_{\ell_2(\mathcal{J})}).\end{aligned}$$

Hence, for α as above and δ sufficiently small, i.e., $\|\mathbf{I}-\alpha D\mathbf{R}(\mathbf{u}^0)^T D\mathbf{R}(\mathbf{u}^0)\|_{\ell_2(\mathcal{J})\to\ell_2(\mathcal{J})} =: b < 1$ and $\mathrm{o}(1) + \mathrm{O}(\epsilon(\delta)) < 1 - b$, one has $\|\mathbf{u}^{n+1} - \mathbf{u}\|_{\ell_2(\mathcal{J})} < \delta$. We can summarize these observations as follows.

Remark 3.4. Under the above assumptions there exist a $\delta_0 > 0$ and a positive α such that for any $\delta \le \delta_0$ and $\mathbf{u}^0 \in B_\delta(\mathbf{u})$ the iteration (3.24) converges to the locally unique solution $\mathbf{u}$ of (2.8). Moreover, there exists some $\rho < 1$ such that

$$\|\mathbf{u}^n - \mathbf{u}\|_{\ell_2(\mathcal{J})} \le \rho\|\mathbf{u}^{n-1} - \mathbf{u}\|_{\ell_2(\mathcal{J})}, \quad n = 1, 2, \ldots. \tag{3.25}$$

4. A perturbed first order iteration scheme. We shall now turn to step (n4) under the assumption that (2.11) gives rise to a fixed error reduction ρ per iteration step. Recall that by (3.17) and (3.25), this is indeed already the case for (L), (SL), and (GNL) for the corresponding stationary choices of $\mathbf{B}_n = \mathbf{B}$. In order to minimize technicalities we shall consider only this case in connection with such first order schemes. In order to arrive at computable versions of these schemes, we have to *approximate* the weighted residuals $\mathbf{BR}(\mathbf{u}^n)$ in each step. Already in the linear case (L), this requires approximating the application of an infinite matrix to a finitely supported vector and approximating the given data $\mathbf{f}$. In the nonlinear cases (SL), (GNL), the additional difficulty is to approximately evaluate the *nonlinear* expressions $\mathbf{R}(\mathbf{u}^j)$.

Our strategy can be outlined as follows. In the present section we shall address only issue (a) from section 1.2, namely, How accurate must these approximations be to be chosen at a given stage of the iteration so as to guarantee convergence to the correct solution? We shall do so at this point under the *assumption* that a subroutine for approximating the weighted residuals $\mathbf{BR}(\mathbf{v})$ with desired accuracy is at our disposal. Once (a) has been clarified for the general scope of problems, we shall in subsequent sections then narrow down step by step the specific requirements on the basic subroutine, develop concrete realizations for the various problem types (L), (SL), and (GNL), and analyze their complexity.

Thus for the time being we assume now that for $\mathbf{R}(\cdot) = \mathbf{F}(\cdot) - \mathbf{f}$ a routine with the following property is given.

$\mathbf{RES}\,[\eta, \mathbf{B}, \mathbf{F}, \mathbf{f}, \mathbf{v}] \to \mathbf{w}_\eta$ *determines for any positive tolerance* η *and any finitely supported input* $\mathbf{v}$ *a finitely supported* $\mathbf{w}_\eta$ *satisfying*

$$\|\mathbf{BR}(\mathbf{v}) - \mathbf{w}_\eta\|_{\ell_2(\mathcal{J})} \le \eta. \tag{4.1}$$

The need for the following further ingredient is at this point less obvious. It will be applied after a certain finite number of perturbed iterations based on the application of **RES**. It will be seen later that this is crucial for controlling the complexity of the scheme.

$\mathbf{CCOARSE}\,[\eta, \mathbf{v}] \to \mathbf{w}_\eta$ *determines for any positive tolerance* η *and any finitely supported input vector* $\mathbf{v}$ *a finitely supported output vector* $\mathbf{w}_\eta$ *such that*

$$\|\mathbf{v} - \mathbf{w}_\eta\|_{\ell_2(\mathcal{J})} \le \eta, \tag{4.2}$$

while the support of $\mathbf{w}_\eta$ *is minimized subject to certain constraints on the distribution of its entries.*

The constraints mentioned in **CCOARSE** will depend on the particular application and will be specified later. A perturbed iteration based on these ingredients requires specifying a suitable initialization.

Initialization. We distinguish the following three cases for the choice of the initial guess.

(L) In the linear case $\mathbf{R}(\mathbf{v}) = \mathbf{A}\mathbf{v} - \mathbf{f}$, we can set $\mathbf{u}^0 := \mathbf{0}$ so that an initial error bound is given, directly in view of (3.8), by $\|\mathbf{u} - \mathbf{u}^0\|_{\ell_2(\mathcal{J})} = \|\mathbf{u}\|_{\ell_2(\mathcal{J})} \le c_A^{-1}\|\mathbf{f}\|_{\ell_2(\mathcal{J})} =: \epsilon_0$. Moreover, in the positive definite case, any fixed $\alpha < 2/C_A$ in $\mathbf{B} = \alpha\mathbf{I}$ and $\alpha < 2/C_A^2$ in $\mathbf{B} = \alpha\mathbf{A}^T$ for the general least squares formulation ensure that $\mathbf{I} - \mathbf{BA}$ is a contraction on $B = \ell_2(\mathcal{J})$.

(SL) In the case of semilinear elliptic problems (3.7), (3.4), we recall from (3.12) that for $\mathbf{u}^0 = \mathbf{0}$ the initial error is bounded by

$$\|\mathbf{u}\|_{\ell_2(\mathcal{J})} \le c_A^{-1}\big(\|\mathbf{G}(\mathbf{0})\|_{\ell_2(\mathcal{J})} + \|\mathbf{f}\|_{\ell_2(\mathcal{J})}\big) =: \epsilon_0. \tag{4.3}$$

We choose $B = B_{2\epsilon_0}(\mathbf{u})$ and $\mathbf{B} := \alpha\mathbf{I}$ for a fixed $\alpha < 2/(C_A + \hat{C}(2\epsilon_0))$, $\alpha \leq 1$, so that (3.16) holds for $\delta_0 = 2\epsilon_0$ and $\rho = \rho(\alpha) < 1$, defined in (3.15).

(GNL) For the locally convergent scheme, we adhere to the assumptions made in section 3.2. For any fixed $\delta < \delta_0$ (the parameter from Remark 3.4) which satisfies $(1+\alpha)\delta < \delta_0$, where α is the constant from (3.23), we choose $\mathbf{u}^0$ according to (3.20). In this case, we have $\mathbf{B} = \alpha D\mathbf{R}(\mathbf{u}^0)^T$, and $\epsilon_0 := \delta$ is a valid initial error bound which ensures that for $\mathbf{v} \in B := B_\delta(\mathbf{u})$ the matrix $\mathbf{I} - \mathbf{B}D\mathbf{R}(\mathbf{v})$ is a contraction.

Thus in all cases (L), (SL), and (GNL) one has under the above premises

$$\|\mathbf{u} - \mathbf{u}^0\|_{\ell_2(\mathcal{J})} \leq \epsilon_0. \tag{4.4}$$

In order to control the perturbations caused by applications of **RES**, it will be convenient to extract the following fact from the above considerations.

Remark 4.1. For each of the above choices of $\mathbf{B}$ in (L), (SL), and (GNL) and the respective neighborhoods B of the exact solution $\mathbf{u}$ specified in the initialization, one has

$$\|(\mathbf{v} - \mathbf{z}) - \mathbf{B}(\mathbf{R}(\mathbf{v}) - \mathbf{R}(\mathbf{z}))\|_{\ell_2(\mathcal{J})} \leq \rho\|\mathbf{v} - \mathbf{z}\|_{\ell_2(\mathcal{J})}, \quad \mathbf{v}, \mathbf{z} \in B, \tag{4.5}$$

where $\rho < 1$ is the respective error reduction rate in (3.17) for the iteration (2.11).

Proof. The linear case (L) is obvious.

In the case (SL) of the semilinear elliptic problem (3.4), respectively, (3.7), one has for $\mathbf{B} = \alpha\mathbf{I}$ (with α specified in the initialization), by the same reasoning used in (3.11),

$$\mathbf{v} - \mathbf{z} - \alpha(\mathbf{R}(\mathbf{v}) - \mathbf{R}(\mathbf{z})) = \left(\mathbf{I} - \alpha\left(\mathbf{A} + \int_0^1 D\mathbf{G}(\mathbf{z} + s(\mathbf{v} - \mathbf{z}))ds\right)\right)(\mathbf{v} - \mathbf{z}).$$

The assertion follows then from (3.16).

Finally, for (GNL) (see section 3.2), the claim follows from (3.22) for $\mathbf{B} = \alpha D\mathbf{R}(\mathbf{u}^0)^T$ and α satisfying (3.23) with $B = B_\delta(\mathbf{u})$. □

The following last prerequisite will allow us to control the number of calls of **RES** before applying a coarsening step.

Remark 4.2. For each of the above choices of $\mathbf{B}$ in (L), (SL), and (GNL) and for the respective neighborhoods B of the exact solution $\mathbf{u}$, there exists a positive finite constant β such that

$$\|\mathbf{u} - \mathbf{v}\|_{\ell_2(\mathcal{J})} \leq \beta\|\mathbf{B}\mathbf{R}(\mathbf{v})\|_{\ell_2(\mathcal{J})}, \quad \mathbf{v} \in B. \tag{4.6}$$

Proof. Of course, in principle, this follows from (4.5) by the triangle inequality. However, β then depends on ρ for which only a poor estimate may be available. For (L) and (SL) one obtains better bounds as follows. From (3.7) and (3.9) we infer that

$$\begin{aligned}\|\mathbf{R}(\mathbf{v})\|_{\ell_2(\mathcal{J})} &= \sup_{\mathbf{z}\neq\mathbf{0}} \frac{\mathbf{z}^T(\mathbf{A}(\mathbf{v} - \mathbf{u}) + \mathbf{G}(\mathbf{v}) - \mathbf{G}(\mathbf{u}))}{\|\mathbf{z}\|_{\ell_2(\mathcal{J})}} \geq \frac{(\mathbf{v} - \mathbf{u})^T\mathbf{A}(\mathbf{v} - \mathbf{u})}{\|\mathbf{v} - \mathbf{u}\|_{\ell_2(\mathcal{J})}} \\ &\geq c_A\|\mathbf{v} - \mathbf{u}\|_{\ell_2(\mathcal{J})}.\end{aligned}$$

Thus (4.6) holds with $\beta = 1/(\alpha c_A)$. Similarly $\beta = 1/(\alpha c_A^2)$ works for the least squares formulation (3.19). For (GNL) we have by our assumptions that $\|\mathbf{B}(\mathbf{R}(\mathbf{v}) - D\mathbf{R}(\mathbf{v}))(\mathbf{v} - \mathbf{u})\|_{\ell_2(\mathcal{J})} \leq o(\|\mathbf{v} - \mathbf{u}\|_{\ell_2(\mathcal{J})})$ so that in a sufficiently small neighborhood of $\mathbf{u}$ (4.6) follows from the well-posedness relation (2.10). □

We can now describe our computable analogue of (2.11). For this we choose any fixed summable sequence $(\omega_j)_{j\in\mathbb{N}_0}$, which, for convenience, we arrange to sum to one $\sum_{j=0}^{\infty}\omega_j = 1$, and a fixed constant C^* which depends on the specific realization of the routine **CCOARSE**; see sections 5.2 and 6.2.

```
SOLVE $[\epsilon, \mathbf{R}, \mathbf{u}^0] \to \bar{\mathbf{u}}(\epsilon)$
  (i) Choose some $\bar{\rho} \in (0,1)$. Set $\bar{\mathbf{u}}^0 = \mathbf{u}^0$ and the corresponding initial bound $\epsilon_0$
      according to the above initialization, and define $j = 0$;
  (ii) If $\epsilon_j \le \epsilon$, stop and output $\bar{\mathbf{u}}(\epsilon) := \bar{\mathbf{u}}^j$; else set $\mathbf{v}^0 := \bar{\mathbf{u}}^j$ and $k = 0$
    (ii.1) Set $\eta_k := \omega_k \bar{\rho}^k \epsilon_j$ and compute
           $$\mathbf{r}^k = \mathbf{RES}\,[\eta_k, \mathbf{B}, \mathbf{F}, \mathbf{f}, \mathbf{v}^k], \qquad \mathbf{v}^{k+1} = \mathbf{v}^k - \mathbf{r}^k.$$
    (ii.2) If
           $$\beta\big(\eta_k + \|\mathbf{r}^k\|_{\ell_2(\mathcal{J})}\big) \le \epsilon_j/(2(1+2C^*)), \tag{4.7}$$
           set $\tilde{\mathbf{v}} := \mathbf{v}^k$ and go to (iii). Else set $k+1 \to k$ and go to (ii.1).
  (iii) CCOARSE $[\frac{2C^*\epsilon_j}{2(1+2C^*)}, \tilde{\mathbf{v}}] \to \bar{\mathbf{u}}^{j+1}$, $\epsilon_{j+1} = \epsilon_j/2$, $j+1 \to j$, go to (ii).
```

Let us confirm first that the choice of accuracy tolerances in **SOLVE** implies convergence.

PROPOSITION 4.3. *The iterates* $\bar{\mathbf{u}}^j$ *produced by the scheme* **SOLVE** *satisfy*

$$\|\mathbf{u} - \bar{\mathbf{u}}^j\|_{\ell_2(\mathcal{J})} \le \epsilon_j \tag{4.8}$$

so that, in particular, $\|\mathbf{u} - \bar{\mathbf{u}}(\epsilon)\|_{\ell_2(\mathcal{J})} \le \epsilon$. *By* (2.5)*, this means*

$$\Bigg\| u - \sum_{\lambda\in\Lambda(\epsilon)} \bar{u}(\epsilon)_\lambda \psi_\lambda \Bigg\|_{\mathcal{H}} \le C_1\epsilon, \tag{4.9}$$

where C_1 *is the constant from* (2.5) *and* $\Lambda(\epsilon) := \operatorname{supp}\mathbf{u}(\epsilon)$.

Moreover, the number of updates in step (ii.1) *prior to a coarsening step is uniformly bounded by some fixed* $K \in N$, *independent of* ϵ *and the data.*

Proof. We assume the above initialization and employ a simple perturbation argument using induction on j. We fix a value of j and let $\mathbf{u}^k := \mathbf{u}^k(\mathbf{v}^0)$ be the exact iterates $\mathbf{u}^{k+1} = \mathbf{u}^k - \mathbf{B}\mathbf{R}(\mathbf{u}^k)$ with initial guess $\mathbf{u}^0 = \mathbf{v}^0 = \bar{\mathbf{u}}^j$. Hence

$$\begin{aligned} \mathbf{v}^{k+1} - \mathbf{u}^{k+1} &= \mathbf{v}^k - \mathbf{u}^k - (\mathbf{r}^k - \mathbf{B}\mathbf{R}(\mathbf{u}^k)) \\ &= \mathbf{v}^k - \mathbf{u}^k - \mathbf{B}(\mathbf{R}(\mathbf{v}^k) - \mathbf{R}(\mathbf{u}^k)) + (\mathbf{B}\mathbf{R}(\mathbf{v}^k) - \mathbf{r}^k). \end{aligned} \tag{4.10}$$

Next we wish to invoke (4.5). To do this we need to make sure that the iterates $\mathbf{v}^k, \mathbf{u}^k$ stay in the neighborhood B mentioned in Remark 4.1. In the linear case (L) there is no constraint, i.e., $B = \ell_2(\mathcal{J})$. Let us look at the semilinear case (SL) next. By the induction assumption we know that $\|\mathbf{u} - \bar{\mathbf{u}}^j\|_{\ell_2(\mathcal{J})} \le \epsilon_j \le \epsilon_0$. Therefore, $\|\mathbf{u} - \mathbf{u}^k\|_{\ell_2(\mathcal{J})} \le \rho^k\|\mathbf{u} - \mathbf{u}^0\|_{\ell_2(\mathcal{J})} \le \rho^k\|\mathbf{u} - \bar{\mathbf{u}}^j\|_{\ell_2(\mathcal{J})} \le \rho^k\epsilon_j$. So $\mathbf{u}^k \in B$ for all $k \le K$. Also $\mathbf{v}^0 = \bar{\mathbf{u}}^j \in B$. Thus suppose that $\mathbf{v}^k$ is in B. We wish to show that then also $\mathbf{v}^{k+1} \in B$. To this end, let ρ_* be the true reduction rate in (2.11) (for which $\rho(\alpha)$ from (3.15) might be a poor estimate) and set $\hat{\rho} := \max\{\rho_*, \bar{\rho}\}$. Then we infer from (4.5), (4.10), and the definition of $\mathbf{r}^k$ in step (ii) that

$$\begin{aligned} \|\mathbf{v}^{k+1} - \mathbf{u}^{k+1}\|_{\ell_2(\mathcal{J})} &\le \rho_*\|\mathbf{v}^k - \mathbf{u}^k\|_{\ell_2(\mathcal{J})} + \omega_k\bar{\rho}^k\epsilon_j \\ &\le \left(\sum_{l=0}^{k} \rho_*^{k-l}\omega_l\bar{\rho}^l\right)\epsilon_j \le \hat{\rho}^k\epsilon_j, \end{aligned} \tag{4.11}$$

where we have used that $\mathbf{u}^0 = \mathbf{v}^0$. Moreover, since

$$\|\mathbf{v}^{k+1} - \mathbf{u}\|_{\ell_2(\mathcal{J})} \leq \hat{\rho}^k \epsilon_j + \|\mathbf{u}^{k+1} - \mathbf{u}\|_{\ell_2(\mathcal{J})} \leq 2\hat{\rho}^k \epsilon_j, \tag{4.12}$$

we see that $\mathbf{v}^{k+1} \in B = B_{2\epsilon_0}(\mathbf{u})$ so that the iteration can be advanced.

For the locally convergent scheme (GNL) with $\mathbf{B} = \alpha \mathbf{R}(\mathbf{u}^0)^T$, the reasoning is analogous. The choice of the initial guess ensures that $(\rho + \alpha)\epsilon_j \leq (\rho + \alpha)\epsilon_0 \leq (1 + \alpha)\epsilon_0 \leq \delta_0$. Then the above arguments for (SL) yield again (4.15) so that all iterates stay in $B = B_{\delta_0}(\mathbf{u})$.

Now note that by (4.12),

$$\|\mathbf{r}^k\|_{\ell_2(\mathcal{J})} \leq \|\mathbf{v}^{k+1} - \mathbf{u}\|_{\ell_2(\mathcal{J})} + \|\mathbf{v}^k - \mathbf{u}\|_{\ell_2(\mathcal{J})} \leq 4\hat{\rho}^{k-1}\epsilon_j$$

so that

$$\beta(\eta_k + \|\mathbf{r}^k\|_{\ell_2(\mathcal{J})}) \leq \beta\epsilon_j(\omega_k\hat{\rho} + 4)\hat{\rho}^{k-1}. \tag{4.13}$$

Hence, choosing

$$K := \min\{k \in \mathbb{N} : \beta(\omega_k\hat{\rho} + 4)\hat{\rho}^{k-1} \leq 1/(2(1 + 2C^*))\}, \tag{4.14}$$

we see that (4.7) is met after at most K steps. Moreover, (4.6) says that

$$\|\mathbf{u} - \tilde{\mathbf{v}}\|_{\ell_2(\mathcal{J})} \leq \frac{\epsilon_j}{2(1 + 2C^*)}. \tag{4.15}$$

Thus in all cases the estimate (4.8) follows now immediately from step (iii) in **SOLVE**, the definition of **CCOARSE**, and (4.2). Finally, (4.9) is an immediate consequence of the norm equivalence (2.5) and (4.8). □

Thus, for an idealized infinite dimensional scheme of order one in (n3), we know how to choose the tolerances in the routines **RES** and **CCOARSE** so as to guarantee convergence. Moreover, the true error reduction rate need not be known, and one can use any (possibly too optimistic) guess $\bar{\rho}$. Of course, choosing $\bar{\rho}$ too small, the intermediate tolerances get perhaps unnecessarily small.

Our paradigm for solving nonlinear problems is built on the availability of numerical algorithms such as **CCOARSE** and **RES**. The remainder of this paper shows how to construct concrete practical realizations of these algorithms in various settings and then shows how, under suitable controls on the computations in these algorithms, we can give complexity estimates for the entire numerical scheme **SOLVE**. More precisely, we wish to determine its *work/accuracy balance*, i.e., given any target accuracy ϵ, how many degrees of freedom $N = N(\epsilon) := \#\Lambda(\epsilon)$, where $\Lambda(\epsilon) := \operatorname{supp} \bar{\mathbf{u}}(\epsilon)$, are needed to achieve it, and what is the associated (asymptotic) computational work. Of course, one hopes to keep the latter quantity proportional to $N(\epsilon)$ so that the number of degrees of freedom is a reasonable complexity measure. In the following section we shall address these issues first for the linear case (3.18). We review quickly the relevant facts from [8, 9] tailored somewhat to the present situation. On one hand, they will serve as building blocks for the general nonlinear case. On the other hand, they also help to bring out some conceptual distinctions.

5. Realization and complexity analysis in the linear case (L). Recall from (3.10) that in the linear case, $\mathbf{BR}(\mathbf{v}) = \alpha(\mathbf{Av} - \mathbf{f})$ (or $\alpha\mathbf{A}^T(\mathbf{Av} - \mathbf{f})$). Thus one part of approximating the residual is to approximate given data, here in the form of the right-hand side $\mathbf{f}$, which, in general, is an infinite sequence.

5.1. Coarsening and best N-term approximation. We will also assume in what follows that all coefficients of $\mathbf{f}$ are *known* and thus in principle accessible. In practice this may require a preprocessing step that computes for some overall target accuracy $\bar{\epsilon}$ (depending on the desired solution accuracy) an approximation $\mathbf{f}_{\bar{\epsilon}}$ satisfying $\|\mathbf{f} - \mathbf{f}_{\bar{\epsilon}}\|_{\ell_2(\mathcal{J})} \leq \bar{\epsilon}$ and then orders the entries by size. Once this has been done, any coarser approximations needed in the course of the iteration process can be produced by the following simplest version of **CCOARSE**, introduced and analyzed in [8].

COARSE $[\eta, \mathbf{v}] \to \mathbf{v}_\eta$ *associates with any finitely supported input* $\mathbf{v}$ *a vector* $\mathbf{v}_\eta$ *such that*

$$\|\mathbf{v} - \mathbf{v}_\eta\|_{\ell_2(\mathcal{J})} \leq \eta, \quad \#\mathrm{supp}\,\mathbf{w} \geq \#\mathrm{supp}\,\mathbf{v}_\eta, \text{ whenever } \|\mathbf{v} - \mathbf{w}\|_{\ell_2(\mathcal{J})} \leq \eta. \tag{5.1}$$

Thus **COARSE** determines for a given finitely supported vector a new vector with the smallest possible support deviating no more than a prescribed tolerance from the input. There is no constraint on the distribution of active indices in this case. Ordering the entries of $\mathbf{v}$ sizewise, this can be realized by summing entries in increasing order until the sum of their squares reaches η^2. For a detailed description of this routine, see [8]. In fact, a strict ordering is not necessary. The same effect is realized by collecting the entries in binary bins, which avoids a log factor at the expense of a fixed factor in the accuracy tolerance [1].

The routine **COARSE** can be used to approximate the data $\mathbf{f}$ as follows:

$$\mathbf{RHS}\,[\eta, \mathbf{f}] := \mathbf{COARSE}\,[\eta - \bar{\epsilon}, \mathbf{f}_{\bar{\epsilon}}], \tag{5.2}$$

whenever $\eta > \bar{\epsilon}$.

Note that **COARSE** is a *nonlinear* process that realizes a given accuracy tolerance at the expense of a minimal number of degrees of freedom. It is therefore a version of *best N-term approximation* in $\ell_2(\mathcal{J})$. In fact, defining

$$\sigma_{N,\ell_2(\mathcal{J})}(\mathbf{u}) := \min_{\#\mathrm{supp}\,\mathbf{v} \leq N} \|\mathbf{u} - \mathbf{v}\|_{\ell_2(\mathcal{J})}, \tag{5.3}$$

one has for any $\mathbf{v} \in \ell_2(\mathcal{J})$

$$\sigma_{N,\ell_2(\mathcal{J})}(\mathbf{v}) = \|\mathbf{v} - \mathbf{v}_N\|_{\ell_2(\mathcal{J})} = \left(\sum_{n>N} |v_n^*|^2\right)^{1/2}, \tag{5.4}$$

where $(v_n^*)_{n\in\mathbb{N}}$ is the *nonincreasing rearrangement* of $\mathbf{v}$. Thus $\mathbf{v}_N$ is obtained by retaining the N largest (in modulus) terms of $\mathbf{v}$ and setting all other entries to zero. Depending on the context, $\mathbf{v}_N$ will be viewed as a sequence in $\ell_2(\mathcal{J})$ or a vector in $\mathbb{R}^N$.

The best N-term approximation sets a lower bound for the complexity that could ever be achieved by a scheme like **SOLVE**. In fact, it will serve as our benchmark in the case of linear variational problems of the form (3.18). In order to make this precise, we introduce the corresponding *approximation classes*

$$\mathcal{A}^s := \{\mathbf{v} \in \ell_2(\mathcal{J}) : \sigma_{N,\ell_2(\mathcal{J})}(\mathbf{v}) \lesssim N^{-s}\},$$

which we endow with the quasi norm $\|\mathbf{v}\|_{\mathcal{A}^s} := \sup_{N\in\mathbb{N}} N^s \sigma_{N,\ell_2(\mathcal{J})}(\mathbf{v})$. Clearly, every $\mathbf{v}$ with finite support belongs to $\mathcal{A}^s$ for any $s > 0$. The question is, Does **SOLVE** produce for any target accuracy ϵ an approximate solution within that tolerance at a computational expense that stays bounded by $C\epsilon^{-1/s}$ whenever the exact solution belongs to $\mathcal{A}^s$, at least for some range of $s > 0$?

5.2. Adaptive application of compressible matrices. It remains to approximate the action of $\mathbf{A}$ on a finitely supported vector $\mathbf{v}$. While the treatment of the right-hand side data has been already seen to comply with best N-term approximation complexity, the question arises whether $\mathbf{Av}$ can be approximated with a similar efficiency. This has been answered affirmatively in [8], and we briefly recall the relevant facts from there.

Due to the *vanishing moment* property of wavelets, the wavelet representation of many operators turns out to be *quasi-sparse.* The following quantification of sparsity is appropriate [8].

We shall use the notation $(\alpha_j)_{j=1}^{\infty}$ to denote a summable sequence of positive numbers: $\sum_{j=1}^{\infty} \alpha_j < \infty$. A matrix $\mathbf{C}$ is said to be s^*-compressible, $\mathbf{C} \in \mathcal{C}_{s^*}$, if for any $0 < s < s^*$ and every $j \in \mathbb{N}$ there exists a matrix $\mathbf{C}_j$ with the following properties: For some summable sequence $(\alpha_j)_{j=1}^{\infty}$, $\mathbf{C}_j$ is obtained by replacing all but the order of $\alpha_j 2^j$ entries per row and column in $\mathbf{C}$ by zero and satisfies

$$\|\mathbf{C} - \mathbf{C}_j\|_{\ell_2(\mathcal{J}) \to \ell_2(\mathcal{J})} \leq C \alpha_j 2^{-js}, \quad j \in \mathbb{N}. \tag{5.5}$$

Specifically, wavelet representations of differential (and also certain singular integral) operators fall into this category. One typically has then estimates of the type

$$|a(\psi_\lambda, \psi_\mu)| \lesssim 2^{-\sigma ||\lambda| - |\mu||}, \tag{5.6}$$

where $\sigma > d/2$ depends on the *regularity* of the wavelets.

In order to describe the essence of an approximate application scheme for compressible matrices, we abbreviate for any finitely supported $\mathbf{v}$ the best 2^j-term approximations by $\mathbf{v}_{[j]} := \mathbf{v}_{2^j}$ and define

$$\mathbf{w}_j := \mathbf{A}_j \mathbf{v}_{[0]} + \mathbf{A}_{j-1}(\mathbf{v}_{[1]} - \mathbf{v}_{[0]}) + \cdots + \mathbf{A}_0(\mathbf{v}_{[j]} - \mathbf{v}_{[j-1]}) \tag{5.7}$$

as an approximation to $\mathbf{Av}$. Obviously this scheme is *adaptive* in that it exploits directly information on $\mathbf{v}$. In fact, if $\mathbf{A} \in \mathcal{C}_{s^*}$, then the triangle inequality together with the above compression estimates yield for any fixed $s < s^*$

$$\|\mathbf{Av} - \mathbf{w}_j\|_{\ell_2(\mathcal{J})} \leq c \left(\underbrace{\|\mathbf{v} - \mathbf{v}_{[j]}\|_{\ell_2(\mathcal{J})}}_{\sigma_{2^j, \ell_2(\mathcal{J})}(\mathbf{v})} + \sum_{l=0}^{j} \alpha_l 2^{-ls} \underbrace{\|\mathbf{v}_{[j-l]} - \mathbf{v}_{[j-l-1]}\|_{\ell_2(\mathcal{J})}}_{\lesssim\ \sigma_{2^{j-l-1}, \ell_2(\mathcal{J})}(\mathbf{v})} \right), \tag{5.8}$$

where $\mathbf{v}_{[-1]} := \mathbf{0}$. One can now exploit the a posteriori information offered by the quantities $\sigma_{2^{j-l-1}, \ell_2(\mathcal{J})}(\mathbf{v})$ to choose the smallest j for which the right-hand side of (5.8) is smaller than a given target accuracy η and set $\mathbf{w}_\eta := \mathbf{w}_j$. Since the sum is finite for each finitely supported input $\mathbf{v}$, such a j does indeed exist. This leads to a concrete multiplication scheme (see [8, 2] for a detailed description, analysis, and implementation) which we summarize as follows.

APPLY $[\eta, \mathbf{A}, \mathbf{v}] \to \mathbf{w}_\eta$ *determines for any finitely supported input* $\mathbf{v}$ *a finitely supported output* $\mathbf{w}_\eta$ *such that*

$$\|\mathbf{Av} - \mathbf{w}_\eta\|_{\ell_2(\mathcal{J})} \leq \eta. \tag{5.9}$$

The complexity of this scheme would be asymptotically optimal if the size of $\operatorname{supp}(\mathbf{w}_\eta)$ and the corresponding work count remain bounded by $C(N_\eta + \#\operatorname{supp}(\mathbf{v}))$, where N_η is the smallest N such that $\sigma_{N,\ell_2(\mathcal{J})}(\mathbf{Av}) \le \eta$. We shall see that this is indeed the case for a certain range of decay rates of $\sigma_{N,\ell_2(\mathcal{J})}(\mathbf{v})$.

The main result concerning **APPLY** can be formulated as follows [8].

THEOREM 5.1. *Suppose that* $\mathbf{C} \in \mathcal{C}_{s^*}$ *and that* $0 < s < s^*$. *Then, in addition to* (5.9)*, for any input vector* $\mathbf{v}$ *with finite support,* $\mathbf{w}_\eta = \mathbf{APPLY}\,[\eta, \mathbf{C}, \mathbf{v}]$ *satisfies*

(i) $\|\mathbf{w}_\eta\|_{\mathcal{A}^s} \lesssim \|\mathbf{v}\|_{\mathcal{A}^s}$*,*

(ii) $\#\mathrm{supp}\,\mathbf{w}_\eta \lesssim \|\mathbf{v}\|_{\mathcal{A}^s}^{1/s}\eta^{-1/s}$ *and* $\#\mathrm{flops} \lesssim \#\mathrm{supp}\,\mathbf{v} + \|\mathbf{v}\|_{\mathcal{A}^s}^{1/s}\eta^{-1/s}$*,*

where the constants in these estimates depend only on s *when* s *is small. Moreover, any* $\mathbf{C} \in \mathcal{C}_{s^*}$ *is bounded on* $\mathcal{A}^s$ *as long as* $s < s^*$*.*

Thus, when dealing with linear problems (3.18), an approximation within the tolerance $\eta > 0$ to the weighted residual $\mathbf{BR}(\mathbf{v}) = \alpha(\mathbf{Av} - \mathbf{f})$ for any finitely supported input $\mathbf{v}$ can be computed as

$$\mathbf{RES}_{\mathrm{lin}}[\eta, \alpha\mathbf{I}, \mathbf{A}, \mathbf{f}, \mathbf{v}] := \alpha\left(\mathbf{APPLY}\left[\frac{\eta}{2\alpha}, \mathbf{A}, \mathbf{v}\right] - \mathbf{RHS}\left[\frac{\eta}{2\alpha}, \mathbf{f}\right]\right), \tag{5.10}$$

where **RHS** is given by (5.2). The same ideas can be used in the least squares case (3.19), where again **RHS** can be composed of **COARSE** and **APPLY**; see [9] for details.

Remark 5.1. Since by Theorem 5.1 $\mathbf{f} \in \mathcal{A}^s$, whenever the solution $\mathbf{u}$ belongs to $\mathcal{A}^s$, the above considerations and analogous facts about **COARSE** from [8] show that the output $\mathbf{f}_\eta$ of $\mathbf{RHS}\,[\eta, \mathbf{f}]$ satisfies $\|\mathbf{f}_\eta\|_{\mathcal{A}^s} \lesssim \|\mathbf{u}\|_{\mathcal{A}^s}$ and $\#\mathrm{supp}\,\mathbf{f}_\eta \lesssim \eta^{-1/s}\|\mathbf{u}\|_{\mathcal{A}^s}^{1/s}$.

These observations provide the following result.

PROPOSITION 5.2. *If the sequence of wavelet coefficients* $\mathbf{u}$ *of the exact solution* u *of* (3.18) *belongs to* $\mathcal{A}^s$ *and if* $\mathbf{A}$ *belongs to* $\mathcal{C}_{s^*}$ *with* $s^* > s$*, then, for any finitely supported input* $\mathbf{v}$*, the output* $\mathbf{w}_\eta$ *of the scheme* $\mathbf{RES}_{\mathrm{lin}}\,[\eta, \alpha\mathbf{I}, \mathbf{A}, \mathbf{f}, \mathbf{v}]$ *satisfies*

$$\begin{aligned} \|\mathbf{w}_\eta\|_{\mathcal{A}^s} &\lesssim (\|\mathbf{v}\|_{\mathcal{A}^s} + \|\mathbf{u}\|_{\mathcal{A}^s}), \\ \#\,\mathrm{supp}\,\mathbf{w}_\eta &\lesssim \eta^{-1/s}\left(\|\mathbf{u}\|_{\mathcal{A}^s}^{1/s} + \|\mathbf{v}\|_{\mathcal{A}^s}^{1/s}\right), \quad \eta > 0, \end{aligned} \tag{5.11}$$

where the constants in these estimates depend only on s*.*

Proposition 5.2 controls the complexity within each iteration block (ii) of perturbed iterations where, however, the constants may build up. To avoid this is exactly the role of step (iii) in **SOLVE**, which is based on the following "coarsening lemma" from [8]. (The following version can be found in [7].)

PROPOSITION 5.3. *Let* a *be some fixed number strictly larger than* 1*. If* $\mathbf{v} \in \mathcal{A}^s$ *and* $\|\mathbf{v} - \mathbf{w}\|_{\ell_2(\mathcal{J})} \le \eta$ *with* $\#\mathrm{supp}\,\mathbf{w} < \infty$*, then* $\bar{\mathbf{w}}_\eta := \mathbf{COARSE}\,[a\eta, \mathbf{w}]$ *satisfies* $\|\mathbf{v} - \bar{\mathbf{w}}_\eta\|_{\ell_2(\mathcal{J})} \le (1 + a)\eta$ *and*

$$\#\mathrm{supp}\,\bar{\mathbf{w}}_\eta \lesssim \|\mathbf{v}\|_{\mathcal{A}^s}^{1/s}\eta^{-1/s}, \quad \|\bar{\mathbf{w}}_\eta\|_{\mathcal{A}^s} \lesssim \|\mathbf{v}\|_{\mathcal{A}^s}, \tag{5.12}$$

where the constants in these estimates depend only on s *when* s *becomes small.*

By Proposition 5.3, the coarsening step (iii), with the above algorithm **COARSE** used as **CCOARSE**, pulls a current approximation to the unknown $\mathbf{u}$ toward its best N-term approximation and controls the $\mathcal{A}^s$-norms of the approximations, *independently* of the possible increase of these norms caused by several preceding applications of **RES**, provided that $2C^* > 1$. Thus, in connection with **COARSE**, one can take any fixed $C^* > 1/2$ in (4.7) and step (iii) of **SOLVE**.

Let us denote by $\mathbf{SOLVE}_{\mathrm{lin}}$ the specification of **SOLVE** obtained by using $\mathbf{RES}_{\mathrm{lin}}$ and **COARSE** in place of **RES**, respectively, **CCOARSE**. We emphasize that adaptivity enters the scheme $\mathbf{SOLVE}_{\mathrm{lin}}$ solely through the adaptive application of $\mathbf{A}$ and the residual check (4.7).

Under the above premises, Propositions 5.2 and 5.3 allow one to show that $\mathbf{SOLVE}_{\rm lin}$ exhibits optimal complexity in the sense that it realizes the Meta-Theorem from section 1.1 with (unconstrained) best N-term approximation as a benchmark.

Remark 5.4. We conclude this section by recalling that $\mathbf{u} \in \mathcal{A}^s$ is, for instance, implied by a certain Besov regularity of u. In fact, when $\mathcal{H} = H^t$ (a Sobolev space of smoothness t), $u \in B_\tau^{t+ds}(L_\tau)$, with $\tau^{-1} = s + 1/2$, implies $\mathbf{u} \in \mathcal{A}^s$. This can be used to identify circumstances under which the adaptive scheme performs asymptotically better than a scheme based on uniform refinements. Recall that $B_\tau^{t+ds}(L_\tau)$ is the "largest" space of smoothness $t + sd$ embedded in H^t.

6. The nonlinear case.

In view of the fact that **SOLVE** has the same structure, regardless of whether the involved operators are linear or nonlinear, our strategy will be to follow closely the above lines also when the variational problem (2.2) is nonlinear. In principle, this will prove successful, although some important modifications of the ingredients will be encountered. The main distinction lies in the sparsity measure in that the role of best (unrestricted) N-term approximation will be replaced by *best tree approximation*. This constraint on the distribution of active coefficients arises naturally when analyzing the approximate evaluation of nonlinear expressions $\mathbf{R}(\mathbf{v})$. Moreover, index sets with tree structure are analogous to locally refined meshes in adaptive finite element methods.

6.1. Tree approximation and coarsening.

Let us explain first what we mean by a tree structure associated to the set of wavelet indices. In the simplest case of a one dimensional basis $\psi_\lambda = \psi_{j,k} = 2^{j/2}\psi(2^j \cdot -k)$, this structure is obvious: each index (j, k) has two children $(j + 1, 2k)$ and $(j + 1, 2k + 1)$. A similar tree structure can be associated to all available constructions of wavelet basis on a multidimensional domain: to each index λ one can assign $m(\lambda) \geq 2$ children μ such that $|\mu| = |\lambda| + 1$, where $m(\lambda)$ might vary from one index to another but is uniformly bounded by some fixed M. We shall use the notation $\mu \prec \lambda$ ($\mu \preceq \lambda$) in order to express that μ is a descendent of λ (or equals λ) in the tree. We also have the property

$$\mu \prec \lambda \Rightarrow S_\mu \subset S_\lambda, \tag{6.1}$$

where we recall that $S_\lambda := \operatorname{supp} \psi_\lambda$. A set $\mathcal{T} \subset \mathcal{J}$ is called a *tree* if $\lambda \in \mathcal{T}$ implies $\mu \in \mathcal{T}$ whenever $\lambda \prec \mu$.

If the tree $\mathcal{T} \subset \mathcal{J}$ is finite, we define the set $\mathcal{L} = \mathcal{L}(\mathcal{T})$ of *outer leaves* as the set of those indices outside the tree whose parent belongs to the tree

$$\mathcal{L} := \{\lambda \in \mathcal{J} : \lambda \notin \mathcal{T},\ \lambda \prec \mu \implies \mu \in \mathcal{T}\}. \tag{6.2}$$

We shall make use of the easily verifiable relation

$$\#\mathcal{T} \sim \#\mathcal{L}(\mathcal{T}), \tag{6.3}$$

where the constants depend only on the number M of children.

Note that $\mathcal{L}(\mathcal{T})$ plays the role of a (locally refined) mesh. Associating to any sequence $\mathbf{v} = (v_\lambda)$ in $\ell_2(\mathcal{J})$, another sequence $\tilde{\mathbf{v}} = (\tilde{v}_\lambda)$ whose entries are defined by

$$\tilde{v}_\lambda := \left(\sum_{\mu \preceq \lambda} |v_\mu|^2 \right)^{1/2}, \tag{6.4}$$

one readily confirms that $\mu \prec \lambda$ implies $\tilde{v}_\lambda \geq \tilde{v}_\mu$ and

$$\|\mathbf{v} - \mathbf{v}|_{\mathcal{T}}\|^2_{\ell_2(\mathcal{J})} = \sum_{\lambda \in \mathcal{L}(\mathcal{T})} \tilde{v}^2_\lambda. \tag{6.5}$$

Recall that in the linear case we have been able to compare the performance of **SOLVE** with the *best (unconstrained) N-term approximation.* We want now to develop sparsity measures that respect tree structure. Once a suitable measure has been identified, one can follow conceptually the lines of section 3. The counterpart for the spaces $\mathcal{A}^s$ are now the analogous spaces defined via *best tree N-term approximation* where the distribution of active coefficients in an approximant has a tree structure; see [10]. In fact, let

$$\sigma^{\text{tree}}_{N,\ell_2(\mathcal{J})}(\mathbf{v}) := \min\{\|\mathbf{v} - \mathbf{w}\|_{\ell_2(\mathcal{J})} :\ \mathcal{T} := \operatorname{supp} \mathbf{w} \ \text{ is a tree and } \ \#\mathcal{T} \leq N\}. \tag{6.6}$$

Any minimizing tree will be called $\mathcal{T}_N(\mathbf{v})$. We define now

$$\mathcal{A}^s_{\text{tree}} := \{\mathbf{v} \in \ell_2(\mathcal{J}) : \sigma^{\text{tree}}_{N,\ell_2(\mathcal{J})}(\mathbf{v}) \ \lesssim \ N^{-s}\}, \tag{6.7}$$

which again becomes a quasi-normed space under the quasi norm

$$\|\mathbf{v}\|_{\mathcal{A}^s_{\text{tree}}} := \sup_{n \in \mathbb{N}} N^s \sigma^{\text{tree}}_{N,\ell_2(\mathcal{J})}(\mathbf{v}). \tag{6.8}$$

One can again relate the membership of $\mathbf{v}$ to $\mathcal{A}^s_{\text{tree}}$ to the regularity of the corresponding expansion v [10].

Remark 6.1. Let $\mathcal{H} = H^t$ for some $t > 0$. If the wavelet expansion v with coefficient sequence $\mathbf{v}$ belongs to $B^{t+sd}_{\tau'}(L_{\tau'})$ for some τ' satisfying $\tau' > (s + 1/2)^{-1}$, then $\mathbf{v} \in \mathcal{A}^s_{\text{tree}}$. Thus, in terms of regularity, $\mathcal{A}^s_{\text{tree}}$ differs from $\mathcal{A}^s$ by a little additional regularity imposed on the respective expansions, due to the stronger metric $L_{\tau'}$, $\tau' > \tau$. Thus a tree approximation rate N^{-s} can still be achieved for much larger spaces than H^{t+sd}, which governs the corresponding rate for uniform refinements.

6.2. Tree coarsening. We shall specify next a coarsening routine **CCOARSE** that preserves *tree structures* and, as before, applies to finitely supported sequences. It will be referred to as **TCOARSE**. Its definition requires some preparation. Given $\mathbf{w}$, a tree $\mathcal{T} = \mathcal{T}^*(\eta, \mathbf{w})$ is called η-best for $\mathbf{w}$ if

$$\|\mathbf{w} - \mathbf{w}|_{\mathcal{T}}\|_{\ell_2(\mathcal{J})} \leq \eta \ \text{ and } \ \#\mathcal{T} = \min\{\#\mathcal{T}' : \|\mathbf{w} - \mathbf{w}|_{\mathcal{T}'}\|_{\ell_2(\mathcal{J})} \leq \eta,\ \mathcal{T}' \text{ a tree}\}.$$

Requiring best trees will be too stringent from a practical point of view. Therefore, we shall be content with the following relaxed version. A tree $\mathcal{T} = \mathcal{T}(\eta, \mathbf{w})$ is called (η, C)-*near best* (or briefly near best when the parameters are clear from the context) if

$$\|\mathbf{w} - \mathbf{w}|_{\mathcal{T}}\|_{\ell_2(\mathcal{J})} \leq \eta \ \text{ and } \ \#\mathcal{T} \leq C\#\mathcal{T}^*(\eta/C, \mathbf{w}).$$

The action of **TCOARSE** can now be described as follows.

TCOARSE $[\eta, \mathbf{w}] \to \bar{\mathbf{w}}_\eta$ *determines for a fixed constant $C^* \geq 1$, any finitely supported input* $\mathbf{w}$*, and any tolerance $\eta > 0$ an (η, C^*)-near best tree $\mathcal{T}(\eta, \mathbf{w})$ and sets* $\bar{\mathbf{w}}_\eta := \mathbf{w}|_{\mathcal{T}(\eta,\mathbf{w})}$.

The realization of this routine can be based on either one of the two algorithms for generating near best tree approximations developed in [3]. To apply the results from [3] in the present situation, the role of the partition P associated in [3] to a tree $\mathcal{T}$ is played here by the set $\mathcal{L}(\mathcal{T})$ of outer leaves; recall (6.5). Moreover, for $\lambda \in \mathcal{L}(\mathcal{T})$, the local error terms for a given $\mathbf{v} \in \ell_2(\mathcal{J})$ are here given by $e(\lambda) := \tilde{v}_\lambda^2$. Obviously, the $e(\lambda)$ are subadditive in the sense of [3]. Hence the results from [3] apply. To use the algorithm from [3], we need to know the values $\tilde{w}_\lambda$, $\lambda \in \mathcal{T}(\operatorname{supp}\mathbf{w})$, the smallest tree containing the support of $\mathbf{w}$. Summing the squares of the entries of $\mathbf{w}$ starting from the leaves of $\mathcal{T}(\operatorname{supp}\mathbf{w})$ and working toward the roots provide these quantities at an expense of $\#\mathcal{T}(\operatorname{supp}\mathbf{w})$ operations. Combining this with Theorem 5.2 from [3] establishes the following fact.

Proposition 6.2. *For any given finitely supported input* $\mathbf{w}$*, the computational cost of the output* $\bar{\mathbf{w}}_\eta$ *produced by* **TCOARSE** $[\eta, \mathbf{w}]$ *remains proportional to* $\#\mathcal{T}(\operatorname{supp}$ $\mathbf{w})$. *The underlying tree* $\mathcal{T}(\eta, \mathbf{w})$ *is* (η, C^*)*-near best, where* C^* *is the constant appearing in the estimate* (5.8) *in Theorem* 5.2 *of* [3].

The routine **TCOARSE** will be used as **CCOARSE** in step (iii) of **SOLVE**. The constant C^* appears in the stopping criterion in step (ii.2) of **SOLVE** and in the coarsening step (iii). In analogy to the linear case, its purpose is to control the $\mathcal{A}^s_{\text{tree}}$-norms of the approximants. This is made precise by the following counterpart to Proposition 5.3.

Proposition 6.3. *If* $\mathbf{v} \in \mathcal{A}^s_{\text{tree}}$ *and* $\|\mathbf{v} - \mathbf{w}\|_{\ell_2(\mathcal{J})} \le \eta$ *with* $\#\operatorname{supp}\mathbf{w} < \infty$*, then* $\bar{\mathbf{w}}_\eta :=$ **TCOARSE** $[2C^*\eta, \mathbf{w}]$ *satisfies*

$$\#\operatorname{supp}\bar{\mathbf{w}}_\eta \lesssim \|\mathbf{v}\|_{\mathcal{A}^s_{\text{tree}}}^{1/s}\eta^{-1/s}, \quad \|\mathbf{v} - \bar{\mathbf{w}}_\eta\|_{\ell_2(\mathcal{J})} \le (1+2C^*)\eta, \tag{6.9}$$

and

$$\|\bar{\mathbf{w}}_\eta\|_{\mathcal{A}^s_{\text{tree}}} \lesssim \|\mathbf{v}\|_{\mathcal{A}^s_{\text{tree}}}, \tag{6.10}$$

where the constants depend only on s *when* $s \to 0$ *and on* C^* *in* **TCOARSE**.

Proof. The second estimate in (6.9) follows from the triangle inequality. As for the first estimate in (6.9), assume that $\mathbf{v} \in \mathcal{A}^s_{\text{tree}}$ and consider the best N-term tree $\mathcal{T}_N(\mathbf{v})$ for $\mathbf{v}$ defined by (6.6). We first note that

$$\begin{aligned} \|\mathbf{w} - \mathbf{w}|_{\mathcal{T}_N(\mathbf{v})}\|_{\ell_2(\mathcal{J})} &\le \|(\mathbf{w} - \mathbf{v})|_{\mathcal{J}\setminus\mathcal{T}_N(\mathbf{v})}\|_{\ell_2(\mathcal{J})} + \|\mathbf{v} - \mathbf{v}|_{\mathcal{T}_N(\mathbf{v})}\|_{\ell_2(\mathcal{J})} \\ &\le \eta + \|\mathbf{v} - \mathbf{v}|_{\mathcal{T}_N(\mathbf{v})}\|_{\ell_2(\mathcal{J})}. \end{aligned} \tag{6.11}$$

According to the definition of the norm $\|\cdot\|_{\mathcal{A}^s_{\text{tree}}}$ by (6.8), we have $\|\mathbf{v} - \mathbf{v}|_{\mathcal{T}_N(\mathbf{v})}\|_{\ell_2(\mathcal{J})} \le \eta$ for some $N \lesssim \|\mathbf{v}\|_{\mathcal{A}^s_{\text{tree}}}^{1/s}\eta^{-1/s}$. Therefore, $\|\mathbf{w} - \mathbf{w}|_{\mathcal{T}_N(\mathbf{v})}\|_{\ell_2(\mathcal{J})} \le 2\eta$ so that by the definition of the $(2C^*\eta, C^*)$-near best tree $\mathcal{T}(2C^*\eta, \mathbf{w})$, we have

$$\#\mathcal{T}(2C^*\eta, \mathbf{w}) \le C^*\#\mathcal{T}^*(2\eta, \mathbf{w}) \le C^*\#\mathcal{T}_N(\mathbf{v}) \le C^*N \lesssim \|\mathbf{v}\|_{\mathcal{A}^s_{\text{tree}}}^{1/s}\eta^{-1/s}, \tag{6.12}$$

which proves the first estimate in (6.9). It remains to prove (6.10). We wish to show that for each $\delta > 0$ there exists a tree $\mathcal{T}_\delta$ such that

$$\|\bar{\mathbf{w}}_\eta - \bar{\mathbf{w}}_\eta|_{\mathcal{T}_\delta}\|_{\ell_2(\mathcal{J})} \le \delta \quad \text{and} \quad \#(\mathcal{T}_\delta) \lesssim \|\mathbf{v}\|_{\mathcal{A}^s_{\text{tree}}}^{1/s}\delta^{-1/s}. \tag{6.13}$$

The existence of such a tree has been already established for $\delta = \eta$. Now consider first the case $\delta \le 2(1+2C^*)\eta$. In this case, we take $\mathcal{T}_\delta := \operatorname{supp}\bar{\mathbf{w}}_\eta = \mathcal{T}(2C^*\eta, \mathbf{w})$. In fact, since then $\|\bar{\mathbf{w}}_\eta - \bar{\mathbf{w}}_\eta|_{\mathcal{T}_\delta}\|_{\ell_2(\mathcal{J})} = 0$, the first relation in (6.13) holds trivially. Moreover, since $\#(\mathcal{T}_\delta) = \#(\mathcal{T}(2C^*\eta, \mathbf{w})) \lesssim \|\mathbf{v}\|_{\mathcal{A}^s_{\text{tree}}}^{1/s}\eta^{-1/s} \lesssim \|\mathbf{v}\|_{\mathcal{A}^s_{\text{tree}}}^{1/s}\delta^{-1/s}$, the second estimate in (6.13) is also valid.

Now consider the case $\delta > 2(1+2C^*)\eta$. Let $\mathcal{T}_\delta := \mathcal{T}_N(\mathbf{v})$ be a best N-term tree for $\mathbf{v}$, where N will be chosen in a moment. Note that

$$\begin{aligned} \|\bar{\mathbf{w}}_\eta - \bar{\mathbf{w}}_\eta|_{\mathcal{T}_\delta}\|_{\ell_2(\mathcal{J})} &\le \|(\bar{\mathbf{w}}_\eta - \mathbf{v})|_{\mathcal{J}\setminus\mathcal{T}_\delta}\|_{\ell_2(\mathcal{J})} + \|\mathbf{v} - \mathbf{v}|_{\mathcal{T}_N(\mathbf{v})}\|_{\ell_2(\mathcal{J})} \\ &\le (1+2C^*)\eta + \|\mathbf{v} - \mathbf{v}|_{\mathcal{T}_N(\mathbf{v})}\|_{\ell_2(\mathcal{J})} \le \delta, \end{aligned} \tag{6.14}$$

provided that $\delta-(1+2C^*)\eta \ge \|\mathbf{v}-\mathbf{v}|_{\mathcal{T}_N(\mathbf{v})}\|_{\ell_2(\mathcal{J})}$. This holds for some $N \lesssim \|\mathbf{v}\|_{\mathcal{A}^s_{\rm tree}}^{1/s}(\delta-(1+2C^*)\eta)^{-1/s} \le 2^{1/s}\|\mathbf{v}\|_{\mathcal{A}^s_{\rm tree}}^{1/s}\delta^{-1/s}$, which confirms (6.13) also in the case $\delta > 2(1+2C^*)\eta$. This finishes the proof. □

6.3. The key requirement. Up to this point, we have not imposed any conditions on the subroutine **RES**, which is used to approximate the residual at each iteration. In later realizations of the routine **RES**, the support of its output will have tree structure which we will therefore assume to hold from now on without further mention. We will now introduce a condition, called s^*-sparsity, motivated by the analysis of the previous section for the linear case; see Proposition 5.2. We will then show that whenever **RES** is s^*-sparse, the algorithm **SOLVE** is optimal in its rate/complexity for the range of error decay rates $\sigma^{\rm tree}_{N,\ell_2(\mathcal{J})}(\mathbf{u}) \lesssim N^{-s}$ with $s < s^*$. The subsequent section will then show how to construct s^*-sparse routines for nonlinear problems.

We say that the scheme **RES** used to approximate residuals is *s^*-sparse* if, in addition to (4.1), the following property holds.

***s^**-sparsity.** *Whenever the exact solution* $\mathbf{u}$ *of* (2.8) *belongs to* $\mathcal{A}^s_{\rm tree}$ *for some* $s < s^*$*, then one has for any finitely supported input* $\mathbf{v}$ *and any tolerance* $\eta > 0$ *that the output* $\mathbf{w}_\eta := \mathbf{RES}\,[\eta, \mathbf{B}, \mathbf{F}, \mathbf{f}, \mathbf{v}]$ *satisfies*

$$\begin{aligned} \#\mathrm{supp}\,\mathbf{w}_\eta &\le C\eta^{-1/s}(\|\mathbf{v}\|_{\mathcal{A}^s_{\rm tree}}^{1/s} + \|\mathbf{u}\|_{\mathcal{A}^s_{\rm tree}}^{1/s} + 1), \\ \|\mathbf{w}_\eta\|_{\mathcal{A}^s_{\rm tree}} &\le C\left(\|\mathbf{v}\|_{\mathcal{A}^s_{\rm tree}} + \|\mathbf{u}\|_{\mathcal{A}^s_{\rm tree}} + 1\right), \end{aligned} \tag{6.15}$$

where C *depends only on* s *when* $s \to s^*$*. Moreover, the number of operations needed to compute* $\mathbf{w}_\eta$ *stays proportional to* $C(\eta^{-1/s}(\|\mathbf{v}\|_{\mathcal{A}^s_{\rm tree}}^{1/s} + \|\mathbf{u}\|_{\mathcal{A}^s_{\rm tree}}^{1/s} + 1) + \#\mathcal{T}(\mathrm{supp}\,\mathbf{v}))$*, where again* $\mathcal{T}(\mathrm{supp}\,\mathbf{v})$ *denotes the smallest tree containing* $\mathrm{supp}\,\mathbf{v}$.

The occurrence of $\|\mathbf{u}\|_{\mathcal{A}^s_{\rm tree}}$ in the above estimates is already plausible from the linear case, as explained in Remark 5.1.

It will be understood in what follows that **TCOARSE** is used as **CCOARSE** and that a proper initialization is used that complies if necessary with the requirements on the quality of the initial guess (see section 5) so that, in particular, the respective variant of the iteration (2.11) satisfies (3.17).

Under these premises we now show that s^*-sparsity implies asymptotically optimal complexity of the scheme **SOLVE**.

THEOREM 6.1. *Assume that the scheme* **RES** *is* s^**-sparse for some* $s^* > 0$*. If the exact solution* $\mathbf{u}$ *of* (2.8) *belongs to* $\mathcal{A}^s_{\rm tree}$ *for some* $s < s^*$*, then the approximations* $\bar{\mathbf{u}}(\epsilon)$ *satisfy for every target accuracy* $\epsilon > 0$

$$\|\mathbf{u} - \bar{\mathbf{u}}(\epsilon)\|_{\ell_2(\mathcal{J})} \le \epsilon, \tag{6.16}$$

while

$$\#\mathrm{supp}\,\bar{\mathbf{u}}(\epsilon) \le C\epsilon^{-1/s}\|\mathbf{u}\|_{\mathcal{A}^s_{\rm tree}}^{1/s}, \quad \|\bar{\mathbf{u}}(\epsilon)\|_{\mathcal{A}^s_{\rm tree}} \le C\|\mathbf{u}\|_{\mathcal{A}^s_{\rm tree}}, \tag{6.17}$$

where the constant C *depends only on* s *when* $s \to s^*$*. Moreover, the number of operations needed to compute* $\bar{\mathbf{u}}(\epsilon)$ *remains bounded by* $C\epsilon^{-1/s}\|\mathbf{u}\|_{\mathcal{A}^s_{\rm tree}}^{1/s}$.

Proof. The first part follows directly from Proposition 4.3. From (4.15) we know that the result $\tilde{\mathbf{v}}$ after at most K perturbed iterations in the $(j+1)$st block of step (ii) in **SOLVE** satisfies $\|\mathbf{u}-\tilde{\mathbf{v}}\|_{\ell_2(\mathcal{J})} \le \epsilon_j/(2(1+2C^*))$. Now Proposition 6.3 ensures that then

$$\|\bar{\mathbf{u}}^{j+1}\|_{\mathcal{A}^s_{\text{tree}}} \le C\|\mathbf{u}\|_{\mathcal{A}^s_{\text{tree}}}, \quad \#\text{supp}\,\bar{\mathbf{u}}^{j+1} \le C\epsilon_j^{-1/s}\|\mathbf{u}\|^{1/s}_{\mathcal{A}^s_{\text{tree}}}. \tag{6.18}$$

Moreover, the computational work required by the routine **TCOARSE** stays, by Proposition 6.2, proportional to the support size of $\tilde{\mathbf{v}}$, since the support of $\tilde{\mathbf{v}}$, as an output of **RES**, has tree structure. Here it is important to note that the constant C is *independent* of the input $\tilde{\mathbf{v}}$ of **TCOARSE**. Thus for $j > 0$ the input of the first application of **RES** in step (ii) of **SOLVE** satisfies (6.18). Since there are only at most a uniformly bounded number K of applications of **RES** in each iteration block, $\tilde{\mathbf{v}}$ also satisfies (6.18) with a constant depending now on the number of updates (ii.1) bounded by K; see (4.14). (Here we have tacitly assumed that the initial guess has been subjected to a **TCOARSE** so that it also satisfies (6.18). Otherwise, we would have to add $\#\text{supp}\,\mathbf{u}^0$ to the above estimates.) The estimate in (6.17) now follows from these estimates for the terminal value of j. This also shows that the number of operations remains proportional to $\#\text{supp}\,\bar{\mathbf{u}}(\epsilon)$. □

We shall discuss below how to obtain schemes **RES** that are s^*-sparse for certain $s^* > 0$ and what limits the value of s^*.

7. Nonlinear evaluation schemes.

Just as the efficient application of compressible matrices $\mathbf{A}$ was pivotal for the adaptive solution in the linear case (L), we need efficient evaluation schemes for $\mathbf{F}(\mathbf{v})$ that allow us to realize the residual approximation in **RES**. Such a scheme has been already proposed in [10] for a class of nonlinearities F that will be described next.

7.1. A class of nonlinear mappings.

We shall be concerned with nonlinear operators of the form

$$V = (v_1, \ldots, v_n) \mapsto w = F(D^{\alpha_1}v_1, \ldots, D^{\alpha_n}v_n), \tag{7.1}$$

acting from $\mathcal{H} \times \cdots \times \mathcal{H}$ to the dual $\mathcal{H}'$. (Here $\alpha_i = (\alpha_{i,1}, \ldots, \alpha_{i,d})$ are multi-indices.) This clearly covers our previous example of a single argument $n = 1$ but also further important cases like the nonlinearity appearing in the Navier–Stokes equations. Although we shall not address variational problems involving nonlinearities of several arguments in this paper, we shall present the evaluation schemes in this somewhat greater generality because they are important for such applications and because we shall apply the case of two arguments later in connection with Newton's scheme.

We shall first describe our requirements on F in the wavelet coordinate domain and point out later circumstances under which these requirements are met.

Denoting by $\mathbf{v}_i = (v_{i,\lambda})$ the arrays of the wavelet coefficients of the function v_i, setting $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_n)$, and defining the corresponding discrete mapping $\mathbf{F}$ by

$$\mathbf{F}(\mathbf{V}) := (\langle \psi_\lambda, F(D^{\alpha_1}v_1, \ldots, D^{\alpha_n}v_n)\rangle)_{\lambda \in \mathcal{J}}, \tag{7.2}$$

we make the following basic assumptions.

Assumption 1. $\mathbf{F}$ *is a Lipschitz map from* $(\ell_2(\mathcal{J}))^n$ *into* $\ell_2(\mathcal{J})$,

$$\|\mathbf{F}(\mathbf{U}) - \mathbf{F}(\mathbf{V})\|_{\ell_2(\mathcal{J})} \le C\sum_{i=1}^{n}\|\mathbf{u}_i - \mathbf{v}_i\|_{\ell_2(\mathcal{J})}, \tag{7.3}$$

with $C = C(\max_i\{\|\mathbf{u}_i\|_{\ell_2(\mathcal{J})}, \|\mathbf{v}_i\|_{\ell_2(\mathcal{J})}\})$*, where* $x \mapsto C(x)$ *is a positive nondecreasing function.*

Bearing the norm equivalences (2.5), (2.6) in mind, we see that property P1 from section 3.1 is a special case of Assumption 1 for $n = 1$.

Assumption 2. *There exists a constant* $\gamma > d/2$ *such that, for any finitely supported* $\mathbf{V}$ *(i.e., with all* $\mathbf{v}_i$ *finitely supported) and* $\mathbf{w} = \mathbf{F}(\mathbf{V})$*, we have for* $\lambda \in \mathcal{J}_\psi$ *the estimate*

$$|w_\lambda| \le C \sup_{\mu\,:\,S_\lambda \cap S_\mu \neq \emptyset} \left(\sum_{i=1}^n |v_{i,\mu}|\right) 2^{-\gamma(|\lambda|-|\mu|)}, \tag{7.4}$$

where $C = C(\max_i \|\mathbf{v}_i\|_{\ell_2(\mathcal{J})})$ *and* $x \mapsto C(x)$ *is a positive nondecreasing function.*

The parameter γ plays a similar role as the compressibility range s^* for the wavelet representation of linear operators.

Nonlinear mappings that satisfy the above assumptions are, for instance, those with polynomial growth. In the special case of a single argument, a typical condition reads

$$|F^{(k)}(v)| \lesssim (1+|v|)^{(p-k)_+}, \quad k = 0, \dots, n^*, \qquad \text{for some} \;\; n^* \ge 1, \tag{7.5}$$

where $a_+ := \max\{0, a\}$.

Remark 7.1. Suppose that $\mathcal{H} = H^t$ is a Sobolev space of smoothness order $t > 0$ (or a closed subspace determined by, e.g., homogeneous boundary conditions). One can show that, for the special case $n = 1$, (7.5) implies Assumptions 1 and 2 and, in particular, property P1, with no condition on p when $t \ge d/2$ and otherwise provided that

$$1 \le p < p^* := \frac{d+2t}{d-2t}; \tag{7.6}$$

see [10].

In the general case, when the nonlinear map F has the form $F(D^{\alpha_1}u_1, \dots, D^{\alpha_n}u_n)$, we impose the growth condition

$$|D^\beta F(x_1, \dots, x_n)| \le C \prod_{i=1}^n (1+|x_i|)^{[p_i-\beta_i]_+}, \;\; |\beta| = 0, 1, \dots, n^*, \tag{7.7}$$

for some $p_i \ge 0$ and n^* a positive integer. The following fact (covering Remark 7.1) has been proven in [10].

Theorem 7.1. *Suppose that* $\mathcal{H} = H^t$ *is a closed subspace (determined by, e.g., homogeneous boundary conditions) of* $H^t(\Omega)$ *for some* $t \ge 0$*. Assume that the growth assumptions* (7.7) *hold at least with* $n^* = 0$*. Then* F *maps* $\mathcal{H} \times \cdots \times \mathcal{H}$ *to* $\mathcal{H}'$ *whenever* $t \ge 0$ *satisfies*

$$\left(\frac{1}{2} - \frac{t}{d}\right)_+ + \sum_{i=1}^n p_i \left(\frac{1}{2} - \frac{t}{d} + \frac{|\alpha_i|}{d}\right)_+ < 1. \tag{7.8}$$

If in addition $n^* = 1$*, then we also have under the same restriction*

$$\|F(u) - F(v)\|_{\mathcal{H}'} \le C \sum_{i=1}^n \|u_i - v_i\|_{\mathcal{H}}, \tag{7.9}$$

where $C = C(\max_i\{\|u_i\|_{\mathcal{H}}, \|v_i\|_{\mathcal{H}}\})$ *and* $x \to C(x)$ *is nondecreasing, and therefore, on account of* (2.5)*, Assumption* 1 *holds.*

For the verification of Assumption 2, we treat separately the polynomial case for which we have the growth condition

$$(7.10) \qquad |D^\beta F(x_1,\dots,x_n)| \le C\prod_{i=1}^{n}(1+|x_i|)^{p_i-\beta_i}, \ \ \beta_i \le p_i,$$

and $D^\beta F = 0$ if $\beta_i > p_i$ for some i, where the p_i are positive integers. We recall the following result from [10].

Theorem 7.2. *Assume that the wavelets belong to* C^m *and have vanishing moments of order* m *(i.e., are orthogonal to* $\mathbb{P}_{m-1}$ *the space of polynomials of total degree at most* $m-1$*) for some positive integer* m*. Then Assumption* 2 *holds for* $\gamma = r + t + d/2$ *with the following values of* r*:*

(i) *If* F *satisfies* (7.7) *with* p *such that* $\sum_{i=1}^{n} p_i(d/2 - t + |\alpha_i|)_+ < d/2 + t$*, then* $r = \lceil \min\{m, n^*, p^*\}\rceil$*, where* $p^* = \min\{p_i \ : \ i \text{ s.t. } d/2 - t + |\alpha_i| > 0\}$.

(ii) *If* F *satisfies* (7.10) *with* p *such that* $\sum_{i=1}^{n} p_i(d/2 - t + |\alpha_i|)_+ < d/2 + t$*, then* $r = m$.

7.2. An adaptive evaluation scheme. The schemes for approximating $\mathbf{F}(\mathbf{V})$ consist of two conceptual steps: (i) the prediction of a possibly small set of indices that covers the significant coefficients of $\mathbf{F}(\mathbf{V})$ using only knowledge of the indices of the significant coefficients of the input $\mathbf{V}$; (ii) a sufficiently accurate computation of those coefficients of $\mathbf{F}(\mathbf{V})$ that correspond to the predicted index set. Once the predicted sets are known, one can invoke the techniques developed in [16] to tackle the latter task. A more detailed treatment of this issue will be given elsewhere. Motivated by the results in [16], we shall work in what follows with the following assumption.

Assumption E. *The entries* $w_\lambda = \mathbf{F}(\mathbf{v})_\lambda$ *can be computed (with sufficient accuracy) on average at unit cost.*

Therefore, we shall concentrate here only on task (i) under the assumption that the computation can be done in linear time; see the discussion of this issue in [16]. We recall now briefly the construction from [10] of good predictions for mappings satisfying Assumptions 1 and 2. For $j = 0, 1, \dots,$ and each component $\mathbf{v}_i$ of $\mathbf{V}$ we define the near best trees introduced in section 6.2

$$(7.11) \qquad \mathcal{T}_{i,j} := \mathcal{T}\left(\frac{2^j\epsilon}{1+j}, \mathbf{v}_i\right),$$

by invoking the thresholding algorithm from [3]. Moreover, it can be shown that for any tree $\mathcal{T}$ there exists an expansion $\tilde{\mathcal{T}}$ such that for some constant C one has $\#\tilde{\mathcal{T}} \le C\#\mathcal{T}$, while for any $\lambda \in \mathcal{L}(\tilde{\mathcal{T}})$ the number of $\mu \in \mathcal{L}(\tilde{\mathcal{T}})$ such that $S_\lambda \cap S_\mu \ne \emptyset$ is bounded by C; see Lemma 3.1 in [10]. We denote these expansions of $\mathcal{T}_{i,j}$ by $\tilde{\mathcal{T}}_{i,j}$ and set

$$(7.12) \qquad \tilde{\mathcal{T}}_j := \bigcup_{i=0}^{n} \tilde{\mathcal{T}}_{i,j} \ \text{ and } \ \Delta_j := \tilde{\mathcal{T}}_j \setminus \tilde{\mathcal{T}}_{j+1}.$$

In order to build a tree which will be adapted to $\mathbf{w} = \mathbf{F}(\mathbf{V})$, we introduce

$$(7.13) \qquad \alpha := \frac{2}{2\gamma - d} > 0,$$

where γ is the constant in (7.4), and for each $\mu \in \Delta_j$, we define the *influence set*

$$\Lambda_{\epsilon,\mu} := \{\lambda \ : \ S_\lambda \cap S_\mu \neq \emptyset \text{ and } |\lambda| \leq |\mu| + \alpha j\}. \tag{7.14}$$

We then define $\mathcal{T}$ by

$$\mathcal{T}_\epsilon(\mathbf{F}, \mathbf{V}) = \mathcal{T} := \mathcal{J}_\phi \cup \left(\cup_{\mu \in \tilde{\mathcal{T}}_0} \Lambda_{\epsilon,\mu} \right). \tag{7.15}$$

The following fact has been shown in [10, Theorem 5.1].

Theorem 7.3. *Assume that F satisfies Assumptions* 1 *and* 2. *Given any $\mathbf{V} \in (\ell_2(\mathcal{J}))^n$, we have the error estimate*

$$\|\mathbf{F}(\mathbf{V}) - \mathbf{F}(\mathbf{V})|_{\mathcal{T}}\|_{\ell_2(\mathcal{J})} \ \lesssim \ \epsilon. \tag{7.16}$$

Moreover, if $\mathbf{V} \in (\mathcal{A}^s_{\rm tree})^n$ for some $s \in \left(0, \frac{2\gamma - d}{2d}\right)$, we have the estimate

$$\#(\mathcal{T}) \ \lesssim \ \|\mathbf{V}\|^{1/s}_{(\mathcal{A}^s_{\rm tree})^n} \epsilon^{-1/s} + \#(\mathcal{J}_\phi). \tag{7.17}$$

We therefore have $\mathbf{F}(\mathbf{V}) \in \mathcal{A}^s_{\rm tree}$ and

$$\|\mathbf{F}(\mathbf{V})\|_{\mathcal{A}^s_{\rm tree}} \ \lesssim \ 1 + \|\mathbf{V}\|_{(\mathcal{A}^s_{\rm tree})^n}. \tag{7.18}$$

The constants in these above inequalities depend only on $\|\mathbf{V}\|_{\ell_2(\mathcal{J})}$, the space dimension d, and the parameter s.

This suggests the following routine for approximating $\mathbf{F}(\mathbf{V})$ for any finitely supported vector $\mathbf{V}$:

EV $[\epsilon, \mathbf{F}, \mathbf{V}] \to \mathbf{w}_\epsilon$. *Given the inputs $\epsilon > 0$ and $\mathbf{V}$ with finite support do:*

Step 1: *Invoke* **TCOARSE** $[2^j\epsilon/(1+j), \mathbf{v}_i]$, $i = 1, \dots, n$, *to compute the trees*

$$\mathcal{T}_{i,j} := \mathcal{T}\left(\frac{2^j \epsilon}{C_0(j+1)}, \mathbf{v}_i \right), \tag{7.19}$$

where $C_0 = C_0(\|\mathbf{v}\|)$ is the constant involved in (7.16) *for $j = 0, \dots, J$, and stop for the smallest J such that $\mathcal{T}_J$ is empty (we always have $J \ \lesssim \ \log_2(\|\mathbf{V}\|_{\ell_2(\mathcal{J})}/\epsilon)$).*

Step 2: *Derive the expanded trees $\tilde{\mathcal{T}}_{i,j}$, the layers Δ_j, and the outcome tree $\mathcal{T} = \mathcal{T}_\epsilon(\mathbf{F}, \mathbf{V})$ according to* (7.15).

Step 3: *Compute the coefficients $\mathbf{F}(\mathbf{V})_\lambda$, $\lambda \in \mathcal{T}$; see* [16].

A more detailed discussion of this scheme can be found in [10].

We can now formulate concrete realizations of the scheme **SOLVE** that are suitable for nonlinear problems. We shall use **TCOARSE** as our version of **CCOARSE**. Moreover, for the semilinear elliptic problem (3.4) we can take

$$\begin{aligned} \mathbf{RES}_{\rm ell}[\eta, \alpha\mathbf{I}, \mathbf{A}, \mathbf{G}, \mathbf{f}, \mathbf{v}] := \alpha\,(&\mathbf{APPLY}\,[\eta/3, \mathbf{A}, \mathbf{v}] \\ &+ \mathbf{EV}\,[\eta/3, \mathbf{G}, \mathbf{v}] - \mathbf{RHS}\,[\eta/3, \mathbf{f}]), \end{aligned} \tag{7.20}$$

where **RHS** is defined here as in (5.2) but with **COARSE** replaced by **TCOARSE**.

For the general nonlinear problem (GNL), we shall now devise a residual approximation for the scheme from section 3.2 with $\mathbf{B}_n := \mathbf{B}$. Suppose that $\|\mathbf{B}\|_{\ell_2(\mathcal{J}) \to \ell_2(\mathcal{J})} \leq C_B$, and set

$$\begin{aligned} \mathbf{RES}_{\rm lc}[\eta, \mathbf{B}, \mathbf{F}, \mathbf{f}, \mathbf{v}] := \mathbf{APPLY}\ \big[&\eta/2, \mathbf{B}, \big(\mathbf{EV}\ [\eta/4C_B, \mathbf{F}, \mathbf{v}] \\ &-\mathbf{RHS}\,[\eta/4C_B, \mathbf{f}]\big)\big]. \end{aligned} \tag{7.21}$$

Of course, we have assumed here that the matrix $\mathbf{B}$ is compressible. In particular, for the stationary choice $\mathbf{B} = D\mathbf{R}(\mathbf{u}^0)$ this might be expected to be the case. We shall return to this issue later.

7.3. Complexity estimates. We have just explained how to obtain concrete realizations of the scheme **SOLVE** in each of the three cases (L), (SL), (GNL). The remainder of this section will be devoted to giving a complexity analysis of these schemes. The following theorem summarizes the properties of Algorithm **EV**; see [10, Theorem 3.4].

THEOREM 7.4. *Given the inputs $\epsilon > 0$, a nonlinear function F such that $\mathbf{F}$ satisfies Assumptions* 1 *and* 2*, and a finitely supported vector $\mathbf{V}$, then the output tree $\mathcal{T}$ has the following properties:*

P1. $\|\mathbf{F}(\mathbf{V}) - \mathbf{F}(\mathbf{V})|_{\mathcal{T}}\| \leq \epsilon$.

P2. *For any $0 < s < \frac{2\gamma-d}{2d}$ (see Theorem* 7.3*),*

$$\#(\mathcal{T}) \leq C\|\mathbf{V}\|^{1/s}_{(\mathcal{A}^s_{\text{tree}})^n}\epsilon^{-1/s} + \#(\mathcal{J}_\phi) =: N_\epsilon \tag{7.22}$$

with C a constant depending only on the constants appearing in Theorem 7.3*.*

P3. *Moreover, the number of computations needed to find $\mathcal{T}$ is bounded by $C(N_\epsilon + \#\mathcal{T}(\mathbf{V}))$, where N_ϵ is the right-hand side of* (7.22) *and $\mathcal{T}(\mathbf{V})$ is the smallest tree containing* $\operatorname{supp}\mathbf{v}$*.*

Recall that in the case of semilinear equations, $\mathbf{R}$ involves a linear and a nonlinear operator as in (3.2) or (3.4). Also recall from Theorem 5.1 that when the wavelet representation $\mathbf{A}$ of the linear operator $\mathcal{A}$ defined by

$$\langle w, \mathcal{A}v\rangle = a(v,w), \quad v,w \in \mathcal{H} = H^t,$$

belongs to $\mathcal{C}_{s^*}$, then $\mathbf{A}$ is bounded on $\mathcal{A}^s$ with $s < s^*$. However, when nonlinear operators are also involved, Theorem 6.1 tells us that the spaces $\mathcal{A}^s_{\text{tree}}$ should now play the role of $\mathcal{A}^s$. Thus we shall prove in Proposition 7.4 below the boundedness of $\mathbf{A}$ with respect to this slightly stronger norm.

To prepare for Proposition 7.4, we make some remarks.

Remark 7.2. It has been shown in [12] that when $\mathcal{A}$ is a local operator, i.e., $\langle v, \mathcal{A}w\rangle = 0$ whenever $|\operatorname{supp} v \cap \operatorname{supp} w| = 0$, and when $\mathcal{A}$ is still bounded as a mapping from H^{t+a} to H^{-t+a} for $a \leq m + d/2$, where m is less than or equal to the order of differentiability and vanishing moments of the wavelets ψ_λ, then one has

$$|a(\psi_\lambda, \psi_\nu)| \lesssim 2^{-\sigma||\lambda|-|\nu||}, \quad \sigma = t + m + d/2. \tag{7.23}$$

Moreover, we know from Proposition 3.4 in [8] that in this case $\mathbf{A}$ belongs to $\mathcal{C}_{s^*}$ with

$$s^* = \frac{t+m}{d} = \frac{2\sigma - d}{2d}. \tag{7.24}$$

Note that σ agrees with the value of γ in Assumption 2 or at least depends in an analogous way on the spatial dimension, the order of the operator, and the order of the vanishing moments and the smoothness of the wavelets; see Theorem 7.2 (ii). Moreover, the condition $s < s^*$ with s^* from (7.24) agrees with the constraint on s from Theorem 7.3 formulated in terms of γ.

Remark 7.3. One can show that for piecewise polynomial wavelets, s^* can be chosen larger than in (7.24); see [2].

By definition one has $\mathcal{A}^s_{\text{tree}} \subset \mathcal{A}^s$. We shall need the following refinement of Theorem 5.1.

PROPOSITION 7.4. *Under the assumptions from Remark* 7.2 *on the linear part $\mathcal{A}$, let*

$$\sigma = m + t + d/2. \tag{7.25}$$

Then one has for $s < \frac{2\sigma - d}{2d}$

$$\|\mathbf{A}\mathbf{v}\|_{\mathcal{A}^s_{\text{tree}}} \lesssim \|\mathbf{v}\|_{\mathcal{A}^s_{\text{tree}}}, \quad \mathbf{v} \in \mathcal{A}^s_{\text{tree}}; \tag{7.26}$$

that is, $\mathbf{A}$ *maps* $\mathcal{A}^s_{\text{tree}}$ *boundedly into itself.*

Proof. By assumption (3.1), $\mathcal{A}$ is a topological isomorphism from $\mathcal{H}$ onto $\mathcal{H}'$ and obviously satisfies Assumption 1. We need to show that Assumption 2 holds for all $\gamma' < \sigma$, defined by (7.25). To this end, note that (7.23) provides

$$\begin{aligned} |(\mathbf{A}\mathbf{v})_\lambda| &\le \sum_{\substack{|\nu| \le |\lambda|, \\ S_\nu \cap S_\lambda \ne \emptyset}} |v_\nu| 2^{-\sigma(|\lambda|-|\nu|)} + \sum_{\substack{|\nu| > |\lambda|, \\ S_\nu \cap S_\lambda \ne \emptyset}} |v_\nu| 2^{-\sigma(|\lambda|-|\nu|)} 2^{-2\sigma(|\nu|-|\lambda|)} \\ &\le \sup_{\substack{|\nu| \le |\lambda|, \\ S_\nu \cap S_\lambda \ne \emptyset}} 2^{-\gamma'(|\lambda|-|\nu|)} |v_\nu| \sum_{\substack{|\nu| \le |\lambda|, \\ S_\nu \cap S_\lambda \ne \emptyset}} 2^{-(\sigma-\gamma')(|\lambda|-|\nu|)} \\ &\quad + \sup_{\substack{|\nu| > |\lambda|, \\ S_\nu \cap S_\lambda \ne \emptyset}} |v_\nu| 2^{-\sigma(|\lambda|-|\nu|)} \sum_{\substack{|\nu| > |\lambda|, \\ S_\nu \cap S_\lambda \ne \emptyset}} 2^{-2\sigma(|\nu|-|\lambda|)}. \end{aligned}$$

We now check that both sums appearing on the right-hand side are bounded independently of λ provided that $\gamma' < \sigma$. Indeed, in the first sum, for any $k > |\lambda|$, there is a bounded number C_0 of indices ν with $|\nu| = k$ such that $S_\nu \cap S_\lambda \ne \emptyset$. Hence this sum is bounded by $C_0 \sum_{j=0}^{\infty} 2^{-j(\sigma-\gamma')}$. For the second sum, note that for $|\nu| = |\lambda| + k$, there are at most $C_0 2^{kd}$ indices ν for which $S_\nu \cap S_\lambda \ne \emptyset$. Since, by (7.25), $2\sigma > d$, this sum can also be bounded by a geometric series. We have thus verified Assumption 2 for all $\gamma' < \sigma$. The assertion now follows from Theorem 7.3 and the restriction on s given in Theorem 7.3. However, in contrast to the general nonlinear case, we can dispense here with the constant term in the right-hand side of (7.26); see Remark 3.2 in [10]. □

Note that the support of the output of the scheme **APPLY** from section 5.2 generally does not have tree structure. In order to ensure that the output of **RES** complies with our previous assumption that its support has tree structure, we shall employ the following modification of **APPLY** while keeping the notation unchanged. First, the original version of **APPLY** is carried out with target accuracy $\eta/2$. We then apply **TCOARSE** to the output with target accuracy $\eta/2$ so that (5.9) is still valid.

We shall make use of the following consequence of Proposition 7.4.

COROLLARY 7.5. *Under the same assumptions as in Proposition* 7.4 *let* $s < s^*$. *Then* $\mathbf{w}_\eta = \mathbf{APPLY}\,[\eta, \mathbf{A}, \mathbf{v}]$ *satisfies*

(i) $\#\text{flops} \lesssim \#\mathcal{T}(\operatorname{supp}\mathbf{v}) + \|\mathbf{v}\|^{1/s}_{\mathcal{A}^s_{\text{tree}}} \eta^{-1/s}$, $\#\operatorname{supp}\mathbf{w}_\eta \lesssim \|\mathbf{v}\|^{1/s}_{\mathcal{A}^s_{\text{tree}}} \eta^{-1/s}$;

(ii) $\|\mathbf{w}_\eta\|_{\mathcal{A}^s_{\text{tree}}} \lesssim \|\mathbf{v}\|_{\mathcal{A}^s_{\text{tree}}}$.

Proof. Let $\hat{\mathbf{w}}$ denote the output of the original version of **APPLY** with target accuracy $\eta/2$. The first estimate in (i) for $\hat{\mathbf{w}}$ follows directly from Theorem 5.1 (ii) and the fact that $\|\cdot\|_{\mathcal{A}^s} \lesssim \|\cdot\|_{\mathcal{A}^s_{\text{tree}}}$ even with $\#\mathcal{T}(\operatorname{supp}\mathbf{v})$ replaced by $\operatorname{supp}\mathbf{v}$. Now the cost of the subsequent application of **TCOARSE** with target accuracy $\eta/2$, yielding $\mathbf{w}_\eta$, is, by Proposition 6.2, proportional to $\#\mathcal{T}(\operatorname{supp}\mathbf{v})$. This confirms the first estimate in (i). Next note that, since the chunks $(\mathbf{v}_{[j]} - \mathbf{v}_{[j-1]})$ have disjoint supports, for each j the vector $\mathbf{w}_j$, defined by (5.7), can be interpreted as $\mathbf{w}_j = \mathbf{C}^{(j)}\mathbf{v}$, where the matrix $\mathbf{C}^{(j)}$ is a compressed version of $\mathbf{A}$ defined as follows. All columns with indices outside $\operatorname{supp}\mathbf{v}_{[j]}$ are zero. The columns of $\mathbf{C}^{(j)}$ whose indices belong to $\operatorname{supp}(\mathbf{v}_{[k]} - \mathbf{v}_{[k-1]})$, $k \le j$, agree with the corresponding columns in the matrix

$\mathbf{A}_{j-k}$. Therefore, since the $\mathbf{A}_j$ are derived from $\mathbf{A}$ by replacing certain entries by zero, we conclude that $\mathbf{C}^{(j)}$ is obtained by replacing certain entries in $\mathbf{A}$ by zero. Thus the $\mathbf{C}^{(j)}$ still satisfy (7.23) uniformly in j. Assumptions 1 and 2 also remain valid. Thus Proposition 7.4 can be applied to $\mathbf{C}^{(j)}$ with constants independent of j, which provides $\|\hat{\mathbf{w}}\|_{\mathcal{A}^s_{\text{tree}}} \lesssim \|\mathbf{v}\|_{\mathcal{A}^s_{\text{tree}}}$. Therefore, one also has $\|\mathbf{w}_\eta\|_{\mathcal{A}^s_{\text{tree}}} \lesssim \|\mathbf{v}\|_{\mathcal{A}^s_{\text{tree}}}$, which is (ii). Since **TCOARSE** produces a near best tree approximation, we conclude that $\#\operatorname{supp}\mathbf{w}_\eta \lesssim \|\hat{\mathbf{w}}\|^{1/s}_{\mathcal{A}^s_{\text{tree}}}\eta^{-1/s}$, which, in view of the previous remark, confirms the second estimate in (i) and finishes the proof. □

We have now collected all the ingredients needed to confirm s^*-sparsity of the residual approximations defined before. We start with (7.20) for the case (SL).

Corollary 7.6. *Let γ be the parameter given in Theorem* 7.2 *for the respective nonlinear mapping G. Suppose that σ, defined by* (7.25), *satisfies $\sigma \geq \gamma$. Then* $\mathbf{RES}_{\text{ell}}$, *defined by* (7.20), *is s^*-sparse with $s^* := \frac{2\gamma-d}{2d}$.*

Proof. We have to verify the validity of (6.15) for $s < s^*$. If $\mathbf{u} \in \mathcal{A}^s_{\text{tree}}$ for some $s < s^*$, then Proposition 7.4 and Theorem 7.4 imply that $\mathbf{f} \in \mathcal{A}^s_{\text{tree}}$ and $\|\mathbf{f}\|_{\mathcal{A}^s_{\text{tree}}} \lesssim \|\mathbf{u}\|_{\mathcal{A}^s_{\text{tree}}}$. Hence, since **TCOARSE** satisfies completely analogous properties with regard to $\mathcal{A}^s_{\text{tree}}$ as **COARSE** with respect to $\mathcal{A}^s$ (see [8]), we conclude that the output $\mathbf{f}_\eta$ of $\mathbf{RHS}\,[\eta, \mathbf{f}]$ satisfies

$$\#\operatorname{supp}\mathbf{f}_\eta \lesssim \eta^{-1/s}\|\mathbf{u}\|^{1/s}_{\mathcal{A}^s_{\text{tree}}}, \quad \|\mathbf{f}_\eta\|_{\mathcal{A}^s_{\text{tree}}} \lesssim \|\mathbf{u}\|_{\mathcal{A}^s_{\text{tree}}}. \tag{7.27}$$

Furthermore, Corollary 7.5 says that the output of **APPLY** remains bounded in $\mathcal{A}^s_{\text{tree}}$, while Theorem 7.4 ensures that the same is true for the output of **EV**. Hence, by (7.20), one has for $\mathbf{w}_\eta := \mathbf{RES}_{\text{ell}}[\eta, \alpha\mathbf{I}, \mathbf{A}, \mathbf{G}, \mathbf{f}, \mathbf{v}]$

$$\|\mathbf{w}_\eta\|_{\mathcal{A}^s_{\text{tree}}} \lesssim \left(\|\mathbf{v}\|_{\mathcal{A}^s_{\text{tree}}} + \|\mathbf{u}\|_{\mathcal{A}^s_{\text{tree}}} + 1\right), \tag{7.28}$$

which is the second estimate in (6.15). The first estimate in (6.15) follows also from (7.27), Theorem 7.4, and Corollary 7.5 (i). Since the supports of the outputs of $\mathbf{RES}_{\text{ell}}$ and **TCOARSE** have tree structure, the required bounds for the operations count follow (under Assumption E) from P3 in Theorem 7.4 and Corollary 7.5 (i). This completes the proof. □

Combining Corollary 7.6 with Theorem 6.1 yields the first main result of this paper.

Theorem 7.5. *Under the same assumptions as in Corollary* 7.6, *suppose that the solution $u = \sum_{\lambda\in\mathcal{J}} u_\lambda\psi_\lambda$ satisfies $\mathbf{u} \in \mathcal{A}^s_{\text{tree}}$ for some $s < s^* := (2\gamma - d)/(2d)$. Then the approximate solution $u(\epsilon) = \sum_{\lambda\in\Lambda(\epsilon)} \bar{u}(\epsilon)_\lambda\psi_\lambda$ produced by* **SOLVE** *(with the initialization for the semilinear problem) after finitely many steps satisfies*

$$\|u - u(\epsilon)\|_{\mathcal{H}} \leq C_1\epsilon.$$

Moreover,

$$\#(\text{flops}),\ \#(\Lambda(\epsilon)) \lesssim \epsilon^{-1/s}\big(1 + \|\mathbf{u}\|^{1/s}_{\mathcal{A}^s_{\text{tree}}}\big), \tag{7.29}$$

and

$$\|\bar{\mathbf{u}}(\epsilon)\|_{\mathcal{A}^s_{\text{tree}}} \lesssim \|\mathbf{u}\|_{\mathcal{A}^s_{\text{tree}}}, \tag{7.30}$$

where the constants depend only on $\|\mathbf{u}\|_{\ell_2(\mathcal{J})}$, on the constants in (2.5), (3.8), (3.3), *and on s when $s \to s^*$ or $s \to 0$.*

Let us briefly discuss now the locally convergent scheme from section 2.3. We shall assume that for the general problem (2.2) the nonlinear map in (2.1) satisfies Assumptions 1 and 2. Moreover, we assume that a sufficiently good initial guess $\mathbf{u}^0$ is given so that the error reduction (3.25) holds and that **SOLVE** is initialized accordingly; see section 5.

Corollary 7.7. *Let γ be the parameter given in Theorem* 7.2 *for the respective nonlinear mapping F. Moreover, assume that the matrix $\mathbf{B} = \mathbf{B}_n$ appearing in* (2.11) *satisfies decay estimates like* (7.23) *for some $\sigma \geq \gamma$. Then the scheme $\mathbf{RES}_{\rm lc}$ defined by* (7.21) *is s^*-sparse for $s^* := (2\gamma - d)/(2d)$.*

Proof. The assertion follows again from Proposition 7.4 and Theorem 7.4. □

Corollary 7.8. *Under the assumptions of Corollary* 7.7, *the assertion of Theorem* 7.5 *remains valid for the locally convergent scheme based on $\mathbf{RES}_{\rm lc}$.*

We end this section with analyzing the compressibility of the special choice $\mathbf{B} = D\mathbf{R}(\mathbf{u}^0)^T$. We consider $\mathcal{H} = H^t$ and only nonlinear maps of a single argument $n = 1$ and the subcritical case $t < d/2$, $p < p^*$; recall (7.6). Recall from (2.9) that the entries of $D\mathbf{R}(\mathbf{u}^0)^T = D\mathbf{R}(\mathbf{u}^0)$ have the form $w_{\lambda,\nu} := \langle \psi_\lambda, \psi_\nu R'(u^0)\rangle$. Since, in view of the $\mathcal{H} = H^t$-normalization of the ψ_λ, one has $\|\psi_\lambda\|_{L_\infty} \sim 2^{(\frac{d}{2}-t)|\lambda|}$, the same arguments as used in the proof of Theorem 4.2 in [10] yield

$$|w_{\lambda,\nu}| \;\lesssim\; \|\psi_\lambda\|_{L_\infty} \inf_{P\in\mathbb{P}_r} \|P - \psi_\nu R'(u^0)\|_{L_1(S_\lambda)} \;\lesssim\; 2^{-(r+\frac{d}{2}+t)|\lambda|} |\psi_\nu R'(u^0)|_{W^r(L_\infty(S_\lambda))},$$

where we assume without loss of generality that $|\lambda| \geq |\nu|$. Moreover, we obtain

$$|\psi_\nu R'(u^0)|_{W^r(L_\infty(S_\lambda)} \;\lesssim\; \max_{l\leq r} |\psi_\nu|_{W^{r-l}(L_\infty(S_\lambda))} |R'(u^0)|_{W^l(L_\infty(S_\lambda))}.$$

Abbreviating as before $\sigma = r + \frac{d}{2} + t$, we can estimate the first factor by

$$2^{(\frac{d}{2}-t)|\nu|} 2^{(r-l)|\nu|} = 2^{\sigma|\nu|} 2^{-(2t+l)|\nu|}, \tag{7.31}$$

while the second factor can be estimated along the lines of the proof of Theorem 4.2 in [10] as

$$\begin{aligned} |R'(u^0)|_{W^l(L_\infty(S_\lambda))} \;&\lesssim\; \max_{k=1,\dots,l} \|\mathbf{u}^0\|_{\ell_2(\mathcal{J})}^{m+k-1} \\ &\times \sup_{\mu: S_\mu\cap S_\lambda\cap S_\nu\neq\emptyset} |u^0_\mu| 2^{(l+(p-1)\epsilon)|\mu|} 2^{(p-1)(\frac{d}{2}-t)|\mu|}, \end{aligned} \tag{7.32}$$

where $\epsilon > 0$, $j_i \in \mathbb{N}$, and

$$m := \begin{cases} (p-l-1)_+ & \text{if} \quad \|\mathbf{u}^0\|_{\ell_2(\mathcal{J})} \geq 1, \\ 0 & \text{if} \quad \|\mathbf{u}^0\|_{\ell_2(\mathcal{J})} < 1. \end{cases}$$

As in [10], we can choose ϵ so that

$$2^{(l+(p-1)\epsilon)|\mu|} 2^{(p-1)(\frac{d}{2}-t)|\mu|} \;\lesssim\; 2^{(l+\frac{d}{2}+t)|\mu|} 2^{-\epsilon|\mu|} 2^{-(\frac{d}{2}-t)|\mu|} = 2^{-\epsilon|\mu|} 2^{(l+2t)|\mu|}.$$

Thus, combining (7.31) and (7.32), we obtain

$$|w_{\lambda,\nu}| \;\lesssim\; C(\|\mathbf{u}^0\|_{\ell_2(\mathcal{J})}) 2^{-\sigma||\lambda|-|\nu||} \sup_{\mu: S_\mu\cap S_\lambda\cap S_\nu\neq\emptyset} |u^0_\mu| 2^{-\epsilon|\mu|} 2^{(r+2t)(|\mu|-|\nu|)}. \tag{7.33}$$

Note that the first factor $C(\|\mathbf{u}^0\|_{\ell_2(\mathcal{J})}) 2^{-\sigma||\lambda|-|\nu||}$ represents the same scalewise decay of the entries as in the matrix $\mathbf{A}$ in (7.23). This ensures that $D\mathbf{R}(\mathbf{u}^0)$ belongs to $\mathcal{C}_{s^*}$

with s^* given by (7.24). However, the entries are weighted by additional u^0-dependent factors that could, in principle, become rather large when the finite expansion u^0 contains basis functions from high scales overlapping S_λ. Nevertheless, these factors depend only on u^0 (and hence on the accuracy δ of the initial guess) but not on the accuracy by which $D\mathbf{R}(\mathbf{u}^0)$ is applied through the scheme **APPLY**. Therefore, in principle, one obtains asymptotically optimal complexity, however, with possibly poor quantitative behavior.

8. Newton's method. In concrete cases, the error reduction ρ obtained in (3.25) or (3.17) may be so close to one that the number K of necessary updates in the perturbed scheme **SOLVE** may become fairly large. So, in spite of its asymptotic optimality, the quantitative performance may be poor. We shall therefore address Newton's method, corresponding to $\mathbf{B}_n = (D\mathbf{R}(\mathbf{u}^n))^{-1}$ in (2.11), as an example where the ideal scheme permits a faster error decay. The adaptive realization of Newton's method, applied to the infinite dimensional problem (2.2) or, better yet, (2.8), does not quite fit into the format of **SOLVE**, explaining its separate treatment in this section.

Note that, for $\mathbf{B}_n = (D\mathbf{R}(\mathbf{u}^n))^{-1}$, (2.11) can be reformulated as follows. Given an initial guess $\mathbf{u}^n$, the next iterate $\mathbf{u}^{n+1}$ is determined by solving

$$D\mathbf{R}(\mathbf{u}^n)\mathbf{w}^n = -\mathbf{R}(\mathbf{u}^n) \tag{8.1}$$

and setting

$$\mathbf{u}^{n+1} := \mathbf{u}^n + \mathbf{w}^n. \tag{8.2}$$

We are not interested here in the weakest assumptions under which the iterative scheme (8.2) converges to a locally unique solution. We are instead content here with the following setting: Recall that the mapping R in the variational problem (2.2) has the form

$$R(v) = F(v) - f, \tag{8.3}$$

where throughout this section we shall confine the discussion again to nonlinear maps F of a single argument satisfying the growth condition (7.5) for some $n^* \geq 1$. (Of course, F can have a linear part as in (3.4).) Therefore, we have, in particular, that R and F have the same Frechét derivative $DR(v) = DF(v)$. Moreover, we assume that for some open ball $\mathcal{U} \subset \mathcal{H}$ one has the following:

(N1) The Frechét derivative $DR(v) : w \mapsto DR(v)w$ is an isomorphism from $\mathcal{H}$ to $\mathcal{H}'$ (see (2.4)), and there exists $\omega > 0$ such that for all $v \in \mathcal{U}$ and y such that $v + y \in \mathcal{U}$, we have

$$\|(DR(v))^{-1}(DR(v+y) - DR(v))y\|_{\mathcal{H}} \leq \omega\|y\|_{\mathcal{H}}^2. \tag{8.4}$$

(N2) There exists a solution $u \in \mathcal{U}$ and an initial guess u^0 in $\mathcal{U}$ such that

$$\|u - u^0\|_{\mathcal{H}} \leq \delta < 2/\omega \quad \text{and} \quad B_\delta(u) \subseteq \mathcal{U} \tag{8.5}$$

with ω from (N1).

Given the validity of (N1) and (N2), standard arguments can be employed to prove that all iterates

$$u^{n+1} = u^n - DR(u^n)^{-1}R(u^n), \tag{8.6}$$

arising from Newton's method formulated in $\mathcal{H}$, remain in $\mathcal{U}$ and satisfy

$$\|u-u^n\|_{\mathcal{H}}<\delta \quad \text{for} \quad n\in\mathbb{N} \quad \text{and} \quad \lim_{n\to\infty}\|u-u^n\|_{\mathcal{H}}=0. \tag{8.7}$$

In fact, one has quadratic convergence

$$\|u-u^{n+1}\|_{\mathcal{H}}\le \frac{\omega}{2}\|u-u^n\|_{\mathcal{H}}^2, \quad n=0,1,2,\dots; \tag{8.8}$$

see, e.g., [18, 19]. Finally, note that, by (2.5), the corresponding iterates in wavelet coordinates satisfy

$$\|\mathbf{u}-\mathbf{u}^{n+1}\|_{\ell_2(\mathcal{J})}\le\tilde{\omega}\|\mathbf{u}-\mathbf{u}^n\|_{\ell_2(\mathcal{J})}^2, \quad n\in\mathbb{N}_0, \quad \tilde{\omega}:=\frac{C_1^2\omega}{2c_1}. \tag{8.9}$$

Let us check that (N1) holds, for instance, in the semilinear case (SL) when F is defined by (3.4) with monotone (scalar valued) G which satisfies (7.5) with $n^*\ge 2$. We have already seen (see Remark 3.1) that $DR(v)=DF(v)$ is an isomorphism from $\mathcal{H}$ to $\mathcal{H}'$. In order to verify (8.4), it therefore suffices to show that

$$\|(G'(v+y)-G'(v))y\|_{\mathcal{H}'}\le\omega\|y\|_{\mathcal{H}}^2. \tag{8.10}$$

For this, we remark that

$$\|(G'(v+y)-G'(v))y\|_{\mathcal{H}'}\le\max_{t\in[0,1]}\|G''(v+ty)y^2\|_{\mathcal{H}'}=\|y\|_{\mathcal{H}}^2\max_{z\in\mathcal{U},\|w\|_{\mathcal{H}}\le 1}\|T(z,w,w)\|_{\mathcal{H}'}$$

with the mapping T defined by

$$T(v,w,z)=G''(v)wz. \tag{8.11}$$

Now, it is easy to check that if G satisfies (7.5) for some p and $n^*\ge 2$, then T satisfies (7.7) with $p_1=[p-2]_+$, $p_2=p_3=1$, and n^* replaced by n^*-2, and therefore, according to Theorem 7.1,

$$\omega:=\max_{z\in\mathcal{U},\|w\|_{\mathcal{H}}\le 1}\|T(z,w,w)\|<\infty, \tag{8.12}$$

as desired. The purpose of this section is to show how the approximate solution of the linear problem (8.1) can be performed again by an iterative scheme along the lines of [9]. By our assumption (N1), we know that $DR(z)$ is an isomorphism from $\mathcal{H}$ to $\mathcal{H}'$ provided that $z\in\mathcal{U}$ and hence $DR(z)$ satisfies (2.10). Given this mapping property (2.10), the adaptive scheme from [9] can actually be applied under fairly general assumptions on the linear isomorphism. For the sake of simplicity, we shall assume that $D\mathbf{R}(\mathbf{z})$ is symmetric positive definite. This is true, for example, in the case (SL) when G is monotone. Recall from Remark 3.3 that one can then find a parameter $\alpha>0$ such that $\mathbf{I}-\alpha D\mathbf{R}(\mathbf{v})$ is a contraction.

The heart of the envisaged adaptive Newton scheme will be to solve the linear problem (8.1) approximately with the aid of a variant, which will be called $\mathbf{SOLVE}_{\mathrm{N}}$, of the scheme $\mathbf{SOLVE}_{\mathrm{lin}}$ discussed in section 3. Before we describe the ingredients of $\mathbf{SOLVE}_{\mathrm{N}}$, let us point out two issues to be addressed when designing these ingredients and analyzing their complexity.

(a) The first point concerns the application of the Jacobian. Approximating at each stage the Jacobian $D\mathbf{R}(\mathbf{u}^n)$ in order to use the **APPLY** scheme based on

(5.7) might be computationally very expensive. (b) The second point concerns the complexity of approximately solving the linear problem (8.1). Recall from Theorem 6.1 that the logic of complexity estimates is to infer from a certain compressibility (or regularity) of the solution a corresponding convergence rate of the adaptive scheme. In the context of the Newton iteration, such a property will be assumed about the solution $\mathbf{u}$ of the original nonlinear problem (2.8) which, however, does not necessarily imply the same property for the solutions of the subproblems (8.1). So it is initially not clear how to derive complexity estimates for the resolution of these subproblems. It will be seen though that the solutions to these subproblems become increasingly closer to elements having the necessary properties, a fact that, as it turns out, can be exploited as long as the subproblems are not solved too accurately. In particular, the question then arises whether the quadratic convergence of the Newton scheme can be preserved.

We now turn to collecting the ingredients of the adaptive Newton scheme. First, the coarsening will be again done by **TCOARSE** even though the problem is linear. More importantly, the **RES** scheme will be of the form $\mathbf{RES}_{\mathrm{lin}}$ from (5.10) but with different schemes playing the roles of **APPLY** and **RHS**.

In view of the above issue (a), we shall pursue here the following approach. Recall from (2.9) that for any $\mathbf{v}, \mathbf{z} \in \ell_2(\mathcal{J})$ and corresponding $v = \sum_{\lambda\in\mathcal{J}} v_\lambda \psi_\lambda$, $z = \sum_{\lambda\in\mathcal{J}} z_\lambda \psi_\lambda \in \mathcal{H}$,

$$D\mathbf{R}(\mathbf{z})\mathbf{v} = (\langle \psi_\lambda, DR(z)v\rangle : \lambda \in \mathcal{J}), \tag{8.13}$$

where $DR(z)$ is the Frechét derivative of R at z. This suggests employing the scheme $\mathbf{EV}[\epsilon, \mathbf{Q}, \mathbf{V}]$ with $\mathbf{V} := (\mathbf{z}, \mathbf{v})$ and $\mathbf{Q}(\mathbf{V}) := D\mathbf{R}(\mathbf{z})\mathbf{v} = D\mathbf{F}(\mathbf{z})\mathbf{v}$. We have seen that this scheme has the optimal complexity provided that $\mathbf{Q}$ fulfills Assumptions 1 and 2.

For F defined by (3.4), the mapping $\mathbf{Q}$ has the form

$$\mathbf{Q}(\mathbf{z}, \mathbf{v}) := (\mathbf{A} + D\mathbf{G}(\mathbf{z}))\mathbf{v}. \tag{8.14}$$

The $\mathbf{A}$ part obviously satisfies Assumption 1. We also have seen that it fulfills Assumption 2 under the hypothesis from Remark 7.2. It remains to verify these assumptions for the part

$$D\mathbf{G}(\mathbf{z})\mathbf{v} = (\langle \psi_\lambda, G'(z)v\rangle : \lambda \in \mathcal{J}). \tag{8.15}$$

For this, we simply remark that if G satisfies (7.5) for some p and $n^* \geq 1$, the mapping $(z, v) \mapsto G'(z)v$ satisfies (7.7) with $p_1 = [p-1]_+$, $p_2 = 1$, and n^* replaced by $n^* - 1$. Hence Theorems 7.1 and 7.2 ensure that the mapping $(\mathbf{z}, \mathbf{v}) \mapsto D\mathbf{G}(\mathbf{z})\mathbf{v}$ satisfies Assumptions 1 and 2 in section 7.1.

This suggests the following routine.

$\mathbf{APPLY}_{\mathrm{N}}[\eta, D\mathbf{R}(\mathbf{z}), \mathbf{v}] \to \mathbf{w}_\eta$ *determines for any tolerance* $\eta > 0$ *and any finitely supported input vectors* $\mathbf{v}$ *and* $\mathbf{z}$ *a finitely supported output vector* $\mathbf{w}_\eta$ *such that*

$$\|D\mathbf{R}(\mathbf{z})\mathbf{v} - \mathbf{w}_\eta\|_{\ell_2(\mathcal{J})} \leq \eta, \tag{8.16}$$

through

$$\mathbf{APPLY}_{\mathrm{N}}[\eta, D\mathbf{R}(\mathbf{z}), \mathbf{v}] := \mathbf{EV}[\eta, \mathbf{Q}, (\mathbf{z}, \mathbf{v})], \tag{8.17}$$

where the routine **EV** *was introduced in section* 7.2.

It remains to specify the routine **RHS**. Here it is issue (b) that calls for some further preparations. The main point is that if the current right-hand sides in (8.1) are not approximated too accurately, then one actually approximates a nearby right-hand side of a problem whose solution is known to be sufficiently sparse and thus can be approximated efficiently by a linear version of **SOLVE**.

Remark 8.1. Suppose that R is twice Frechét differentiable and that $\mathbf{u}$ is the exact solution of (2.8). Then there exists a constant $\hat{C}$ such that, for any $\mathbf{z}$ such that $z = \sum_{\lambda\in\mathcal{J}} z_\lambda \psi_\lambda \in \mathcal{U}$,

$$\|D\mathbf{R}(\mathbf{z})(\mathbf{u}-\mathbf{z}) + \mathbf{R}(\mathbf{z})\|_{\ell_2(\mathcal{J})} \leq \hat{C}\|\mathbf{u}-\mathbf{z}\|^2_{\ell_2(\mathcal{J})}. \tag{8.18}$$

Proof. One has

$$-\mathbf{R}(\mathbf{z}) = \mathbf{R}(\mathbf{u}) - \mathbf{R}(\mathbf{z}) = D\mathbf{R}(\mathbf{z})(\mathbf{u}-\mathbf{z}) + \mathcal{O}\left(\|\mathbf{u}-\mathbf{z}\|^2_{\ell_2(\mathcal{J})}\right),$$

which confirms the claim. □

We shall employ the following routine in which $\hat{C}$ is the constant of Remark 8.1.

$\mathbf{RHS}_{\mathrm{N}}[\eta, \mathbf{R}, \mathbf{z}] \to \mathbf{r}_\eta(\mathbf{z})$ *is defined for any finitely supported* $\mathbf{z}$ *with* $z \in \mathcal{U}$ *such that* $\|\mathbf{u}-\mathbf{z}\|_{\ell_2(\mathcal{J})} \leq \xi$ *and for any* $\eta/2 > \hat{C}\xi^2$ *by*

$$\mathbf{RHS}_{\mathrm{N}}\,[\eta, \mathbf{R}, \mathbf{z}] := -\left(\mathbf{EV}\left[\frac{\eta}{2} - \hat{C}\xi^2, \mathbf{F}, \mathbf{z}\right] - \mathbf{RHS}\,[\eta/2, \mathbf{f}]\right), \tag{8.19}$$

where **RHS** *is defined by* (5.2) *but with* **TCOARSE** *used as* **CCOARSE**.

The role of the above conditions on $\mathbf{z}$ and η will become clear later.

We are now prepared to describe the version of **SOLVE** to be used for the approximate solution of the Newton systems (8.1) as follows.

$\mathbf{SOLVE}_{\mathrm{N}}[\eta, \mathbf{R}, \mathbf{z}] \to \mathbf{w}_\eta$ *determines for a given* $\mathbf{z} \in \ell_2(\mathcal{J})$*, such that* $z \in \mathcal{U}$*, an approximate solution* $\mathbf{w}_\eta$ *of the system* $D\mathbf{R}(\mathbf{z})\mathbf{w} = -\mathbf{R}(\mathbf{z})$ *satisfying*

$$\|\mathbf{w} - \mathbf{w}_\eta\|_{\ell_2(\mathcal{J})} \leq \eta, \tag{8.20}$$

by invoking $\mathbf{SOLVE}_{\mathrm{lin}}[\eta, D\mathbf{R}(\mathbf{z}), -\mathbf{R}(\mathbf{z})] \to \mathbf{w}_\eta$*, where, under the above assumptions on* $\mathbf{z}$ *and* η*, in step* (ii) *of* **SOLVE** *the residual approximation*

$$\begin{aligned}&\mathbf{RES}_{\mathrm{N}}[\eta, \alpha\mathbf{I}, D\mathbf{R}(\mathbf{z}), -\mathbf{R}(\mathbf{z}), \mathbf{v}]\\ &\qquad := \alpha\left(\mathbf{APPLY}_{\mathrm{N}}\left[\frac{\eta}{2\alpha}, D\mathbf{R}(\mathbf{z}), \mathbf{v}\right] - \mathbf{RHS}_{\mathrm{N}}\left[\frac{\eta}{2\alpha}, \mathbf{R}, \mathbf{z}\right]\right),\end{aligned} \tag{8.21}$$

and in step (iii) **TCOARSE** *is used.*

Note that, in view of (8.3), the evaluation of $\mathbf{R}$ also requires the approximation of the data $\mathbf{f}$ as stated explicitly in (8.19). From Theorems 7.3 and 7.4 and Proposition 6.3 we infer, as in Remark 5.1, that $\mathbf{u} \in \mathcal{A}^s_{\mathrm{tree}}$ implies $\mathbf{f} \in \mathcal{A}^s_{\mathrm{tree}}$, and its η-accurate tree approximation satisfies estimates of the form $\#\mathrm{supp}\,\mathbf{f}_\eta \lesssim \eta^{-1/s}\|\mathbf{u}\|^{1/s}_{\mathcal{A}^s_{\mathrm{tree}}}$, $\|\mathbf{f}_\eta\|_{\mathcal{A}^s_{\mathrm{tree}}} \lesssim \|\mathbf{u}\|_{\mathcal{A}^s_{\mathrm{tree}}}$.

Moreover, whenever $\mathbf{Q}(\mathbf{z}, \mathbf{v}) := D\mathbf{R}(\mathbf{z})\mathbf{v}$ satisfies Assumptions 1 and 2, we can apply Theorems 7.3 and 7.4 to conclude that the output $\mathbf{w}_\eta$ of $\mathbf{APPLY}_{\mathrm{N}}\,[\eta, D\mathbf{R}(\mathbf{z}), \mathbf{v}]$ satisfies

$$\begin{aligned}\#\mathrm{supp}\,\mathbf{w}_\eta &\lesssim \eta^{-1/s}\left(1 + \|\mathbf{z}\|^{1/s}_{\mathcal{A}^s_{\mathrm{tree}}} + \|\mathbf{v}\|^{1/s}_{\mathcal{A}^s_{\mathrm{tree}}} + \|\mathbf{u}\|^{1/s}_{\mathcal{A}^s_{\mathrm{tree}}}\right),\\ \|\mathbf{w}_\eta\|_{\mathcal{A}^s_{\mathrm{tree}}} &\lesssim 1 + \|\mathbf{z}\|_{\mathcal{A}^s_{\mathrm{tree}}} + \|\mathbf{v}\|_{\mathcal{A}^s_{\mathrm{tree}}} + \|\mathbf{u}\|_{\mathcal{A}^s_{\mathrm{tree}}}.\end{aligned} \tag{8.22}$$

Likewise, the output $\mathbf{r}_\eta(\mathbf{z})$ of $\mathbf{RHS}_{\rm N}\,[\eta, \mathbf{R}, \mathbf{z}]$ satisfies

$$\begin{array}{rcl}\#\mathrm{supp}\,\mathbf{r}_\eta(\mathbf{z}) & \lesssim & \eta^{-1/s}\left(1 + \|\mathbf{z}\|^{1/s}_{\mathcal{A}^s_{\rm tree}}\right), \\ \|\mathbf{r}_\eta(\mathbf{z})\|_{\mathcal{A}^s_{\rm tree}} & \lesssim & 1 + \|\mathbf{z}\|_{\mathcal{A}^s_{\rm tree}}.\end{array} \tag{8.23}$$

Recalling from (6.15) the definition of s^*-sparsity, we can infer from (8.22) the following consequence.

Remark 8.2. Let $s^* := (2\gamma - d)/(2d)$, where γ is the parameter associated with F by Theorem 7.2. Then the scheme $\mathbf{RES}_{\rm N}$, defined by (8.21), is s^*-sparse whenever $\|\mathbf{z}\|_{\mathcal{A}^s_{\rm tree}} \lesssim 1$.

We can now formulate an adaptive Newton iteration as follows.

$\mathbf{NEWTON}\,[\epsilon, R, \bar{\mathbf{u}}^0] \to \bar{\mathbf{u}}(\epsilon)$ *determines, for any finitely supported initial guess* $\bar{\mathbf{u}}^0$ *whose corresponding expansion* u_0 *satisfies* (8.5), *an approximate solution* $\bar{\mathbf{u}}(\epsilon)$ *satisfying*

$$\|\mathbf{u} - \bar{\mathbf{u}}(\epsilon)\|_{\ell_2(\mathcal{J})} \le \epsilon, \tag{8.24}$$

by the following steps:

(i) *Set* $\epsilon_0 := c_1^{-1}\delta$, $j = 0$.

(ii) *If* $\epsilon_j \le \epsilon$, *stop and output* $\bar{\mathbf{u}}(\epsilon) := \bar{\mathbf{u}}^j$. *Otherwise, choose some* $\eta_j > 0$ *(see* (8.27) *below), and perform*

$$\mathbf{SOLVE}_{\rm N}\,[\eta_j, \mathbf{R}, \bar{\mathbf{u}}^j, \mathbf{0}] \to \bar{\mathbf{w}}^j.$$

(iii) *Let (see* (8.9)*)*

$$\hat{\mathbf{u}} := \bar{\mathbf{u}}^j + \bar{\mathbf{w}}^j, \quad \hat{\eta}_j := \tilde{\omega}\epsilon_j^2 + \eta_j, \quad \bar{\mathbf{u}}^{j+1} := \mathbf{TCOARSE}\,[2C^*\hat{\eta}_j, \hat{\mathbf{u}}]$$

(where C^* *is the constant from section* 6.2*), and set* $\epsilon_{j+1} := (1 + 2C^*)\hat{\eta}_j$, $j + 1 \to j$, *and go to* (ii).

The choice of the dynamic tolerance η_j in step (ii) is yet to be specified. The first requirement is to keep the iterates $\bar{\mathbf{u}}^j$ in the right neighborhood of the solution, which means that the corresponding expansions $\bar{u}^j$ lie in $B_\delta(u)$. For this we shall use the following lemma.

LEMMA 8.1. *Fix a positive number* $\beta < 1$, *and assume that* $\delta > 0$ *is chosen sufficiently small to ensure that, in addition to* (8.5),

$$\delta < \frac{c_1^3}{(1 + 2C^*)C_1^3\omega} \tag{8.25}$$

and

$$\frac{(1 + 2C^*)\tilde{\omega}\delta}{c_1} < \beta. \tag{8.26}$$

Then the condition $\eta_j \le \eta_0 < \delta/(2(1 + 2C^*)C_1)$ *implies that* $\bar{u}_j \in B_\delta(u)$ *for all subsequent approximate iterates. Moreover, if*

$$\eta_j \le \frac{\epsilon_j\,(\beta - (1 + 2C^*)\tilde{\omega}\epsilon_j)}{1 + 2C^*}, \quad j = 0, 1, \ldots, \tag{8.27}$$

one has for $\hat{\eta}_j$ *defined in step* (iii) *of* **NEWTON**

$$\epsilon_{j+1} = (1 + 2C^*)\hat{\eta}_j \le \beta\epsilon_j, \quad j = 0, 1, \ldots. \tag{8.28}$$

Proof. Denoting by $\mathbf{u}^1$ the exact solution of $D\mathbf{R}(\bar{\mathbf{u}}^0)\mathbf{u}^1 = -\mathbf{R}(\bar{\mathbf{u}}^0)$ and recalling from step (i) that $\|\mathbf{u}-\bar{\mathbf{u}}^0\|_{\ell_2(\mathcal{J})} \le c_1^{-1}\delta = \epsilon_0$, the vector $\hat{\mathbf{u}}$ produced in steps (ii) and (iii) satisfies, by (8.9) and (8.20),

$$\|\mathbf{u}-\hat{\mathbf{u}}\|_{\ell_2(\mathcal{J})} \le \|\mathbf{u}-\mathbf{u}^1\|_{\ell_2(\mathcal{J})} + \|\mathbf{u}^1-\hat{\mathbf{u}}\|_{\ell_2(\mathcal{J})} \le \tilde{\omega}c_1^{-2}\delta^2 + \eta_0.$$

Thus, taking the coarsening step into account, we infer from (2.5) and (8.9) that

$$\begin{aligned}\|u-\bar{u}^1\|_{\mathcal{H}} &\le C_1\|\mathbf{u}-\bar{\mathbf{u}}^1\|_{\ell_2(\mathcal{J})} \le (1+2C^*)C_1(\tilde{\omega}c_1^{-2}\delta^2+\eta_0)\\ &= \frac{(1+2C^*)C_1^3\omega\delta^2}{2c_1^3} + (1+2C^*)C_1\eta_0.\end{aligned}$$

Thus when, e.g., $\frac{(1+2C^*)C_1^3\omega\delta^2}{2c_1^3} < \delta/2$, which is (8.25), it suffices to take $\eta_0 < \delta/(2(1+2C^*)C_1)$ at the initial stage $j=0$ to ensure that

$$\|\bar{u}^1 - u\|_{\mathcal{H}} < \|u-u^0\|_{\mathcal{H}},$$

which verifies that $\bar{u}^1 \in B_\delta(u)$. We can now iterate this result; e.g., using $\bar{u}^1$ in place of $\bar{u}^0$, we obtain that $\bar{u}^2 \in B_\delta(u)$, and so on. Now when (8.26) holds, we have $\beta > (1+2C^*)\tilde{\omega}\epsilon_0$ so that the condition (8.27) on η_j is feasible for $j=0$. Moreover, (8.27) implies that $\epsilon_{j+1} = (1+2C^*)\hat{\eta}_j \le \beta\epsilon_j$, which is (8.28). This ensures that the error bounds ϵ_j decay to zero so that (8.27) remains feasible for all j. This completes the proof. □

Proposition 8.3. *Assume that δ and η_j satisfy* (8.25), (8.26)*, and* (8.27)*, respectively. Then the scheme* **NEWTON** *terminates after finitely many steps and produces a finitely supported vector $\bar{\mathbf{u}}(\epsilon)$ satisfying*

$$\|\mathbf{u}-\bar{\mathbf{u}}(\epsilon)\|_{\ell_2(\mathcal{J})} \le \epsilon. \tag{8.29}$$

Thus, by (2.5)*,*

$$\left\| u - \sum_{\lambda\in\operatorname{supp}\bar{\mathbf{u}}(\epsilon)} u(\epsilon)_\lambda \psi_\lambda \right\|_{\mathcal{H}} \le C_1\epsilon.$$

Proof. We employ a simple perturbation argument as in the proof of Lemma 8.1. Let $\mathbf{u}^{j+1}$ denote the exact Newton iteration $\mathbf{u}^{j+1} = \bar{\mathbf{u}}^j - D\mathbf{R}(\bar{\mathbf{u}}^j)^{-1}\mathbf{R}(\bar{\mathbf{u}}^j)$. By step (i) we know that $\|\mathbf{u}-\bar{\mathbf{u}}^0\|_{\ell_2(\mathcal{J})} \le \epsilon_0$. Then, supposing that $\|\mathbf{u}-\bar{\mathbf{u}}^j\|_{\ell_2(\mathcal{J})} \le \epsilon_j$, we infer from (8.9) that

$$\|\mathbf{u}-\mathbf{u}^{j+1}\|_{\ell_2(\mathcal{J})} \le \tilde{\omega}\epsilon_j^2. \tag{8.30}$$

Hence, denoting by $\mathbf{w}^j := \mathbf{u}^{j+1} - \bar{\mathbf{u}}^j$ the exact solution of $D\mathbf{R}(\bar{\mathbf{u}}^j)\mathbf{w} = -\mathbf{R}(\bar{\mathbf{u}}^j)$, we obtain, according to step (iii),

$$\begin{aligned}\|\mathbf{u}-\bar{\mathbf{u}}^{j+1}\|_{\ell_2(\mathcal{J})} &\le \|\mathbf{u}-\mathbf{u}^{j+1}\|_{\ell_2(\mathcal{J})} + \|\mathbf{u}^{j+1}-\hat{\mathbf{u}}\|_{\ell_2(\mathcal{J})} + \|\hat{\mathbf{u}}-\bar{\mathbf{u}}^{j+1}\|_{\ell_2(\mathcal{J})}\\ &\le \tilde{\omega}\epsilon_j^2 + \|\mathbf{w}^j - \bar{\mathbf{w}}^j\|_{\ell_2(\mathcal{J})} + 3\hat{\eta}_j \le \tilde{\omega}\epsilon_j^2 + \eta_j + 2C^*\hat{\eta}_j\\ &= (1+2C^*)\hat{\eta}_j = \epsilon_{j+1},\end{aligned} \tag{8.31}$$

which advances the induction assumption. By (8.28), this finishes the proof. □

It remains to analyze the work/accuracy rate of **NEWTON**. So far the only condition on the tolerances η_j in step (ii) of **NEWTON** is (8.27), which ensures only that the error bounds ϵ_j decay to zero. This would allow us to keep η_j proportional to ϵ_j, which would result in an overall first order error reduction rate. On the other hand, choosing η_j proportional to ϵ_j^2, the error bounds ϵ_j decay, by step (iii), quadratically. However, according to the earlier discussion of issue (b), the subproblem (8.1) should not be resolved too accurately, as reflected by the above right-hand side scheme $\mathbf{RHS}_{\rm N}$; see (8.19). The following main result of this section says that within these constraints on the tolerances η_j one can still realize an outer convergence rate, ranging from first to second order, in such a way that the overall scheme exhibits optimal rate/complexity.

THEOREM 8.2. *Suppose that* (N1), (N2), *and the above hypotheses on* F *hold. Assume that* δ *satisfies* (8.25) *and* (8.26) *for some fixed* $\beta < 1$. *Moreover, assume that at the* j*th stage of* **NEWTON** *the tolerance* η_j *is in addition to* (8.27) *subjected to the condition*

$$\eta_j \rho^K \geq \zeta \hat{C} \epsilon_j^2 \quad \textit{for some fixed} \;\; \zeta > 1, \tag{8.32}$$

where ρ, K *are the constants from* (3.17) *and* (4.14). *Then, for any target accuracy* $\epsilon > 0$, **NEWTON** *outputs after finitely many steps a finitely supported vector* $\bar{\mathbf{u}}(\epsilon)$ *satisfying* $\|\mathbf{u} - \bar{\mathbf{u}}(\epsilon)\|_{\ell_2(\mathcal{J})}$, *and hence*

$$\left\| u - \sum_{\lambda \in \operatorname{supp} \bar{\mathbf{u}}(\epsilon)} \bar{u}(\epsilon)_\lambda \psi_\lambda \right\|_{\mathcal{H}} \leq C_1 \epsilon.$$

Moreover, if $\mathbf{Q}(\mathbf{z}, \mathbf{v}) = D\mathbf{R}(\mathbf{z})\mathbf{v}$ *fulfills Assumptions* 1 *and* 2, *and if the solution* $\mathbf{u}$ *of* (2.8) *belongs to* $\mathcal{A}^s_{\rm tree}$ *for some* $s < s^* = (2\gamma - d)/(2d)$, *the output* $\mathbf{u}(\epsilon)$ *of* **NEWTON** *has the following properties:*

$$\|\mathbf{u}(\epsilon)\|_{\mathcal{A}^s_{\rm tree}} \;\lesssim\; \|\mathbf{u}\|_{\mathcal{A}^s_{\rm tree}}, \qquad \#\operatorname{supp} \mathbf{u}(\epsilon) \;\lesssim\; \|\mathbf{u}\|^{1/s}_{\mathcal{A}^s_{\rm tree}} \epsilon^{-1/s}. \tag{8.33}$$

Under Assumption E the number of floating point operations needed to compute $\bar{\mathbf{u}}(\epsilon)$ *remains proportional to* $\#\operatorname{supp} \mathbf{u}(\epsilon)$.

It is understood that in the final step η_j is chosen within the above constraints as large as possible so as to attain the target accuracy. Note that, as indicated before, within the above constraints on η_j (see (8.27), (8.32)) the convergence rate of the outer inexact Newton iteration can be chosen to vary between linear and quadratic convergence.

Proof of Theorem 8.2. First, observe that in step (iii) one has at the jth stage, by the first part of (8.31),

$$\|\mathbf{u} - \hat{\mathbf{u}}\|_{\ell_2(\mathcal{J})} \leq \|\mathbf{u} - \mathbf{u}^{j+1}\|_{\ell_2(\mathcal{J})} + \|\mathbf{u}^{j+1} - \hat{\mathbf{u}}\|_{\ell_2(\mathcal{J})} \leq \tilde{\omega} \epsilon_j^2 + \eta_j = \hat{\eta}_j. \tag{8.34}$$

Hence, by Proposition 6.3 and step (iii), we obtain

$$\|\bar{\mathbf{u}}^{j+1}\|_{\mathcal{A}^s_{\rm tree}} \;\lesssim\; \|\mathbf{u}\|_{\mathcal{A}^s_{\rm tree}}, \quad \#\operatorname{supp} \bar{\mathbf{u}}^{j+1} \;\lesssim\; \|\mathbf{u}\|^{1/s} \epsilon_{j+1}^{-1/s}. \tag{8.35}$$

Moreover, by Proposition 6.2, the computational work required by the application of **TCOARSE** remains proportional to $\#\operatorname{supp} \hat{\mathbf{u}}$ since the support of $\hat{\mathbf{u}}$ already has tree structure.

Now let us discuss the computational complexity in step (ii) encountered between the coarsening steps. Here it is important that the Newton updates (8.1) are, in view of (8.32), not computed too accurately. In fact, under this constraint the approximation of $\mathbf{R}(\bar{\mathbf{u}}^j)$, computed in $\mathbf{SOLVE}_{\rm N}$, is incidentally also a sufficiently accurate approximation to $\mathbf{G}(\bar{\mathbf{u}}^j)$; see Remark 8.1. To explain this, recall (8.19) and set

$$\mathbf{Y}^j := \mathbf{EV}\,[\eta_j - \hat{C}\epsilon_j^2, -\mathbf{R}(\bar{\mathbf{u}}^j)].$$

Then one has, by Remark 8.1,

$$\begin{aligned}\|\mathbf{G}(\bar{\mathbf{u}}^j) - \mathbf{Y}^j\|_{\ell_2(\mathcal{J})} &\le \|\mathbf{G}(\bar{\mathbf{u}}^j) + \mathbf{R}(\bar{\mathbf{u}}^j)\|_{\ell_2(\mathcal{J})} + \|\mathbf{R}(\bar{\mathbf{u}}^j) + \mathbf{Y}^j\|_{\ell_2(\mathcal{J})}\\ &\le \hat{C}\epsilon_j^2 + \eta_j - \hat{C}\epsilon_j^2 = \eta_j.\end{aligned}$$

Thus, within the accuracy range permitted by (8.32), the routine $\mathbf{RHS}_{\rm N}$ invoked by $\mathbf{SOLVE}_{\rm N}$ satisfies the accuracy requirements for the perturbed equation $D\mathbf{R}(\bar{\mathbf{u}}^j)\hat{\mathbf{w}} = \mathbf{G}(\bar{\mathbf{u}}^j)$, whose solution is, by definition of $\mathbf{G}(\bar{\mathbf{u}}^j)$, just $\hat{\mathbf{w}}^j = \mathbf{u} - \bar{\mathbf{u}}^j$. Now we infer from Remark 8.2, Theorem 6.1, and (8.35) that

$$\#\mathrm{supp}\,\bar{\mathbf{w}}^j \;\lesssim\; \eta_j^{-1/s}\|\mathbf{u} - \bar{\mathbf{u}}^j\|^{1/s}_{\mathcal{A}^s_{\rm tree}} \;\lesssim\; \eta_j^{-1/s}\|\mathbf{u}\|^{1/s}_{\mathcal{A}^s_{\rm tree}},$$

that the computational complexity has the same bound, and that $\|\bar{\mathbf{w}}^j\|_{\mathcal{A}^s_{\rm tree}} \lesssim \|\mathbf{u}\|_{\mathcal{A}^s_{\rm tree}}$. The same bounds hold for $\hat{\mathbf{u}}$. By definition of η_j, the computational complexity of step (ii) in **NEWTON** and, by the previous remarks, also of **TCOARSE** in step (iii) therefore remains bounded by $C\|\mathbf{u}\|^{1/s}\epsilon_{j+1}^{-1/s}$, which completes the proof. □

REFERENCES

[1] A. Barinka, *Fast Evaluation Tools for Adaptive Wavelet Schemes*, Ph.D. Dissertation, Rheinisch Westfälische Technische Hochschule Aachen, Aachen, Germany, 2002, in preparation.

[2] A. Barinka, T. Barsch, P. Charton, A. Cohen, S. Dahlke, W. Dahmen, and K. Urban, *Adaptive wavelet schemes for elliptic problems—implementation and numerical experiments*, SIAM J. Sci. Comput., 23 (2001), pp. 910–939.

[3] P. Binev and R. DeVore, *Fast computation in adaptive tree approximation*, Numer. Math., to appear.

[4] P. Binev, W. Dahmen, and R. DeVore, *Adaptive finite element methods with convergence rates*, Numer. Math., to appear.

[5] A. Canuto, A. Tabacco, and K. Urban, *The wavelet element method, part* I: *Construction and analysis*, Appl. Comput. Harmon. Anal., 6 (1999), pp. 1–52.

[6] A. Canuto, A. Tabacco, and K. Urban, *The wavelet element method, part* II: *Realization and additional features in* 2*D and* 3*D*, Appl. Comput. Harmon. Anal., 8 (2000), pp. 123–165.

[7] A. Cohen, *Numerical Analysis of Wavelet Methods*, Studies in Mathematics and Its Applications 32, North–Holland, Amsterdam, 2003.

[8] A. Cohen, W. Dahmen, and R. DeVore, *Adaptive wavelet methods for elliptic operator equations: Convergence rates*, Math. Comp., 70 (2001), pp. 27–75.

[9] A. Cohen, W. Dahmen, and R. DeVore, *Adaptive wavelet methods* II: *Beyond the elliptic case*, Found. Comput. Math., 2 (2002), pp. 203–245.

[10] A. Cohen, W. Dahmen, and R. DeVore, *Sparse evaluation of compositions of functions using multiscale expansions*, SIAM J. Math. Anal., 35 (2003), pp. 279–303.

[11] A. Cohen and R. Masson, *Wavelet adaptive methods for second order elliptic problems, boundary conditions and domain decomposition*, Numer. Math., 86 (2000), pp. 193–238.

[12] S. Dahlke, W. Dahmen, R. Hochmuth, and R. Schneider, *Stable multiscale bases and local error estimation for elliptic problems*, Appl. Numer. Math., 23 (1997), pp. 21–47.

[13] S. Dahlke, W. Dahmen, and K. Urban, *Adaptive wavelet methods for saddle point problems: Optimal convergence rates*, SIAM J. Numer. Anal., 40 (2002), pp. 1230–1262.

[14] W. Dahmen and R. Schneider, *Composite wavelet bases for operator equations*, Math. Comp., 68 (1999), pp. 1533–1567.

[15] W. Dahmen and R. Schneider, *Wavelets on manifolds* I: *Construction and domain decomposition*, SIAM J. Math. Anal., 31 (1999), pp. 184–230.

[16] W. Dahmen, R. Schneider, and Y. Xu, *Nonlinear functions of wavelet expansions: Adaptive reconstruction and fast evaluation*, Numer. Math., 86 (2000), pp. 49–101.

[17] W. Dahmen, K. Urban, and J. Vorloeper, *Adaptive wavelet methods: Basic concepts and applications to the Stokes problem*, in Wavelet Analysis, D.-X. Zhou, ed., World Scientific, River Edge, NJ, 2002, pp. 39–80.

[18] P. Deuflhard and M. Weiser, *Local inexact Newton multilevel FEM for nonlinear elliptic problems*, in Computational Science for the 21st Century (Tours, France), M.-O. Bristeau, G. Etgen, W. Fitzigibbon, J.-L. Lions, J. Periaux, and M. Wheeler, eds., Wiley-Interscience-Europe, London, 1997, pp. 129–138.

[19] P. Deuflhard and M. Weiser, *Global inexact Newton multilevel FEM for nonlinear elliptic problems*, in Multigrid Methods, Lecture Notes in Comput. Sci. Engrg. 3, W. Hackbusch and G. Wittum, eds., Springer-Verlag, New York, 1998, pp. 71–89.

[20] R. DeVore, *Nonlinear approximation*, in Acta Numerica, Acta Numer. 7, Cambridge University Press, Cambridge, UK, 1998, pp. 51–150.

[21] J. Pousin and J. Rappaz, *Consistency, stability, a-priori and a-posteriori errors for Petrov-Galerkin methods applied to nonlinear problems*, Numer. Math., 69 (1994), pp. 213–231.

[22] M. Renardy and R.C. Rogers, *An Introduction to Partial Differential Equations*, Texts Appl. Math. 13, Springer-Verlag, New York, 1993.

[23] R. Verfürth, *A Review of A Posteriori Error Estimation and Adaptive Mesh-Refinement Techniques*, Wiley-Teubner, New York, 1996.

[24] E. Zeidler, *Nonlinear Functional Analysis and Its Applications* III: *Variational Methods and Optimization*, Springer-Verlag, New York, 1985.

SIAM J. Numer. Anal.
Vol. 41, No. 5, pp. 1824–1850

DISCRETE ABSORBING BOUNDARY CONDITIONS FOR SCHRÖDINGER-TYPE EQUATIONS. CONSTRUCTION AND ERROR ANALYSIS[*]

ISAÍAS ALONSO-MALLO[†] AND NURIA REGUERA[‡]

Abstract. When a partial differential equation in an unbounded domain is solved numerically, it is necessary to introduce artificial boundary conditions. In this paper, a general class of absorbing boundary conditions is constructed for one-dimensional Schrödinger-type equations discretized in space by finite differences. For this, rational approximations to the transparent boundary conditions are used. We study the simplest case in detail, obtaining an estimate for the full discrete error and showing that the discrete problem is weakly unstable. Moreover, we show numerically that the discrete problems associated to higher order absorbing boundary conditions are more unstable. Several numerical experiments confirm the results previously obtained.

Key words. Schrödinger equation, transparent boundary conditions, absorbing boundary conditions

AMS subject classifications. Primary, 65M12, 65M20; Secondary, 65M99

DOI. 10.1137/S0036142902412658

1. Introduction.

Let us consider the initial value problem for a Schrödinger-type equation given by

$$\begin{cases} \partial_t u(x,t) = \dfrac{-i}{c}\left(\partial_{xx}u(x,t) + Vu(x,t)\right), \quad x \in \mathbf{R}, \quad t \geq 0, \\ u(x,0) = u_0(x), \quad x \in \mathbf{R}, \end{cases} \tag{1.1}$$

where c is a real constant. (In this paper, we suppose without loss of generality that $c > 0$; the case $c < 0$ is analogous.) Moreover, we suppose that the potential V is constant, and we assume that when $x \notin (x_l, x_r)$, the initial value $u_0(x)$ vanishes. We remark that, to obtain the transparent boundary conditions (TBCs) below, it suffices that V is constant when $x \notin (x_l, x_r)$. The study of (1.1) arises in a wide variety of applications. Two well-known cases are the one-dimensional time dependent Schrödinger equation for a particle with mass m [15],

$$i\hbar\partial_t\Psi = -\frac{\hbar^2}{2m}\partial_x^2\Psi + V\Psi,$$

and the Fresnel equation for the evolution of a paraxial electrical field E along the z-direction in a Cartesian coordinate system [12, 21],

$$2in_0k_0\partial_z E = \partial_x^2 E + (n^2 - n_0^2)k_0^2 E. \tag{1.2}$$

For practical purposes, the numerical approximation is computed only in a finite subdomain. Therefore, it is necessary to introduce suitable boundary conditions. For

[*] Received by the editors August 5, 2002; accepted for publication (in revised form) March 18, 2003; published electronically October 28, 2003. This research was supported by MCYT BFM 2001-2013 and JCYL VA025/01.

http://www.siam.org/journals/sinum/41-5/41265.html

[†] Departamento de Matemática Aplicada y Computación, Universidad de Valladolid, 47005 Valladolid, Spain (isaias@mac.cie.uva.es).

[‡] Departamento de Matemáticas y Computación, Universidad de Burgos, Burgos, Spain (nreguera@ubu.es).

example, if *transparent*, i.e., reflection-free, boundary conditions are used, then the solution in the finite subdomain is exactly the original solution. However, such conditions are nonlocal in time for Schrödinger-type equations [5, 23], and therefore their computational cost is high. An interesting idea is to consider fast evaluations of these TBCs [4, 14]. Another possibility is to use local *absorbing* boundary conditions (ABCs), allowing only small reflections. These ABCs have been studied in several contributions [2, 6, 7] for (1.1). We have proved in [2] that the spatial semidiscretizations of the problems obtained with these ABCs may be weakly ill-posed and that they are worse-posed for higher order ABCs.

As an alternative, in the literature we can find some works where TBCs or ABCs are obtained for a semidiscrete or fully discrete problem [9, 14, 22, 23]. These approaches are specific for the chosen discretization, but a better absorption is usually obtained.

In this paper, we are going to consider ABCs for the semidiscrete problem in space [1] (we will call them SABCs). This way, we obtain much more absorption than with the ABCs proposed in [2] with the same value of the spatial stepsize. Our idea is similar to the one proposed by Halpern in [9], where a systematic method to obtain ABCs for the one-dimensional wave equation semidiscretized in space with finite differences is explained. We have also considered finite differences for the discretization in space of (1.1), and we have obtained the TBCs for this problem. As they are nonlocal, we construct a class of local ABCs by using interpolatory rational approximations to the Fourier symbol of the TBCs. Our first result is that the discrete problems obtained with these SABCs are weakly unstable in a similar way to those with ABCs in [2], with an instability that can only be compensated when the absorption is high. It seems that this instability may probably appear if other ABCs for Schrödinger-type equations are obtained in a similar way.

In the literature, several implementations for ABCs similar to the one considered in this paper have been used for other equations. For instance, Higdon [11] proposes, in the case of the dispersive wave equation, to choose the interpolatory nodes for the ABCs in a manual and approximated way. On the other hand, Trefethen and Halpern [25], for the wave equation, and Halpern and Rauch [10], for diffusion equations, propose to use several fixed nodes so that the rational approximation is optimal in a certain norm. Nevertheless, none of these implementations are useful in this case for the Schrödinger equation due to the instability (see [3]).

In this paper, we study one of the simplest cases of SABCs in detail, and we make a complete analysis of the error of the full discrete problem with these SABCs. We remark that, to our best knowledge, the study of the full discrete error has not been done in other works of the literature on ABCs for Schrödinger-type equations. For example, this study of the error is an important difference from [2, 6, 7]. There are three different terms in this error. When the spatial discretization is refined, the first term, which depends only on the spatial discretization, decreases. However, the second and third term may grow. The growth of the third term, associated to the discretization in time, is due to the instability and can be compensated for by taking smaller stepsizes or by using integrators in time of higher order. Finally, the second term is associated to the capacity of absorption of the SABCs, and it may be a small value only when the SABC is suitable for the absorption of the solution of (1.1). When this term is not small enough, it is possible to use SABCs with a higher order of absorption. However, it is necessary to take into account that, in this way, we increase the instability, and the third term of the error may behave worst.

The paper is organized as follows. The TBCs for the semidiscrete problem are obtained in section 2. In section 3, we construct one of the simplest SABCs for this problem. In section 4, we prove that the discrete problem is unstable. The reflection coefficient for these SABCs is studied in section 5. In section 6, we obtain an expression for the full discrete error. In section 7, other SABCs are constructed and studied numerically, showing a similar behavior, although the discrete problems are more unstable. Finally, in section 8, we present several numerical experiments showing the results previously obtained.

2. TBCs for the problem semidiscretized in space. Let us express the initial value problem (1.1) in an abstract way. Let us consider the space $X = L^2(\mathbf{R})$ and the dense subspace $D(A) = H^2(\mathbf{R})$. Let us define the linear operator $A : D(A) \subset X \to X$ by

$$Au = -\frac{i}{c}(u_{xx} + Vu), \quad u \in D(A).$$

PROPOSITION 2.1. *The initial value problem* (1.1) *may be written in an abstract way as*

$$\begin{cases} u'(t) = Au(t), \\ u(0) = u_0 \in X, \end{cases} \tag{2.1}$$

and its generalized solution $u(t) := \exp(tA)u_0$ *satisfies*

$$\|u(t)\| = \|u_0\|, \quad t \geq 0. \tag{2.2}$$

Proof. Since iA is linear and self-adjoint, we deduce from Stone's theorem [16] that A is the infinitesimal generator of a C_0-group of unitary operators, $\exp(tA)$, on X. Therefore, (2.1) is well-posed, its generalized solution is $u(t) = \exp(tA)u_0$, and (2.2) is satisfied. □

Let us consider now the spatial discretization of the initial value problem (1.1). Let $(x^j)_{j\in\mathbf{Z}}$ be a uniform mesh of $\mathbf{R}$, with $x^j = x_l + jh$, and denote by $u^j(t)$ an approximation of $u(x^j, t)$. For the discretization of (1.1), we have used finite differences:

$$\begin{cases} \dfrac{d}{dt}u^j = \dfrac{-i}{c}\left(\dfrac{u^{j+1} - 2u^j + u^{j-1}}{h^2} + Vu^j\right), \quad j \in \mathbf{Z}, \quad t \geq 0, \\ u^j(0) = u_0(x^j), \quad j \in \mathbf{Z}. \end{cases} \tag{2.3}$$

Let us consider the space

$$X_h = l^2(\mathbf{Z}) = \left\{ u_h = (u^j)_{j\in\mathbf{Z}} \,:\, \|u_h\|_h = \left(h \sum_{j\in\mathbf{Z}} |u^j|^2 \right)^{1/2} < \infty \right\}.$$

The spaces X and X_h are related through the linear operators

$$\begin{array}{rcl} r_h : Z \subset X & \to & X_h, \\ u & \to & r_h u = (u(x^j))_{j\in\mathbf{Z}}, \end{array} \tag{2.4}$$

where Z is a subspace such that, for $u \in Z$, we have $r_h u \in X_h$. Notice that this condition is clearly satisfied when we consider the initial value u_0 in (1.1) because $u_0(x)$ vanishes for $x \notin (x_l, x_r)$.

The operator A is approximated by the operators

$$A_h : X_h \to X_h,$$
$$u_h = (u^j)_{j\in\mathbf{Z}} \to A_h u_h = \left(\frac{-i}{c}\left(\frac{u^{j+1} - 2u^j + u^{j-1}}{h^2} + Vu^j\right)\right)_{j\in\mathbf{Z}}.$$

Since for each h fixed, A_h is a linear and bounded operator and iA_h is self-adjoint, we obtain the following result with a proof similar to that of Proposition 2.1.

PROPOSITION 2.2. *Suppose that $u_0 \in X$ is such that $r_h u_0 \in X_h$. Then the initial value problem* (2.3) *may be written in an abstract way as*

$$\text{(2.5)} \qquad \begin{cases} u_h'(t) = A_h u_h(t), \\ u_h(0) = r_h u_0 \in X_h, \end{cases}$$

and its solution $u_h(t) := \exp(tA_h)u_h(0)$ satisfies

$$\|u_h(t)\|_h = \|r_h u_0\|_h, \quad t \geq 0.$$

Let us see now what the dispersion relation of this semidiscrete problem is. Let us consider a wave solution $(u^j(t))_{j\in\mathbf{Z}} = (\exp(i(\eta j - \omega(\eta)t)))_{j\in\mathbf{Z}}$ of the problem discretized in space. Then $\omega(\eta)$ should satisfy the dispersion relation

$$\text{(2.6)} \qquad \omega(\eta) = \frac{2}{ch^2}(\cos(\eta) - 1) + \frac{V}{c}.$$

In practice, we are not going to solve (2.3) for $j \in \mathbf{Z}$ but for $0 \leq j \leq N$ with $h = L/N$, where $L = x_r - x_l$, so $(x_l, x_r) = (x^0, x^N)$. Let us obtain the TBC for this problem.

Let us define the operators $A_{h,0} : X_{h,0} \to X_{h,0}$, where

$$X_{h,0} = l^2(N, \infty) = \left\{ u_h = (u^N, u^{N+1}, \dots) : \|u_h\|_h^2 = h \sum_{j=N}^{\infty} |u^j|_h^2 < \infty \right\},$$

by

$$(A_{h,0}u_h)^N = \frac{-i}{c}\left(\frac{-2u^N + u^{N+1}}{h^2} + Vu^N\right),$$
$$(A_{h,0}u_h)^{N-1+j} = \frac{-i}{c}\left(\frac{u^{N-2+j} - 2u^{N-1+j} + u^{N+j}}{h^2} + Vu^{N-1+j}\right), \quad j \geq 2.$$

PROPOSITION 2.3. *Let us consider the problem*

$$\text{(2.7)} \qquad \begin{cases} \dfrac{d}{dt}u_{h,b}(t) = A_{h,0}u_{h,b}(t) + \phi_h(t), \\ u_{h,b}(0) = 0, \end{cases}$$

where $\phi_h(t) = [-i\phi(t)/(ch^2), 0, 0, \dots]$ and $\phi(t) \in L^1_{\text{loc}}(0, \infty)$. Then the solution of (2.7) *is*

$$\text{(2.8)} \qquad u_{h,b}(t) = \int_0^t \exp((t-s)A_{h,0})\phi_h(s)ds.$$

Moreover, the Laplace transform $\tilde{u}_{h,b}(i\omega) = \int_0^\infty \exp(-i\omega t) u_{h,b}(t)dt$, *with* $\Im(\omega) < 0$, *satisfies*

$$r_+^j(V + c\omega, h)\tilde{u}_b^{N-1+j}(i\omega) = \tilde{\phi}(i\omega), \quad j \geq 1, \tag{2.9}$$

with

$$r_+(\eta, h) = 1 - \frac{h^2}{2}\eta + i\sqrt{\frac{h^2}{2}\eta\left(2 - \frac{h^2}{2}\eta\right)},$$

where $\sqrt{}$ *denotes the squared root with positive real part. Finally,* (2.9) *can be analytically extended to* $\mathbf{C} - \{\omega \in \mathbf{C} : \Im(\omega) > 0, \Re(\omega) = -V/c, \Re(\omega) = -V/c + 4/ch^2\}$ *and, when* $\omega \in \mathbf{R}$,

$$r^j(V + c\omega, h)\tilde{u}_b^{N-1+j}(i\omega) = \tilde{\phi}(i\omega), \quad j \geq 1, \tag{2.10}$$

where

$$r(\eta, h) = \begin{cases} 1 - \frac{h^2}{2}\eta + \sqrt{\frac{h^2}{2}\eta\left(\frac{h^2}{2}\eta - 2\right)} & \text{if} \quad \frac{h^2}{2}\eta < 0, \\ 1 - \frac{h^2}{2}\eta + i\sqrt{\frac{h^2}{2}\eta\left(2 - \frac{h^2}{2}\eta\right)} & \text{if} \quad 0 \leq \frac{h^2}{2}\eta \leq 2, \\ 1 - \frac{h^2}{2}\eta - \sqrt{\frac{h^2}{2}\eta\left(\frac{h^2}{2}\eta - 2\right)} & \text{if} \quad \frac{h^2}{2}\eta > 2. \end{cases} \tag{2.11}$$

Proof. Since $A_{h,0}$ is a linear and bounded operator, the solution of (2.7) is (2.8). Moreover, $iA_{h,0}$ is self-adjoint, and we deduce that $\exp(tA_{h,0})$ is a group of unitary operators. Therefore, the type of $\exp(tA_{h,0})$ is 0. Taking the Laplace transform of (2.8) with $\Im(\omega) < 0$, we have $\tilde{u}_{h,b}(i\omega) = (i\omega - A_{h,0})^{-1}\tilde{\phi}_h(i\omega)$, and then

$$(A_{h,0} - i\omega)\tilde{u}_{h,b}(i\omega) = -\tilde{\phi}_h(i\omega).$$

This way, we can obtain the values $(\tilde{u}_b^{N-1+j})_{j\geq 1}$ from the solution of the recurrence relation

$$\begin{cases} \frac{-i}{c}\left(\frac{\tilde{u}_b^{N-2+j} - 2\tilde{u}_b^{N-1+j} + \tilde{u}_b^{N+j}}{h^2} + (V + c\omega)\tilde{u}_b^{N-1+j}\right) = 0, \ j \geq 1, \\ \tilde{u}_b^{N-1} = \tilde{\phi}, \end{cases} \tag{2.12}$$

with $\lim_{j\to\infty} \tilde{u}^{N-1+j} = 0$. The characteristic polynomial of the previous equation is

$$r^2 - 2\left(1 - \frac{h^2}{2}\eta\right)r + 1,$$

with $\eta = V + c\omega$. The roots of this polynomial are

$$r_\pm(\eta, h) = 1 - \frac{h^2}{2}\eta \pm i\sqrt{\frac{h^2}{2}\eta\left(2 - \frac{h^2}{2}\eta\right)},$$

and then the solution of (2.12) is

$$\tilde{u}_b^{N-1+j} = A_1 r_-^j + A_2 r_+^j, \quad j \geq 0.$$

We have that $|r_+| > 1$ and $|r_-| < 1$, and then, since $\tilde{u}_b^{N-1+j}$ goes to 0 as j goes to ∞, $A_2 = 0$. Therefore, $\tilde{u}_b^{N-1+j} = A_1 r_-^j$ and since for $j = 0$, $A_1 = \tilde{u}_b^{N-1} = \tilde{\phi}$, we have that

$$\tilde{u}_b^{N-1+j} = \tilde{\phi} r_-^j,$$

which is equivalent to (2.9).

Now, $r_+^j(V + c\omega, h)$ has two singularities at $\omega = -V/c$ and $\omega = -V/c + 4/ch^2$, and we obtain that (2.10) holds with $r(\eta, h)$ given by (2.11). □

Next, we use that $u^j(t)$, $j \in \mathbf{Z}$, is a tempered distribution in order to obtain an expression of the TBC in terms of its Fourier transform. For this, let us denote by $u_h(0) = (u^j(0))_{j\in\mathbf{Z}} \in l^2(-\infty, \infty)$ the initial condition of problem (2.3), and let us consider the function given by the Fourier series

$$U(\eta) = \sum_{j=-\infty}^{\infty} u^j(0) \exp(-ij\eta) \in L^2(-\pi, \pi).$$

In fact, we have supposed that $u^j(0) = 0$ for $j \notin [0, N]$, and therefore, $U(\eta) = \sum_{j=0}^{N} u^j(0) \exp(-ij\eta)$ is an analytic function of $\eta \in [-\pi, \pi]$. From $U(\eta)$ we can obtain $(u^j(0))_{j\in\mathbf{Z}}$ by means of

$$u^j(0) = \frac{1}{2\pi} \int_{-\pi}^{\pi} U(\eta) \exp(ij\eta) d\eta.$$

Proposition 2.4. *Keeping the above notation, let $(u^j(t))_{j\in\mathbf{Z}}$ be the solution of* (2.3). *Then, in the sense of the distributions,*

$$u^j(t) = \frac{1}{2\pi} \int_{\omega_1}^{\omega_2} \hat{u}^j(\omega) \exp(i\omega t) d\omega, \tag{2.13}$$

where $\omega_1 = -V/c$, $\omega_2 = 4/(ch^2) - V/c$, and $\hat{u}^j(\omega)$ is the Fourier transform of the tempered distribution $u^j(t)$.

Proof. We consider the solution of (2.3) expressed as

$$u^j(t) = \frac{1}{2\pi} \int_{-\pi}^{\pi} U(\eta) \exp(i(j\eta - \omega(\eta)t)) d\eta, \quad j \in \mathbf{Z},$$

where $\omega(\eta)$ is given by (2.6). This way we have that $u^j(t) = (I_1 + I_2)/(2\pi)$, where

$$I_1 = \int_{-\pi}^{0} U(\eta) \exp(i(j\eta - \omega(\eta)t)) d\eta, \quad I_2 = \int_{0}^{\pi} U(\eta) \exp(i(j\eta - \omega(\eta)t)) d\eta.$$

Let us make the change of variable $\omega = -\omega(\eta)$ in I_1; that is,

$$\eta = \arccos\left(1 - \frac{h^2}{2}(V + c\omega)\right) \in (-\pi, 0). \tag{2.14}$$

Then we have that

$$I_1 = \int_{\omega_1}^{\omega_2} \exp(i\omega t)\hat{u}_1(\omega) d\omega,$$

where

$$\hat{u}_1(\omega) = U(\arccos(\varphi(\omega))) \exp(ij \arccos(\varphi(\omega))) \frac{ch^2}{2\sqrt{1-\varphi(\omega)^2}},$$

with $\varphi(\omega) = 1 - h^2(V + c\omega)/2$. Similarly, $I_2 = \int_{\omega_1}^{\omega_2} \exp(i\omega t)\hat{u}_2(\omega) d\omega$, where $\hat{u}_2$ has the same expression as $\hat{u}_1$ with $\arccos(\theta) \in (0, \pi)$. Therefore, we get (2.13) with $\hat{u}^j(\omega) = \hat{u}_1(\omega) + \hat{u}_2(\omega)$. That is, the extension by zero to $\mathbf{R}$ of $\hat{u}^j(\omega)$ is the Fourier transform of $u^j(t)$ (also extended by zero for $t < 0$) for $\omega \in \mathbf{R}$. □

THEOREM 2.5 (TBCs). *Let $u_h(t)$ be the solution of* (2.3)*, and assume that $u_0(x)$, the initial value of* (1.1)*, vanishes when $x \notin (x_l, x_r) = (x^0, x^N)$, and V is constant. Then, when $h^2(V + c\omega)/2 \notin (0, 2)$, $\hat{u}^j(\omega) = 0$, $j \in \mathbf{Z}$, and when $h^2(V + c\omega)/2 \in (0, 2)$, the TBCs are given by*

$$\hat{u}^{N-1}(\omega) = r(V + c\omega, h)\hat{u}^N(\omega), \tag{2.15}$$

$$\hat{u}^1(\omega) = r(V + c\omega, h)\hat{u}^0(\omega), \tag{2.16}$$

where

$$r(\eta, h) = 1 - \frac{h^2}{2}\eta + i\sqrt{\frac{h^2}{2}\eta\left(2 - \frac{h^2}{2}\eta\right)}. \tag{2.17}$$

Proof. Let us obtain the right TBC (2.15). (The proof for the left TBC is similar.) We take $\phi(t) = u^{N-1}(t)$ in Proposition 2.3. Then $(u^j(t))_{j \geq N}$, the restriction of the solution of (2.3), is the solution of (2.7). The result is a straightforward consequence of Proposition 2.4. □

3. The simplest ABC. As we have already mentioned, the TBCs (2.15) and (2.16) are nonlocal, and thus, for practical purposes, we are going to obtain local ABCs. For this, we will consider different approximations to $r(s, h)$ which, under the conditions of Proposition 2.4, is given by (2.17).

Similarly to what we did in [2] for the continuous problem, we are going to consider approximations

$$r(V + c\omega, h) \approx q(V + c\omega, h), \tag{3.1}$$

where $q(s, h)$ is a rational function in s that interpolates $r(s, h)$. We will use the notation SABC(j_1, j_2) for the ABCs obtained when we consider $q = p_1/p_2$, with p_1 and p_2 relatively prime polynomials in s with degrees j_1 and j_2, respectively. Following the notation in [2], ABC(j_1, j_2) denotes the ABCs obtained there for the continuous problem. Moreover, we also call order of absorption the number $j_1 + j_2 + 1$.

Let us study now which should be the choice for the interpolatory nodes. Let us consider a wave solution $(u^j(t))_{j\in\mathbf{Z}} = (\exp(i(\eta j - \omega(\eta)t)))_{j\in\mathbf{Z}}$, where $\omega(\eta)$ is given by the dispersion relation (2.6). We are considering the SABCs obtained by using the approximation (3.1), and therefore we should choose the interpolatory nodes in such a way that $r(V + c\omega, h) - q(V + c\omega, h)$ is small when $\omega = -\omega(\eta)$, that is, when

$V + c\omega = (2/h^2)(1-\cos\eta)$. Therefore, the approximation should be good when at least one of the interpolatory nodes is

$$s_1 = \frac{2}{h^2}(1-\cos\eta). \tag{3.2}$$

Moreover, when the velocity of the solution is a unique known value, it is reasonable to take all the interpolatory nodes as equal.

Let us examine in more detail SABC(1,0), one of the simplest choices for the approximation (3.1). Let us take $q(s,h) = \alpha_0 + \alpha_1 s$, the polynomial that interpolates (2.17) at the points s_1 and s_2, with $s_1, s_2 \in (0, 4/h^2)$, where

$$\alpha_0 = 1 + i\frac{s_2c_1 - s_1c_2}{s_2 - s_1}, \qquad \alpha_1 = \frac{-h^2}{2} - i\frac{c_1 - c_2}{s_2 - s_1},$$

with

$$c_j = h\sqrt{s_j\left(1 - \frac{h^2}{4}s_j\right)}, \quad j = 1, 2.$$

(We omit the dependence on s_1, s_2, and h of the coefficients.) Since $h^2 \leq 4/s_j$ for $j = 1, 2$, $c_j \in \mathbf{R}$. Notice that if we take the limit when s_1 and s_2 go to a unique positive number b, the approximation we are considering is the Taylor expansion of first order of $r(s,h)$ at $s = b$.

Considering this approximation in (2.15) and taking the inverse Fourier transform, we obtain

$$\frac{d}{dt}v^N(t) = \widetilde{\alpha}v^N(t) + \widetilde{\beta}v^{N-1}(t),$$

where

$$\widetilde{\alpha} = \frac{\left(1 - \dfrac{h^2V}{2}\right)(s_1 - s_2) + ic_1(V - s_2) - ic_2(V - s_1)}{c(c_2 - c_1) - \dfrac{ich^2}{2}(s_1 - s_2)},$$

$$\widetilde{\beta} = \frac{s_2 - s_1}{c(c_2 - c_1) - \dfrac{ich^2}{2}(s_1 - s_2)}.$$

Similarly for the left boundary, we have

$$\frac{d}{dt}v^0(t) = \widetilde{\alpha}v^0(t) + \widetilde{\beta}v^1(t).$$

This way we have obtained a first order ordinary differential system

$$v_h'(t) = M(h)v_h(t), \tag{3.3}$$

where $v_h(t) = [v^0(t), v^1(t), \ldots, v^{N-1}(t), v^N(t)]^T$ and

$$M(h) = \begin{bmatrix} \widetilde{\alpha} & \widetilde{\beta} & 0 & 0 & \cdots & 0 \\ \widetilde{m}_1 & \widetilde{m}_2 & \widetilde{m}_1 & 0 & \cdots & 0 \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & \widetilde{m}_1 & \widetilde{m}_2 & \widetilde{m}_1 \\ 0 & \cdots & 0 & 0 & \widetilde{\beta} & \widetilde{\alpha} \end{bmatrix} \in \mathcal{M}_{(N+1)\times(N+1)}, \tag{3.4}$$

with $\widetilde{m}_1 = -i/ch^2$ and $\widetilde{m}_2 = i(2 - Vh^2)/ch^2$. We have omitted in the notation the dependence on h and on the interpolatory nodes of the elements of $M(h)$.

4. The full discrete problem with SABC(1,0). We are going to study the system (3.3) obtained when SABC(1,0) is considered. In order to add these SABCs, recall that we suppose that the initial value has a support included in the finite interval $(x_l, x_r) = (x^0, x^N)$, where $N \in \mathbf{N}$. Therefore, we take the norm

$$\|(u^j)_{j=0}^N\|_{[x^0,x^N],h} = \left(h \sum_{j=0}^{N} |u^j|^2 \right)^{1/2}. \tag{4.1}$$

In section 2, we have considered the abstract problem (2.1), which is approximated by (2.5). We take $Y_h = \mathbf{C}^{N+1}$ with the norm (4.1) and denote by P_h the projection operator

$$\begin{array}{rccl} P_h : & X_h & \rightarrow & Y_h, \\ & u_h = (u^j)_{j\in\mathbf{Z}} & \rightarrow & P_h u_h = (u^j)_{0\le j\le N}. \end{array} \tag{4.2}$$

Notice that

$$\|P_h u_h\|_{[x^0,x^N],h} \le \|u_h\|_h, \quad u_h \in X_h.$$

Then we use SABC(1,0) and approximate the solution of (2.5) through the solution of the ordinary differential system

$$\begin{cases} v_h'(t) = M(h)v_h(t), \\ v_h(0) = P_h r_h u_0 \in Y_h, \end{cases} \tag{4.3}$$

where we have used the notation of section 3 for the matrix of coefficients $M(h)$. The solution of (4.3) is

$$v_h(t) = \exp(tM(h))P_h r_h u_0, \quad 0 \le t \le T.$$

The last step is the time discretization of (4.3). Since (4.3) is stiff and has very large eigenvalues close to the imaginary axis, it is necessary to consider an A-stable time integration method. In this paper, we use A-stable implicit Runge–Kutta methods (cf. [2]).

Let us take a time stepsize $k > 0$, and let $t_n = kn$ for $n > 0$. We consider a Runge–Kutta method of order p whose stability function is given by $s(z)$. Then the integration in time of (4.3) is given by

$$\begin{cases} v_{h,n+1} = s(kM(h))v_{h,n}, \\ \quad v_{h,0} = P_h r_h u_0 \in Y_h, \end{cases} \tag{4.4}$$

which provides the numerical approximations

$$v_{h,n} = s(kM(h))v_{h,n-1} = s^n(kM(h))P_h r_h u_0, \quad 0 \le t_n \le T,$$

to the values $v_h(t_n)$. Now, we are interested in the boundedness of $\|s^n(kM(h))\|_{[x_0,x_N],h}$ for $n \in \mathbb{N}$, that is, in the stability analysis of this time approximation.

We denote by $\mu_2(M(h))$ the logarithmic norm of $M(h)$ (see [8]). From a well-known result (see, e.g., [8, Theorem IV.11.2]), if $\mu_2(M(h)) \leq 0$ and the Runge–Kutta method is A-stable, then

$$\|s^n(kM(h))\|_{[x_0,x_N],h} \leq 1 \quad \text{for } n \in \mathbb{N}.$$

However, this result of stability is not applicable. A straightforward calculation shows that $\mu_2(M(h)) = 1/(2ch^2) + O(h^{-1})$, and we cannot reject a possible exponential instability of the discrete L^2 norm of the solution when h goes to 0. This catastrophic behavior is not observed in the numerical experiments of section 8 because $M(h)$ is nonnormal and the estimate provided by the logarithmic norm is not suitable.

If $M(h)$ has a complete eigensystem, then we can study the stability applying the scalar case (see, e.g., [13]). For this, a necessary condition is given by the fact that the matrices $M(h)$ are stable, i.e., its eigenvalues have negative real part. Of course, this fact does not imply the stability of (4.4), but it will allow us to obtain a more realistic estimate. This result is similar to the situation studied in [2], where we proved that an analogous problem to (3.3) is only weakly unstable. The following theorem is a straightforward consequence of Theorem 4.3 of [2] (see also [17]).

THEOREM 4.1. *Let us consider the matrix*

(4.5)
$$M_N(h) = \begin{bmatrix} \alpha(h) & \beta(h) & 0 & 0 & \cdots & 0 \\ -i & i(2-\delta(h)) & -i & 0 & \cdots & 0 \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & -i & i(2-\delta(h)) & -i \\ 0 & \cdots & 0 & 0 & \beta(h) & \alpha(h) \end{bmatrix} \in \mathcal{M}_{(N+1)\times(N+1)},$$

whose coefficients satisfy

$$\alpha(h) = h\alpha_0 + h^2\alpha_1(h), \quad \textit{where} \quad \alpha_0 < 0, \quad \alpha_1(0) = ia_1, \tag{4.6}$$

$$\beta(h) = -h\alpha_0 + h^2\beta_1(h), \quad \textit{where} \quad \beta_1(0) = ib_1, \tag{4.7}$$

$$\delta(h) = h^2\delta_1(h) \in \mathbf{R}, \quad \textit{where} \quad \delta_1(0) = d_1, \tag{4.8}$$

$$0 > a_1 + b_1 + d_1, \tag{4.9}$$

$$0 > \alpha_0(1-N)(a_1+b_1+d_1) + 2\Re(\alpha_1'(0) + \beta_1'(0)). \tag{4.10}$$

We shall use the notation $\alpha_r(h) = \Re(\alpha(h))$, $\alpha_i(h) = \Im(\alpha(h))$, $\beta_r(h) = \Re(\beta(h)) \neq 0$, $\beta_i(h) = \Im(\beta(h)) \neq 0$, *where these coefficients satisfy one of the following properties for* $0 < h < h_0$*:*

$$|\beta(h)| \leq |\alpha_r(h)| \quad \textit{or} \tag{4.11}$$

$$|\beta(h)| > |\alpha_r(h)| \quad \textit{and} \tag{4.12}$$

$$\sqrt{|\beta(h)|^2 - \alpha_r(h)^2} < \delta(h) + \alpha_i(h) < 4 - \sqrt{|\beta(h)|^2 - \alpha_r^2(h)}. \tag{4.13}$$

We will suppose that for $0 < h < h_0$,

$$\delta(h) + \alpha_i(h) < \frac{\alpha_r(h)\beta_i(h)}{\beta_r(h)}, \tag{4.14}$$

$$(\delta(h) + \alpha_i(h))(2 + \beta_i(h)) < -\alpha_r(h)\beta_r(h) - 2\beta_i(h), \tag{4.15}$$

where $\gamma(h) = 2 - \delta(h) - \alpha_i(h)$.

Then, for every $h \in (0, h_0)$, all the eigenvalues of $M_N(h)$ have negative real part.

Making use of Theorem 4.1, we are going to prove the stability of the matrix $M(h)$, given by (3.4).

Theorem 4.2. *We suppose that $s_1 = s_2 = b > 0$ is fixed. Let $h_0 = \sqrt{2/b}$. Then all the eigenvalues of the matrix* (3.4) *have negative real part for $0 < h < h_0$.*

Proof. Notice that since $c > 0$, it suffices to prove that all the eigenvalues of $ch^2M(h)$ have negative real part. The matrix $ch^2M(h)$ is $M_N(h)$ with

$$\alpha(h) = -2\sqrt{1-a^2} + i(b-V)h^2, \quad \beta(h) = 2\sqrt{1-a^2}(a - i\sqrt{1-a^2}),$$

where $a = 1 - h^2b/2$. Then, hypotheses (4.6)–(4.8) are satisfied for

$$\alpha_0 = -2\sqrt{b} < 0, \quad a_1 = b - V \in \mathbf{R}, \quad b_1 = -2b \in \mathbf{R}, \quad d_1 = V.$$

We also have

$$\alpha_1'(0) = \frac{b^{3/2}}{4}, \quad \beta_1'(0) = \frac{-5b^{3/2}}{4}.$$

Therefore, hypotheses (4.9) and (4.10) are satisfied, since

$$a_1 + b_1 + d_1 = -b < 0,$$

$$\alpha_0(1-N)(a_1+b_1+d_1) + 2\Re(\alpha_1'(0) + \beta_1'(0)) = -2Nb^{3/2} < 0.$$

On the other hand, since $h < 2/\sqrt{b}$,

$$\alpha_r(h) = -2\sqrt{1-a(h)^2}, \quad \alpha_i(h) = (b-V)h^2,$$

$$\beta_r(h) = 2a(h)\sqrt{1-a(h)^2}, \quad \beta_i(h) = -2(1-a(h)^2).$$

We have then $|\beta| = |\alpha_r|$, so (4.11) is satisfied.

Because we are assuming that $h < \sqrt{2/b}$,

$$\delta(h) + \alpha_i(h) - \frac{\alpha_r(h)\beta_i(h)}{\beta_r(h)} = \frac{2bh^2}{-2+bh^2} < 0,$$

$$(\delta(h) + \alpha_i(h))(2 + \beta_i(h)) + \alpha_r(h)\beta_r(h) + 2\beta_i(h) = 2bh^2(-3 + bh^2) < 0,$$

so hypotheses (4.14) and (4.15) of Theorem 4.1 are satisfied.

Therefore, from Theorem 4.1 we can conclude that for $h \in (0, h_0)$ (with $h_0 = \sqrt{2/b}$), all the eigenvalues of $M(h)$ have negative real part. □

In a similar way, we have obtained the following more general result (see [17, 20]).

Theorem 4.3. *Let us consider the semidiscrete problem obtained when $SABC$(1,0) is used with $s_1 \neq s_2$. Let $h_0 = 2/\sqrt{s_1 + s_2}$. Then all the eigenvalues of matrix* (3.4) *have negative real part for $0 < h < h_0$.*

We can conclude the stability from Theorem 4.2 (or Theorem 4.3) when $M(h)$ has a complete eigensystem. Nevertheless, as $M(h)$ is nonnormal, the bound of the powers of $M(h)$ depends on h in the general case, as is proved in the following proposition.

Proposition 4.4. *Let $L(h)$ be an invertible matrix such that $M(h)$=$L(h)J(h)L(h)^{-1}$, where $J(h)$ is in Jordan form. Then we have the stability bound*

$$\|s^n(kM(h))\|_{[x_0,x_N],h} \leq \kappa_h \, \|s^n(kJ(h))\|_{[x^0,x^N],h} \tag{4.16}$$

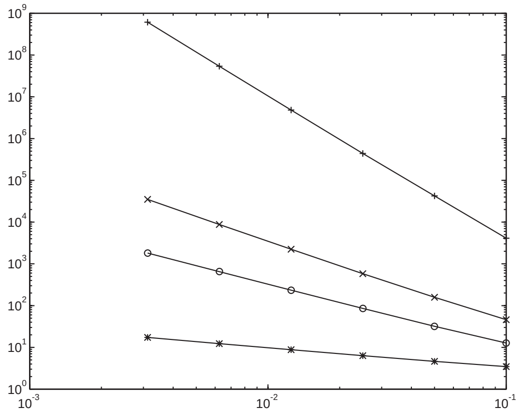


FIG. 1. *Condition number κ_h as a function of h. $*$ SABC(1,0), $\circ$ SABC(1,1), $\times$ SABC(2,1), $+$ SABC(3,2).*

for $n \in \mathbb{N}$, where κ_h is the condition number of $L(h)$. Moreover, if $M(h)$ has a complete eigensystem for $h > 0$, we have

$$\|s^n(kM(h))\|_{[x_0,x_N],h} \le \kappa_h. \tag{4.17}$$

Proof. First,

$$\begin{aligned}\|s^n(kM(h))\|_{[x^0,x^N],h} &= \left\|L(h)s^{n-j}(kJ(h))L(h)^{-1}\right\|_{[x^0,x^N],h} \\ &\le \|L(h)\|_{[x^0,x^N],h} \|s^n(kJ(h))\|_{[x^0,x^N],h} \left\|L(h)^{-1}\right\|_{[x^0,x^N],h} \\ &= \kappa_h \|s^n(kJ(h))\|_{[x^0,x^N],h},\end{aligned}$$

and we have obtained (4.16).

If $M(h)$ has a complete eigensystem, that is, $J(h)$ is the diagonal matrix of eigenvalues, then by Theorems 4.2 and 4.3, we obtain (4.17) from (4.16). □

With this stability result, we will prove in section 6 a result of convergence. Now, we are interested in the study of the behavior of the condition number κ_h. We have checked this numerically (see Figure 1), obtaining that $\kappa_h = O(h^{-1/2})$, similarly to the case of ABC(1,0) in [2]. Therefore, the possible instability of SABC(1,0) is mild, although the absorption for this SABC is not very high.

To study this weak instability, it is also possible to make an analysis of the ε-pseudoeigenvalues of $M(h)$ [24]. The ε-pseudospectrum may be calculated using random perturbations of the matrix $M(h)$. We have computed, for an example of section 8, the spectrum of $M(h) + E$ (with $h = 1/80$), where E is a random matrix of norm ϵ. In the second column of Table 1, we can see the maximum of the real part of the eigenvalues computed this way. We observe that the nonnormality of $M(h)$ is mild for SABC(1,0) because this maximum varies slowly with ϵ.

Finally, we remark that the instability arising when SABC(1,0) are used is very weak, and we have not been able to obtain numerical evidence. However, the instability may be clearly observed when other SABCs with higher orders of absorption are used because the discrete problems are more stable (see sections 7 and 8).

TABLE 1
Maximum of the real part of the ϵ-pseudoeigenvalues.

ϵ	SABC(1,0)	SABC(1,1)	SABC(2,1)	SABC(3,2)
1.0d−1	1.1254d−3	1.3378d−1	3.7731d−1	8.1664d+2
1.0d−3	−4.7348d−4	−9.1666d−4	−1.1637d−3	1.8658
1.0d−6	−5.6131d−4	−8.7425d−4	−1.1226d−3	−3.8555d−3
1.0d−9	−5.6126d−4	−8.7427d−4	−1.1225d−3	−3.8831d−3
1.0d−12	−5.6126d−4	−8.7427d−4	−1.1225d−3	−3.8832d−3

5. Reflection coefficient.

As we have remarked in section 3, the absorption of SABCs depends strongly on the choice of the interpolatory nodes. There exists a suitable node to absorb the component of the solution traveling with a given group velocity. Therefore, the absorption depends on the initial data, as is well displayed in the numerical experiments of section 8. In this section we are going to measure the absorption of SABC(1,0) by means of the reflection coefficient in a similar way to the study in [9]. The analytical results obtained confirm the previous remark.

We will consider only the case $s_1 = s_2 = b \in (0, 4/h^2)$, and we will use the notation $x = h^2(V + c\omega)/2$, $x^* = h^2 b/2$, so that to obtain SABC(1,0) we have considered the approximation (3.1) with

$$q(V + c\omega, h) = \alpha_0 + \alpha_1(x - x^*), \tag{5.1}$$

where

$$\alpha_0 = 1 - x^* + i\sqrt{1 - (1 - x^*)^2}, \quad \alpha_1 = -1 + \frac{i(1 - x^*)}{\sqrt{1 - (1 - x^*)^2}}.$$

5.1. One boundary.

First, we will study the problem with just one boundary at x^N. This case is used later to analyze the more practical case of two artificial boundary conditions. We remark that this case can also be used if the original problem is semi-infinite with a unique known boundary, for example, Dirichlet or Neumann.

Recall that $u_h(t) = (u^j(t))_{j\in\mathbf{Z}}$ is the solution of the semidiscretized problem in space (2.3) with an initial condition whose support is included in (x^J, x^L) with $L < N$. This solution satisfies the TBC (2.15). In Fourier variables,

$$\hat{u}^{j+1}(\omega) - 2\left(1 - \frac{h^2}{2}(V + c\omega)\right)\hat{u}^j(\omega) + \hat{u}^{j-1}(\omega) = 0, \quad j \in \mathbf{Z},$$

$$r(V + c\omega, h)\hat{u}^N(\omega) - \hat{u}^{N-1}(\omega) = 0.$$

On the other hand, let $v_h(t) = (v^j(t))_{j\le N}$ be the solution of the semidiscretized problem, defined for $j \le N$, with the same initial condition as for the previous problem and the SABC at x^N obtained when considering the approximation (3.1). Its Fourier transform $(\hat{v}^j(\omega))_{j\le N}$ satisfies

$$\hat{v}^{j+1}(\omega) - 2\left(1 - \frac{h^2}{2}(V + c\omega)\right)\hat{v}^j(\omega) + \hat{v}^{j-1}(\omega) = 0, \quad j \le N,$$

$$q(V + c\omega, h)\hat{v}^N(\omega) - \hat{v}^{N-1}(\omega) = 0.$$

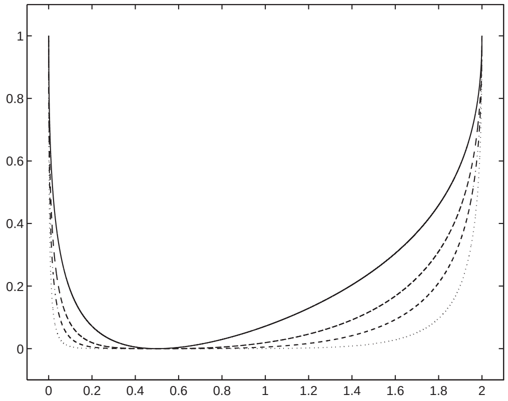


FIG. 2. *Reflection coefficient* $|R|$ *for one boundary as a function of* $x = h^2(V + c\omega)/2$ *when* $x^* = 0.5$. – *SABC*(1,0), – – *SABC*(1,1), – · *SABC*(2,1), · · · *SABC*(3,2).

Finally, let us define $w^j = v^j - u^j$, $j \leq N$, the reflected part of the solution caused by the ABC. Its Laplace transform $(\tilde{w}^j)_{j\leq N}$ satisfies

$$\hat{w}^{j+1}(\omega) - 2\left(1 - \frac{h^2}{2}(V + c\omega)\right)\hat{w}^j(\omega) + \hat{w}^{j-1}(\omega) = 0, \quad j \leq N,$$

$$q(V + c\omega, h)\hat{w}^N(\omega) - \hat{w}^{N-1}(\omega) + q(V + c\omega, h)\hat{u}^N(\omega) - \hat{u}^{N-1}(\omega) = 0.$$

For each ω and h, the reflection coefficient is defined as

$$\hat{w}^N(\omega) = R(V + c\omega, h)\hat{u}^N(\omega). \tag{5.2}$$

Note that it is enough to consider $R(V + c\omega, h)$ for $0 \leq h^2(V + c\omega)/2 \leq 2$ because of the result in Proposition 2.4.

The ideal situation is $R \equiv 0$, which leads to $\hat{w}^N(\omega) \equiv 0$. That is, there is no reflection, and the boundary condition is transparent. However, $R \not\equiv 0$ when we use ABCs, and the size of R measures the capacity of absorption of the ABCs. In Figure 2, we can observe the reflection coefficient when $x^* = 0.5$.

Since the initial value vanishes for $x = x^j$, $j > L$, and by using the TBCs (2.15) and (2.16), we have

$$\hat{u}^j(\omega) = r^{L-j}(V + c\omega, h)\hat{u}^L(\omega) \quad \text{for} \quad j > L. \tag{5.3}$$

Similarly, since the reflection vanishes for $x = x^j$, $j < N$,

$$\hat{w}^j(\omega) = r^{j-N}(V + c\omega, h)\hat{w}^N(\omega) \quad \text{for} \quad j < N. \tag{5.4}$$

Taking into account these relations,

$$(q - r^{-1})\hat{w}^N = q\hat{w}^N - \hat{w}^{N-1} = -q\hat{u}^N + \hat{u}^{N-1} = (r-q)\hat{u}^N,$$

where we have omitted the dependence on ω and h. Therefore, (5.2) is satisfied with

$$R(V + c\omega, h) = \frac{r(V + c\omega, h) - q(V + c\omega, h)}{q(V + c\omega, h) - r^{-1}(V + c\omega, h)}. \tag{5.5}$$

THEOREM 5.1. *Keeping the above notation, let $R(V + c\omega, h)$ be the reflection coefficient when SABC(1,0) with $s_1 = s_2 = b \in (0, 4/h^2)$ is considered. Then, we have*

$$|R(V + c\omega, h)| \leq \min\left(1, \frac{(x - x^*)^2}{\min((x^*)^2, (2 - x^*)^2)}\right)$$

for $0 \leq x \leq 2$, where $x = h^2(V + c\omega)/2$ and $x^* = h^2 b/2 \in (0, 2)$.

Proof. In this case q is given by (5.1). We will use the notation $\alpha_j^r = \mathrm{Re}(\alpha_j)$, $\alpha_j^i = \mathrm{Im}(\alpha_j)$. Taking into account (2.17),

$$|R| = \frac{|\alpha_0^i + \alpha_1^i(x - x^*) - \sqrt{x(2-x)}|}{|\alpha_0^i + \alpha_1^i(x - x^*) + \sqrt{x(2-x)}|} = \frac{\left|\frac{x(1-x^*)+x^*}{\sqrt{x^*(2-x^*)}} - \sqrt{x(2-x)}\right|}{\left|\frac{x(1-x^*)+x^*}{\sqrt{x^*(2-x^*)}} + \sqrt{x(2-x)}\right|}.$$

With a straightforward calculus, we have

$$|R| = \frac{\frac{x(1-x^*)+x^*}{\sqrt{x^*(2-x^*)}} - \sqrt{x(2-x)}}{\frac{x(1-x^*)+x^*}{\sqrt{x^*(2-x^*)}} + \sqrt{x(2-x)}} \leq 1$$

for $x \in [0, 2]$, having the identity only for $x = 0, 2$. Let us see now

$$|R(V + c\omega, h)| \leq (x - x^*)^2 / \min((x^*)^2, (2 - x^*)^2).$$

This bound is a consequence of

$$|R(V + c\omega, h)| \leq (x - x^*)^2/(x^*)^2$$

for $0 < x^* \leq 1$ and

$$|R(V + c\omega, h)| \leq (x - x^*)^2/(2 - x^*)^2$$

for $1 < x^* < 2$. Both estimates are easily obtained with a straightforward calculation. □

5.2. Two boundaries. In practice we do not have only one boundary as considered previously but two, and then, the reflected part of the solution caused by the SABCs will have two components, one traveling to the left and another one to the right.

As we prove below, the reflection coefficient has in this case two singularities, when the spatial parameter goes to 0, at the values $x = 0$ and $x = 2$. (A similar situation arises in the case of the wave equation [9].) These values correspond to the values of the Fourier components of the solution with group velocity very small or very high because of the aliasing brought about by the spatial discretization.

In a practical situation, the components of the solution with a very small velocity take a long time to arrive at the boundary and, on the other hand, the components with a very large velocity are usually small. As a consequence, in many practical cases these singularities are not important. However, when we obtain an estimate of the error in section 6 using the bound of the reflection coefficient, these singularities bring about a poor estimate of the part of error depending on the absorption.

Let us consider $u_h(t) = (u^j(t))_{j\in\mathbf{Z}}$ as in the previous section. It satisfies both TBCs

$$r(V + c\omega, h)\tilde{u}^0(i\omega) - \tilde{u}^1(i\omega) = 0,$$

$$r(V + c\omega, h)\tilde{u}^N(i\omega) - \tilde{u}^{N-1}(i\omega) = 0.$$

Let $v_h(t) = (v^j(t))_{0\le j\le N}$ be the solution of the semidiscretized problem with the same initial condition and SABCs in both boundaries; then

$$\hat{v}^{j+1}(\omega) - 2\left(1 - \frac{h^2}{2}(V + c\omega)\right)\hat{v}^j(\omega) + \hat{v}^{j-1}(i\omega) = 0, \quad 0 \le j \le N,$$

$$q(V + c\omega, h)\hat{v}^0(\omega) - \hat{v}^1(\omega) = 0,$$

$$q(V + c\omega, h)\hat{v}^N(\omega) - \hat{v}^{N-1}(\omega) = 0,$$

where we have omitted the dependence on ω and h of q and v^j. As in the previous section, $w^j(t) = v^j(t) - u^j(t)$, $0 \le j \le N$, will denote the reflected part of the solution caused by both SABCs. Therefore, omitting the dependence on ω and h,

$$\hat{w}^{j+1} - 2\left(1 - \frac{h^2}{2}(V + c\omega)\right)\hat{w}^j + \hat{w}^{j-1} = 0, \quad 0 \le j \le N, \tag{5.6}$$

$$q\hat{w}^0 - \hat{w}^1 = -q\hat{u}^0 + \hat{u}^1, \tag{5.7}$$

$$q\hat{w}^N - \hat{w}^{N-1} = -q\hat{u}^N + \hat{u}^{N-1}. \tag{5.8}$$

Solving this difference equation, we have

$$\hat{w}^j = A_1 r^{-j} + A_2 r^{j-N} \quad \text{for} \quad 0 \le j \le N, \tag{5.9}$$

and due to the TBC (2.15) and (2.16),

$$\hat{u}^j = r^{L-j}\hat{u}^L \quad \text{for} \quad j > L, \quad \text{and} \quad \hat{u}^j = r^{j-J}\hat{u}^J \quad \text{for} \quad j < J. \tag{5.10}$$

From (5.9) for $j = 0, N$, we get

$$\hat{w}^0 = A_1 + A_2 r^{-N}, \quad \hat{w}^N = A_1 r^{-N} + A_2,$$

and solving this system for A_1 and A_2, we obtain

$$A_1 = \beta\hat{w}^0 + \gamma\hat{w}^N, \quad A_2 = \gamma\hat{w}^0 + \beta\hat{w}^N, \tag{5.11}$$

with $\beta = 1/(1 - r^{-2N})$, $\gamma = -r^{-N}/(1 - r^{-2N})$. On the other hand, taking into account (5.9) and (5.10) in relations (5.7) and (5.8), we get

$$q\hat{w}^0 - A_1 r^{-1} - A_2 r^{1-N} = (r - q)\hat{u}^0,$$

$$q\hat{w}^N - A_1 r^{1-N} - A_2 r^{-1} = (r - q)\hat{u}^N.$$

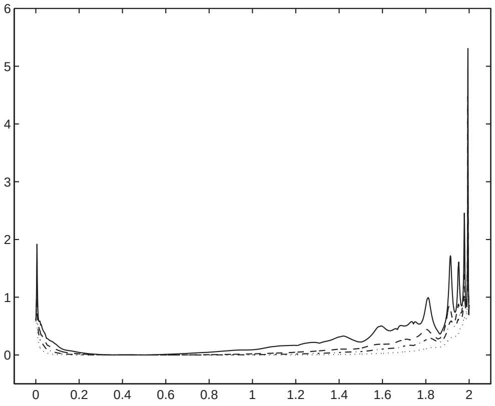


FIG. 3. *Reflection coefficients* $\max(|K_1(12)|, |K_2(12)|)$ *for two boundaries as a function of* $x = h^2(V + c\omega)/2$ *when* $x^* = 0.5$, $N = 30$, $L = 20$, $J = 10$. – *SABC*(1,0), – – *SABC*(1,1), – · *SABC*(2,1), · · · *SABC*(3,2).

Taking into account (5.11), we obtain expressions for $\hat{w}^0$, $\hat{w}^N$ in terms of $\hat{u}^0$, $\hat{u}^N$, and finally, from (5.9), we have

$$\hat{w}^j = K_1(j)\hat{u}^{L+j} + K_2(j)\hat{u}^{J-j},$$

where the reflection coefficients

$$K_1(j) = \frac{-r^L R(R + r^{2j})}{R^2 - r^{2N}}, \quad K_2(j) = \frac{-r^{-J} R(Rr^{2j} + r^{2N})}{R^2 - r^{2N}} \tag{5.12}$$

also depend on $V + c\omega$ and h, but this dependence is not displayed in the notation.

In Figure 3 we can observe $\max(|K_1(12)|, |K_2(12)|)$ when $N = 30$, $L = 20$, $J = 10$, and $x^* = 0.5$. As we have previously mentioned, we can see two possible singularities at $x = 0$ and $x = 2$. Let us analyze $|K_1(j)|$ and $|K_2(j)|$ in more detail.

THEOREM 5.2. *Let us consider* $x^* \in (0, 2)$ *and let* $K_1(j)$, $K_2(j)$ *be the reflection coefficients given by* (5.12) *when using SABC*(1,0) *with* $s_1 = s_2 = b \in (0, 4/h^2)$. *Then, for* $x = h^2(V + c\omega)/2 \in (0, 2)$ *and* $j = 0, \ldots N$,

$$\max(|K_1(j)|, |K_2(j)|) \le (N + 1)\frac{(x - x^*)^2}{\min((x^*)^2, (2 - x^*)^2)}. \tag{5.13}$$

Proof. We will suppose that $x^* \in (0, 1)$. (The proof for $1 \le x^* < 2$ is similar.) First, let us study $|K_1(j)|$. Recall that $R = -q_2(x, x^*)/q_1(x, x^*) < 0$, where

$$q_{1,2}(x, x^*) = x(1 - x^*) + x^* \pm \sqrt{x(2 - x)x^*(2 - x^*)}. \tag{5.14}$$

Taking into account that for $x \in (0, 2)$, $|r| = 1$ and $|R| < 1$, we get

$$|K_1(j)| = \frac{|R||R + r^{2j}|}{|R^2 - r^{2N}|} \le \frac{|R||R + r^{2j}|}{1 - |R|^2}.$$

Using now again that $|r| = 1$ and that R is real and negative,

$$|R + r^{2j}|^2 = (R + \Re(r^{2j}))^2 + \Im(r^{2j})^2 = (1 - |R|)^2 + 2R(\Re(r^{2j}) - 1).$$

Thus we have that

$$|K_1(j)|^2 \leq \frac{|R|^2}{(1+|R|)^2}\left(1+\frac{2|R|(1-\Re(r^{2j}))}{(1-|R|)^2}\right),$$

and using that $|R|<1$,

$$|K_1(j)|^2 \leq |R|^2\left(1+\frac{f(x)q_1(x,x^*)^2}{2x^*(2-x^*)}\right),$$

where $q_1(x,x^*)$ is given by (5.14) and

$$f(x)=\frac{1-\Re(r^{2j})}{x(2-x)}.$$

Finally, taking into account Lemmas 5.3 and 5.4 below, we get

$$|K_1(j)| \leq |R|\left(1+\frac{j^2 q_1(x,x^*)^2}{x^*(2-x^*)}\right)^{1/2} \leq (j+1)\frac{(x-x^*)^2}{(x^*)^2}.$$

Similarly, for $K_2(j)$ we obtain

$$|K_2(j)| \leq (N-j+1)\frac{(x-x^*)^2}{(x^*)^2},$$

and we deduce (5.13). □

In the proof of Theorem 5.2, we have used the following technical lemmas, whose proof can be found in [17].

Lemma 5.3. *Let*

$$f(x)=\frac{1-\Re(r^{2j})}{x(2-x)},\quad x\in(0,2),\quad j\in\mathbf{N}.$$

Then $f(x)\leq 2j^2$.

Lemma 5.4. *Let* $x^*\in(0,1)$ *and* R *the reflection coefficient for only one boundary* (5.5) *when considering SABC*(1,0) *with* $s_1=s_2=b$. *Then*

$$|R|^2\left(1+\frac{j^2 q_1(x,x^*)^2}{x^*(2-x^*)}\right) \leq (j+1)^2\frac{(x-x^*)^4}{(x^*)^4}, \tag{5.15}$$

where $q_1(x,x^*)$ *is given by* (5.14).

6. Error analysis.

In this section, our goal is to obtain an expression for the error of the full discretization with SABC(1,0). This error is given in the following definition.

Definition 6.1. *The full discrete global error is given by*

$$e_{h,n}=P_h r_h u(t_n)-v_{h,n},\quad 0\leq t_n\leq T, \tag{6.1}$$

where $v_{h,n}$ *is the solution of* (4.4), $u(t)$ *is the solution of* (2.1), r_h *is defined in* (2.4), *and* P_h *is defined in* (4.2).

The full discrete problem with SABC(1,0) has been obtained with three consecutive steps associated with three kinds of error. First, we discretize (2.1) in space obtaining the problem (2.5). The error made in this step is given by the spatial error

$$e^1_{h,n}=P_h r_h u(t_n)-P_h u_h(t_n),\quad 0\leq t_n\leq T, \tag{6.2}$$

where $u_h(t)$ is the solution of (2.5).

In the second step, we add the SABC(1,0) and we obtain (4.3). The error is now given by the error of absorption

$$e^2_{h,n} = P_h u_h(t_n) - v_h(t_n), \quad 0 \leq t_n \leq T, \tag{6.3}$$

where $v_h(t)$ is the solution of (4.3).

Finally, we use the time discretization introduced in section 4, and the error is the time global truncation error defined as

$$e^3_{h,n} = v_h(t_n) - v_{h,n}, \quad 0 \leq t_n \leq T, \tag{6.4}$$

where $v_{h,n}$ is the solution of (4.4).

The estimate of the error obtained in Theorem 6.2 is separated into three terms given by $e^i_{h,n}$, $i = 1, 2, 3$. The error may grow when h goes to zero due to two of these terms which behave as $e^3_{h,n} = O(k^p \kappa_h)$ and $\|e^2_{h,n}\|_{[x^0,x^N],h} \leq \epsilon h^{-2}$. This is a crucial difference from the wave equation (cf. [9]).

Nevertheless, notice that the growth is compensated for, in the case of $e^3_{h,n}$, by the presence of the factor k^p, which decreases when k decreases, or when p, the order of the integrator in time, grows, as is well displayed in the numerical experiments of section 8. Therefore, the use of methods of high order for the integration in time is crucial when we use SABCs with high orders of absorption, and the behavior of κ_h is worse, as we will see in section 7.

On the other hand, $\|e^2_{h,n}\|_{[x^0,x^N],h} \leq \epsilon h^{-2}$, where ϵ measures the capacity of absorption at the boundary and depends only on the solution of the semidiscrete problem. Nevertheless, we will see at the end of this section that ϵ can overestimate the error due to the absorption, and we will analyze better bounds for our numerical experiments. If the SABCs are suitable to absorb the solution arriving at the boundary, the numerical experiments indicate that the absorption is very high and the error absorption is small. If this is not the case, the alternative is to consider the higher order SABCs that are studied in next section. However, the instability is increased with these higher order SABCs and the term $e^3_{h,n}$ may grow. Moreover, due to the bad absorption of solutions traveling with a velocity distinct to the one associated with the interpolatory nodes, it is necessary to consider a distinct implementation (see [3]). Finally, we remark that the factor h^{-2} is, at least partially, a consequence of the singularities observed in the reflection coefficient in section 5, affecting components of the solution traveling with velocities very small or very large.

Then, we have the following result.

Theorem 6.2. *We assume that in the initial value problem* (2.1)*, the initial value* $u_0 \in D(A^2) = H^4(\mathbf{R})$*, and that the hypotheses of Propositions* 2.4 *and* 4.4 *are satisfied. Then the full discrete global error defined in* (6.1) *can be expressed as*

$$e_{h,n} = e^1_{h,n} + e^2_{h,n} + e^3_{h,n},$$

where $e^1_{h,n}$*, defined in* (6.2)*, satisfies*

$$\|e^1_{h,n}\|_{[x^0,x^N],h} = O(h^2) \tag{6.5}$$

and depends only on the spatial discretization. Moreover, for a fixed spatial discretization, the term $e^2_{h,n}$*, defined in* (6.3)*, depends only on SABC*(1,0)*, and, if* $s_1 = s_2 = b$*,*

$$\|e^2_{h,n}\|_{[x^0,x^N],h} \leq \epsilon h^{-2}, \tag{6.6}$$

where ϵ measures the capacity of absorption of the solution. Finally, the term $e^3_{h,n}$, defined in (6.4)*, depends only on the discretization in time and*

$$\|e^3_{h,n}\|_{[x^0,x^N],h} = O(k^p \kappa_h). \tag{6.7}$$

Proof. With the definition (6.1)

$$\begin{aligned} e_{h,n} &= P_h r_h u(t_n) - v_{h,n} \\ &= P_h r_h u(t_n) - P_h u_h(t_n) + P_h u_h(t_n) - v_h(t_n) + v_h(t_n) - v_{h,n} \\ &= e^1_{h,n} + e^2_{h,n} + e^3_{h,n}. \end{aligned}$$

The value $e^3_{h,n} = v_h(t_n) - v_{h,n}$ is the global truncation error arisen when the problem (4.3) is time discretized with an A-stable Runge–Kutta method. Then the proof the estimate (6.7) is very similar to the one used in section 5 of [2].

On the other hand, the proof of (6.5) is a classical result. Finally, we prove the estimate (6.6) in Lemma 6.3. □

LEMMA 6.3. *With the hypotheses and notation of Theorem* 6.2, (6.7) *is satisfied.*

Proof. Notice that $P_h u_h(t_n)$ is the restriction of the solution of (2.5) to $[x^0, x^N]$, that is, with TBCs, and that $v_h(t_n)$ is the approximation obtained when we use SABC(1,0). Therefore, $e^2_{h,n} = v_h(t_n) - P_h u_h(t_n)$ depends only on the absorption of the solution of (2.5) with SABC(1,0).

Let us suppose that $0 < h^2 b/2 \le 1$. (The proof for $1 < h^2 b/2 < 2$ is similar.) With the notation of section 5, we have

$$\|e^2_{h,n}\|_{[x^0,x^N],h} = \left(h \sum_{j=0}^{N} |w^j(t_n)|^2 \right)^{1/2}.$$

By Proposition 2.4,

$$w^j(t) = \frac{1}{2\pi} \int_{\omega_1}^{\omega_2} \exp(i\omega t) \hat{w}^j(\omega) d\omega,$$

where $\omega_1 = -V/c$, and $\omega_2 = -V/c + 4/(ch^2)$, and we deduce that

$$\begin{aligned} |w^j(t)|^2 &\le \frac{1}{ch^2\pi^2} \int_{\omega_1}^{\omega_2} |\hat{w}^j(\omega)|^2 d\omega \\ &= \frac{1}{ch^2\pi^2} \int_{\omega_1}^{\omega_2} |K_1(j)\hat{u}^{L+j}(\omega) + K_2(j)\hat{u}^{J-j}(\omega)|^2 d\omega. \end{aligned} \tag{6.8}$$

Using now (5.13),

$$|w^j(t)|^2 \le \frac{2(N+1)^2 c^3}{h^2 b^4 \pi^2} \int_{\omega_1}^{\omega_2} \left(|(\omega-\omega^*)^2 \hat{u}^{L+j}(\omega)|^2 + |(\omega-\omega^*)^2 \hat{u}^{J-j}(\omega)|^2 \right) d\omega.$$

Making now the change of variable $\tau = \omega - \omega^*$ and taking into account that $N = O(h^{-1})$,

$$\begin{aligned} |w^j(t)|^2 &\le \frac{2(N+1)^2 c^3}{h^2 b^4 \pi^2} \int_{\omega_1-\omega^*}^{\omega_2-\omega^*} \left(|\tau^2 \hat{u}^{L+j}(\tau+\omega^*)|^2 + |\tau^2 \hat{u}^{J-j}(\tau+\omega^*)|^2 \right) d\tau \\ &= \frac{C^2}{h^4 b^4} \int_{\omega_1-\omega^*}^{\omega_2-\omega^*} \left(|\tau^2 \hat{u}^{L+j}_{\omega^*}(\tau)|^2 + |\tau^2 \hat{u}^{J-j}_{\omega^*}(\tau)|^2 \right) d\tau, \end{aligned}$$

Table 2
Value of $\tilde{\epsilon}$.

SABC(1,0)	SABC(1,1)	SABC(2,1)	SABC(3,2)
1.2300d−2	5.5623d−4	3.0115d−5	1.3161d−7

where $\hat{u}^j_{\omega^*}(\tau) = \hat{u}^j(\tau + \omega^*)$. Finally, taking into account that $J < L$, we obtain (6.6) with

$$\epsilon = \frac{C}{b^2}\left(h\sum_{j=0}^{N}\int_{\omega_1-\omega^*}^{\omega_2-\omega^*}\left(|\tau^2\hat{u}^{L+j}_{\omega^*}(\tau)|^2 + |\tau^2\hat{u}^{J-j}_{\omega^*}(\tau)|^2\right)d\tau\right)^{1/2}. \qquad \square$$

We also note that it is possible, although only numerically, to obtain an estimate of the absorption using the last term of (6.8) from which we deduce a bound

$$\|e^2_{h,n}\|_{[x^0,x^N],h} \le \tilde{\epsilon}h^{-1},$$

where $\tilde{\epsilon}$ is given in terms of the reflection coefficients $K_1(j)$, $K_2(j)$. This bound obviously will be smaller than the previous one and has the advantage that we can use it for other SABCs of higher order. We have calculated numerically the value for $\tilde{\epsilon}$ for the experiment considered in section 8 for the initial datum with $\alpha = 20°$, $\sigma = 10$ (as in Figure 4), $L = 100$, $t_n = 500$, and $h = 0.1$, obtaining the results of Table 2. It is plain that these values are small pointing out that the absorption is high. Moreover, the value of $\tilde{\epsilon}$ decreases when the order of absorption increases.

7. Other ABCs. In section 3 we studied the procurement of SABC(1,0). Let us consider now other different choices for the approximation (3.1).

SABC(0,0). Let $q = \alpha_0$ be the function that interpolates $r(s,h)$ at s_1. The SABC obtained this way is

$$v^{N-1}(t) = \alpha_0 v^N(t).$$

In this case, we obtain a system $v_h' = Mv_h$ with $v_h = [v^1,\dots,v^{N-1}]^T$ and $M \in \mathcal{M}_{(N-1)\times(N-1)}$ such that $\mu_2(M) < 0$. Therefore, the problem is stable. Nevertheless, as the absorption of these SABCs is too small, we will not consider them for the numerical experiments of section 8.

SABC(1,1). Let $q(s,h) = (\alpha_0 + \alpha_1 s)/(1 + \alpha_2 s)$ be a rational function that interpolates $r(s,h)$ at s_1, s_2, and s_3. An approximation of (2.15) is then

$$(1 + \alpha_2(V + c\omega))\hat{v}^{N-1}(\omega) = (\alpha_0 + \alpha_1(V + c\omega))\hat{v}^N(\omega),$$

obtaining this way

$$\beta_0 v^{N-1}(t) + \beta_1\frac{d}{dt}v^{N-1}(t) = \beta_2 v^N(t) + \beta_3\frac{d}{dt}v^N(t)$$

for certain coefficients β_j depending on s_1, s_2, and s_3. Finally, taking into account the spatial discretization (2.3) we are considering for dv^{N-1}/dt, we obtain

$$\frac{d}{dt}v^N(t) = \gamma_0 v^N(t) + \gamma_1 v^{N-1}(t) + \gamma_2 v^{N-2}(t).$$

A similar expression is obtained for the left boundary.

Notice that in the limit case when $s_1 = s_2 = s_3 = b$, the approximation we are considering in (3.1) is the Padé (1,1) expansion of $r(V + c\omega, h)$ at $\omega = \omega^*$, with $\omega^* = (b - V)/c$.

This way, we have reduced the problem to the resolution of a system $v_h' = Mv_h$ with $v_h = [v^0, v^1, \dots, v^{N-1}, v^N]^T$ and $M \in \mathcal{M}_{(N+1)\times(N+1)}$.

SABC(2,1). Continuing this approach, let $q(s, h) = (\alpha_0 + \alpha_1 s + \alpha_2 s^2)/(1 + \alpha_3 s)$ be a rational function that interpolates $r(s, h)$ at s_1, s_2, s_3, and s_4. This gives rise to

$$\beta_0 v^{N-1} + \beta_1 \frac{d}{dt} v^{N-1} = \beta_2 v^N + \beta_3 \frac{d}{dt} v^N + \beta_4 \frac{d^2}{dt^2} v^N.$$

Let us define now the new function $z^N(t) = dv^N(t)/dt$. If in the previous expression, we also take into account (2.3) for $j = N - 1$, we get

$$\frac{d}{dt} z^N = \gamma_0 z^N + \gamma_1 v^N + \gamma_2 v^{N-1} + \gamma_3 v^{N-2}$$

for certain coefficients γ_j depending on s_1, s_2, s_3, and s_4. Therefore, we have obtained a system $v_h' = Mv_h$ with $v_h = [z^0, v^0, \dots, v^N, z^N]^T$ and $M \in \mathcal{M}_{(N+2)\times(N+2)}$.

SABC(3,2). Let us consider now the rational function $q = p_1/q_1$ with p_1 and q_1 polynomials of degree 3 and 2 in s, respectively, that interpolates $r(s, h)$ at s_j, $j = 1, \dots, 6$. We obtain this way the following ABC for the right boundary:

$$\beta_0 v^{N-1} + \beta_1 \frac{d}{dt} v^{N-1} + \beta_2 \frac{d^2}{dt^2} v^{N-1} = \beta_3 v^N + \beta_4 \frac{d}{dt} v^N + \beta_5 \frac{d^2}{dt^2} v^N + \beta_6 \frac{d^3}{dt^3} v^N. \tag{7.1}$$

Taking (2.3) into account for $j = N - 1$ and taking the derivative with respect to t, we get

$$\frac{d^2}{dt^2} v^{N-1} = \widetilde{m}_1 \frac{d}{dt} v^N + \widetilde{m}_1 \widetilde{m}_2 v^N + (\widetilde{m}_1^2 + \widetilde{m}_2^2) v^{N-1} + 2\widetilde{m}_1 \widetilde{m}_2 v^{N-2} + \widetilde{m}_1^2 v^{N-3}.$$

On the other hand, let us define the functions $z^N(t) = dv^N(t)/dt$ and $w^N(t) = d^2 v^N(t)/dt^2$. In this way, (7.1) gives rise to

$$\frac{d}{dt} w^N = \gamma_0 w^N + \gamma_1 z^N + \gamma_2 v^N + \gamma_3 v^{N-1} + \gamma_4 v^{N-2} + \gamma_5 v^{N-3}$$

for certain coefficients γ_j depending on the points of interpolation. In this case, we have obtained a system $v_h' = Mv_h$ with $v_h = [w^0, z^0, v^0, \dots, v^N, z^N, w^N]^T$ and $M \in \mathcal{M}_{(N+3)\times(N+3)}$.

Regarding the stability, the case of SABC(1,1) has been studied in [18, 19], where it is proved that the eigenvalues of the matrices of the associated discrete problems have a negative real part when the three interpolatory are equal.

In the other cases, we have checked numerically that all the eigenvalues of the matrices associated to the previous SABCs have negative real part. All of them are nonnormal and a study of the ϵ-pseudospectra shows that the rate of nonnormality is higher for the most ABCs. In this way, Table 1 shows, for an example of what is discussed in section 8, the maximum of the real part of the eigenvalues of $M + E$, where E is a random matrix of norm ϵ and M is the matrix associated to the indicated SABC.

In Figure 1 we can see the behavior of the condition number κ_h for the different SABCs. It is $O(h^{-3/2})$ for SABC(1,1), $O(h^{-2})$ for SABC(2,1), and $O(h^{-7/2})$ for

SABC(3,2). Therefore, for the same stepsize k, the behavior of the error term $e^1_{h,n}$ (6.7) is worse for higher order ABCs.

On the other hand, we have checked numerically (see Figure 2) that in the case $s_1 = s_2 = s_3 = b$, the reflection coefficient for one boundary satisfies

$$|R(V + c\omega, h)| \leq \min\left(1, \frac{(x - x^*)^m}{\min((x^*)^m, (2 - x^*)^m)}\right),$$

where m is the order of the ABCs used. We have also checked the following bound for the reflection coefficients in the case of two boundaries (see Figure 3):

$$\max\left(|K_1(j)|, |K_2(j)|\right) \leq (N + 1)\frac{(x - x^*)^m}{\min((x^*)^m, (2 - x^*)^m)}.$$

This way we see that for a fixed value of h, the higher the order of the ABCs is, the better the reflection coefficient behaves.

The numerical experiments of section 8 confirm the previous estimates.

SABC(2,0). Finally, if we consider the polynomial of order two that interpolates $r(s, h)$ at the points s_1, s_2, and s_3, we obtain a system $v'_h = Mv_h$ with $M \in \mathcal{M}_{(N+1)\times(N+1)}$. Nevertheless, we have checked numerically that this matrix has eigenvalues with positive real part, giving rise to an unstable problem. This result is similar to the one obtained in [2] for ABC(2,0).

8. Numerical experiments.

We are going to consider the Fresnel equation (1.2) with $n = 1$, $n_0 = \cos(21.8°)$, $\lambda = 0.832$, and $k_0 = 2\pi/\lambda$. The experiments we are going to present in this section are similar to those in [2] for the continuous case. We will consider, as in [2, 7, 22, 23], the following kind of initial conditions for the experiments:

$$u_0(x) = \exp\left(-\left(\frac{\bar{x}}{\sigma}\right)^2\right)\exp(i\tau\bar{x}), \quad x \in [0, L], \tag{8.1}$$

with $\bar{x} = x - L/2$ and $\tau = -n_0 k_0 \tan(\alpha)$. This initial condition, which is discretized by taking $r_h u_0(x) = (u_0(jh))_{j\in\mathbf{Z}}$, will give rise to a solution traveling with a velocity $\tan\alpha$. In fact, the exact solution can be calculated explicitly and is given by

$$u(x, t) = \left(1 - \frac{4it}{c\sigma^2}\right)^{-1/2} \exp\left(\frac{i}{c}(\tau^2 - V)t\right)\exp\left(i\tau\bar{x} - \frac{\left(\bar{x} + \frac{2\tau t}{c}\right)^2}{\sigma^2 - \frac{4it}{c}}\right). \tag{8.2}$$

The dispersion relation for the numerical solution is given by

$$\omega(\tau h) = \frac{2}{ch^2}(\cos(\tau h) - 1) + \frac{V}{c}. \tag{8.3}$$

Notice that, although this value is a good approximation to the dispersion relation of the theoretical solution when h is small, in general they are different.

We are going to study numerically the behavior of the SABCs obtained in this paper and compare them with the ABCs previously obtained in [2]. The interpolatory nodes s_j are considered equal to a unique positive number b. We draw in each case the relative L^2 norm error of the numerical solution with respect to the analytical solution given by (8.2), that is,

$$\frac{\|v_{h,n} - P_h r_h u(t_n)\|_{[x^0,x^N],h}}{\|P_h r_h u_0\|_{[x^0,x^N],h}} = \frac{\|v_{h,n} - P_h r_h u(t_n)\|_{[x^0,x^N],h}}{\|r_h u(t_n)\|_h}.$$

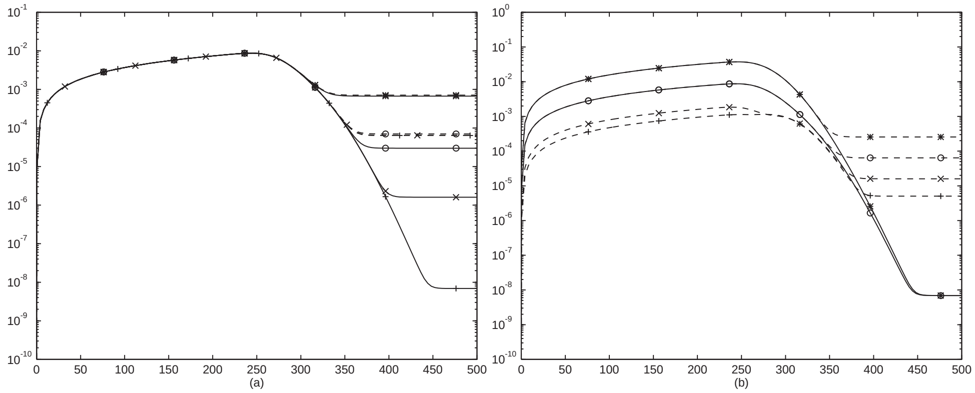


FIG. 4. *Error as a function of time.* (a) – –$*$ *ABC*(1,0), – –$\circ$ *ABC*(1,1), – –$\times$ *ABC*(2,1), – – + *ABC*(3,2), –$*$ *SABC*(1,0), –$\circ$ *SABC*(1,1), –$\times$ *SABC*(2,1), –+ *SABC*(3,2). (b) *ABC*(3,2)*:* – –$*$ $N = 8000$, – –$\circ$ $N = 16000$, – –$\times$ $N = 32000$, – –+ $N = 64000$*; SABC*(3,2)*:* –$*$ $N = 8000$, –$\circ$ $N = 16000$.

Let us consider the Fresnel equation along with the initial condition (8.1) with $\alpha = 20°$, $\sigma = 10$, and $L = 200$. Notice that, numerically, this initial condition is 0 at the boundary. (This assumption is made in the deduction of the TBC in section 2.) We have chosen the optimal value for b given by (3.2) with $\eta = \tau h$, obtained by using the discrete dispersion relation (8.3) and a spatial stepsize $h = .0125$. We have carried out the integration in time with the implicit mid-point rule (IMPR) with a stepsize $k = 0.2$. In Figure 4(a) we can observe the results in terms of relative error for the SABCs previously obtained in this paper (SABC(j_1,j_2)) and for the ABCs obtained in [2] for the continuous problem (ABC(j_1,j_2)). We can observe that the error for $t \leq 300$ is approximately the same for every ABC. This is due to the fact that until that time, the solution is traveling through the interior domain and it has not arrived to the boundary. Therefore, until that time, the error is caused by the discretization in the interior domain, which is the same for every ABC. In this paper, we are interested in the error caused by the different ABCs when the solution reaches the boundary, which for the example of Figure 4(a) happens for $t > 300$ approximately. We see that in every case the results obtained with the SABCs are better than the ones for the ABCs of the same order. This difference is bigger when the order of the boundary condition is higher. Notice that the reflections for ABC(1,1), ABC(2,1), and ABC(3,2) are almost the same. This does not happen for the SABCs for the semidiscrete problem. Let us try to explain this fact.

The ABC(j_1,j_2) has been built as an approximation to the TBC for the continuous problem and not for the problem semidiscretized in space. In this way, the reflection caused by the ABC(j_1,j_2) depends not only on the order of absorption ($j_1 + j_2 + 1$) but also on the difference between the TBC for the continuous problem and that for the semidiscretized one. Thus it is expected that in Figure 4(a), the reflection for ABC(j_1,j_2) will be smaller if h decreases. This can be seen in Figure 4(b) for ABC(3,2) and SABC(3,2). We have considered the same initial condition and the same stepsize k as in Figure 4(a), but now we have taken different spatial stepsizes. We observe that while for SABC(3,2) the reflection does not change with h, for ABC(3,2) the results are better the smaller h is. This way, in order to obtain with ABC(3,2) a similar result to that of SABC(3,2) with $N = 8000, 16000$, we should consider a much

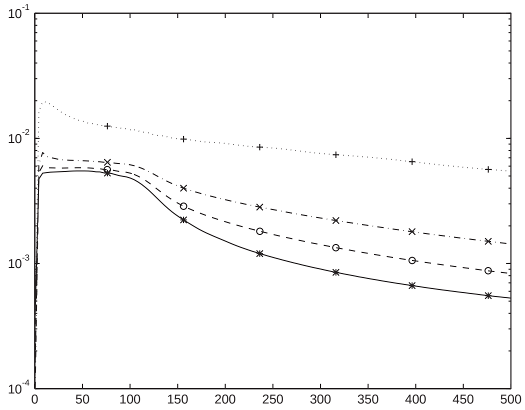


Fig. 5. *Error as a function of time.* –∗ *SABC*(1,0), – –○ *SABC*(1,1), – · –× *SABC*(2,1), · · · + *SABC*(3,2).

smaller value for h. Therefore, the use of the SABCs for the semidiscretized problem allows us to use bigger values of h.

In Figure 4(a) we also observe that the results obtained for the SABC(j_1,j_2) are due to the order of the different SABCs. We have shown in section 4 that the discrete problems associated to these SABC(j_1,j_2) are weakly unstable; nevertheless, this is not visible in Figure 4(a). This is due to the kind of initial condition we are considering, which is zero at the boundary and gives rise to a solution traveling with a fixed velocity. Since we are taking the optimal choice for the interpolatory nodes, there is a great absorption when the solution reaches the boundary. This cancels the bad behavior due to the weak instability of these problems.

It is essential that the initial condition is zero at the boundary as we can see in Figure 5, where we have taken the same initial condition as in the previous experiments but with $L = 40$. Now the value of the initial condition at the boundary is approximately 0.018. Of course, we are breaking a basic hypothesis used in section 2 for the construction of SABCs. We have considered the same stepsizes as in Figure 4(a) and the optimal value for b. The results now are quite different from those observed in Figure 4(a). The instability of the discrete problems is now visible, and we can observe a large initial growth of the error. Moreover, the higher the order of absorption of SABC is, the bigger the growth is. As a consequence, the absorption is worse for higher order SABCs. Similar experiments can be done considering initial conditions that are not regular, obtaining similar results to those in [2].

In Theorem 6.2, we have seen that there exist two elements of the error that can grow when h goes to zero. These terms measure the capacity of absorption of the SABC and the error of the discretization in time. This behavior can be observed by taking a value for b different from the optimal so that these elements are big enough. Let us consider the initial condition (8.1) with $\alpha = 30°$, $\sigma = 3$, and $L = 36$. (Notice it is zero at the boundary.) In Figure 6 we observe the results obtained when we use SABC(3,2), the worst case of instability studied in this paper, with a value for b very different from the optimal. The integration in time is carried out with the IMPR with $k = 0.1$ and values of h decreasing. We see that for $h = 36/80000$, the smaller value considered, the error grows.

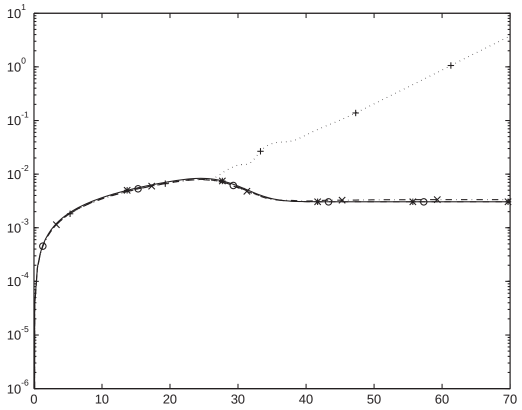


FIG. 6. *Error as a function of time. SABC*(3,2) *IMPR.* $k = 0.1$, $-*$ $N = 10000$, $--\circ$ $N = 20000$, $-\cdot-\times$ $N = 40000$, $\cdots+$ $N = 80000$.

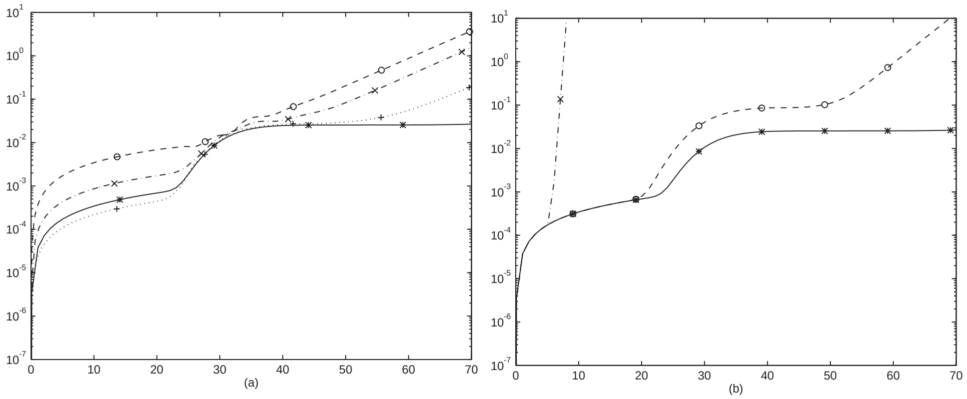


FIG. 7. *Error as a function of time. SABC*(3,2) (a) $N = 80000$*;* $-*$ *DIRK*3 $k = 0.1$*; IMPR:* $-\circ$ $k = 0.1$, $-\cdot-\times$ $k = 0.05$, $\cdots+$ $k = 0.025$. (b) *DIRK*3 $k = 0.1$*;* $-*$ $N = 80000$, $--\circ$ $N = 1600000$, $-\cdot-\times$ $N = 1320000$.

This behavior can be lessened by decreasing k or by increasing the order of the integrator in time. This is seen in Figure 7(a), where we have carried out the integration in time first with a diagonally implicit Runge–Kutta method (DIRK3) of order 3 with $h = 36/80000$, $k = 0.1$, and second with the IMPR for decreasing values of k. With this experiment, the influence of the error term, which is $O(\kappa_h k^p)$ (see section 6), is checked numerically. Nevertheless, we emphasize that the use of an integrator of high order in time will only improve the behavior of this term, but the problem is still unstable. This is observed in Figure 7(b), where we have used DIRK3 with $k = 0.1$ and where we have considered smaller values for h than in Figure 7(a). As expected, the unstable behavior again arises. On the other hand, we remark that the error of absorption is quite large because the value for b is very distinct from the optimal.

The previous experiments have been carried out considering initial conditions that give rise to solutions traveling with a quite small velocity. We have obtained similar conclusions with other numerical experiments made with higher velocities.

REFERENCES

[1] I. Alonso-Mallo and N. Reguera, *Condiciones frontera transparentes y absorbentes para la ecuación de Schrödinger en una dimensión*, in Proceedings of the 16th Congress on Differential Equations and Applications and the 6th Congress on Applied Mathematics, Universidad de Las Palmas de Gran Canaria, Las Palmas de Gran Canaria, Spain, 1999, pp. 467–473.

[2] I. Alonso-Mallo and N. Reguera, *Weak ill-posedness of spatial discretizations of absorbing boundary conditions for Schrödinger-type equations*, SIAM J. Numer. Anal., 40 (2002), pp. 134–158.

[3] I. Alonso-Mallo and N. Reguera, *Discrete absorbing boundary conditions for Schrödinger-type equations. Practical implementation*, Math. Comp., to appear.

[4] B. Alpert, L. Greengard, and T. Hagstrom, *Rapid evaluation of nonreflecting boundary kernels for time-domain wave propagation*, SIAM J. Numer. Anal., 37 (2000), pp. 1138–1164.

[5] V. A. Baskakov and A. V. Popov, *Implementation of transparent boundaries for numerical solution of the Schrödinger equation*, Wave Motion, 14 (1991), pp. 123–128.

[6] L. Di Menza, *Transparent and absorbing boundary conditions for the Schrödinger equation in a bounded domain*, Numer. Funct. Anal. Optim., 18 (1997), pp. 759–775.

[7] T. Fevens and H. Jiang, *Absorbing boundary conditions for the Schrödinger equation,* SIAM J. Sci. Comput., 21 (1999), pp. 255–282.

[8] E. Hairer and G. Wanner, *Solving Ordinary Differential Equations* II, Springer-Verlag, Berlin, 1991.

[9] L. Halpern, *Absorbing boundary conditions for the discretization schemes of the one-dimensional wave equation*, Math. Comp., 38 (1982), pp. 415–429.

[10] L. Halpern and J. Rauch, *Absorbing boundary conditions for diffusion equations*, Numer. Math., 71 (1995), pp. 185–224.

[11] R. L. Higdon, *Radiation boundary conditions for dispersive waves*, SIAM J. Numer. Anal., 31 (1994), pp. 64–100.

[12] W. Huang, C. Xu, S. Chu, and S. Chaudhuri, *The finite-difference vector beam propagation method: Analysis and assessment*, J. Lightwave Technology, 10 (1992), pp. 295–305.

[13] D. Levy and E. Tadmor, *From semidiscrete to fully discrete: Stability of Runge–Kutta schemes by the energy method*, SIAM Rev., 40 (1998), pp. 40–73.

[14] C. Lubich and A. Schädle, *Fast convolution for nonreflecting boundary conditions*, SIAM J. Sci. Comput., 24 (2002), pp. 161–182.

[15] A. Messiah, *Quantum Mechanics, Vol.* I, Interscience, New York, 1961.

[16] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, New York, 1983.

[17] N. Reguera, *Condiciones de frontera transparentes y absorbentes para ecuaciones de tipo Schrödinger*, Ph.D. thesis, Universidad de Valladolid, Valladolid, Spain, 2001.

[18] N. Reguera, *Stability of a class of matrices with applications to absorbing boundary conditions for Schrödinger-type equations*, Appl. Math. Lett., to appear.

[19] N. Reguera, *Analysis of a third order absorbing boundary condition for the Schrödinger equation discretized in space*, Appl. Math. Lett., to appear.

[20] N. Reguera, *Analysis of the Eigenvalues of Matrices Associated to a Discretization of a Schrödinger-Type Equation with Absorbing Boundary Conditions*, Applied Mathematics and Computation Report 5, Universidad de Valladolid, Valladolid, Spain, 2001.

[21] F. Schmidt, *An adaptive approach to the numerical solution of Fresnel's wave equation*, J. Lightwave Technology, 11 (1993), pp. 1425–1434.

[22] F. Schmidt and P. Deuflhard, *Discrete transparent boundary conditions for numerical solutions of Fresnel's equation*, Comput. Math. Appl., 29 (1995), pp. 53–76.

[23] F. Schmidt and D. Yevick, *Discrete transparent boundary conditions for Schrödinger-type equations*, J. Comput. Phys., 134 (1997), pp. 96–107.

[24] L. N. Trefethen, *Pseudospectra of matrices*, in Proceedings of the 14th Dundee Biennial Conference on Numerical Analysis, Pitman Res. Notes Math. Ser. 260, D. F. Griffiths and G. A. Watson, eds., Longman Sci. Tech., Harlow, UK, 1992, pp. 234–266.

[25] L. N. Trefethen and L. Halpern, *Wide-angle one way wave equations*, J. Acoust. Soc. Amer., 84 (1998), pp. 1397–1404.

SIAM J. Numer. Anal.
Vol. 41, No. 5, pp. 1851–1869

ERROR EXPANSION FOR AN UPWIND SCHEME APPLIED TO A TWO-DIMENSIONAL CONVECTION-DIFFUSION PROBLEM*

NATALIA KOPTEVA[†]

Abstract. We consider a singularly perturbed convection-diffusion problem in a rectangular domain. It is solved numerically using a first-order upwind finite-difference scheme on a tensor-product piecewise-uniform Shishkin mesh with $O(N)$ mesh points in each coordinate direction. It is known [G. I. Shishkin, *Grid Approximations of Singularly Perturbed Elliptic and Parabolic Equations*, Russian Academy of Sciences, Ural Branch, Ekaterinburg, Russia, 1992 (in Russian)] that the error is almost-first-order accurate in the maximum norm. We decompose the error into a sum of continuous almost-first-order terms and the almost-second-order residual under the assumption $\varepsilon \le CN^{-1}$, where ε is the singular perturbation parameter and C is a constant. This error expansion is applied to obtain maximum-norm error estimates for the Richardson extrapolation technique and derive bounds on the errors in approximating the derivatives of the true solution by divided differences of the computed solution. The analysis uses a decomposition of the true solution requiring fewer compatibility conditions than in earlier publications. Numerical results are presented that support our theoretical results.

Key words. convection-diffusion, upwind scheme, singular perturbation, error expansion, Richardson extrapolation, approximation of derivatives, Shishkin mesh

AMS subject classifications. 65N06, 65N15, 35C20

DOI. 10.1137/S003614290241074X

1. Introduction.

The main result of this paper is a certain error expansion for the singularly perturbed two-dimensional convection-diffusion problem

$$(1.1)\qquad \begin{aligned} Lu := -\varepsilon\triangle u + b_1u_x + b_2u_y + cu &= f \qquad \text{in } \Omega = (0,1)\times(0,1),\\ u &= 0 \qquad \text{on } \partial\Omega. \end{aligned}$$

Here ε is a small parameter that satisfies $0 < \varepsilon \ll 1$, while $b_1(x,y)$, $b_2(x,y)$, $c(x,y)$ are smooth functions with

$$(1.2a)\qquad b_1(x,y) > \beta_1 > 0, \quad b_2(x,y) > \beta_2 > 0, \quad c(x,y) \ge 0 \quad \text{for all } (x,y) \in \bar\Omega,$$

where β_1, β_2 are positive constants. To simplify the presentation we assume that

$$(1.2b)\qquad \beta_1 = \beta_2 = \beta > 0.$$

Note that all the results of this paper also hold true for the general case (1.2a); see Remarks 1.1 and 4.5.

The solution of problem (1.1) has exponential layers at the outflow boundaries $x = 1$ and $y = 1$ (see [8, 10]). We are interested in ε-uniform numerical methods that resolve the boundary layers. One approach is using layer-adapted highly nonuniform meshes.

Problem (1.1) is discretized using the standard first-order upwind scheme

$$(1.3)\qquad \begin{aligned} L^Nu^N := \big(-\varepsilon(\delta_x^2+\delta_y^2) + b_{1,ij}D_x^- + b_{2,ij}D_y^- + c_{ij}\big)u_{ij}^N &= f_{ij} \qquad \text{in } \Omega^N,\\ u_{ij}^N &= 0 \qquad \text{on } \partial\Omega^N. \end{aligned}$$

*Received by the editors July 9, 2002; accepted for publication (in revised form) January 20, 2003; published electronically October 28, 2003. This work was supported by the Russian Foundation for Basic Research under grants N 02-01-01089, 02-01-06122.

http://www.siam.org/journals/sinum/41-5/41074.html

[†]Faculty of Computational Mathematics and Cybernetics, Lomonosov Moscow State University, Vorob'evy gory, RU-119992 Moscow, Russia (kopteva@cs.msu.su).

Here δ_x^2, δ_y^2, D_x^-, D_y^- are the standard finite-difference differentiation operators; see notation (2.1). Note from [11, 8, 10] that this scheme satisfies the maximum principle.

We discretize on the mesh $\bar\Omega^N = \bar\omega_{\sigma,N} \times \bar\omega_{\sigma,N} = \{(x_i, y_j) \in \bar\Omega :\ i, j = 0, \dots, N\}$ that is the tensor-product of two equal piecewise-uniform meshes. Each of these one-dimensional meshes $\bar\omega_{\sigma,N}$ is constructed by dividing each of the subintervals $[0, 1-\sigma]$ and $[1-\sigma, 1]$ into $N/2$ equal subintervals of width H and h, respectively:

$$x_i = y_i = \begin{cases} iH & \text{for} \quad i = 0, \dots, N/2, \quad \text{where} \quad H = 2(1-\sigma)/N, \\ (1-2\sigma) + ih & \text{for} \quad i = N/2, \dots, N, \quad \text{where} \quad h = 2\sigma/N. \end{cases} \tag{1.4}$$

Shishkin [13] was the first to suggest such piecewise-uniform meshes for problems like (1.1) with the mesh transition parameter $\sigma := \min\{(2/\beta)\,\varepsilon \ln N,\, 1/2\}$. For simplicity we assume that

$$\varepsilon \le CN^{-1}, \tag{1.5}$$

which is not a restriction in practical situations. This assumption implies that

$$\sigma = \frac{2}{\beta}\,\varepsilon \ln N. \tag{1.6}$$

Note also that $N^{-1} < H < 2N^{-1}$ and

$$\frac{h}{\varepsilon} = \frac{4}{\beta} N^{-1} \ln N, \qquad e^{-\beta(1-x_{N/2})/\varepsilon} = e^{-\beta(1-y_{N/2})/\varepsilon} = e^{-\beta\sigma/\varepsilon} = N^{-2}. \tag{1.7}$$

Further, let $\partial\Omega^N$ be the set of mesh points on the boundary, i.e., $\partial\Omega^N = \bar\Omega^N \cap \partial\Omega$, while $\Omega^N = \bar\Omega^N \backslash \partial\Omega^N$ is the set of the internal mesh points.

Thus the domain $\bar\Omega$ is dissected by the transition lines $x = 1-\sigma$ and $y = 1-\sigma$ into four parts

$$\begin{aligned} \Omega_2 &:= [0, 1-\sigma] \times (1-\sigma, 1], & \Omega_{12} &:= (1-\sigma, 1] \times (1-\sigma, 1], \\ \bar\Omega_0 &:= [0, 1-\sigma] \times [0, 1-\sigma], & \Omega_1 &:= (1-\sigma, 1] \times [0, 1-\sigma]. \end{aligned}$$

The restriction of the mesh $\bar\Omega^N$ to each of them is a rectangular uniform mesh.

Remark 1.1. The analogue of $\bar\Omega^N$ for $\beta_1 \ne \beta_2$ is the tensor-product rectangular mesh $\bar\omega_{\sigma_1,N} \times \bar\omega_{\sigma_2,N}$, where σ_k, H_k, h_k, $\bar\omega_{\sigma_k,N}$, for $k = 1, 2$, are defined similarly to σ, H, h, $\bar\omega_{\sigma,N}$ with β_k used instead of β; see, e.g., [8, p. 101].

The paper is organized as follows. Most of the notation is collected in section 2. In section 3 we analyze a decomposition of the true solution into an asymptotic expansion of order one and its residual. This decomposition and our estimates of its components require fewer compatibility conditions than in earlier publications [13, 7].

In section 4 we present a certain error expansion for the upwind scheme (1.3) on the Shishkin mesh (1.4). Shishkin [13] gave an ε-uniform almost-first-order estimate of the error in the discrete maximum norm, which was slightly improved in [5, Remark 3.3] to

$$\|u_{ij}^N - u(x_i, y_j)\| \le CN^{-1} \ln N.$$

We decompose the error into a sum of continuous almost-first-order terms and the almost-second-order residual (Theorem 4.1). This error expansion is applied in subsection 4.1 to obtain maximum-norm error estimates for the Richardson extrapolation technique, and in subsection 4.2 to derive bounds on the errors in approximating the derivatives. Section 5 is devoted to the proof of Theorem 4.1.

Similar error expansions were constructed in [9, 4] for one-dimensional convection-diffusion problems. These error expansions were used there to analyze the Richardson extrapolation technique. We mainly follow the analysis in [9], extending it to the two-dimensional problem. In subsection 4.2 we obtain a two-dimensional analogue of the one-dimensional estimates [1, 3, 4]. We follow the approach of [4], where, to analyze a defect correction method, the error expansion was also used to obtain bounds on the differences of the error in two adjoining nodes.

Richardson extrapolation applied to singularly perturbed problems was also studied in earlier publications of Shishkin [12, 14], where ε-uniform maximum-norm error estimates were obtained for a one-dimensional parabolic problem and a two-dimensional elliptic problem in an infinite strip.

Numerical results supporting our theory are presented in section 6.

2. Notation. Throughout the paper we use the following notation. Let k be a nonnegative integer and $\alpha \in (0,1]$. The standard notation $C^k(\bar\Omega)$ is used for the space of functions whose derivatives up to order k are continuous on $\bar\Omega$, with the norm

$$\|v\|_k = \sum_{0\le l\le k}\sum_{i+j=l}\max_{(x,y)\in\bar\Omega}\left|\frac{\partial^{i+j}}{\partial x^i\partial y^j}v(x,y)\right|.$$

As usual, we simply write $C(\bar\Omega)$ and $\|v\|$ when $k=0$. The notation $C^{k,\alpha}(\bar\Omega)$ is used for the space of Hölder continuous functions with the norm

$$\|v\|_{0,\alpha} = \sup_{x,x'\in\bar\Omega,\ x\neq x'}\frac{|v(x)-v(x')|}{\|x-x'\|_e^\alpha}, \qquad \|v\|_{k,\alpha} = \|v\|_k + \sum_{i+j=k}\left\|\frac{\partial^{i+j}}{\partial x^i\partial y^j}v\right\|_{0,\alpha},$$

where $\|\cdot\|_e$ is the Euclidean norm in R^2. Further, we shall use the notation $C^{1,1}(\bar\Omega)$ when $\alpha=1$, and $C^{1,\alpha}(\bar\Omega)$ only when $\alpha\in(0,1)$.

Let v be a discrete function defined on $\tilde\Omega^N\subset\bar\Omega^N$. By $\|v\|_{\tilde\Omega^N} = \max_{\tilde\Omega^N}|v_{ij}|$ we denote the discrete maximum norm of v on $\tilde\Omega^N$. Sometimes we shall simply write $\|v\|$ when $\tilde\Omega^N=\bar\Omega^N$.

The finite-difference operators are defined in a standard manner by

$$\begin{aligned} &h_i := x_i - x_{i-1}, \qquad D_x^- v_{ij} := \frac{v_{ij}-v_{i-1,j}}{h_i}, \qquad \delta_x^2 v_{ij} := \frac{D_x^- v_{i+1,j} - D_x^- v_{ij}}{(h_i+h_{i+1})/2},\\ &h_j := y_j - y_{j-1}, \qquad D_y^- v_{ij} := \frac{v_{ij}-v_{i,j-1}}{h_j}, \qquad \delta_y^2 v_{ij} := \frac{D_y^- v_{i,j+1} - D_y^- v_{ij}}{(h_j+h_{j+1})/2}.\end{aligned} \tag{2.1}$$

Here v_{ij} is any discrete function. Note that when it is clear that $v(x,y)$ is a continuous function, we shall sometimes use the notation $v_{ij} := v(x_i,y_j)$, while when it is clear that v_{ij} is a discrete function, we shall sometimes use the notation $v(x_i,y_j) := v_{ij}$.

For an arbitrary $\tilde\Omega\subset\bar\Omega$ and arbitrary constants a, b on $\bar\Omega^N$ define the function

$$\mathcal{E}_{ij}(a,\tilde\Omega;\ b) = \begin{cases} a & \text{for } (x_i,y_j)\in\tilde\Omega,\\ b & \text{for } (x_i,y_j)\in\bar\Omega^N\backslash\tilde\Omega.\end{cases}$$

We shall also use other similar notation, e.g., $\mathcal{E}_{ij}(a, i\le N/2;\ b)$.

Throughout the paper, C, sometimes subscripted, denotes a generic positive constant that is independent of ε and any mesh used.

3. Decomposition of the solution. In this section we decompose the solution into an asymptotic expansion of order one and its residual. We estimate the components of this decomposition and their derivatives.

THEOREM 3.1. *Let $\alpha \in (0,1)$, and β be from* (1.2). *Suppose that $f \in C^{3,1}(\bar{\Omega})$ and satisfies the compatibility conditions*

$$f(0,0) = f(0,1) = f(1,1) = f(1,0) = 0, \tag{3.1a}$$

$$\left(\frac{f}{b_1}\right)_y (0,0) = \left(\frac{f}{b_2}\right)_x (0,0), \tag{3.1b}$$

$$\left(\frac{1}{b_1}\left(b_1\frac{\partial}{\partial x} - b_2\frac{\partial}{\partial y} - c\right)\left[\frac{f}{b_1}\right]\right)_y (0,0) = \left(\frac{f}{b_2}\right)_{xx} (0,0). \tag{3.1c}$$

Then the boundary-value problem (1.1) *has a classical solution $u \in C^{3,\alpha}(\bar{\Omega})$, and this solution can be decomposed as*

$$u = (u_0 + v_0 + w_0 + z_0) + \varepsilon(u_1 + v_1 + w_1 + z_1) + \varepsilon^2 R,$$

where $u_0 \in C^{3,1}(\bar{\Omega})$, $u_1 \in C^{1,1}(\bar{\Omega})$, $\frac{\partial^k}{\partial x^k}v_1$, $\frac{\partial^k}{\partial y^k}w_1 \in C^{1,1}(\bar{\Omega})$ for $k \geq 0$, $z_1 \in C^3(\bar{\Omega})$, $R \in C^{1,1}(\bar{\Omega})$, and

$$\|u_0\|_{3,1} + \|u_1\|_{1,1} \leq C, \quad u_0(x,0) = u_0(0,y) = 0, \quad u_{0,xx}(0,0) = u_{0,yy}(0,0) = 0, \tag{3.2}$$

$$\begin{aligned} &v_0(x,y) = -u_0(1,y)e^{-b_1(1,y)(1-x)/\varepsilon}, \qquad w_0(x,y) = -u_0(x,1)e^{-b_2(x,1)(1-y)/\varepsilon}, \\ &z_0(x,y) = u_0(1,1)e^{-b_1(1,1)(1-x)/\varepsilon - b_2(1,1)(1-y)/\varepsilon}, \end{aligned} \tag{3.3}$$

$$\begin{aligned} &\left\|\frac{\partial^k}{\partial x^k}v_1(x,\cdot)\right\|_{1,1,[0,1]} \leq C\varepsilon^{-k}e^{-\beta(1-x)/\varepsilon}, \\ &\left\|\frac{\partial^k}{\partial y^k}w_1(\cdot,y)\right\|_{1,1,[0,1]} \leq C\varepsilon^{-k}e^{-\beta(1-y)/\varepsilon} \qquad \textit{for } \ 0 \leq k \leq 3, \end{aligned} \tag{3.4a}$$

$$\left|\frac{\partial^{k+m}}{\partial x^k \partial y^m} z_1(x,y)\right| \leq C\varepsilon^{-(k+m)}e^{-\beta((1-x)+(1-y))/\varepsilon} \qquad \textit{for } \ 0 \leq k+m \leq 3, \tag{3.4b}$$

$$\|R\| \leq C, \qquad |LR(x,y)| \leq C\big(1 + \varepsilon^{-1}e^{-\beta(1-x)/\varepsilon} + \varepsilon^{-1}e^{-\beta(1-y)/\varepsilon}\big). \tag{3.5}$$

Remark 3.1. In (3.4a) by $\|\frac{\partial^k}{\partial x^k}v_1(x,\cdot)\|_{1,1,[0,1]}$ we denote the norm of the function $\frac{\partial^k}{\partial x^k}v_1(x,y)$ as a function of the variable y in the space $C^{1,1}[0,1]$ of Hölder continuous functions. The second line in (3.4a) should be understood similarly.

Remark 3.2. Note that $C^{1,1}(\bar{\Omega}) = W^{2,\infty}(\Omega)$, and for any function in $C^{1,1}(\bar{\Omega})$ its second partial derivatives exist almost everywhere [2, pp. 151, 154]. Hence, since $R \in C^{1,1}(\bar{\Omega})$, in (3.5) the second inequality is to be understood in the sense that it holds true almost everywhere.

Remark 3.3. Shishkin [13, Theorem III.2.1] decomposed the solution into a smooth part and a layer part so that the layer part lay in the null space of L. A similar decomposition was constructed by Linß and Stynes [7]. They presented a full analysis and the explicit compatibility conditions. The solution was decomposed into an asymptotic expansion of order one and its residual. Then the residual was combined with the smooth part of the solution so that the layer part "almost" lay in the null space of L. Note that the hypotheses of our theorem are weaker than those of [7, Theorem 5.1]. In particular, since we do not combine the smooth part with the residual $\varepsilon^2 R$ and do not estimate the derivatives of the latter, our decomposition is useful only for small values of ε, e.g., under our assumption (1.5), but we require *fewer compatibility conditions* at the corner $(0,0)$.

Proof. We mainly follow the proof and the notation of [7, Theorem 5.1], but omit certain parts of this proof that are unnecessary for our decomposition, and combine certain terms in a different manner.

By [7, Lemma 2.1], the compatibility conditions (3.1a) combined with $f \in C^{3,1}(\bar{\Omega})$ imply that $u \in C^{3,\alpha}(\bar{\Omega})$.

We decompose u as in [7]. Thus, u_0 and u_1 are the solutions of the reduced problems [7, (5.2)]. Note that the boundary conditions $u_0(x,0) = u_0(0,y) = 0$ for u_0 yield $u_{0,xx}(0,0) = u_{0,yy}(0,0) = 0$ in (3.2), while the first estimate in (3.2) is obtained applying [7, Theorem 4.1] twice. First, $u_0 \in C^{3,1}(\bar{\Omega})$ since $f \in C^{3,1}(\bar{\Omega})$, while (3.1) implies the compatibility conditions [7, (4.8a), (4.8b), (4.8c)]. Second, $u_1 \in C^{1,1}(\bar{\Omega})$ since $\triangle u_0 \in C^{1,1}(\bar{\Omega})$, while the compatibility condition $\triangle u_0(0,0) = 0$ corresponds to [7, (4.8a)].

Furthermore, v_1 and w_1 are given explicitly by [7, (5.11b), (5.15b)], while z_1 is the solution of the problem [7, (5.17b), (5.17c)]. By [7, Lemma 5.2], the compatibility condition $f(1,1) = 0$ implies that there exists $z_1 \in C^3(\bar{\Omega})$ satisfying (3.4b).

Estimates (3.5) are derived similarly to [7, (5.31)] and the argument that follows it. Note that in [7] $R \in C^{2,\alpha}(\bar{\Omega})$, while we have $R \in C^{1,1}(\bar{\Omega})$; see Remark 3.2. The first estimate in (3.5) follows from the second by the maximum/comparison principle extended to functions in the Sobolev space $W^{1,2}(\Omega)$ (see [2, section 8.1]). □

4. Error expansion and its applications. In this section we present an expansion of the error of the upwind scheme (1.3) on the Shishkin mesh (1.4), (1.6) into a sum of continuous first-order terms and the second-order residual. This error expansion is applied in subsection 4.1 to obtain ε-uniform maximum-norm error estimates for the Richardson extrapolation technique, and in subsection 4.2 to derive bounds on the errors in approximating the derivatives.

THEOREM 4.1. *Suppose that* (1.5) *and the conditions of Theorem* 3.1 *are satisfied. Let u^N be the solution of the discrete problem* (1.3) *on the mesh* (1.4), (1.6). *Then*

$$u_{ij}^N - u(x_i, y_j) = H\Phi(x_i, y_j) + \left(\frac{h}{\varepsilon}\right)\Psi(x_i, y_j) + \mathcal{R}_{ij}^N, \tag{4.1}$$

where $\Phi(x,y)$ and $\Psi(x,y)$ are defined in terms of u_0, v_0, w_0, and z_0 from Theorem 3.1, *and $\varphi(x,y) \in C^{1,1}(\bar{\Omega})$ such that*

$$\|\varphi\|_{1,1} \le C, \tag{4.2}$$

as follows:

$$\begin{aligned}\Phi(x,y) = \varphi(x,y) &- \varphi(1,y)e^{-b_1(1,y)(1-x)/\varepsilon} - \varphi(x,1)e^{-b_2(x,1)(1-y)/\varepsilon} \\ &+ \varphi(1,1)e^{-b_1(1,1)(1-x)/\varepsilon - b_2(1,1)(1-y)/\varepsilon},\end{aligned} \tag{4.3}$$

$$\begin{aligned}\Psi(x,y) = &\ \varepsilon^{-1}(1-x)\frac{\big(b_1^2(1,y)v_0 + b_1^2(1,1)z_0\big)}{2} \\ &+ \varepsilon^{-1}(1-y)\frac{\big(b_2^2(x,1)w_0 + b_2^2(1,1)z_0\big)}{2},\end{aligned} \tag{4.4}$$

while the residual $\mathcal{R}^N_{ij}$ *satisfies*

$$|\mathcal{R}^N_{ij}| \le CN^{-2}\,\mathcal{E}_{ij}(1,\bar{\Omega}^N_0;\,\ln^2 N). \tag{4.5}$$

Proof. The whole of section 5 is devoted to the proof of this theorem; see also Remark 4.1. □

Remark 4.1. A careful inspection of the proof of Theorem 4.1 shows that

$$L^N(u^N-u) = H\frac{(b_1u_{0,xx}+b_2u_{0,yy})}{2} + \left(\frac{h}{\varepsilon}\right)\left[\frac{\varepsilon b_1(v_{0,xx}+z_{0,xx})}{2}+\frac{\varepsilon b_2(w_{0,yy}+z_{0,yy})}{2}\right]+\dots,$$

where ... denotes the terms whose contribution to the error is of almost-second order; see (5.1), (5.2), (5.7), (5.16), (5.18). The standard approach is to define the auxiliary continuous problems

$$L\bar{\Phi} = \frac{(b_1u_{0,xx}+b_2u_{0,yy})}{2} \text{ in } \Omega, \qquad \bar{\Phi} = 0 \text{ on } \partial\Omega, \tag{4.6a}$$

$$L\bar{\Psi}_1 = \frac{\varepsilon b_1(v_{0,xx}+z_{0,xx})}{2} \text{ in } \Omega, \qquad \bar{\Psi}_1 = 0 \text{ on } \partial\Omega, \tag{4.6b}$$

$$L\bar{\Psi}_2 = \frac{\varepsilon b_2(w_{0,yy}+z_{0,yy})}{2} \text{ in } \Omega, \qquad \bar{\Psi}_2 = 0 \text{ on } \partial\Omega, \tag{4.6c}$$

and derive the error expansion

$$u^N_{ij}-u(x_i,y_j) = H\bar{\Phi}(x_i,y_j) + \left(\frac{h}{\varepsilon}\right)[\bar{\Psi}_1(x_i,y_j)+\bar{\Psi}_2(x_i,y_j)]+\cdots,$$

where ... denotes almost-second-order terms; see, e.g., [9] for the one-dimensional case. Our proof mainly follows the analysis of [9], extending it to the two-dimensional case, but, as we shall see, the solutions of the two-dimensional auxiliary problems (4.6) are only in $C^{1,\alpha}(\bar{\Omega})$ since the first-order compatibility conditions are violated. Since the solutions of (4.6) do not exhibit enough smoothness for our analysis, our error expansion (4.1) uses their asymptotic expansions of order zero; see Remarks 4.2–4.4.

Remark 4.2. $\varphi(x,y)$ used in Theorem 4.1 is the solution of the reduced problem

$$b_1\varphi_x + b_2\varphi_y + c\varphi = \frac{\big(b_1u_{0,xx}+b_2u_{0,yy}\big)}{2} \text{ in } \Omega, \qquad \varphi(x,y)=0 \text{ if } x=0 \text{ or } y=0, \tag{4.7}$$

where u_0 is from Theorem 3.1.

Remark 4.3. $\Phi(x,y)$ in (4.3) is an asymptotic expansion of order zero for the solution $\bar{\Phi}(x,y)$ of problem (4.6a). We chose to use $\Phi(x,y)$ instead of $\bar{\Phi}(x,y)$ since, as we shall prove in Lemma 5.7, $\Phi(x,y) \in C^{1,1}(\bar{\Omega})$, while by [7, Lemma 2.1], we have $\bar{\Phi}(x,y) \in C^{1,\alpha}(\bar{\Omega})$ for $\alpha \in (0,1)$. Note that generally $\bar{\Phi}(x,y) \notin C^{2,\alpha}(\bar{\Omega})$ for any $\alpha \in (0,1)$, since the right-hand side $(b_1u_{0,xx}+b_2u_{0,yy})/2$ does not generally vanish at $(1,1)$ and thus does not satisfy one of the compatibility conditions.

Remark 4.4. Decompose $\Psi(x,y)$ in (4.4) as $\Psi = (\Psi_1+\tilde{\Psi}_1)+(\Psi_2+\tilde{\Psi}_2)$; see (5.8) for details. Note that $\Psi_1(x,y)+\tilde{\Psi}_1(x,y)$ and $\Psi_2(x,y)+\tilde{\Psi}_2(x,y)$ are asymptotic expansions of order zero for the solutions $\bar{\Psi}_1(x,y)$ and $\bar{\Psi}_2(x,y)$ of problems (4.6b), (4.6c). By (3.3), one can easily check that Ψ_1, $\tilde{\Psi}_1$, Ψ_2, and $\tilde{\Psi}_2$ are chosen so that

$$-\varepsilon\Psi_{1,xx}+b_1(1,y)\Psi_{1,x} = \frac{b_1(1,y)\varepsilon v_{0,xx}}{2}, \qquad -\varepsilon\tilde{\Psi}_{1,xx}+b_1(1,1)\tilde{\Psi}_{1,x} = \frac{b_1(1,1)\varepsilon z_{0,xx}}{2},$$

$$-\varepsilon\Psi_{2,yy}+b_2(x,1)\Psi_{2,y} = \frac{b_2(x,1)\varepsilon v_{0,yy}}{2}, \qquad -\varepsilon\tilde{\Psi}_{2,yy}+b_2(1,1)\tilde{\Psi}_{2,y} = \frac{b_1(1,1)\varepsilon z_{0,yy}}{2}.$$

Remark 4.5. If $\beta_1 \neq \beta_2$ and the mesh $\bar\omega_{\sigma_1,N} \times \bar\omega_{\sigma_2,N}$ described in Remark 1.1 is used, then we have a slightly different error expansion:

$$\begin{aligned}u_{ij}^N - u(x_i, y_j) &= H_1\Phi_1(x_i, y_j) + H_2\Phi_2(x_i, y_j)\\ &\quad + \left(\frac{h_1}{\varepsilon}\right)[\Psi_1 + \tilde\Psi_1](x_i, y_j) + \left(\frac{h_2}{\varepsilon}\right)[\Psi_2 + \tilde\Psi_2](x_i, y_j) + \mathcal{R}_{ij}^N.\end{aligned}$$

Here $\Psi_1 + \tilde\Psi_1$ and $\Psi_2 + \tilde\Psi_2$ are the first and the second terms on the right-hand side in (4.4)—see Remark 4.4 and (5.8)—while Φ_1 and Φ_2 are defined by (4.3) with Φ and φ replaced by Φ_k and φ_k for $k = 1, 2$. These functions φ_1 and φ_2 are the solutions of the reduced problem (4.7) with the right-hand side $(b_1 u_{0,xx} + b_2 u_{0,yy})/2$ replaced by $b_1 u_{0,xx}/2$ and $b_2 u_{0,yy}/2$, respectively.

4.1. Richardson extrapolation. Now we shall see that the error expansion given by Theorem 4.1 immediately implies ε-uniform maximum-norm error estimates for the Richardson extrapolation technique.

In this subsection for the mesh $\bar\Omega^N$ we shall use the slightly different notation $\bar\Omega_{\sigma,N} := \bar\Omega^N = \bar\omega_{\sigma,N} \times \bar\omega_{\sigma,N}$. We shall also use the tensor-product rectangular mesh $\bar\Omega_{\sigma,2N} := \bar\omega_{\sigma,2N} \times \bar\omega_{\sigma,2N} = \{(\tilde x_i, \tilde y_j) \in \bar\Omega :\ i, j = 0, \ldots, 2N\}$. Here $\bar\omega_{\sigma,2N}$ is a piecewise-uniform mesh with the meshsizes $h/2$ and $H/2$ obtained uniformly bisecting the original mesh $\bar\omega_{\sigma,N}$. Note that $\bar\omega_{\sigma,2N}$ is also described by (1.4) with the same mesh transition parameter σ (1.6) and N replaced by $2N$. The two rectangular meshes are nested; i.e., $\Omega_{\sigma,N} = \{(x_i, y_j)\} \subset \Omega_{\sigma,2N} = \{(\tilde x_i, \tilde y_j)\}$, and $(x_i, y_j) = (\tilde x_{2i}, \tilde y_{2j})$.

Let $\tilde u_{ij}^{2N} = \tilde u^{2N}(\tilde x_i, \tilde y_j)$ be the solution of the discrete problem (1.3) on the mesh $\Omega_{\sigma,2N}$. Then under the conditions of Theorem 4.1, in addition to (4.1) we have

$$\tilde u^{2N}(\tilde x_i, \tilde y_j) - u(\tilde x_i, \tilde y_j) = \frac{1}{2} H\Phi(\tilde x_i, \tilde y_j)\frac{1}{2}\left(\frac{h}{\varepsilon}\right)\Psi(\tilde x_i, \tilde y_j) + \tilde{\mathcal{R}}^{2N}(\tilde x_i, \tilde y_j).$$

Hence

$$[\, 2\tilde u^{2N}(x_i, y_j) - u_{ij}^N\,] - u(x_i, y_j) = 2\tilde{\mathcal{R}}^{2N}(x_i, y_j) - \mathcal{R}_{ij}^N,$$

and we arrive at the following.

COROLLARY 4.2. *Under the conditions of Theorem* 4.1, *we have*

$$\begin{aligned}\left|[\, 2\tilde u^{2N}(x_i, y_j) - u_{ij}^N\,] - u(x_i, y_j)\right| &\leq CN^{-2}\,\mathcal{E}_{ij}(1, \bar\Omega_0; \ln^2 N)\\ &= C\begin{cases} N^{-2} & \text{in } \bar\Omega^N \cap \bar\Omega_0,\\ N^{-2}\ln^2 N & \text{in } \bar\Omega^N \backslash \bar\Omega_0.\end{cases}\end{aligned}$$

Thus, while the two computed solutions u_{ij}^N and $\tilde u_{ij}^{2N}$ are almost-first-order accurate, their linear combination $[2\tilde u^{2N}(x_i, y_j) - u_{ij}^N]$ is almost-second-order accurate ε-uniformly.

4.2. Approximation of derivatives. In this subsection we apply the error expansion given by Theorem 4.1 to derive bounds on the errors in approximating the derivatives of the true solution by divided differences of the computed solution.

COROLLARY 4.3. *Under the conditions of Theorem* 4.1, *we have*

$$\text{(4.8a)}\quad \left|D_x^- e_{ij}^N\right| + \left|D_x^- u_{ij}^N - u_x(x_{i-1/2}, y_j)\right| \leq C\begin{cases} N^{-1} & \text{in } \bar\Omega^N \cap \bar\Omega_0,\\ N^{-1}\ln^2 N & \text{in } \bar\Omega^N \cap \Omega_2,\\ N^{-1}\ln N/\varepsilon & \text{in } \bar\Omega^N \cap (\Omega_1 \cup \Omega_{12}),\end{cases}$$

$$\text{(4.8b)}\quad \left|D_y^- e_{ij}^N\right| + \left|D_y^- u_{ij}^N - u_y(x_i, y_{j-1/2})\right| \le C \begin{cases} N^{-1} & \textit{in } \bar{\Omega}^N \cap \bar{\Omega}_0, \\ N^{-1}\ln^2 N & \textit{in } \bar{\Omega}^N \cap \Omega_1, \\ N^{-1}\ln N/\varepsilon & \textit{in } \bar{\Omega}^N \cap (\Omega_2 \cup \Omega_{12}), \end{cases}$$

where $e_{ij}^N = u_{ij}^N - u(x_i, y_j)$ *is the error, while* $x_{i-1/2}$ *and* $y_{j-1/2}$ *are the midpoints of the segments* $[x_{i-1}, x_i]$ *and* $[y_{j-1}, y_j]$.

Proof. Since (4.8a) and (4.8b) are similar, we shall prove only bound (4.8a). By Theorem 3.1 and (1.5), (2.1), (1.4), the second inequality (4.8a) follows from the bound on $|D_x^- e_{ij}^N|$, so we need prove only the first bound (4.8a).

By Theorem 4.1, we have

$$\text{(4.9)}\qquad D_x^- e_{ij}^N = H D_x^- \Phi(x_i, y_j) + \left(\frac{h}{\varepsilon}\right) D_x^- \Psi(x_i, y_j) + D_x^- \mathcal{R}_{ij}^N.$$

First, using (4.5), (2.1), (1.4), we get estimate (4.8a) for $|D_x^- \mathcal{R}_{ij}^N|$.

Further in this proof and later throughout the paper, we shall use the inequalities

$$\text{(4.10)}\qquad \begin{aligned} \varepsilon^{-1}(1-x)e^{-b_1(1,y)(1-x)/\varepsilon} &\le Ce^{-\beta(1-x)/\varepsilon}, \\ \varepsilon^{-1}(1-y)e^{-b_2(x,1)(1-y)/\varepsilon} &\le Ce^{-\beta(1-y)/\varepsilon}. \end{aligned}$$

Define

$$\hat{\Phi}(x,y) = \varphi(x,y) - \varphi(x,1)e^{-b_2(x,1)(1-y)/\varepsilon}, \qquad \hat{\Psi}(x,y) = \frac{\varepsilon^{-1}(1-y)b_2^2(x,1)w_0(x,y)}{2}.$$

Since $|D_x^- \tilde{\Phi}_{ij}| \le \max_{\bar{\Omega}} |\hat{\Phi}_x|$, then by (4.2) we have $|D_x^- \hat{\Phi}_{ij}| \le C$. This implies estimate (4.8a) also for $|H D_x^- \hat{\Phi}_{ij}|$.

Similarly, we get $|D_x^- \hat{\Psi}_{ij}| \le C$. Note that in $\bar{\Omega}_0 \cup \Omega_1$ we have the sharper estimate $|D_x^- \hat{\Psi}_{ij}| \le CN^{-2}$, since (3.3) and (1.7) imply that $\max_{\bar{\Omega}_0 \cup \Omega_1} |\hat{\Psi}_x| \le CN^{-2}$. Hence, $|(h/\varepsilon) D_x^- \hat{\Phi}_{ij}|$ also satisfies inequality (4.8a).

We proceed similarly with $D_x^-(\Phi - \hat{\Phi})_{ij}$ and $D_x^-(\Psi - \hat{\Psi})_{ij}$. Using (4.3), (4.4), (3.3), we obtain $|D_x^-(\Phi - \hat{\Phi})_{ij}| + |D_x^-(\Psi - \hat{\Psi})_{ij}| \le 1/\varepsilon$. However, in $\bar{\Omega}_0 \cup \Omega_2$ we need sharper estimates. By (2.1), we have $|D_x^-(\Phi - \hat{\Phi})_{ij}| \le (2/H) \max_{\bar{\Omega}_0 \cup \Omega_2} |\Phi - \hat{\Phi}|$. Combining this with (3.3), (1.7), we get $|D_x^-(\Phi - \hat{\Phi})_{ij}| \le CN^{-1}$ in $\bar{\Omega}_0 \cup \Omega_2$. Similarly, $|D_x^-(\Psi - \hat{\Psi})_{ij}| \le CN^{-1}$ in $\bar{\Omega}_0 \cup \Omega_2$. Hence, $|H D_x^-(\Phi - \hat{\Phi})_{ij}|$ and $|(h/\varepsilon) D_x^-(\Psi - \hat{\Psi})_{ij}|$ also satisfy inequality (4.8a).

Combining the estimates that we derived for the right-hand terms in (4.9), we obtain the first bound (4.8a). This completes the proof. □

5. Proof of Theorem 4.1.

5.1. Discrete maximum/comparison principle and its corollaries. In this subsection we state the comparison lemmas that will be used to prove Theorem 4.1.

It is well known that the upwind scheme (1.3) satisfies the *discrete maximum/comparison principle*, which implies the following comparison lemma.

Lemma 5.1. *Let* $\tilde{\Omega}^N$ *be a connected submesh of* Ω^N.

(i) *If* $|L^N v_{ij}| \le L^N B_{ij}$ *in* $\tilde{\Omega}^N$ *and* $|v_{ij}| \le B_{ij}$ *on* $\partial\tilde{\Omega}^N$, *then* $|v_{ij}| \le B_{ij}$ *in* $\tilde{\Omega}^N$.

(ii) *If* $v_{ij} = 0$ *on* $\partial\tilde{\Omega}^N$, *then* $\|v\|_{\tilde{\Omega}^N} \le \beta^{-1}\|L^N v\|_{\tilde{\Omega}^N}$.

(iii) *If* $L^N v_{ij} = 0$ *in* $\tilde{\Omega}^N$, *then* $\|v\|_{\tilde{\Omega}^N} \le \|v\|_{\partial\tilde{\Omega}^N}$.

Proof. See [11, Chapter IV] and [8, Chapter 13]. □

The following three lemmas follow from Lemma 5.1(i). We defer their proofs to Appendix A.

LEMMA 5.2. *If $L^N v_{ij} = 0$ in Ω^N and $|v_{ij}| \le e^{-\beta(1-x_i)/\varepsilon}$ on $\partial\Omega^N$, then $|v_{ij}| \le CN^{-2}$ for $i \le N/2$.*

LEMMA 5.3. (i) *If $|L^N v_{ij}| \le e^{-\beta(1-x_i)/\varepsilon}$ in Ω^N and $v_{ij} = 0$ on $\partial\Omega^N$, then $|v_{ij}| \le CN^{-1}$ in $\bar\Omega^N$, and $|v_{ij}| \le CN^{-2}$ for $i \le N/2$.*

(ii) *If $|L^N v_{ij}| \le \mathcal{E}_{ij}(\varepsilon^{-1}e^{-\beta(1-x_i)/\varepsilon}, i > N/2;\ 0)$ in Ω^N and $v_{ij} = 0$ on $\partial\Omega^N$, then $|v_{ij}| \le C\mathcal{E}_{ij}(N^{-1}, i \le N/2;\ 1)$.*

(iii) *Let $|L^N v_{ij}| \le \varepsilon^{-1}e^{-\beta(1-x_i)/\varepsilon}$ for $i > N/2$, where v_{ij} is defined for $i = N/2, \dots, N$, $j = 0, \dots, N$, and $v_{ij} = 0$ on the boundary of this submesh, i.e., if $i = N/2, N$ or $j = 0, N$. Then $|v_{ij}| \le C$ for $i \ge N/2$.*

Remark 5.1. Clearly, the analogues of Lemmas 5.2 and 5.3, with x, i replaced by y, j, also hold true.

LEMMA 5.4. *If $|L^N v_{ij}| \le \mathcal{E}_{ij}(0, \bar\Omega_0;\ 1)$ and $v_{ij} = 0$ on $\partial\Omega^N$, then $|v_{ij}| \le C\varepsilon\, \mathcal{E}_{ij}(1, \bar\Omega_0;\ \ln N) \le CN^{-1}\, \mathcal{E}_{ij}(1, \bar\Omega_0;\ \ln N)$.*

5.2. Error and truncation error. We shall derive a representation of the error

$$e_{ij}^N := u_{ij}^N - u(x_i, y_j).$$

One can easily check that $L^N e_{ij}^N = -L^N u_{ij} + (Lu)_{ij} =: r_{ij}[u]$. Here for the truncation error we have used the notation

$$r_{ij}[v] := -L^N v_{ij} + (Lv)_{ij}. \tag{5.1}$$

Recalling the decomposition of u given by Theorem 3.1, we have

$$L^N e^N = r[u_0 + v_0 + w_0 + z_0] + \varepsilon r[u_1 + v_1 + w_1 + z_1] + \varepsilon^2 r[R]. \tag{5.2}$$

Furthermore, we study the contributions to the error of each of the right-hand side terms separately.

In this section and the related appendices we shall use the *notation*

$$\begin{aligned} L_1 v &:= -\varepsilon v_{xx} + b_1(x,y)v_x, & L_2 v &:= -\varepsilon v_{yy} + b_2(x,y)v_y,\\ L_1^N v &:= -\varepsilon\delta_x^2 v + b_1(x,y)D_x^- v, & L_2^N v &:= -\varepsilon\delta_y^2 v + b_2(x,y)D_y^- v, \end{aligned} \tag{5.3}$$

$$r_{1,ij}[v] := -L_1^N v_{ij} + (L_1 v)_{ij}, \qquad r_{2,ij}[v] := -L_2^N v_{ij} + (L_2 v)_{ij}, \tag{5.4}$$

so that $L = L_1 + L_2 + c$, $L^N = L_1^N + L_2^N + c$, and $r[v] = r_1[v] + r_2[v]$.

5.3. Contribution of $\varepsilon^2 r[R]$ in the maximum norm. The contribution to the error of this component of the right-hand side in (5.2) is described by the following result.

LEMMA 5.5. *If $L^N w_{ij} = r_{ij}[R]$ in Ω^N, where R is from (3.5), and $w_{ij} = 0$ on $\partial\Omega^N$, then $\|w\| \le C\varepsilon^{-1}N^{-1}$.*

Proof. Obviously,

$$\|w\| \le \|w + R\| + \|R\|. \tag{5.5}$$

Since $r_{ij}[R] = -L^N R_{ij} + (LR)_{ij}$, we have $L^N[w + R]_{ij} = (LR)_{ij}$. Recalling (3.5) and applying Lemmas 5.1(ii), 5.3(i), and 5.1(iii), we get

$$\|w + R\| \le C(1 + \varepsilon^{-1}N^{-1}) + \max_{\partial\Omega^N} |R_{ij}|.$$

Combining this with (5.5), (3.5) and observing that (1.5) implies $1 \le C\varepsilon^{-1}N^{-1}$, we complete the proof. $\square$

Remark 5.2. The proof of this lemma does not use any estimates of the derivatives of R and thus allows us to use a decomposition of the solution requiring fewer compatibility conditions; see Remark 3.3.

Now, by (1.5), we have the following.

COROLLARY 5.6. *If $L^N v_{ij} = \varepsilon^2 r_{ij}[R]$ in Ω^N, where R is from* (3.5), *and $v_{ij} = 0$ on $\partial\Omega^N$, then $|v| \le CN^{-2}$.*

5.4. Contribution of $r[u_0]$. The contribution to the error of this component of the right-hand side in (5.2) is described by the following two lemmas.

LEMMA 5.7. (i) *The reduced problem* (4.7) *has a solution $\varphi(x,y) \in C^{1,1}(\bar\Omega)$ such that $\|\varphi\|_{1,1} \le C$, and thus $\Phi(x,y)$ from* (4.3) *using this function φ is also in $C^{1,1}(\bar\Omega)$.*

(ii) *If w satisfies*

$$L^N w_{ij} = \frac{\big(b_1 u_{0,xx} + b_2 u_{0,yy}\big)_{ij}}{2} \text{ in } \Omega^N, \quad w_{ij} = 0 \quad \text{on } \partial\Omega^N, \tag{5.6}$$

then

$$|w_{ij} - \Phi(x_i, y_j)| \le CN^{-1}\, \mathcal{E}_{ij}(1, \bar\Omega_0;\ \ln N).$$

Proof. (i) Note that φ is the solution of the reduced problem (4.7) with the right-hand side $\big(b_1 u_{0,xx} + b_2 u_{0,yy}\big)/2$, which, by (3.2), is in $C^{1,1}(\bar\Omega)$ and vanishes at the corner $(0,0)$, i.e., satisfies the compatibility condition [7, (4.8a)]. Hence, applying [7, Theorem 4.1], we have $\varphi(x,y) \in C^{1,1}(\bar\Omega)$. This implies that $\Phi(x,y) \in C^{1,1}(\bar\Omega)$.

(ii) This part of the proof is given in Appendix B. $\square$

LEMMA 5.8. *If $L^N v_{ij} = r_{ij}[u_0]$ in Ω^N, where u_0 is from Theorem* 3.1, *and $v_{ij} = 0$ on $\partial\Omega^N$, then*

$$|v_{ij} - H\Phi(x_i, y_j)| \le CN^{-2}\, \mathcal{E}_{ij}(1, \bar\Omega_0;\ \ln N).$$

Proof. Recalling (5.1) and using Taylor series expansions and (3.2), we obtain

$$\big|r_{ij}[u_0] - (h_i b_1 u_{0,xx} + h_j b_2 u_{0,yy})_{ij}/2\big| \le C(\varepsilon N^{-1} + N^{-2})\|u_0\|_3 \le CN^{-2}.$$

Furthermore, since $h_i = h_j = H$ for $(x_i, y_j) \in \bar\Omega_0$, we have

$$\big|r_{ij}[u_0] - H\,(b_1 u_{0,xx} + b_2 u_{0,yy})_{ij}\big| \le C[N^{-1}\, \mathcal{E}_{ij}(0, \bar\Omega_0;\ 1) + N^{-2}]. \tag{5.7}$$

Combining this with $L^N(v_{ij} - Hw_{ij}) = r_{ij}[u_0] - H(b_1 u_{0,xx} + b_2 u_{0,yy})_{ij}$, where w_{ij} is from Lemma 5.7, we get

$$\big|L^N(v_{ij} - Hw_{ij})\big| \le C[N^{-1}\, \mathcal{E}_{ij}(0, \bar\Omega_0;\ 1) + N^{-2}].$$

Now, applying Lemmas 5.4 and 5.1(ii), we have

$$|v_{ij} - Hw_{ij}| \le CN^{-2}\, \mathcal{E}_{ij}(1, \bar\Omega_0;\ \ln N).$$

By Lemma 5.7, this yields the statement of the lemma. $\square$

5.5. Contribution of $r[v_0 + w_0 + z_0]$. Now we shall study the contribution to the error of the component $r[v_0 + w_0 + z_0]$ of the right-hand side in (5.2).

The main result of this subsection is the following.

LEMMA 5.9. *If $L^N v_{ij} = r_{ij}[v_0 + w_0 + z_0]$ in Ω^N, where v_0, w_0, z_0 are from Theorem* 3.1 *and $v_{ij} = 0$ on $\partial\Omega^N$, then*

$$\left|v_{ij} - \left(\frac{h}{\varepsilon}\right) \Psi(x_i, y_j)\right| \leq CN^{-2}\, \mathcal{E}_{ij}(1, \bar{\Omega}_0;\ \ln^2 N),$$

where Ψ is from (4.4).

The whole subsection is devoted to the *proof* of this lemma.

Decompose Ψ from (4.4) as $\Psi = \Psi_1 + \Psi_2 + \tilde{\Psi}_1 + \tilde{\Psi}_2$, where

$$\begin{aligned} \Psi_1(x,y) &:= \frac{\varepsilon^{-1}(1-x)b_1^2(1,y)v_0}{2}, & \tilde{\Psi}_1(x,y) &:= \frac{\varepsilon^{-1}(1-x)b_1^2(1,1)z_0}{2}, \\ \Psi_2(x,y) &:= \frac{\varepsilon^{-1}(1-y)b_2^2(x,1)w_0}{2}, & \tilde{\Psi}_2(x,y) &:= \frac{\varepsilon^{-1}(1-y)b_2^2(1,1)z_0}{2}. \end{aligned} \tag{5.8}$$

Regarding the components of this decomposition, see Remark 4.4.

Now decompose v from Lemma 5.9 as $v_{ij} = V_{ij} + W_{ij} + Z_{ij}$, where

$$L^N V = r[v_0] \text{ in } \Omega^N, \qquad \left|V - \left(\frac{h}{\varepsilon}\right)\Psi_1\right| \leq CN^{-2} \text{ on } \partial\Omega^N, \tag{5.9a}$$

$$L^N W = r[w_0] \text{ in } \Omega^N, \qquad \left|W - \left(\frac{h}{\varepsilon}\right)\Psi_2\right| \leq CN^{-2} \text{ on } \partial\Omega^N, \tag{5.9b}$$

$$L^N Z = r[z_0] \text{ in } \Omega^N, \qquad \left|Z - \left(\frac{h}{\varepsilon}\right)(\tilde{\Psi}_1 + \tilde{\Psi}_2)\right| \leq CN^{-2} \text{ on } \partial\Omega^N. \tag{5.9c}$$

Note that such a decomposition of the boundary condition $v_{ij} = 0$ on $\partial\Omega^N$ is possible. Indeed, if $x = 1$ or $y = 1$, we have $\Psi_1(x,y) + \tilde{\Psi}_1(x,y) = \Psi_2(x,y) + \tilde{\Psi}_2(x,y) = 0$, while if $x = 0$ or $y = 0$, we have $|\Psi_1| + |\Psi_2| + |\tilde{\Psi}_1| + |\tilde{\Psi}_2| \leq C\varepsilon^{-2} \leq CN^{-2}$. Hence, $|\Psi_1 + \Psi_2 + \tilde{\Psi}_1 + \tilde{\Psi}_2| \leq CN^{-2}$ on $\partial\Omega^N$.

Since $(x_i, y_j) \in \bar{\Omega}_0$ if both $i, j \leq N/2$, Lemma 5.9 follows from (5.10):

$$\left|V_{ij} - \left(\frac{h}{\varepsilon}\right)\Psi_1(x_i, y_j)\right| \leq CN^{-2}\, \mathcal{E}_{ij}\left(\ln^2 N, i > \frac{N}{2};\ 1\right), \tag{5.10a}$$

$$\left|W_{ij} - \left(\frac{h}{\varepsilon}\right)\Psi_2(x_i, y_j)\right| \leq CN^{-2}\, \mathcal{E}_{ij}\left(\ln^2 N, j > \frac{N}{2};\ 1\right), \tag{5.10b}$$

$$\left|Z_{ij} - \left(\frac{h}{\varepsilon}\right)[\tilde{\Psi}_1(x_i, y_j) + \tilde{\Psi}_2(x_i, y_j)]\right| \leq CN^{-2}\, \mathcal{E}_{ij}]\left(\ln^2 N,\ i, j > \frac{N}{2};\ 1\right). \tag{5.10c}$$

Further, we shall prove that (5.10) follows from the following two lemmas.

LEMMA 5.10. *For V, W, Z from* (5.9) *we have*

$$|V_{ij}| \leq CN^{-2} \quad \textit{for } i \leq \frac{N}{2}, \tag{5.11a}$$

$$|W_{ij}| \leq CN^{-2} \quad \textit{for } j \leq \frac{N}{2}, \tag{5.11b}$$

$$|Z_{ij}| \leq CN^{-2} \quad \textit{if} \quad i \leq \frac{N}{2} \quad \textit{or} \quad j \leq \frac{N}{2}. \tag{5.11c}$$

Proof. We defer the proof of this lemma to Appendix C. □

Define the auxiliary discrete functions $\psi_{1,ij}$ for $i = N/2,\dots,N$, $j = 0,\dots,N$; $\psi_{2,ij}$ for $i = 0,\dots,N$, $j = N/2,\dots,N$; and $\tilde\psi_{1,ij}$, $\tilde\psi_{2,ij}$ for $i, j = N/2,\dots,N$ as follows. Let them satisfy the discrete equations

$$(5.12a)\qquad (L^N\psi_1)_{ij} = \frac{\varepsilon(b_1 v_{0,xx})_{ij}}{2} \quad \text{for } i = \frac{N}{2}+1,\dots,N-1,\ j = 1,\dots,N-1,$$

$$(5.12b)\qquad (L^N\psi_2)_{ij} = \frac{\varepsilon(b_2 w_{0,yy})_{ij}}{2} \quad \text{for } i = 1,\dots,N-1,\ j = \frac{N}{2}+1,\dots,N-1,$$

$$(5.12c)\qquad (L^N\tilde\psi_1)_{ij} = \frac{\varepsilon(b_1 z_{0,xx})_{ij}}{2} \quad \text{for } i,j = \frac{N}{2}+1,\dots,N-1,$$

$$(5.12d)\qquad (L^N\tilde\psi_2)_{ij} = \frac{\varepsilon(b_2 z_{0,yy})_{ij}}{2} \quad \text{for } i,j = \frac{N}{2}+1,\dots,N-1,$$

and the following conditions on the boundaries of the submeshes, where they are defined:

$$(5.13a)\qquad \psi_{1,ij} = \Psi_1(x_i,y_j) \quad \text{if } \ i = \frac{N}{2},\, N \quad \text{or } j = 0,\, N,$$

$$(5.13b)\qquad \psi_{2,ij} = \Psi_2(x_i,y_j) \quad \text{if } \ i = 0,\, N \quad \text{or } j = \frac{N}{2},\, N,$$

$$(5.13c)\qquad \tilde\psi_{k,ij} = \tilde\Psi_k(x_i,y_j) \quad \text{if } \ i = \frac{N}{2},\, N \quad \text{or } j = \frac{N}{2},\, N, \qquad k = 1,2.$$

Lemma 5.11. *For ψ_1, ψ_2, $\tilde\psi_1$, $\tilde\psi_1$ defined by* (5.12), (5.13) *and Ψ_1, Ψ_2, $\tilde\Psi_1$, $\tilde\Psi_1$ from* (5.8) *we have*

$$(5.14a)\qquad |\psi_{1,ij} - \Psi_1(x_i,y_j)| \le C\left(\frac{h}{\varepsilon}\right) \quad \textit{for } i = \frac{N}{2}+1,\dots,N,$$

$$(5.14b)\qquad |\psi_{2,ij} - \Psi_2(x_i,y_j)| \le C\left(\frac{h}{\varepsilon}\right) \quad \textit{for } j = \frac{N}{2}+1,\dots,N,$$

$$(5.14c)\qquad |\tilde\psi_{k,ij} - \tilde\Psi_k(x_i,y_j)| \le C\left(\frac{h}{\varepsilon}\right) \quad \textit{for } i,j = \frac{N}{2}+1,\dots,N, \quad k = 1,2.$$

Proof. This lemma is proved in Appendix C. □

Lemma 5.12. *Estimates* (5.10) *follow from Lemmas* 5.10 *and* 5.11.

Proof. To get the statement of this Lemma, it suffices to prove that

(a) estimate (5.10a) follows from (5.11a) and (5.14a),

(b) estimate (5.10b) follows from (5.11b) and (5.14b),

(c) estimate (5.10c) follows from (5.11c) and (5.14c).

(a) By (5.8), (3.3), (4.10), (1.7), we have $|\Psi_1(x_i,y_j)| \le CN^{-2}$ for $i \le N/2$. Combining this with (5.11a), we get (5.10a) for $i \le N/2$. Since we have (5.14a), then to obtain (5.10a) for $i > N/2$, it suffices to prove that

$$(5.15)\qquad \left|V_{ij} - \left(\frac{h}{\varepsilon}\right)\psi_{1,ij}\right| \le C\left(\frac{h}{\varepsilon}\right)^2 \quad \text{for } i > \frac{N}{2},$$

where ψ_1 is defined by (5.12a), (5.13a). Recalling the notation (5.4) and using Taylor series expansions and (3.2), (3.3), for $i > N/2$ we get

$$\left| r_{1,ij}[v_0] - \frac{h(b_1 v_{0,xx})_{ij}}{2} \right| \le Ch^2 \varepsilon^{-3} e^{-\beta(1-x_{i+1})/\varepsilon}, \tag{5.16}$$

$$\big| r_{2,ij}[v_0] \big| \le (2\varepsilon + b_{2,ij} N^{-1}) \max_{y\in[0,1]} |v_{0,yy}(x_i, y)| \le CN^{-1} e^{-\beta(1-x_i)/\varepsilon}.$$

Note that, by (1.7), (1.5), we have $N^{-1} \le Ch^2\varepsilon^{-3}$ and $e^{-\beta(1-x_{i+1})/\varepsilon} \le Ce^{-\beta(1-x_i)/\varepsilon}$, while $L^N[V - (h/\varepsilon)\psi_1] = r_1[v_0] + r_2[v_0] - hb_1 v_{0,xx}/2$. Hence,

$$\left| L^N \left[V_{ij} - \left(\frac{h}{\varepsilon}\right) \psi_{1,ij} \right] \right| \le C \left(\frac{h}{\varepsilon}\right)^2 \varepsilon^{-1} e^{-\beta(1-x_i)/\varepsilon} \quad \text{for } i > \frac{N}{2}.$$

Note that (5.9a), (5.13a) imply $|V - (h/\varepsilon)\psi_1| \le CN^{-2}$ on $\partial\Omega^N$, while (5.10a), (5.13a) imply $|V_{ij} - (h/\varepsilon)\psi_{1,ij}| = |V_{ij} - (h/\varepsilon)\Psi_1(x_i, y_j)| \le CN^{-2}$ for $i = N/2$. Now, applying Lemmas 5.3(iii) and 5.1(iii), we obtain (5.15). This completes part (a) of the proof.

(b) This part of the proof is analogous to part (a).

(c) Since this part of the proof is similar to part (a), we skip certain details. By (5.8), (3.3), (4.10), (1.7), we have $|\tilde\Psi_1(x_i, y_j)| + |\tilde\Psi_2(x_i, y_j)| \le CN^{-2}$ if $i \le N/2$ or $j \le N/2$. Combining this with (5.11c), we get (5.10c) if $i \le N/2$ or $j \le N/2$. Since we have (5.14c), then, to obtain (5.10c) for $i, j > N/2$, it suffices to prove that

$$\left| Z_{ij} - \left(\frac{h}{\varepsilon}\right) (\tilde\psi_{1,ij} + \tilde\psi_{2,ij}) \right| \le C \left(\frac{h}{\varepsilon}\right)^2 \quad \text{for } i, j > \frac{N}{2}. \tag{5.17}$$

Note that $L^N[Z - (h/\varepsilon)(\tilde\psi_1 + \tilde\psi_2)] = (r_1[z_0] - hb_1 z_{0,xx}/2) + (r_2[z_0] - hb_2 z_{0,yy}/2)$. Hence, using Taylor series expansions and (3.2), (3.3), for $i, j > N/2$ we get

$$\left| \left(r_1[z_0] - \frac{hb_1 z_{0,xx}}{2} + r_2[z_0] - \frac{hb_2 z_{0,yy}}{2} \right)_{ij} \right| \le Ch^2 \varepsilon^{-3} (e^{-\beta(1-x_i)/\varepsilon} + e^{-\beta(1-y_j)/\varepsilon}), \tag{5.18}$$

which yields

$$\left| L^N \left[Z_{ij} - \left(\frac{h}{\varepsilon}\right) (\tilde\psi_{1,ij} + \tilde\psi_{2,ij}) \right] \right| \le C \left(\frac{h}{\varepsilon}\right)^2 \varepsilon^{-1} (e^{-\beta(1-x_i)/\varepsilon} + e^{-\beta(1-y_j)/\varepsilon}).$$

Combining this with the boundary conditions from (5.9c), (5.10c), (5.13c) and applying Lemmas 5.3(iii) and 5.1(iii), we obtain (5.17). This completes the proof. □

Proof of Lemma 5.9. By Lemmas 5.10, 5.11, and 5.12, we have (5.10), which yields the statement of Lemma 5.9. □

5.6. Contribution of $\varepsilon r[u_1 + v_1 + w_1 + z_1]$.

The contribution to the error of this component of the right-hand side in (5.2) is described by the following lemma.

LEMMA 5.13. *If* $L^N v_{ij} = \varepsilon r_{ij}[u_1 + v_1 + w_1 + z_1]$ *in* Ω^N, *where* u_1, v_1, w_1, z_1 *are from Theorem* 3.1, *and* $v_{ij} = 0$ *on* $\partial\Omega^N$, *then*

$$|v_{ij}| \le CN^{-2}\, \mathcal{E}_{ij}(1, \bar\Omega_0;\, \ln N).$$

Proof. Since this result is very close to the well-known theorem by Shishkin [13, Theorem 2.3], [8, Theorem 13.2], while the argument is standard, we shall only sketch the proof. Note that it simplifies the argument that the truncation error in the right-hand side is multiplied by ε. By (3.2), (1.5), we have $|\varepsilon r[u_1]| \le C\varepsilon N^{-1} \le CN^{-2}$. By (3.4), (1.7), we get $|\varepsilon r_{1,ij}[v_1 + z_1]| \le C(h/\varepsilon)e^{-\beta(1-x_i)/\varepsilon}$ for $i > N/2$, and $|\varepsilon r_{1,ij}[v_1 + z_1]| \le Ce^{-\beta(1-x_{i+1})/\varepsilon} \le CN^{-2}$ for $i \le N/2$. The term $\varepsilon r_2[w_1 + z_1]$ is estimated similarly. We have to be careful with Taylor series expansions of v_1 and w_1 since $\frac{\partial^k}{\partial x^k}v_1$ and $\frac{\partial^k}{\partial y^k}w_1$ are generally in $C^{1,1}(\bar{\Omega})$. By (3.4a) and Remark 3.1, we estimate as follows:

$$\begin{aligned}
\big|r_{2,ij}[v_1]\big| &\le C\big\|v_1(x_i,\cdot)\big\|_{1,1,[0,1]} \le Ce^{-\beta(1-x_i)/\varepsilon},\\
\big|r_{1,ij}[w_1]\big| &\le C\big\|w_1(\cdot,y_j)\big\|_{1,1,[0,1]} \le Ce^{-\beta(1-y_j)/\varepsilon}.
\end{aligned}$$

Combining our estimates of all the components of the right-hand side and applying Lemmas 5.1(ii) and 5.3(i),(ii), we get the statement of the lemma. □

5.7. Proof of Theorem 4.1. The statement of the theorem is obtained by recalling (5.1), (5.2) and combining Corollary 5.6 and Lemmas 5.8, 5.9, 5.13. □

6. Numerical results. In this section we present numerical results illustrating our estimates for the Richardson extrapolation technique (Corollary 4.2) and on the errors in approximating the derivatives (Corollary 4.3).

We study the performance of the upwind scheme and the Richardson extrapolation technique when applied to the test problem from [6] in which $b_1 = 2$, $b_2 = 3$, $c = 1$,

$$u(x,y) = 2\sin x\,(1 - e^{-2(1-x)/\varepsilon})\,y^2(1 - e^{-3(1-y)/\varepsilon}),$$

and the right-hand side f is chosen so that (1.1) is satisfied. This problem was solved numerically using the upwind scheme (1.3) on the tensor-product piecewise-uniform Shishkin mesh from Remark 1.1 with $\beta_1 = 1.9$, $\beta_2 = 2.9$.

In Table 6.1 we present the errors before and after the Richardson extrapolation. The odd rows contain the maximum nodal errors $e^N := \|u^N_{ij} - u(x_i, y_j)\|$ in the specified subdomains of $\bar{\Omega}$, while the even rows contain the rates of convergence computed by the standard formula $r(e^N) = \log_2(e^N/e^{2N})$. Clearly, the Richardson extrapolation technique decreases the nodal errors and increases the rates of convergence. Note that the errors are very similar for $\varepsilon = 10^{-6}$ and $\varepsilon = 10^{-8}$, which confirms that our estimates are ε-uniform. The rates of convergence are slightly worse than predicted by Corollary 4.2. However, since our rates of convergence are consistent with those for the analogous one-dimensional problems [9, 4], we expect the rates of convergence to increase as N increases, similarly to [9, 4].

TABLE 6.1
Maximum nodal errors before and after Richardson extrapolation.

	$\varepsilon = 10^{-6}$				$\varepsilon = 10^{-8}$			
	Before extrapolation		After extrapolation		Before extrapolation		After extrapolation	
N	$\bar{\Omega}_0$	$\bar{\Omega}\backslash\bar{\Omega}_0$	$\bar{\Omega}_0$	$\bar{\Omega}\backslash\bar{\Omega}_0$	$\bar{\Omega}_0$	$\bar{\Omega}\backslash\bar{\Omega}_0$	$\bar{\Omega}_0$	$\bar{\Omega}\backslash\bar{\Omega}_0$
32	4.944e-2	1.430e-1	1.069e-3	1.404e-2	4.944e-2	1.430e-1	1.069e-3	1.404e-2
	0.901	0.623	1.727	1.265	0.901	0.623	1.727	1.265
64	2.649e-2	9.288e-2	3.230e-4	5.842e-3	2.649e-2	9.288e-2	3.229e-4	5.842e-3
	0.944	0.690	1.782	1.412	0.944	0.690	1.782	1.412
128	1.377e-2	5.759e-2	9.388e-5	2.195e-3	1.377e-2	5.759e-2	9.391e-5	2.195e-3
	0.978	0.748	1.832	1.517	0.978	0.748	1.832	1.517
256	6.990e-3	3.429e-2	2.638e-5	7.669e-4	6.990e-3	3.429e-2	2.638e-5	7.669e-4
	0.991	0.790			0.991	0.790		
512	3.518e-3	1.984e-2			3.518e-3	1.984e-2		

Table 6.2

Maximum nodal errors in approximating the derivatives.

	$\|D_x^- u^N - u_x\|$				$\|D_y^- u^N - u_y\|$			
	$\varepsilon = 10^{-6}$		$\varepsilon = 10^{-8}$		$\varepsilon = 10^{-6}$		$\varepsilon = 10^{-8}$	
N	$\bar\Omega_0$	Ω_2	$\bar\Omega_0$	Ω_2	$\bar\Omega_0$	Ω_1	$\bar\Omega_0$	Ω_1
64	3.841e-2	8.811e-2	3.841e-2	8.811e-2	5.199e-2	1.819e-1	5.199e-2	1.819e-1
	0.938	0.711	0.938	0.711	1.001	0.811	1.001	0.811
128	2.005e-2	5.384e-2	2.005e-2	5.384e-2	2.598e-2	1.037e-1	2.598e-2	1.037e-1
	0.961	0.764	0.961	0.764	0.991	0.824	0.991	0.824
256	1.030e-2	3.171e-2	1.030e-2	3.171e-2	1.307e-2	5.856e-2	1.307e-2	5.856e-2
	0.974	0.805	0.974	0.805	0.996	0.838	0.996	0.838
512	5.241e-3	1.815e-2	5.241e-3	1.815e-2	6.554e-3	3.277e-2	6.554e-3	3.277e-2

Table 6.3

Maximum nodal errors in approximating ε-weighted derivatives.

	$\varepsilon\,\|D_x^- u^N - u_x\|$				$\varepsilon\|D_y^- u^N - u_y\|$			
	$\varepsilon = 10^{-6}$		$\varepsilon = 10^{-8}$		$\varepsilon = 10^{-6}$		$\varepsilon = 10^{-8}$	
N	Ω_1	Ω_{12}	Ω_1	Ω_{12}	Ω_2	Ω_{12}	Ω_2	Ω_{12}
64	2.524e-1	3.115e-1	2.524e-1	3.115e-1	3.739e-1	4.661e-1	3.739e-1	4.661e-1
	0.475	0.517	0.475	0.517	0.479	0.521	0.479	0.521
128	1.816e-1	2.176e-1	1.816e-1	2.176e-1	2.684e-1	3.248e-1	2.684e-1	3.248e-1
	0.616	0.641	0.616	0.641	0.618	0.644	0.618	0.644
256	1.185e-1	1.395e-1	1.185e-1	1.395e-1	1.749e-1	2.079e-1	1.749e-1	2.079e-1
	0.712	0.728	0.713	0.728	0.714	0.729	0.714	0.729
512	7.233e-2	8.427e-2	7.233e-2	8.427e-2	1.066e-1	1.254e-1	1.066e-1	1.254e-1

Tables 6.2 and 6.3 are clear illustrations of Corollary 4.2. In these tables we present the maximum nodal errors in approximating the derivatives and their rates of convergence computed as in Table 6.1.

In summary, our numerical results confirm our theoretical results.

Appendix A. Proof of Lemmas 5.2, 5.3, and 5.4 from subsection 5.1. If the conditions of Lemma 5.1(i) are satisfied, we say that B_{ij} is a *barrier function* for v_{ij}. Define the auxiliary discrete functions

$$B_i := \begin{cases} 2\left(1+\dfrac{\alpha h}{\varepsilon}\right)^{-N/2}\left(1+\dfrac{\alpha H}{\varepsilon}\right)^{-(N/2-i)}, & i = 0,\dots,\dfrac{N}{2}, \\ \left(1+\dfrac{\alpha h}{\varepsilon}\right)^{-(N-i)} + \left(1+\dfrac{\alpha h}{\varepsilon}\right)^{-N/2}, & i = \dfrac{N}{2},\dots,N, \end{cases} \tag{A.1}$$

$$\bar B_i := \begin{cases} 2\left(\dfrac{\varepsilon}{\beta}\right)\left(1+\dfrac{\beta H}{\varepsilon}\right)^{-(N/2-i)}, & i = 0,\dots,\dfrac{N}{2}, \\ 2\left(\dfrac{\varepsilon}{\beta}\right) + \sigma - (N-i)h, & i = \dfrac{N}{2},\dots,N. \end{cases} \tag{A.2}$$

It is assumed here that $\{x_i\}_{i=0}^N$ are the nodes of the mesh (1.4), (1.6). Furthermore, we shall use B_i and $\bar B_i$ normalized in different manners as discrete barrier functions.

Lemma A.1. *For any positive α the discrete function B_i from* (A.1) *is such that* $e^{-\alpha(1-x_i)/\varepsilon} < B_i \le C\mathcal{E}(N^{-2\alpha/\beta}, i \le N/2;\ 1)$ *and* $(-\varepsilon\delta_x^2 + \alpha D_x^-)B_i \ge 0$.

Proof. The lower bound for B_i follows from the inequality $e^{-t} \le (1+t)^{-1}$, which holds true for $t \ge 0$, with $t := \alpha h_i/\varepsilon$. The upper bound for B_i is obvious for $i > N/2$. For $i \le N/2$, it follows from $(1+t)^{-1} \le e^{-t+t^2}$, which we have for $t > 0$. Setting $t := \alpha h/\varepsilon$, we get $B_i \le 2(1+\alpha h/\varepsilon)^{-N/2} \le 2e^{-\alpha\sigma/\varepsilon + (\alpha h/\varepsilon)^2 N/2}$. Further, (1.7) implies $e^{-\alpha\sigma/\varepsilon} \le N^{-2\alpha/\beta}$ and $e^{(\alpha h/\varepsilon)^2 N/2} \le e$. This proves the upper bound for B_i.

The second inequality is checked using (1.4) and (2.1). In fact, $(-\varepsilon\delta_x^2+\alpha D_x^-)B_i = 0$ for $i \neq N/2$ and $(-\varepsilon\delta_x^2+\alpha D_x^-)B_i > 0$ for $i = N/2$. ∎

Proof of Lemma 5.2. Use B_i from Lemma A.1 with $\alpha := \beta$ as a barrier function for v_{ij}. Note that $L^N B_i \geq (b_{1,ij} - \beta)D_x^- B_i \geq 0$. ∎

Lemma A.2. *The discrete function B_i from* (A.1) *with $\alpha := \beta/2$ is such that $B_i \leq C\mathcal{E}(N^{-1}, i \leq N/2;\ 1)$ and $L^N B_i \geq Ce^{-\beta(1-x_i)/\varepsilon}\mathcal{E}(N, i \leq N/2;\ \varepsilon^{-1})$.*

Proof. This lemma follows from Lemma A.1. The first property is obvious. To prove the second, note that $L^N B_i \geq (b_{1,ij} - \beta/2)D_x^- B_i \geq (\beta/2)D_x^- B_i$. By (2.1), (1.5), (1.4), calculations show that $D_x^- B_i = (h_i + 2\varepsilon/\beta)^{-1}B_i$ and $(h_i + 2\varepsilon/\beta)^{-1} \geq C\mathcal{E}(N, i \leq N/2;\ \varepsilon^{-1})$. Recalling that $B_i > e^{-(\beta/2)(1-x_i/\varepsilon)} \geq e^{-\beta(1-x_i/\varepsilon)}$, we complete the proof. ∎

Proof of Lemma 5.3. This lemma follows from Lemma A.2.

(i) By (1.5), use $CN^{-1}B_i$ as a barrier function for v_{ij}.

(ii), (iii) Use CB_i as a barrier function for v_{ij}. ∎

Proof of Lemma 5.4. By (1.4), (1.6), for the discrete function $\bar{B}_i$ defined in (A.2) we have $0 < \bar{B}_i \leq C\varepsilon\ \mathcal{E}_{ij}(1, i \leq N/2;\ \ln N)$. Combining this with the analogous estimate for $\bar{B}_j$ and (1.5), we get

$$0 < \bar{B}_i + \bar{B}_j \leq C\varepsilon\ \mathcal{E}_{ij}(1, \bar{\Omega}_0;\ \ln N) \leq CN^{-1}\ \mathcal{E}_{ij}(1, \bar{\Omega}_0;\ \ln N).$$

By (2.1), (1.4), calculations show that $D_x^-\bar{B}_i = 1$ for $i > N/2$, while $D_x^-\bar{B}_i \geq 0$ for $i \leq N/2$. In particular, $D_x^-\bar{B}_{N/2} = 2(1 + \beta H/\varepsilon)^{-1}$. Further, $L^N\bar{B}_i \geq b_{1,ij} \geq \beta$ for $i > N/2$, while $L^N\bar{B}_i \geq (-\varepsilon\delta_x^2 + \beta D_x^-)\bar{B}_i = 0$ for $i < N/2$. For $i = N/2$ we also have $L^N\bar{B}_i \geq 0$, which follows from

$$L^N\bar{B}_i \geq \big(-\varepsilon\delta_x^2 + \beta D_x^-\big)\bar{B}_i = \big[\beta + 2\varepsilon(h+H)^{-1}\big]D_x^-\bar{B}_i - \big[2\varepsilon(h+H)^{-1}\big]D_x^-\bar{B}_{i+1},$$

where $i = N/2$. These imply that $L^N\bar{B}_i \geq \beta\ \mathcal{E}_{ij}\,(0, i \leq N/2;\ 1)$. Combining this estimate with its analogue for $L^N\bar{B}_j$, we obtain

$$L^N(\bar{B}_i + \bar{B}_j) \geq \beta\ \mathcal{E}_{ij}(0, \bar{\Omega}_0;\ 1).$$

Hence, $(B_i + B_j)/\beta$ is a barrier function for v_{ij}. ∎

Appendix B. Proof of Lemma 5.7(ii).

Proof. Note that (4.3) implies that $\Phi(x,y) = 0$ if $x = 1$ or $y = 1$. Further, $|\Phi(x,y)| \leq C\varepsilon \leq CN^{-1}$ on $\partial\Omega$. Hence,

$$|w_{ij} - \Phi_{ij}| \leq CN^{-1} \quad \text{on } \partial\Omega^N. \tag{B.1}$$

To study $L^N(w-\Phi)$, note that, by (4.7), (5.6), we have $L^N w_{ij} = (b_1\varphi_x + b_2\varphi_y + c\varphi)_{ij}$. Hence

$$L^N(w - \Phi) = (b_1\varphi_x + b_2\varphi_y + c\varphi - L^N\varphi) + L^N(\varphi - \Phi). \tag{B.2}$$

Using Taylor series expansions, (1.5), and (4.2), which was proved in Lemma 5.7(i), we obtain for the first term on the right-hand side that

$$|(b_1\varphi_x + b_2\varphi_y + c\varphi)_{ij} - L^N\varphi_{ij}| \leq C(N^{-1} + \varepsilon)\|\varphi\|_{1,1} \leq CN^{-1}. \tag{B.3}$$

To estimate $L^N(\varphi - \Phi)$, we define

$$\Phi_1(x,y) := \varphi(1,y)e^{-b_1(1,y)(1-x)/\varepsilon}, \qquad \Phi_2(x,y) := \varphi(x,1)e^{-b_2(x,1)(1-y)/\varepsilon},$$
$$\Phi_{12}(x,y) := \varphi(1,1)e^{-b_1(1,1)(1-x)/\varepsilon - b_2(1,1)(1-y)/\varepsilon},$$

so that $\varphi - \Phi = \Phi_1 + \Phi_2 - \Phi_{12}$. Thus, recalling the notation (5.3), we have

$$L^N(\varphi - \Phi) = L_1^N(\Phi_1 - \Phi_{12}) + L_2^N(\Phi_2 - \Phi_{12}) + L_2^N\Phi_1 + L_1^N\Phi_2 + c(\Phi_1 + \Phi_2 - \Phi_{12}). \tag{B.4}$$

For $i \le N/2$, using (1.3), (2.1), (1.4), and (1.7), we obtain

$$|L_1^N(\Phi_1 - \Phi_{12})_{ij}| \le CNe^{-\beta(1-x_{i+1})/\varepsilon} \le CNe^{-\beta(\sigma-h)/\varepsilon} \le CN^{-1}. \tag{B.5}$$

Consider $i > N/2$. First note that

$$-\varepsilon\Phi_{1,xx} + b_1(1,y)\Phi_{1,x} = 0, \qquad -\varepsilon\Phi_{12,xx} + b_1(1,1)\Phi_{12,x} = 0,$$

while the left-hand sides here are slightly different from $L_1\Phi_1$ and $L_1\Phi_{12}$. Hence,

$$L_1^N(\Phi_1 - \Phi_{12})_{ij} = (L_1^N - L_1)(\Phi_1 - \Phi_{12})_{ij} + [b_1(x_i, y_j) - b_1(1, y_j)]\Phi_{1,x}(x_i, y_j) - [b_1(x_i, y_j) - b_1(1,1)]\Phi_{12,x}(x_i, y_j).$$

Using Taylor series expansions to estimate the first term on the right-hand side, and the inequalities $|b_1(x,y) - b_1(1,y)| \le C(1-x)$ and $|b_1(x,y) - b_1(1,1)| \le C[(1-x) + (1-y)]$ combined with (4.10) to estimate the other terms, we obtain

$$|L_1^N(\Phi_1 - \Phi_{12})| \le C(h\varepsilon^{-2}e^{-\beta(1-x_{i+1})/\varepsilon} + e^{-\beta(1-x_i)/\varepsilon}) \le C(h\varepsilon^{-2}e^{\beta h/\varepsilon} + 1)e^{-\beta(1-x_i)/\varepsilon}.$$

Combining this with (B.5) and noting that, by (1.7), (1.5), $h\,\varepsilon^{-2} \ge C$ and $e^{\beta h/\varepsilon} \le C$, we get

$$|L_1^N(\Phi_1 - \Phi_{12})_{ij}| \le C\left[\left(\frac{h}{\varepsilon}\right)\mathcal{E}_{ij}\left(\varepsilon^{-1}e^{-\beta(1-x_i)/\varepsilon}, i > \frac{N}{2};\ 0\right) + N^{-1}\right]. \tag{B.6}$$

Furthermore, one can easily see that

$$|L_2^N\Phi_{1,ij}| \le C\|\varphi\|_{1,1}e^{-\beta(1-x_i)/\varepsilon} \le Ce^{-\beta(1-x_i)/\varepsilon}. \tag{B.7}$$

Combining (B.4) with (B.6), (B.7), and their analogues for $L_2^N(\Phi_2 - \Phi_{12})$ and $L_1^N\Phi_2$, and then with (B.2), (B.3), we finally get the estimate

$$\begin{aligned}|L^N(w - \Phi)_{ij}| \le C\Bigg[&\left(\frac{h}{\varepsilon}\right)\mathcal{E}_{ij}\left(\varepsilon^{-1}e^{-\beta(1-x_i)/\varepsilon}, i > \frac{N}{2};\ 0\right)\\ &+ \left(\frac{h}{\varepsilon}\right)\mathcal{E}_{ij}\left(\varepsilon^{-1}e^{-\beta(1-y_j)/\varepsilon}, j > \frac{N}{2};\ 0\right)\\ &+ e^{-\beta(1-x_i)/\varepsilon} + e^{-\beta(1-y_j)/\varepsilon} + N^{-1}\Bigg].\end{aligned}$$

Combining this with (B.1) and applying Lemmas 5.1(ii),(iii) and 5.3(i),(ii), we obtain

$$\begin{aligned}|w_{ij} - \Phi_{ij}| &\le C\left[\left(\frac{h}{\varepsilon}\right)\mathcal{E}_{ij}\left(N^{-1}, i \le \frac{N}{2};\ 1\right) + \left(\frac{h}{\varepsilon}\right)\mathcal{E}_{ij}\left(N^{-1}, j \le \frac{N}{2};\ 1\right) + N^{-1}\right]\\ &\le C\left[\left(\frac{h}{\varepsilon}\right)\mathcal{E}_{ij}(N^{-1}, \bar{\Omega}_0;\ 1) + N^{-1}\right].\end{aligned}$$

By (1.7), this yields the statement of Lemma 5.7(ii). □

Appendix C. Proof of Lemmas 5.10 and 5.11.

Proof of Lemma 5.10. (a) Obviously,

$$|V| \le |V+v_0| + |v_0|, \tag{C.1}$$

where v_0 is defined in (3.3). Since $r[v_0] = -L^N v_0 + Lv_0$, we have $L^N[V+v_0] = Lv_0$. One can easily check that $-\varepsilon v_{0,xx} + b_1(1,y)v_{0,x} = 0$ holds true and implies that $L_1 v_0 = [b_1(x,y) - b_1(1,y)]v_{0,x}$. Combining this with $|b_1(x,y) - b_1(1,y)| \le C(1-x)$ and (4.10), we get $|L_1 v_0| \le Ce^{-\beta(1-x)/\varepsilon}$, while $|(L_2+c)v_0| \le Ce^{-\beta(1-x)/\varepsilon}$. Hence, $|Lv_0| \le Ce^{-\beta(1-x)/\varepsilon}$, which yields

$$\left|L^N[V+v_0]_{ij}\right| \le Ce^{-\beta(1-x_i)/\varepsilon} \quad \text{in } \Omega^N.$$

Combining this with the boundary condition

$$|(V+v_0)_{ij}| \le \left(\frac{h}{\varepsilon}\right)|\Psi_{1,ij}| + CN^{-2} + |v_{0,ij}| \le C(e^{-\beta(1-x_i)/\varepsilon} + N^{-2}) \quad \text{on } \partial\Omega^N,$$

and applying Lemmas 5.1(ii), 5.2, 5.3(i), we get $|(V+v_0)_{ij}| \le CN^{-2}$ for $i \le N/2$. Combining this with (C.1), (3.3), and (1.7), we complete part (a) of the proof.

(b) This part of the proof is analogous to part (a).

(c) Since this part of the proof is similar to part (a), we skip certain details. Again, we have $|Z| \le |Z+z_0| + |z_0|$, where z_0 is defined in (3.3), which implies $L^N[Z+z_0] = Lz_0$. Further, $-\varepsilon z_{0,xx} + b_1(1,1)z_{0,x} = 0$ and $-\varepsilon z_{0,yy} + b_2(1,1)z_0 = 0$ imply $Lz_0 = [b_1(x,y) - b_1(1,1)]z_{0,x} + [b_2(x,y) - b_2(1,1)]z_{0,y} + cz_0$. By (3.3), this yields $|Lz_0| \le Ce^{-\beta[(1-x)+(1-y)]/\varepsilon}$. Hence, $\left|L^N[Z+z_0]_{ij}\right| \le Ce^{-\beta[(1-x_i)+(1-y_j)]/\varepsilon}$ in Ω^N, while $|(Z+z_0)_{ij}| \le C(e^{-\beta[(1-x_i)+(1-y_j)]/\varepsilon} + N^{-2})$ on $\partial\Omega^N$. Applying Lemmas 5.1(ii), 5.2, 5.3(i), we get $|(Z+z_0)_{ij}| \le CN^{-2}$ for $i \le N/2$, and $|(Z+z_0)_{ij}| \le CN^{-2}$ for $j \le N/2$. Combining these two estimates, we proceed similarly to part (a). ∎

Proof of Lemma 5.11. (a) By (5.13a), we have $\psi_1 - \Psi_1 = 0$ on the boundary of the submesh $\{(x_i, y_j) : i = N/2, \dots, N,\ j = 0, \dots, N\}$ where ψ_1 is defined.

In this part of the proof we consider only $i > N/2$. Recalling the notation (5.3), we introduce the following decomposition:

$$L^N(\psi_1 - \Psi_1) = (L^N\psi_1 - L_1\Psi_1) - (L_1^N\Psi_1 - L_1\Psi_1) - (L_2^N + c)\Psi_1.$$

Using Taylor series expansions and (5.8), (3.3), (4.10), we have

$$|L_1^N\Psi_1 - L_1\Psi_1| \le Ch\,\varepsilon^{-2}e^{-\beta(1-x_{i+1})/\varepsilon}, \qquad |(L_2^N + c)\Psi_1| \le Ce^{-\beta(1-x_i)/\varepsilon}.$$

In addition, we claim that

$$|L^N\psi_{1,ij} - (L_1\Psi_1)_{ij}| \le Ce^{-\beta(1-x_i)/\varepsilon}. \tag{C.2}$$

Since (1.5), (1.7) imply that $h\,\varepsilon^{-2} \ge C$ and $e^{-\beta(1-x_{i+1})/\varepsilon} \le Ce^{-\beta(1-x_i)/\varepsilon}$, we have

$$|L^N(\psi_1 - \Psi_1)_{ij}| \le C(h/\varepsilon)\,\varepsilon^{-1}e^{-\beta(1-x_i)/\varepsilon}.$$

Further, by Lemmas 5.1(iii) and 5.3(iii), we get $|\psi_{1,ij} - \Psi_1(x_i, y_j)| \le C(h/\varepsilon + N^{-2})$, which yields statement (a) of the lemma.

To prove our claim (C.2), it suffices to check that

$$\left|\frac{b_1(x,y)\varepsilon v_{0,xx}}{2} - L_1\Psi_1\right| \le Ce^{-\beta(1-x)/\varepsilon}. \tag{C.3}$$

By Remark 4.4, we have $-\varepsilon\Psi_{1,xx} + b_1(1,y)\Psi_{1,x} = b_1(1,y)\varepsilon v_{0,xx}/2$, which implies

$$L_1\Psi_1 = \frac{b_1(x,y)\varepsilon v_{0,xx}}{2} + [b_1(x,y) - b_1(1,y)]\left(\frac{\varepsilon v_{0,xx}}{2} - v_{0,x}\right).$$

Furthermore, using (3.3), (4.10), and $|b_1(x,y) - b_1(1,y)| \le C(1-x)$, we obtain (C.3) and thus complete part (a) of the proof.

(c) This part of the proof is slightly different from part (a); namely, we have to estimate $L_2^N\tilde{\Psi}_1$ more carefully. Note that we consider only $i,j > N/2$ in part (c). Using the notation (5.4), we have $L_2^N\tilde{\Psi}_1 = -r_2[\tilde{\Psi}_1] - L_2\tilde{\Psi}_1$. Further, (5.8), (3.3), (4.10) imply that $|L_2\tilde{\Psi}_1| \le Ce^{-\beta(1-x)/\varepsilon}$ and $|r_{2,ij}[\tilde{\Psi}_1]| \le Ch\,\varepsilon^{-2}e^{-\beta(1-y_{j+1})/\varepsilon}$. Combining these two estimates, we proceed as in part (a).

(b), (d) These parts of the proof are analogous to parts (a) and (c), respectively. ∎

REFERENCES

[1] A. Fröhner, T. Linß, and H.-G. Roos, *Defect correction on Shishkin-type meshes*, Numer. Algorithms, 26 (2001), pp. 281–299.

[2] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, Berlin, 1998.

[3] N. Kopteva and M. Stynes, *Approximation of derivatives in a convection-diffusion two-point boundary value problem*, Appl. Numer. Math., 39 (2001), pp. 47–60.

[4] T. Linß, *Error expansion for a first-order upwind difference scheme*, IMA J. Numer. Anal., to appear.

[5] T. Linß and M. Stynes, *A hybrid difference scheme on a Shishkin mesh for linear convection-diffusion problems*, Appl. Numer. Math., 31 (1999), pp. 255–270.

[6] T. Linß and M. Stynes, *Numerical methods on Shishkin meshes for linear convection-diffusion problems*, Comput. Methods Appl. Mech. Engrg., 190 (2001), pp. 3527–3542.

[7] T. Linß and M. Stynes, *Asymptotic analysis and Shishkin-type decomposition for an elliptic convection-diffusion problem*, J. Math. Anal. Appl., 261 (2001), pp. 604–632.

[8] J. J. H. Miller, E. O'Riordan, and G. I. Shishkin, *Solution of Singularly Perturbed Problems with ε-uniform Numerical Methods—Introduction to the Theory of Linear Problems in One and Two Dimensions*, World Scientific, Singapore, 1996.

[9] M. C. Natividad and M. Stynes, *Richardson extrapolation for a convection-diffusion problem using a Shishkin mesh*, Appl. Numer. Math., 45 (2003), pp. 315–329.

[10] H.-G. Roos, M. Stynes, and L. Tobiska, *Numerical Methods for Singularly Perturbed Differential Equations*, Springer-Verlag, Berlin, 1996.

[11] A. A. Samarski, *Theory of Difference Schemes*, Nauka, Moscow, 1989 (in Russian).

[12] G. I. Shishkin, *Increasing the accuracy of solutions of difference schemes for parabolic equations with a small parameter multiplying the highest derivative*, Zh. Vychisl. Mat. Mat. Fiz., 24 (1984), pp. 864–875 (in Russian).

[13] G. I. Shishkin, *Grid Approximations of Singularly Perturbed Elliptic and Parabolic Equations*, Russian Academy of Sciences, Ural Branch, Ekaterinburg, Russia, 1992 (in Russian).

[14] G. I. Shishkin, *Grid approximations for singularly perturbed elliptic equations*, Zh. Vychisl. Mat. Mat. Fiz., 38 (1998), pp. 1989–2001 (in Russian); English translation in Comput. Math. Math. Phys., 38 (1998), pp. 1909–1921.

SIAM J. NUMER. ANAL.
Vol. 41, No. 5, pp. 1870–1892

TRUNCATED QUADRATURE RULES OVER $(0, \infty)$ AND NYSTRÖM-TYPE METHODS*

G. MASTROIANNI† AND G. MONEGATO‡

Abstract. We propose replacing the classical Gauss–Laguerre quadrature formula by a truncated version of it, obtained by ignoring the last part of its nodes. This has the effect of obtaining optimal orders of convergence. Corresponding quadrature rules with kernels are then considered and optimal error estimates are derived also for them. These rules are finally used to define stable Nyström-type interpolants for a second kind of integral equation on the real semiaxis whose solutions decay exponentially at ∞.

Key words. Gauss–Laguerre rules, integral equations, Nyström interpolants

AMS subject classifications. 65D30, 65R20

DOI. 10.1137/S0036142901391475

1. Introduction.

Consider the classical Gauss–Laguerre quadrature formula

$$\int_0^\infty w_\alpha(x) f(x) dx = \sum_{i=1}^{m} \lambda_{mi}^\alpha f(x_{mi}^\alpha) + R_m(f), \tag{1}$$

where we have set $w_\alpha(x) = x^\alpha e^{-x}$, $\alpha > -1$, and $x_{m1}^\alpha < x_{m2}^\alpha < \cdots < x_{mm}^\alpha$. Then, associate with it the "truncated" rule

$$\int_0^\infty w_\alpha(x) f(x) dx = \sum_{0 < x_{mi}^\alpha \le 4\theta m} \lambda_{mi}^\alpha f(x_{mi}^\alpha) + R_m^\theta(f), \tag{2}$$

where $0 < \theta < 1$ is arbitrarily chosen. For notational convenience, in the following we will set $\lambda_i := \lambda_{mi}^\alpha$ and $x_i := x_{mi}^\alpha$.

Let us denote by $AC = AC(\mathbb{R}^+)$ the set of all real functions which are absolutely continuous on any bounded subinterval of $(0, \infty)$. In [13] we have shown that for functions $f \in W_r^1(w_\alpha)$, with

$$W_r^1(w_\alpha) = W_r^1 = \{f^{(r-1)} \in AC : \|f^{(r)} \varphi^r w_\alpha\|_{L^1} < \infty\},$$

$r \ge 1$ an integer, and $\varphi(x) = \sqrt{x}$, the following estimate holds:

$$|R_m^\theta(f)| \le c \left(\frac{\|f^{(r)} \varphi^r w_\alpha\|_{L^1}}{m^{r/2}} + e^{-am} \|f w_\alpha\|_{L^1} \right), \tag{3}$$

where the constants c and a are independent of f and m.

Thus we can expect to obtain the same accuracy given by (1) using only a fraction of its nodes. For instance, in the examples considered in [13], using (2) with $\theta = \frac{1}{4}$ and

*Received by the editors June 27, 2001; accepted for publication (in revised form) May 4, 2003; published electronically November 19, 2003. This research was supported by the Ministero dell'Università e della Ricerca Scientifica e Tecnologica of Italy, and by the Italian Research Council under contract n.98.00648.CT11.

http://www.siam.org/journals/sinum/41-5/39147.html

†Dipartimento di Matematica, Università della Basilicata, 85100 Potenza, Italy (mastroianni@unibas.it).

‡Dipartimento di Matematica, Politecnico di Torino, 10129 Torino, Italy (monegato@polito.it).

$m = 4, 8, 16, 32, 64, 128$, i.e., taking only the first 2,5,10,19,39,78 nodes, respectively, we had the same accuracy which was given by the complete rule (1).

In section 2 we obtain, using a simpler proof, a new error estimate more convenient than (3) (see Theorem 2.7), which includes (3) as a particular case. Both these estimates, apart from an $O(e^{-an})$ term, a being a positive constant, are of the same order of the best L^1-polynomial approximation of functions in $W^1_r(w_\alpha)$. Further, we show that an error estimate of this quasi-optimal order cannot hold for the (nontruncated) Gauss–Laguerre rule (1).

In the same section we introduce a special sequence of (truncated) Lagrange interpolation operators, based on a fraction of the Gauss–Laguerre nodes. As a major result, we prove the uniform boundedness of this sequence in proper subspaces of $L^2_{w_\alpha}$ and derive an optimal interpolation error estimate. As a consequence of this estimate we also obtain a new bound for $R^\theta_m(f)$, which does not require the absolute continuity of the function f. Thus our truncation procedure combines an optimal error bound with an $O(m)$ function evaluation saving.

Our new error bounds are needed in section 3 to derive a similar error estimate for corresponding product-type rules for integrals of the form

$$\int_0^\infty w_\alpha(x)k(x,y)f(x)dx,$$

with kernels $k(x, y)$ which may have weak singularities.

Then, in section 4, we use the above truncated product rules to define Nyström-type interpolants for a class of integral equations whose solutions decay exponentially to a constant at ∞ and prove stability and convergence estimates for them. These latter results hold under weaker conditions than those required in [12]. Nevertheless, these new conditions are not satisfied yet, for example, by the well-known linear transport equation considered in [12]. For the latter, to prove stability we are forced to modify the interpolant in a neighborhood of infinity, where anyway we know that the solution is practically constant.

The numerical results we present in section 4 confirm the benefits given by the truncated rule. Furthermore, truncation greatly reduces the magnitude of the condition number associated with the linear system generated by the Nyström method.

2. Truncated Lagrange interpolation and Gauss–Laguerre rules.

To prove the results of this section we introduce some notation and recall some well-known results.

We denote by L^2 the set of all real measurable functions in $\mathbb{R}^+ = (0,\infty)$ which are square integrable and by $\|f\|_{L^2} = (\int_0^\infty f^2(x)dx)^{1/2}$ its usual norm. Having set $w_\alpha(x) = x^\alpha e^{-x}$, $x > 0$, $\alpha > -1$, we shall write $f \in L^2_{w_\alpha}$ if $f\sqrt{w_\alpha} \in L^2$ and $\|f\|_{L^2_{w_\alpha}} = \|f\sqrt{w_\alpha}\|_{L^2}$. For each $f \in L^2_{w_\alpha}$, the polynomial $S_m(f;x) = \sum_{k=0}^{m-1} c_k(f)p_k(x)$, of degree $m-1$, is the mth Fourier sum of f associated with the system of orthonormal Laguerre polynomials $\{p_m\}$ and $c_k(f) = \int_0^\infty w(x)p_k(x)f(x)dx$.

For notational convenience, here and in the following we omit in the representation of p_k, c_k, S_m the dependence upon w_α.

Finally, we denote by $E_m(f)_{L^2_{w_\alpha}} = \|f - S_m(f)\|_{L^2_{w_\alpha}}$ the error of the best polynomial approximation in $L^2_{w_\alpha}$.

Next we introduce the scale of subspaces [12]

$$L^2_{w_\alpha,s} = \{f \in L^2_{w_\alpha} : \|f\|_{L^2_{w_\alpha,s}} < \infty\}, \quad s \geq 0 \quad \text{real},$$

where $\|f\|_{L^2_{w_\alpha,s}} = \left(\sum_{i=0}^{\infty}(i+1)^s c_i^2(f)\right)^{1/2}$. Moreover, we recall the following estimate [3]:

$$E_m(f)_{L^2_{w_\alpha}} \le c\omega_\varphi^r\left(f, \frac{1}{\sqrt{m}}\right)_{L^2_{w_\alpha}}, \ r < m, \tag{4}$$

where the positive constant c is independent of f and m,

$$\begin{aligned}\omega_\varphi^r(f,t)_{L^2_{w_\alpha}} = \Omega_\varphi^r(f,t)_{L^2_{w_\alpha}} &+ \inf_{p\in\mathbb{P}_{r-1}} \|f-p\|_{L^2_{w_\alpha}(0,4r^2t^2)} \\ &+ \inf_{p\in\mathbb{P}_{r-1}} \|f-p\|_{L^2_{w_\alpha}\left(\frac{1}{t^2},\infty\right)}\end{aligned}$$

with $4r^2t^4 < 1$,

$$\Omega_\varphi^r(f,t)_{L^2_{w_\alpha}} = \sup_{0<h\le t} \|\Delta_{h\varphi}^r f\|_{L^2_{w_\alpha}(I_h)}, \ \varphi(x) = \sqrt{x}, \ I_h = [4r^2h^2, 1/h^2],$$

and

$$\Delta_{h\varphi}^r f(x) = \sum_{k=0}^{r}(-1)^k \begin{pmatrix} r \\ k \end{pmatrix} f\left(x + \frac{h\sqrt{x}}{2}(r-2k)\right).$$

By using the previous modulus of continuity we have the equivalence

$$\|f\|_{L^2_{w_\alpha,s}} \sim \|f\|_{L^2_{w_\alpha}} + \left(\int_0^1 \left[\frac{\Omega_\varphi^r(f,t)_{L^2_{w_\alpha}}}{t^{s+1/2}}\right]^2 dt\right)^{1/2}, \tag{5}$$

which is true for $r \ge s \in \mathbb{R}^+$ (see [3]). We recall that $A \sim B$ means that $c^{-1} \le A/B \le c$ for some constant c independent of the parameters A and B.

Finally, we recall that (see [12]) $L^2_{w_\alpha,0} = L^2_{w_\alpha}$ and, for a positive integer s, we have

$$L^2_{w_\alpha,s} = W_s^2 := \{f \in L^2_{w_\alpha} : f^{(s-1)} \in AC \text{ and } \|f^{(s)}\varphi^s\|_{L^2_{w_\alpha}} < \infty\}$$

and

$$\|f\|_{L^2_{w_\alpha,s}} \sim \|f\|_{W_s^2} = \|f\|_{L^2_{w_\alpha}} + \|f^{(s)}\varphi^s\|_{L^2_{w_\alpha}},$$

where the equality between the spaces of functions is in the sense of the norm equivalence. In [12] a criteria for detecting an estimate of the real s, such that $f \in L^2_{w_\alpha,s}$, has been given.

If $f \in C^0(A)$, $A = [a, a+\delta]$, $\delta > 0$, and $\omega^k(f,t)_{L^2(A)}$ denotes the ordinary modulus of continuity of order k in $L^2(A)$ defined in [21], then we have (see [8])

$$\sqrt{\delta}\max_{x\in A}|f(x)| \le c\left[\|f\|_{L^2(A)} + \sqrt{\delta}\int_0^\delta \frac{\omega^k(f,t)_{L^2(A)}}{t^{3/2}}dt\right]. \tag{6}$$

Here and in the following c, c_1, c_2 denote constants which may take different values on different occurrences. Incidentally we notice that when $f \in C^0(A)$, the exponent 3/2 in (6) cannot be replaced by a smaller number [8].

Another estimate that we shall need is the following one (see [3]):

$$\left(\int_{2m\frac{(1+\delta)}{\lambda}}^{\infty} |P_m(x)x^{\beta}e^{-\lambda x}|^p dx\right)^{1/p} \le c_1 e^{-c_2 m}\left(\int_0^{\infty} |P_m(x)x^{\beta}e^{-\lambda x}|^p dx\right)^{1/p}, \tag{7}$$

which holds for any polynomial $P_m(x)$ of degree m and real $p, \delta, \lambda > 0, \beta > -1/p$, with constants c_1 and c_2 independent of m and P_m and depending only on δ.

Taking for example $\delta = \frac{1}{4}, \lambda = \frac{1}{2}, p = 2$, and $\beta = \frac{\alpha}{2}$ in (7), we also have

$$\begin{aligned} \|f\sqrt{w_\alpha}\|_{L^2(5m,\infty)} &\le \|(f - S_m(f))\sqrt{w_\alpha}\|_{L^2(5m,\infty)} + \|S_m(f)\sqrt{w_\alpha}\|_{L^2(5m,\infty)} \\ &\le E_m(f)_{L^2_{w_\alpha}} + c_1 e^{-c_2 m}\|S_m(f)\|_{L^2_{w_\alpha}} \\ &\le E_m(f)_{L^2_{w_\alpha}} + c_1 e^{-c_2 m}\|f\|_{L^2_{w_\alpha}}. \end{aligned} \tag{8}$$

Using (8), we further obtain

$$\|f\|_{L^2_{w_\alpha}} \le \frac{1}{1 - c_1 e^{-c_2 m}}[\|f\sqrt{w_\alpha}\|_{L^2(0,5m)} + E_m(f)_{L^2_{w_\alpha}}]. \tag{9}$$

From (9) we conclude that, in order to approximate a function $f \in L^2_{w_\alpha}$ by using polynomials, it is sufficient to approximate a "finite-section" of f, say, f_j, that is equal to f in a "large" interval containing the origin, is equal to zero otherwise, and has the same smoothness of f.

By using the zeros $x_1 < \cdots < x_m$ of the Laguerre polynomial $p_m(x)$, the construction of an f_j is described next. We recall that all these zeros are contained in the interval $(0, 4m)$ and that $x_m \sim 4m - m^{\frac{1}{3}}$ (see [20]).

Choosing a real $0 < \theta < 1$ and m sufficiently large, we define the integer $j = j(m)$ as

$$x_j := \min_k\{x_k : x_k \ge 4\theta m\}.$$

We take a nondecreasing function $\psi \in C^\infty(\mathbb{R})$, with $\psi(x) = 0$ when $x \le 0$ and $\psi(x) = 1$ for $x \ge 1$, and define

$$\psi_{j(m)}(x) = \psi_j(x) = \psi\left(\frac{x - x_j}{x_{j+1} - x_j}\right).$$

Finally, we consider the function

$$f_{j(m)} = f_j = f - \psi_j f, \tag{10}$$

which has the same degree of smoothness of f. Notice that $f_j(x) = f(x)$ for $x \le x_j$ and $f_j(x) = 0$ for $x \ge x_{j+1}$.

Then, by using (7) with $\lambda = \frac{1}{2}$, $p = 2$, and $\beta = \frac{\alpha}{2}$, it is simple to prove that, if we set $M = \lfloor \frac{\theta}{1+\delta} m \rfloor \sim m$,

$$\left.\begin{array}{c} \|f - f_j\|_{L^2_{w_\alpha}} \\ E_m(f_j)_{L^2_{w_\alpha}} \\ \|f_j - S_M(f)\|_{L^2_{w_\alpha}} \end{array}\right\} \le c_1[E_M(f)_{L^2_{w_\alpha}} + e^{-c_2 m}\|f\|_{L^2_{w_\alpha}}], \tag{11}$$

where the constants c_1, c_2 are independent of f and m.

Remark 2.1. Bounds very similar to (11) hold also when the space $L^2_{w_\alpha}$ is replaced by $L^p_w, 1 \leq p \leq \infty$, with weight $w(x) = \sqrt{w_\alpha(x)}x^a(1+x)^b \in L^p$, norm $\|fw\|_{L^p}$, and $\lim_{x\to 0,\infty}(fw)(x) = 0$ when $p = \infty$.

By $L_m(f)$ we denote the Lagrange polynomial based on the zeros of $p_m(x)$ and associated with the function f. Thus we consider

$$L_m(f_j;x) = \sum_{k=1}^{j} l_k(x)f(x_k),$$

where $\{l_k(x)\}$ denotes the fundamental Lagrange polynomials. The following result then holds.

LEMMA 2.2. *Whenever $f \in C^0(\mathbb{R}^+)$ satisfies the condition*

$$\int_0^1 \frac{\omega^r_\varphi(f,t)_{L^2_{w_\alpha}}}{t^{3/2}}dt < \infty,$$

for some $r \geq 1$, we have

$$\|L_m(f_j)\|_{L^2_{w_\alpha}} \leq c\left[\|f\|_{L^2_{w_\alpha}} + \frac{1}{\sqrt[4]{m}}\int_0^{1/\sqrt{m}} \frac{\omega^r_\varphi(f,t)_{L^2_{w_\alpha}}}{t^{3/2}}dt\right], \tag{12}$$

where the constant c is independent of m and f.

Proof. We start from the equality

$$\|L_m(w_\alpha,f_j)\|^2_{L^2_{w_\alpha}} = \sum_{k=1}^{j}\lambda_k f^2(x_k).$$

Using (6) with $\delta = \Delta x_k = x_k - x_{k-1},\ a = x_{k-1},\ x_0 = 0$, and $A = [x_{k-1}, x_k] \equiv I_k, k = 1, 2, \ldots, j$, we have

$$\sqrt{w_\alpha(x_k)}|f(x_k)|\Delta x_k \leq c\left[\|f\|_{L^2(I_k)} + \sqrt{\Delta x_k}\int_0^{\Delta x_k}\frac{\omega^r(f,t)_{L^2(I_k)}}{t^{3/2}}dt\right]\sqrt{w_\alpha(x_k)}.$$

Further,

$$\sqrt{w_\alpha(x_k)}\|f\|_{L^2(I_k)} \leq c\|f\sqrt{w_\alpha}\|_{L^2(I_k)}$$

and

$$\sqrt{w_\alpha(x_k)}\omega^r(f,t)_{L^2(I_k)} = \sqrt{w_\alpha(x_k)}\sup_{h\leq t}\left(\int_{x_{k-1}}^{x_k}[\Delta^r_h f(x)]^2dx\right)^{\frac{1}{2}}$$

$$\leq c\sup_{h\leq t}\left(\int_{x_{k-1}}^{x_k}w_\alpha(x)[\Delta^r_h f(x)]^2dx\right)^{\frac{1}{2}} =: c\tilde{\omega}^r(f,t)_{L^2_{w_\alpha}(I_k)}.$$

Thus we have

$$\sqrt{w_\alpha(x_k)}|f(x_k)|\sqrt{\Delta x_k} \leq c\left[\|f\sqrt{w_\alpha}\|_{L^2(I_k)} + \sqrt{\Delta x_k}\int_0^{\Delta x_k}\frac{\tilde{\omega}^r(f,t)_{L^2_{w_\alpha}(I_k)}}{t^{3/2}}dt\right]. \tag{13}$$

Recalling (see [14]) that

$$\lambda_k \sim w_\alpha(x_k)\Delta x_k,$$

taking the square of (13) and summing the corresponding terms for $k = 1, 2, \ldots, j$, we obtain

$$\sum_{k=1}^{j} \lambda_k f^2(x_k) \le c\left[\|f\|^2_{L^2_{w_\alpha}} + \sum_{k=1}^{j}\left(\sqrt{\Delta x_k}\int_0^{\Delta x_k} \frac{\tilde\omega^r(f,t)_{L^2_{w_\alpha}(I_k)}}{t^{3/2}}dt\right)^2\right]. \tag{14}$$

To estimate the last sum, we remark first that (see [14]) $\Delta x_k \sim \sqrt{\frac{x_k}{m}}, k = 1, \ldots, j$, uniformly with respect to m and k. Then we replace t by $\sqrt{x_k}t$ in the integral. The kth element of this sum is then dominated by

$$\frac{1}{\sqrt{m}}\left[\int_0^{1/\sqrt{m}} \frac{\tilde\omega^r(f,\sqrt{x_k}t)_{L^2_{w_\alpha}(I_k)}}{t^{3/2}}dt\right]^2.$$

For $t \le m^{-1/2}$ let $\Gamma(x) = \Gamma_t(x)$ such that $\Gamma^{(r-1)} \in AC$ and

$$\begin{gathered}\|(\Gamma - f)\sqrt{w_\alpha}\|_{L^2} + t^r\|\Gamma^{(r)}\varphi^r\sqrt{w_\alpha}\|_{L^2}\\ \le 2\inf_g\{\|(f-g)\sqrt{w_\alpha}\|_{L^2} + t^r\|g^{(r)}\varphi^r\sqrt{w_\alpha}\|_{L^2}\}\\ =: 2K_\varphi(f,t^r)_{L^2_{w_\alpha}} \le c\omega^r_\varphi(f,t)_{L^2_{w_\alpha}}.\end{gathered}$$

The last bound is obtained using Theorem 2.1 in [3].

Then, setting $I_{kr} = [x_{k-1}, x_k + r\Delta x_k]$,

$$\begin{aligned}\tilde\omega^r(f,\sqrt{x_k}t)_{L^2_{w_\alpha}(I_k)} &\le \tilde\omega^r(f-\Gamma,\sqrt{x_k}t)_{L^2_{w_\alpha}(I_k)} + \tilde\omega^r(\Gamma,\sqrt{x_k}t)_{L^2_{w_\alpha}(I_k)}\\ &\le c(\|(f-\Gamma)\sqrt{w_\alpha}\|_{L^2(I_{kr})} + (\sqrt{x_k}t)^r\|\Gamma^{(r)}\sqrt{w_\alpha}\|_{L^2(I_{kr})})\\ &\le c(\|(f-\Gamma)\sqrt{w_\alpha}\|_{L^2(I_{kr})} + t^r\|\Gamma^{(r)}\varphi^r\sqrt{w_\alpha}\|_{L^2(I_{kr})}) =: \tilde K_\varphi(f,t)_{L^2_{w_\alpha}(I_{kr})}.\end{aligned}$$

Using the Minkovski inequality (see (6.13.8) in [7]), from (14) we have

$$\begin{aligned}&\|L_m(f_j)\sqrt{w_\alpha}\|_{L^2}\\ &\le c\left(\|f\sqrt{w_\alpha}\|_{L^2} + \tfrac{1}{\sqrt[4]{m}}\int_0^{1/\sqrt{m}}\left[\sum_{k=1}^{j}\tilde K_\varphi(f,t)^2_{L^2_{w_\alpha}(I_k)}\right]^{1/2} t^{-3/2}dt\right)\\ &\le c\left(\|f\|_{L^2_{w_\alpha}} + \tfrac{1}{\sqrt[4]{m}}\left[\int_0^{1/\sqrt{m}}\frac{K_\varphi(f,t^r)_{L^2_{w_\alpha}}}{t^{3/2}}dt\right]^2\right);\end{aligned} \tag{15}$$

hence (12) follows. □

THEOREM 2.3. *Under the same assumptions of Lemma* 2.2 *we have*

$$\|f - L_m(f_j)\|_{L^2_{w_\alpha}} \le c_1\left[\frac{1}{m^{1/4}}\int_0^{1/\sqrt{m}}\frac{\omega^r_\varphi(f,t)_{L^2_{w_\alpha}}}{t^{3/2}}dt + e^{-c_2 m}\|f\|_{L^2_{w_\alpha}}\right], \tag{16}$$

where c_1 and c_2 are two positive constants independent of m and f.

Moreover, if $f \in L^2_{w_\alpha,s}$ with $s > \frac{1}{2}$, then

$$\|f - L_m(f_j)\|_{L^2_{w_\alpha}} \le \frac{c}{n^{s/2}} \|f\|_{L^2_{w_\alpha,s}}. \tag{17}$$

Proof. Having set $P_M = S_M(f)$, $M = \lfloor \frac{\theta}{1+\delta} m \rfloor < m$, we can write

$$\|f - L_m(f_j)\|_{L^2_{w_\alpha}} \le \|f - f_j\|_{L^2_{w_\alpha}} + \|f_j - P_M\|_{L^2_{w_\alpha}}$$

$$+\|L_m((f - P_M)_j)\|_{L^2_{w_\alpha}} + \|L_m(\psi_j P_M)\|_{L^2_{w_\alpha}}, \tag{18}$$

where

$$P_{M_j} = P_M - \psi_j P_M.$$

By (11) the sum of first two terms is dominated by

$$c_1[E_M(f)_{L^2_{w_\alpha}} + e^{-c_2 m}\|f\|_{L^2_{w_\alpha}}] \le c_1 \omega^r_\varphi\left(f, \frac{1}{\sqrt{m}}\right) + e^{-c_2 m}\|f\|_{L^2_{w_\alpha}}$$
$$\le \frac{c}{m^{1/4}} \int_0^{1/\sqrt{m}} \frac{\omega^r_\varphi(f,t)_{L^2_{w_\alpha}}}{t^{3/2}} dt + e^{-cm}\|f\|_{L^2_{w_\alpha}}.$$

About the third term, by using Lemma 2.2 and recalling the property

$$\omega^r_\varphi(f + g, t) \le c[\omega^r_\varphi(f,t) + \omega^r_\varphi(g,t)],$$

which follows directly from the definition of ω^r_φ, we have

$$\|L_m((f - P_M)_j)\|_{L^2_{w_\alpha}} \le \frac{c}{m^{1/4}} \int_0^{1/\sqrt{m}} \frac{\omega^r_\varphi(f - P_M, t)_{L^2_{w_\alpha}}}{t^{3/2}} dt$$
$$+ c\|f - P_M\|_{L^2_{w_\alpha}} \le \frac{c}{m^{1/4}} \int_0^{1/\sqrt{m}} \frac{\omega^r_\varphi(f,t)_{L^2_{w_\alpha}}}{t^{3/2}} dt + \frac{c}{(\sqrt{m})^r} \|P^{(r)}_M \varphi^r w_\alpha\|_2$$
$$+ c\|f - P_m\|_{L^2_{w_\alpha}} \le \frac{c}{m^{1/4}} \int_0^{1/\sqrt{m}} \frac{\omega^r_\varphi(f,t)_{L^2_{w_\alpha}}}{t^{3/2}} dt + c\omega^r_\varphi(f,t)_{L^2_{w_\alpha}}.$$

The last bound is obtained by using Theorem 3.7 in [3]. Therefore,

$$\|L_m((f - P_M)_j)\|_{L^2_{w_\alpha}} \le \frac{c}{m^{1/4}} \int_0^{1/\sqrt{m}} \frac{\omega^r_\varphi(f,t)_{L^2_{w_\alpha}}}{t^{3/2}} dt.$$

Finally,

$$\|L_m(\psi_j P_M)\|^2_{L^2_{w_\alpha}} = \sum_{k=j+1}^{m} \lambda_k P^2_M(x_k) \sim \sum_{k=j+1}^{m} w_\alpha(x_k) \Delta x_k P^2_M(x_k).$$

By (13) with $f = P_M$ and $r = 1$, we get

$$w_\alpha(x_k) P^2_M(x_k) \Delta x_k \le c(\|P_M \sqrt{w_\alpha}\|^2_{L^2(I_k)} + (\Delta x_k)^2 \|P'_M \sqrt{w_\alpha}\|^2)$$
$$\le c\left(\|P_M\|^2_{L^2_{w_\alpha}(I_k)} + \frac{m^{2/3}}{\sqrt{m}} \|P'_M \varphi\|^2_{L^2_{w_\alpha}(I_k)}\right)$$

since for $k > j$ we have $\Delta x_k \le cm^{1/3}\sqrt{\frac{x_k}{m}}$ (see [14]).

Summing on $k > j$, it follows that

$$\|L_m(\psi_j P_M)\|_{L^2_{w_\alpha}} \le c\|P_M w_\alpha\|_{L^2(4\theta m,\infty)} + \frac{m^{1/3}}{\sqrt{m}}\|P'_M \varphi w_\alpha\|_{L^2(4\theta m,\infty)}.$$

Now we apply (7) and then the Bernstein inequality (see [3, Theorem 3.2]), and we obtain

$$\|L_m(\psi_j P_M)\|_{L^2_{w_\alpha}} \le c_1 e^{-c_2 m}\|P_M\|_{L^2_{w_\alpha}} \le c_1 e^{-c_2 m}\|f\|_{L^2_{w_\alpha}},$$

and the proof is completed.

If $f \in L^2_{w_\alpha,s}$ with $r > s > \frac{1}{2}$, then we also have

$$\frac{1}{m^{1/4}}\int_0^{1/\sqrt{m}} \frac{\omega^r_\varphi(f,t)_{L^2_{w_\alpha}}}{t^{3/2}}dt = \frac{1}{m^{1/4}}\int_0^{1/\sqrt{m}} \frac{t^{s-1}\omega^r_\varphi(f,t)_{L^2_{w_\alpha}}}{t^{s+1/2}}dt$$

$$\le \frac{1}{\sqrt{2s-1}}\frac{1}{m^{s/2}}\left(\int_0^{1/\sqrt{m}}\left[\frac{\omega^r_\varphi(f,t)_{L^2_{w_\alpha}}}{t^{s+1/2}}\right]^2 dt\right)^{1/2} \le \frac{c}{m^{s/2}}\|f\|_{L^2_{w_\alpha,s}}.$$

Bound (17) then follows from (16). □

Remark 2.4. Proceeding as in [15, p. 285], it can be shown that when $f \in L^2_{w_\alpha,s}$, $s > \frac{1}{2}$, for any real $0 \le t \le s$ we have

$$\|f - L_m(f_j)\|_{L^2_{w_\alpha,t}} \le \frac{c}{m^{\frac{s-t}{2}}}\|f\|_{L^2_{w_\alpha,s}}. \tag{19}$$

From this it follows that the operator

$$L^{w_\alpha}_m : f \to L_m(f_j)$$

is uniformly bounded in $L^2_{w_\alpha,s}$, $s > \frac{1}{2}$, in the sense that

$$\sup_m \|L^{w_\alpha}_m\|_{L^2_{w_\alpha,s}\to L^2_{w_\alpha,s}} < \infty. \tag{20}$$

COROLLARY 2.5. *If in* (2) $f \in L^2_{w_\alpha,s}$ *with* $s > \frac{1}{2}$, *then*

$$|R^\theta_m(f)| \le \frac{c}{m^{s/2}}\|f\|_{L^2_{w_\alpha,s}}. \tag{21}$$

Proof. The bound follows immediately from (17) and Schwarz's inequality. □

Remark 2.6. Bounds (16) and (20) are of the same order as that of the best polynomial approximation error in $L^2_{w_\alpha,s}$ (see [12]). Moreover, (20) holds also when f is not absolutely continuous.

As stated by (19), the new operator $L^{w_\alpha}_m$ is uniformly bounded in $L^2_{w_\alpha,s}$.

Finally, we remark that bounds (19) and (20) have been obtained by assuming $f \in L^2_{w_\alpha,s}$, where $w_\alpha(x) = x^\alpha e^{-x}$. In [12], where the function f is interpolated by $L_m(f)$, similar upper bounds were derived by assuming $f \in L^2_{w^q_\alpha,s}$, with $w^q_\alpha = x^\alpha e^{-qx}, 0 < q < 1$.

As mentioned in the introduction, in [13] a corresponding estimate, given in terms of weighted L^1 norm, were obtained. This can, however, be derived with a much simpler proof, as we shall do in the next theorem, where we consider functions $f \in W_1^1(w_\alpha)$.

In the following we define

$$L_w^1 = \left\{ f : \int_0^\infty w(x)|f(x)|dx < \infty \right\}.$$

THEOREM 2.7. *For any $f \in W_1^1(w_\alpha)$ we have*

$$\left|R_m^\theta(f)\right| \le c_1 \left[\frac{E_{M-1}(f')_{L^1_{\varphi w_\alpha}}}{\sqrt{m}} + e^{-c_2 m} \|f\|_{L^1_{w_\alpha}} \right],$$

where c_1 and c_2 are independent of m and f.

Proof. First, we recall that the estimates in (11) are true also in $L^1_{w_\alpha}$.

Taking $f_j = f - \psi_j f$ as in (10), we have

$$\begin{aligned} R_m^\theta(f) &= \int_0^\infty [f(x) - f_j(x)] w_\alpha(x) dx \\ &\quad + \int_0^\infty f_j(x) w_\alpha(x) dx - \sum_{k=1}^m \lambda_k(w_\alpha) f_j(x_k) = F_m(f) + R_m(f_j), \end{aligned}$$

where $R_m(f)$ is the remainder term of the Gauss–Laguerre rule (1). By (11) and Favard's theorem (see [3, Theorem 3.3]) we have

$$\begin{aligned} |F_m(f)| &\le \int_{4\theta m}^\infty |f(x) w_\alpha(x)| dx \\ &\le E_M(f)_{L^1_{w_\alpha}} + c_1 e^{-c_2 m} \|f w_\alpha\|_{L^1} \\ &\le \frac{c}{\sqrt{m}} E_{M-1}(f')_{L^1_{w_\alpha \varphi}} + c_1 e^{-c_2 m} \|f w_\alpha\|_{L^1}. \end{aligned} \tag{22}$$

For $R_m(f_j)$ we use Peano's theorem (see [2]) and write

$$R_m(f_j) = \int_0^\infty R_m((\cdot - t)_+^0) f_j'(t) dt, \quad (x - t)_+^0 = \begin{cases} 1, & t < x, \\ 0, & t \ge x. \end{cases}$$

For $0 \le t \le x_{m,1}$,

$$\begin{aligned} |R_m((\cdot - t)_+^0)| &= \left| \int_0^\infty (x - t)_+^0 w_\alpha(x) dx - \sum_{k=1}^m \lambda_k(w_\alpha)(x_k - t_0)_+^0 \right| \\ &= \left| - \int_0^t w_\alpha(x) dx \right| \le c \sqrt{\frac{t}{m}} w_\alpha(t), \end{aligned}$$

while for $t > x_{m,1}$, using the result given in [5, p. 105], we have

$$|R_m((\cdot - t)_+^0)| \le \lambda_m(w_\alpha, t).$$

Therefore,

$$|R_m(f_j)| \le \int_0^{x_{j+1}} |R_m((\cdot - t))| |f_j'(t)| dt.$$

However, when $t \le x_{j+1}$ we also have $\lambda_m(w_\alpha, t) \le c\sqrt{\frac{t}{m}}w_\alpha(t)$ and

$$\begin{aligned}|R_m(f_j)| &\le \frac{c}{\sqrt{m}}\int_0^{x_{j+1}} |f_j'(x)\varphi(x)w_\alpha(x)|dx\\ &\le c\left(\frac{1}{\sqrt{m}}\|f'\varphi w_\alpha\|_{L^1} + \|fw_\alpha\|_{L^1(4\theta m,\infty)}\right),\end{aligned}$$

since $\psi_j'(x) = 0$ if $x \le x_j$ and $\psi_j'(x) \le (\Delta x_j)^{-1}\|\psi'\|_\infty \sim \frac{\sqrt{m}}{\varphi(x_j)}$ for $x \in (x_j, x_{j+1})$.

Thus, for each polynomial Q of degree $M = \lfloor \frac{\theta m}{1+\theta} \rfloor \sim m$ such that

$$\|(f-Q)w_\alpha\|_{L^1} \le cE_M(f)_{w_\alpha,1},$$

we also have

$$|R_m((f-Q)_j)| \le c\left(\frac{\|(f-Q)'\varphi w_\alpha\|_{L^1}}{\sqrt{m}} + \|(f-Q)w_\alpha\|_{L^1(4\theta m,\infty)}\right).$$

Here and in the next lines the constant c is independent of f and m.

From the inequality (see [3])

$$\|(f-Q)'\varphi w_\alpha\|_{L^1} \le c(\sqrt{m}\|(f-Q)w_\alpha\|_{L^1} + E_{M-1}(f')_{L^2_{\varphi w_\alpha,1}}),$$

using the Favard theorem we obtain the bound

$$|R_m((f-Q)_j)| \le \frac{c}{\sqrt{m}}E_{M-1}(f')_{L^2_{\varphi w_\alpha,1}}.$$

Since

$$|R_m(f_j)| = |R_m(f_j - Q)| \le |R_m((f-Q)_j)| + |R_m(\psi_j Q)|$$

and

$$|R_m(\psi_j Q)| \le \int_{4\theta m}^\infty |Q(x)|w_\alpha(x)dx + \sum_{k=j+1}^m \lambda_k|Q(x_k)|,$$

using (7) and proceeding as in the proof of Theorem 2.3 we have

$$|R_m(\psi_j Q)| \le ce^{-Am}\|fw_\alpha\|_{L^1},$$

where c and A are independent of f and m. The theorem then easily follows. □

We remark that from this estimate, using the Favard theorem we easily obtain (3). The same estimate can also be applied to wider classes of functions—for example, to Besov spaces (see [3]).

We also notice that while the remainder of the truncated rule satisfies the bound given in Theorem 2.7, the full Gauss–Laguerre rule (1) can only satisfy an error estimate of type

$$|R_m(f)| \le \frac{c}{m^{1/6}}\|f'\varphi w_\alpha\|_{L^1}, \tag{23}$$

which has been derived in [13]. That is, the exponent 1/6 of m in (23) is optimal. This is confirmed by the following result, which we prove in the case $\alpha = 0$.

THEOREM 2.8. *For any integer $m \geq 2$ there exists a function $f_m(x)$ such that $f_m \in AC$, with $\|f_m' \varphi w_0\|_{L^1} < \infty$, and*

$$R_m(f_m) \sim \frac{c}{m^{1/6}} \|f_m' \varphi w_0\|_{L^1}, \tag{24}$$

where the constant c is independent of m and f_m.

Proof. Set $a_m = x_m - 2$ and notice that for $m \geq 2$ we have $x_{m-1} < a_m < x_m$. Then define the function

$$f_m(x) = \begin{cases} 0 & \text{in } [0, a_m), \\ x - a_m & \text{in } [a_m, x_m), \\ 2 & \text{in } [x_m, \infty). \end{cases}$$

Obviously $f_m \in AC$ and $\|f_m' \varphi w_0\|_{L^1} < \infty$. Recalling Peano's error representation

$$R_m(f) = \int_0^\infty \left[\int_0^\infty (x-t)_+^0 w_0(x) dx - \sum_{i=1}^m \lambda_i (x_i - t)_+^0 \right] f'(t) dt \tag{25}$$

and replacing f by f_m defined above, we have

$$R_m(f_m) = \int_{a_m}^{x_m} [e^{-t} - \lambda_m] f_m'(t) dt. \tag{26}$$

Further,

$$\sqrt{m} R_m(f_m) = \sqrt{m} \int_{a_m}^{x_m} f_m'(t) e^{-t} dt - \sqrt{m} \lambda_m \int_{a_m}^{x_m} f_m'(t) dt =: A_m - B_m.$$

For the first term A_m we have

$$A_m = \sqrt{m} \int_{a_m}^{x_m} \frac{1}{\sqrt{t}} f_m'(t) \sqrt{t} e^{-t} dt \sim \|f_m' \varphi w_0\|_{L^1}.$$

For the second one,

$$B_m = \sqrt{m} \lambda_m \int_{a_m}^{x_m} f_m'(t) \sqrt{t} e^{-t} (t^{-\frac{1}{2}} e^t) dt,$$

after noticing that

$$e^{-2} \frac{e^{x_m}}{\sqrt{x_m - 2}} \leq \frac{e^t}{\sqrt{t}} \leq \frac{e^{x_m}}{\sqrt{x_m}}$$

and (see [9])

$$\lambda_m \sim c m^{\frac{1}{3}} e^{-x_m},$$

we have

$$B_m \sim m^{\frac{1}{3}} \|f_m' \varphi w_0\|_{L^1}.$$

Estimate (24) then follows. □

3. Error bounds for the truncated product rule. In this section we apply the truncation, previously defined for the Gauss–Laguerre formulas, to product integration rules for integrals which exhibit a kernel $k_y(x) = k(x,y)$. This is in view of the construction of Nyström-type interpolants for the numerical solution of corresponding integral equations. This application will be considered in the final section.

The product rules we consider have the form

$$\int_0^\infty w_\alpha(x)k(x,y)f(x)dx = \sum_{i=1}^m w_i(y)f(x_i) + R_m(k_y;f), \tag{27}$$

with

$$w_i(y) = w_i(k;y) = \int_0^\infty w_\alpha(x)k(x,y)l_i(x)dx = \lambda_i S_m(k_y;x_i).$$

They are of interpolatory type; that is, they are obtained by replacing $f(x)$ by its Lagrange interpolation polynomial $L_m(f;x)$ associated with the zeros $\{x_i\}$ of the Laguerre polynomial $p_m(x)$. For the construction of the coefficients $w_i(y)$ of these rules and convergence properties, see [11], [12].

As in the case of (1), we associate with (27) its "truncation"

$$\int_0^\infty w_\alpha(x)k(x,y)f(x)dx = \sum_{0<x_i\le 4\theta m} w_i(y)f(x_i) + R_m^\theta(k_y;f), \tag{28}$$

where $0<\theta<1$ is arbitrarily chosen.

For this rule we derive some estimates which are fundamental for proving stability and convergence properties for our Nyström interpolants.

As for the remainder of (2), using (17) it is straightforward to derive the following bound for the remainder term $R_m^\theta(k_y;f)$ in (28).

Theorem 3.1. *If for a given y we have $\|k_y\|_{L^2_{w_\alpha}} < \infty$ and $f \in L^2_{w_\alpha,s}$, $s>\frac{1}{2}$, then*

$$|R_m^\theta(k_y;f)| \le \frac{c\|k_y\|_{L^2_{w_\alpha}}}{m^{s/2}}\|f\|_{L^2_{w_\alpha,s}},$$

where the constant c is independent of $m, f,$ and k_y.

By making a weaker assumption on the kernel $k_y(x)$, we are also able to derive a bound for $R_m^\theta(k_y;f)$ given in terms of a weighted L^∞ norm of f. The function f could not be in $L^2_{w_\alpha}$. Although this bound will not be used in section 4, we think that it is of interest on its own and could have some applications.

Let $u(x) = \sqrt{w_\alpha(x)}(\frac{x}{1+x})^a(1+x)^b$, $v(x) = \sqrt{w_\alpha(x)}(\frac{x}{1+x})^A(1+x)^B$, and $W(g;x) = 1+\log^+ x+\log^+|g(x)|$. Moreover, let the real constants a, b, A, B satisfy the conditions $A \le a$, $a > -1+\max\left(-\frac{\alpha}{2},\frac{1}{4}\right)$, $A \le \min\left(\frac{\alpha}{2},-\frac{1}{4}\right)$, $b<-1/4$, $B \ge -7/12$, and $B \ge b+1/6$. Then (see [17]) the following bound holds:

$$\int_0^\infty |S_m(F;x)u(x)|dx \le c\left[1+\int_0^\infty |F(x)|v(x)W(F;x)dx\right], \tag{29}$$

where the constant c is independent of m and F.

The next result then follows.

Lemma 3.2. *Let $b < -1/4$, and assume the conditions*

$$\|f\sqrt{w_\alpha(x)}(1+\cdot)^{-b}\|_\infty < \infty \text{ and}$$

$$B(k_y) := \sup_{y\geq 0}\int_0^\infty e^{-\frac{x}{2}}x^\gamma(1+x)^{-A}|k_y(x)|W(k_y;x)dx < \infty, \tag{30}$$

$$\gamma = \frac{\alpha}{2} + \min\left(\frac{\alpha}{2}, -\frac{1}{4}\right).$$

Then

$$\int_0^\infty |L_m(f_j;x)k(x,y)w_\alpha(x)|dx \leq c(1+B(k_y))\|f\sqrt{w_\alpha}(1+\cdot)^{-b}\|_\infty, \tag{31}$$

where the constant c is independent of m and f.

Proof. Set $g_m(x) = \text{sgn } L_m(f_j;x)$. Then

$$\Delta = \int_0^\infty |L_m(f_j;x)k(x,y)w_\alpha(x)|dx = \sum_{i=1}^m \lambda_i S_m(k_y g_m;x_i)f_j(x_i)$$
$$= \sum_{i=1}^j \lambda_i S_m(k_y g_m;x_i)f(x_i)(1+x_i)^{-b}(1+x_i)^b;$$

hence, recalling (13),

$$\Delta \leq c\|f\sqrt{w_\alpha}(1+\cdot)^{-b}\|_\infty \sum_{i=1}^j \sqrt{w_\alpha(x_i)}(1+x_i)^b\Delta x_i|P(x_i)|,$$

where $P = S_m(k_y g_m)$.

At this point it is sufficient to show that the last sum is uniformly bounded with respect to m. To this end, notice that, because of (13) with $f = P$ and $r = 1$,

$$\sqrt{w_\alpha(x_i)}(1+x_i)^b\Delta x_i|P(x_i)| \leq c\int_{x_{i-1}}^{x_i}|P(x)w_\alpha(x)(1+x)^b|dx$$
$$+\frac{c}{\sqrt{m}}\int_{x_{i-1}}^{x_i}|P'(x)|\sqrt{x}w_\alpha(x)(1+x)^b|dx,$$

since $\Delta x_i \leq c\sqrt{\frac{x_i}{m}}$. Therefore,

$$\Delta \leq c\|f_j\sqrt{w_\alpha}(1+\cdot)^{-b}\|_\infty\left[\int_0^{x_j}|P(x)\sqrt{w_\alpha(x)}(1+x)^b|dx\right.$$
$$\left.+\frac{1}{\sqrt{m}}\int_0^{x_j}|P'(x)\sqrt{x}\sqrt{w_\alpha(x)}(1+x)^b|dx\right].$$

By applying a Bernstein inequality (see [16]) to the last integral, we obtain

$$\Delta \leq c\|f_j\sqrt{w_\alpha}(1+\cdot)^{-b}\|_\infty\int_0^\infty|S_m(k_y g_m,x)\sqrt{w_\alpha(x)}(1+x)^b|dx. \tag{32}$$

Now we recall (29) with $A = \min(\frac{\alpha}{2}, -\frac{1}{4}), a = 0$, and $B = 0$. The integral in (32) is dominated by the corresponding bound given by (31).

The proof is then completed. □

Further, by considering the quantity

$$\sum_{i=1}^{j} \lambda_i S_m(k_y; x_i) f(x_i)$$

and proceeding as we have done to bound Δ, we obtain

$$\sup_m \sum_{i=1}^{j} \lambda_i \frac{|S_m(k_y, x_i)|}{\sqrt{w_\alpha(x_i)}(1+x_i)^{-b}} \leq c < \infty. \tag{33}$$

Theorem 3.3. *Under the same assumptions of Lemma* 3.2 *we have*

$$\sup_{y\geq 0} |R_m^\theta(k_y; f)| \leq c_1[E_{M-1}(f)_{L^\infty_\sigma} + e^{-c_2 m}\|f\|_{L^\infty_\sigma}], \quad \sigma(x) = \sqrt{w_\alpha(x)}(1+x)^{-b}, \tag{34}$$

where the constant c is independent of m and f.

Proof. Write

$$R_m^\theta(k_y; f) = \int_0^\infty [f(x) - f_j(x)] k(x,y) w_\alpha(x) dx \tag{35}$$

$$+ \left[\int_0^\infty f_j(x) k(x,y) w_\alpha(x) dx - \sum_{i=1}^{m} w_i(y) f_j(x_i) \right].$$

For the first term on the right-hand side of (35) we have

$$\Gamma := \left| \int_0^\infty (f(x) - f_j(x)) k(x,y) w_\alpha(x) dx \right| = \left| \int_{x_j}^\infty \psi_j(x) f(x) k(x,y) w_\alpha(x) dx \right|.$$

Thus

$$\begin{aligned} \Gamma &\leq \int_{4\theta m}^\infty |f(x)\sqrt{w_\alpha(x)}(1+x)^{-b} k(x,y)\sqrt{w_\alpha(x)}(1+x)^{b}| dx \\ &\leq \left\{ \sup_{[4\theta m,\infty)} |f(x)\sqrt{w_\alpha(x)}(1+x)^{-b}| \right\} \int_{4\theta m}^\infty |k(x,y)| \sqrt{w_\alpha(x)} dx \quad (b<0) \\ &\leq c_1[E_M(f)_{L^\infty_\sigma} + e^{-c_2 m}\|f\|_{L^\infty_\sigma}] \sup_{y\geq 0} E_M(k_y)_{L^1_{\sqrt{w_\alpha}}}, \end{aligned}$$

that is,

$$\Gamma \leq c_1[E_{M-1}(f)_{L^\infty_\sigma} + e^{-c_2 m}\|f\|_{L^\infty_\sigma}],$$

since

$$E_M(f)_{L^\infty_\sigma} \leq E_{M-1}(f)_{L^\infty_\sigma}$$

and

$$\sup_{y\geq 0} E_M(K_y)_{L^1_{\sqrt{w_\alpha}}} \leq B(k).$$

The second term in (35) is dominated by

$$\int_0^\infty |f_j - q_{m-1}|(x)|k(x,y)|w_\alpha(x)dx + \sum_{i=1}^m \lambda_i |S_m(k_y;x_i)||f_j(x_i) - q_{m-1}(x_i)| \\ =: A_1 + A_2$$

for each $q_{m-1} \in \mathbb{P}_{m-1}$. Moreover,

$$A_1 \leq \left(\int_0^\infty |k(x,y)|\sqrt{w_\alpha(x)}(1+x)^b dx \right) \|(f_j - q_{m-1})\sigma\|_\infty, \\ A_2 \leq c\|(f_j - q_{m-1})\sigma\|_\infty.$$

The bound of A_2 follows from (33). Taking the infimum over q_{m-1} and the sup over y, it follows that

$$A_1 + A_2 \leq cE_{m-1}(f_j)_{L^\infty_\sigma} \leq c_1[E_{M-1}(f)_{L^\infty_\sigma} + e^{-c_2 m}\|f\|_{L^\infty_\sigma}];$$

hence (34) follows. □

In the next section, to prove the stability of our Nyström interpolants, we shall need to consider the so-called (see [19]) companion rule, associated with (28), that is, obtained by taking in (28) the absolute values of $k_y(x)$ and $w_i(y)$. This is

$$(36) \qquad \int_0^\infty w_\alpha(x)|k_y(x)|h(x)dx = \sum_{0<x_i\leq 4\theta m} |w_i(y)|h(x_i) + R_m^{c,\theta}(k_y;h).$$

In particular, we shall need to prove its convergence, as $m \to \infty$, for a certain class of functions $h(x)$. To this end, Theorem 3.4 and Proposition 3.5 are of importance.

THEOREM 3.4. *If for a given y we have $\|k_y\|_{L^2_{w_\alpha}} < \infty$ and $h \in L^2_{w_\alpha,1}$, then*

$$|R_m^{c,\theta}(k_y;h)| \leq c\|h\|_{L^2_{w_\alpha,1}} \left[\omega_\varphi\left(k_y, \frac{1}{\sqrt{m}}\right)_{L^2_{w_\alpha}} + \frac{1}{\sqrt{m}}\|k_y\|_{L^2_{w_\alpha}} \right],$$

where the constant c is independent of k_y and m.

Proof. Writing

$$|R_m^{c,\theta}(k_y;h)| = \left| \int_0^\infty |k_y(x)|h(x)w_\alpha(x)dx - \sum_{i=1}^j \lambda_i|S_m(k_y;x_i)|h(x_i) \right|$$

$$\leq \int_0^\infty |k_y(x) - S_m(k_y;x)|h(x)w_\alpha(x)dx \quad (=: A_1)$$

$$+ \left| \int_0^\infty |S_m(k_y;x)|h(x)w_\alpha(x)dx - \sum_{i=1}^j \lambda_i|S_m(k_y;x_i)|h(x_i) \right| \quad (=: A_2)$$

and recalling (4), for A_1 we have

$$A_1 \leq \|h\|_{L^2_{w_\alpha}}\|k_y - S_m(k_y)\|_{L^2_{w_\alpha}} \leq c\|h\|_{L^2_{w_\alpha}}\omega_\varphi\left(k_y, \frac{1}{\sqrt{m}}\right)_{L^2_{w_\alpha}}.$$

Moreover, if in Theorem 2.7 we take $f(x) = |S_m|h(x) = |S_m(k_y;x)|h(x)$ and recall that

$$E_{M-1}(f')_{L^1_{\varphi w_\alpha}} \le \|f'\varphi w_\alpha\|_{L^1},$$

then we have

$$A_2 = |R^\theta_m(|S_m|h)| \le c_1 \left[e^{-c_2 m}\|S_m h w_\alpha\|_{L^1} + \frac{\|(S_m h)'\varphi w_\alpha\|_{L^1}}{\sqrt{m}} \right].$$

By applying Cauchy's inequality we further have

$$\begin{aligned} A_2 &\le c_1\|h\|_{L^2_{w_\alpha,1}} \left[\left(\frac{1}{\sqrt{m}} + e^{-c_2 m} \right) \|S_m - k_y + k_y\|_{L^2_{w_\alpha}} + \frac{1}{\sqrt{m}}\|S'_m\varphi\|_{L^2_{w_\alpha}} \right] \\ &\le c\|h\|_{L^2_{w_\alpha,1}} \left[E_m(k_y)_{L^2_{w_\alpha}} + \frac{1}{\sqrt{m}}\|S'_m\varphi\|_{L^2_{w_\alpha}} + \frac{1}{\sqrt{m}}\|k_y\|_{L^2_{w_\alpha}} \right]. \end{aligned}$$

Because of Theorem 3.7 in [3], the sum of the first two terms in the last line is bounded by $\omega_\varphi(k_y, \frac{1}{\sqrt{m}})$. Thus the bound for $|R^{c,\theta}_m(k_y;h)|$ is proved. □

The next result is needed to define the final behavior of the remainder terms $R^{c,\theta}_m(k_y;h)$ as $m \to \infty$.

Proposition 3.5. *If* $\sup_{y\ge 0}\rho(y)\|k_y\|_{L^2_{w_\alpha}} < \infty$ *for a bounded nonnegative weight function* $\rho(y)$, *then*

$$\lim_{m\to\infty}\sup_{y\ge 0}\rho(y)\omega^r_\varphi\left(k_y, \frac{1}{\sqrt{m}}\right)_{L^2_{w_\alpha}} = 0 \qquad \textit{with } r \ge 1.$$

Proof. Setting $d_k = \int_0^\infty k_y(x)p_k(x)w_\alpha(x)dx$, from the assumption made it follows that

$$\sup_{y\ge 0}\rho(y)\sum_{k=0}^{\infty} d^2_k(k_y) = \sup_{y\ge 0}\rho(y)\|k_y\|_{L^2_{w_\alpha}} < \infty;$$

hence

$$\lim_{m\to\infty}\sup_{y\ge 0}\rho(y)\sum_{k>m} d^2_k(k_y) = 0, \qquad \text{that is,} \qquad \lim_{m\to\infty}\sup_{y\ge 0}\rho(y)E_m(k_y)_{L^2_{w_\alpha}} = 0.$$

However (see [3]),

$$\begin{aligned} \rho(y)\omega^r_\varphi\left(k_y, \frac{1}{\sqrt{m}}\right)_{L^2_{w_\alpha}} &\le \frac{c_r}{(\sqrt{m})^r}\sum_{k=0}^{m}(1+k)^{\frac{r}{2}-1}\rho(y)E_k(k_y)_{L^2_{w_\alpha}} \\ &\sim \frac{\sum_{k=0}^{m}(1+k)^{\frac{r}{2}-1}\rho(y)E_k(k_y)_{L^2_{w_\alpha}}}{\sum_{k=0}^{m}(1+k)^{\frac{r}{2}-1}}. \end{aligned}$$

Since the latter expression tends to zero as $m \to \infty$, uniformly with respect to y, we have

$$\lim_{m\to\infty}\sup_{y\ge 0}\rho(y)\omega^r_\varphi\left(k_y, \frac{1}{\sqrt{m}}\right)_{L^2_{w_\alpha}} = 0. \qquad □$$

Remark 3.6. If in (36) we take $h \equiv 1$ and consider the full sum, i.e., $i = 1, \dots, m$, then, using arguments similar to those of Theorem 3.4, it is possible to show, assuming $\sup_{y \geq 0} \|k_y\|_{L^2_{w_\alpha}} < \infty$, that the corresponding quadrature remainder term tends to zero, as $m \to \infty$. This in turn implies that $\sup_{y \geq 0} \sum_{i=1}^m |w_i(y)| < \infty$. In [12] this property was proved under the stronger assumption that $\sup_{y \geq 0} \|k_y\|_{L^2_{w_\alpha,s}} < \infty$ for some $s > 1/3$.

4. A Nyström-type interpolant. As in [12], here we consider integral equations of the form

$$u(y) - \mu \int_0^\infty k(x,y)u(x)dx = f(y), \tag{37}$$

whose solutions decay at least as $e^{-\beta x}, \beta = \frac{1}{2} + \epsilon,\ \epsilon > 0$, for $x \to \infty$, to a finite value $u(\infty)$.

A class of integral equations of this type arises naturally from the equations of Mellin type defined on a finite interval (see [18]). Thus consider

$$u(s) - \mu \int_0^1 k\left(\frac{t}{s}\right) \frac{u(t)}{t} dt = f(s),\ 0 < s \leq 1, \tag{38}$$

and set $t = e^{-x},\ s = e^{-y}$; we have

$$v(y) - \mu \int_0^\infty k(e^{-(y-x)})v(x)dx = f(e^{-y}), \tag{39}$$

with

$$v(x) = u(e^{-x}).$$

We recall (see [4]) that often $u(t) \sim t^\sigma, \sigma > \frac{1}{2}$, near the origin, so that the (weakly) singular behavior at $t = 0$ is transformed, by the above change of variable, into the exponential decay $e^{-\sigma x}$. Thus the method we examine in this section could be used to solve the above Mellin equations.

A test equation of type (37) we have considered in [11], [12] is

$$u(y) - \frac{1}{4} \int_0^\infty E_1(|x-y|)u(x)dx = \frac{1}{2}, \tag{40}$$

where $E_1(x) = \int_x^\infty \frac{e^{-t}}{t} dt$ is the well-known exponential integral. This equation arises, for example, in the modeling of the neutron transfer in an infinite slab. The solution $u(x)$ has the behavior $1 + o(e^{-\beta x})$ as $x \to \infty$, with $\beta = 0.9575$, and has a weak singularity at the origin, of the form $\frac{\sqrt{2}}{2} + O(x^\epsilon)$ with $\epsilon > 0$ as small as we like (see, for example, [6]).

The Nyström interpolant we propose requires us to preliminarily rewrite (37) in the form

$$v(y) - \mu e^y \int_0^\infty e^{-x} k(x,y)v(x)dx = g(y),$$

where

$$v(x) = e^x[u(x) - u(\infty)]$$

and

$$g(y) = e^y \left\{ f(y) - u(\infty) \left[1 - \mu \int_0^\infty k(x,y)dx \right] \right\},$$

that is,

$$(I - \mu K)v = g. \tag{41}$$

To do this one needs to know a priori the value $u(\infty)$, which is, however, often the case. This assumption is also made, for example, in the product integration method proposed in [6]. Then, using the corresponding truncated product rule of form (26), we obtain

$$(I - \mu K_m)v_m = g \tag{42}$$

with

$$K_m v_m(y) = \sum_{i=1}^{j} \bar{w}_i(y) v_m(x_i), \qquad \bar{w}_i(y) = e^y w_i(y),$$

that is,

$$v_m(y) = g(y) + \mu K_m v_m(y). \tag{43}$$

Notice that applying our truncated quadrature rule does not mean performing a finite section of (41).

Both (41) and (42) will be examined in the weighted space

$$X := \{v \in L^\infty(0,\infty) : \|v\|_X := \|\rho(x)v(x)\|_\infty < \infty\},$$

where $\rho(x) = e^{(-\frac{1}{2}+\epsilon_1)x}$, with some $0 < \epsilon_1 \leq \epsilon$.

The values $\{v_m(x_i)\}$ are obtained by solving the linear system (hopefully nonsingular)

$$v_m(x_l) - \mu K_m v_m(x_l) = g(x_l), \ l = 1, \ldots, j. \tag{44}$$

Recalling Remark 3.6, it is now possible to claim that the method is indeed stable in X, whenever the kernel $k(x,y)$ satisfies certain assumptions.

Lemma 4.1. *If*

$$\|K\|_{X\to X} = \sup_{y\geq 0} \rho(y) \|\bar{k}_y \rho^{-1}\|_{L^1_{w_0}} < \infty, \ \bar{k}_y = e^y k(x,y), \tag{45}$$

and

$$\sup_{y\geq 0} \rho(y) \|\bar{k}_y\|_{L^2_{w_0}} < \infty, \tag{46}$$

then

$$\overline{\lim}_{m\to\infty} \|K_m\|_{X\to X} \leq \|K\|_{X\to X}. \tag{47}$$

Proof. In the case of assumption (46), property (47) follows immediately from Theorem 3.4 and Proposition 3.5, since

$$\|K_m\|_{X\to X} \leq \sup_{y\geq 0} \rho(y) \sum_{i=1}^{j} |\bar{w}_i(y)|\rho^{-1}(x_i). \qquad \square$$

Notice that to prove this lemma we have used Theorem 3.4, which requires $h(x) \equiv \rho^{-1}(x) \in L^2_{w_\alpha,1}$.

If the operator K in (41) satisfies the condition

$$\|K\|_{X\to X} < |\mu|^{-1}, \tag{48}$$

standard analysis (see [1]) and Theorem 3.1 give the following main result.

THEOREM 4.2. *If conditions* (46) *and* (48) *are verified, then for all m sufficiently large the operator $I - \mu K_m$ is invertible and*

$$\|(I-\mu K_m)^{-1}\|_{X\to X} \leq c.$$

Moreover,

$$\|v - v_m\|_X \leq c\|(K-K_m)v\|_X \leq c\frac{\|v\|_{L^2_{w_0,s}}}{m^{\frac{s}{2}}} \tag{49}$$

whenever $v \in L^2_{w_0,s}$, $s > \frac{1}{2}$.

Notice that if we define $u_m(x) = e^{-x}v_m(x) + u(\infty)$, then

$$\|v - v_m\|_X = \sup_{x\geq 0} |e^{(\frac{1}{2}+\epsilon_1)x}[u(x) - u_m(x)]|.$$

Unfortunately, it is possible for condition (46) to be not satisfied by the kernel of the integral equation one has to solve. If, for example, we consider (40) and rewrite it in the form (see [11])

$$v(y) - \frac{1}{4}e^y \int_0^\infty e^{-x}k(x,y)v(x)dx = -\frac{1}{4}e^y[e^{-y} - yE_1(y)], \tag{50}$$

where

$$k(x,y) = E_1(|x-y|), \quad v(x) = e^x[u(x) - 1], \tag{51}$$

and (see [12]) $v \in L^2_{w_0,s}$ with s not smaller than $\frac{3}{2} - \delta, \delta > 0$, as close as we like to 0, then it is not difficult to verify, by direct calculation, that assumption (46) of Lemma 4.1 does not hold because of the factor $e^{\epsilon_1 y}$ in $\rho(y)$, while condition (48) is satisfied. Incidentally, we notice that if the kernel were $e^{-\epsilon y}E_1(|x-y|)$, with $\epsilon > 0$ as small as we like, then also condition (46) would be satisfied.

On the other hand, in the weight $\rho(y)$ we have taken $\epsilon_1 > 0$ to be allowed to apply Theorem 3.4. This happens also if in (37) we take the Picard kernel $k(x,y) = e^{-|x-y|}$. Therefore, for these two equations we cannot apply Theorem 4.2 and hence claim stability, although practical computation shows that our interpolants are indeed stable. In our opinion this is due to the (weighted) L^2 estimates we have used to prove Theorem 3.4. A proof of stability using more general tools is needed. Ours is based on the use of the companion rule (36) and Theorem 3.4, and this seems to be a bit restrictive. However, at present this question remains unanswered.

To be able to prove stability in the latter two cases, we need to modify our interpolant when $y > C$, C being a positive constant arbitrarily large. In particular, for simplicity we define the trivially modified operator

$$\bar{K}_m v(y) = \begin{cases} K_m v(y) & \text{if } 0 \leq y \leq \log A, \\ K v(y) & \text{if } y > \log A, \end{cases} \tag{52}$$

where A is a positive constant that can be chosen as large as one likes. Of course this is not satisfactory from a theoretical point of view: when $y > \log A$, we should modify $K_m v(y)$ by introducing a discrete operator which can be explicitly evaluated. However, for simplicity and because the constant A can be chosen arbitrarily large, we have made the choice (52). We recall that the solution of our equation decays exponentially at infinity, to a constant we assume to know a priori.

In this case we have

$$\|\bar{K}_m\|_{X\to X} \leq \max\left\{ \sup_{0\leq y\leq \log A} \rho(y) \sum_{i=1}^{j} |\bar{w}_i(y)|\rho^{-1}(x_i), \|K\|_{X\to X} \right\},$$

that is,

$$\lim_{m\to\infty} \|\bar{K}_m\|_{X\to X} \leq \|K\|_{X\to X}.$$

Thus, if $\|K\|_{X\to X} < |\mu|^{-1}$, then also $\|\bar{K}_m\|_{X\to X} < |\mu|^{-1}$ for all m sufficiently large. However, this implies stability and the convergence estimate (49).

Instead of applying our method to an artificial equation constructed in order to satisfy the conditions required by Theorem 4.2, we have preferred to consider a case of practical interest, taken by many authors, such as (40). For this equation, the modified version of the method described above is stable and convergent with order not smaller than $O(m^{-3/2+\epsilon})$ (see [12]). In the case of (37) with the Picard kernel, $\mu = 1/4$ and $f(y) = 1/2$, whose solution is $u(x) = 1 - (1 - 2^{-\frac{1}{2}})e^{-\frac{x}{\sqrt{2}}}$, the order of convergence is higher than any negative power of m.

In [12] the Nyström interpolant based on the nontruncated rule (27) has been examined from the numerical point of view and compared with a product integration method proposed in [6]. In particular, numerical evidence on the efficiency of the first method was given. Since in [12] this method has already shown to be competitive, here we do not present any further comparisons. We simply show the improvement that the truncation we propose generates with respect to the nontruncated version.

In the following tables we report some numerical results obtained by applying our unmodified Nyström interpolant (43) to the test equation (40), after having rewritten it in the form (41). In particular, in Table 1 we report a sample of relative errors $\{e_m(t_l)\}$ obtained by using the nontruncated product rule, i.e., choosing $j = m$ in (42). In the following three tables we list the corresponding relative errors generated by our truncated rule, applied with $\theta = 1/8, 1/16, 1/32$. These errors are very similar. The choice $\theta = 1/32$ allows to obtain an accuracy similar to that given by the complete rule, using about 1/5 of its abscissas.

While it is known that $u(0) = \frac{\sqrt{2}}{2}$, to compute the relative errors at the other points, we have taken, as reference values, those obtained with $m = 180$ in Table 1 and with $m = 512$ in Tables 2, 3, and 4. The integer $m = 180$ could not be increased because of overflow problems due to the Matlab exponential function. All computation has been performed on a PC using Matlab.

Table 1

$j=m$						
m	$e_m(0)$	$e_m(0.1)$	$e_m(0.5)$	$e_m(1)$	$e_m(5)$	$e_m(10)$
4	3.59e-3	3.46e-3	1.53e-4	4.55e-4	4.05e-6	8.88 e-8
8	1.28e-3	8.96e-4	2.29e-4	4.45e-5	4.09e-6	2.28e-7
16	4.02e-4	1.05e-4	3.23e-5	4.74e-5	4.49e-7	2.25e-8
32	1.18e-4	1.7 e-5	2.23e-5	1.08e-7	2.63e-7	1.28e-8
64	3.34e-5	7.47e-8	4.28e-7	1.62e-7	3.41e-8	1.88e-9
128	9.19e-6	2.7 e-6	2.7 e-7	1.4 e-7	1.26e-8	1.81e-10

Table 2

θ=1/8							
m	j	$e_m(0)$	$e_m(0.1)$	$e_m(0.5)$	$e_m(1)$	$e_m(5)$	$e_m(10)$
4	2	4.73e-3	4.65e-3	6.35e-4	1.72e-3	6.9 e-4	6.61e-6
8	3	8.09e-4	6.34e-4	5.93e-4	1.005e-4	7.47e-4	3.95e-6
16	7	3.88e-4	1.14e-4	2.69e-5	3.36e-5	1.5 e-6	5.33e-6
32	14	1.18e-4	1.62e-5	2.29e-5	7.6 e-9	3.01e-7	2.08e-8
64	28	3.34e-5	1.01e-6	8.9 e-7	4.71e-8	3.51e-8	1.78e-9
128	56	9.19e-6	3.79e-6	2.02e-7	6.71e-8	1.12e-8	1.06e-10
256	113	2.49e-6	1.17e-8	3.7 e-8	9.48e-8	5.89e-10	7.63e-11

Table 3

θ=1/16							
m	j	$e_m(0)$	$e_m(0.1)$	$e_m(0.5)$	$e_m(1)$	$e_m(5)$	$e_m(10)$
4	1	1.45e-3	1.42e-3	9.67e-3	1.57e-2	8.26e-4	8.3 e-6
8	2	3.37e-3	2.47e-3	1.56e-3	3.25e-3	7.4 e-4	1.03e-5
16	5	2.85e-4	1.77e-4	1.77e-5	6.52e-5	4.62e-4	7.46e-6
32	10	1.25e-4	2.95e-5	3.12e-5	5.6 e-6	3.87e-6	5.63e-6
64	20	3.35e-5	9.6 e-7	8.31e-7	7.87e-8	6.55e-8	5.63e-9
128	40	9.19e-6	3.79e-6	2.02e-7	6.72e-8	1.12e-8	1.06e-10
256	80	2.49e-6	1.17e-8	3.7 e-8	9.48e-8	5.9 e-10	7.62e-11

Table 4

θ=1/32							
m	j	$e_m(0)$	$e_m(0.1)$	$e_m(0.5)$	$e_m(1)$	$e_m(5)$	$e_m(10)$
4	1	1.44e-3	1.42e-3	9.67e-3	1.57e-2	8.26e-4	8.3 e-6
8	2	3.37e-3	2.47e-3	1.56e-3	3.25e-3	7.39e-4	1.03e-5
16	3	5.39e-4	1.26e-3	8.26e-4	1.46e-3	7.71e-4	8.39e-6
32	7	6.64e-5	1.07e-4	4.42e-5	6.37e-5	5.81e-4	5.7 e-6
64	14	3.66e-5	1.88e-6	2.81e-6	7.12e-6	2.52e-6	4.87e-6
128	28	9.29e-6	3.88e-6	1.08e-7	2.44e-8	3.05e-8	4.94e-10
256	57	2.49e-6	1.16e-8	3.71e-8	9.48e-8	5.9 e-10	7.56e-11

Table 5

2-norm condition numbers				
m	$j=m$	$\theta=1/8$	$\theta=1/16$	$\theta=1/32$
4	3.07E00	1.36E00	1.00E00	1.00E00
8	9.84E03	1.56E00	1.30E00	1.30E00
16	1.83E15	2.37E00	1.85E00	1.44E00
32	3.17E39	6.73E00	2.49E00	1.92E00
64		1.57E06	5.48E00	2.52E00
128		7.69E18	7.36E05	4.26E00
256		1.76E43	3.56E18	3.16E05

Truncation also reduces significantly the magnitude of the condition numbers of the final linear systems (see Table 5).

However, we have also noticed that in all cases considered, in spite of the values of the condition numbers, no error propagation has shown up. Actually, following the idea suggested in [10], we could replace, for example, the linear system generated by the nontruncated rule by the equivalent (preconditioned) one

$$\lambda_l^{1/2} e^{x_l} \sum_{i=1}^{m} a_i \left[\lambda_i^{-1/2} e^{-x_i} \delta_{il} - \frac{\lambda_i^{-1/2}}{4} w_i(x_l) \right] = \lambda_l^{1/2} g(x_l), \quad l = 1, \ldots, m, \tag{53}$$

where $a_i = \lambda_i^{1/2} v_m(x_i)$ and $g(x) = -\frac{e^x}{4}[e^{-x} - xE_1(x)]$, which turns out to be perfectly conditioned. The corresponding results however are very similar to those produced by the original system.

Aknowledgments. The authors are grateful to Dr. C. Frammartino for performing all the numerical computations reported in section 4. They are also indebted to the two referees for their valuable comments.

REFERENCES

[1] K. E. Atkinson, *The Numerical Solution of Integral Equations of the Second Kind*, Cambridge University Press, Cambridge, UK, 1997.

[2] P. J. Davis, *Interpolation & Approximation*, Dover, New York, 1975.

[3] M. C. De Bonis, G. Mastroianni, and M. Viggiani, *K−functionals, moduli of smoothness and weighted best approximation on the semi-axis*, in Functions, Series, Operators, Alexits Memorial Conference, Budapest, 1999, L. Leindler, F. Schipp, and J. Szabados, eds., János Bolyai Math. Soc., Budapest, 2002, pp. 181–211.

[4] J. Elschner, *On spline approximation for a class of non-compact integral equations*, Math. Nachr., 146 (1990), pp. 271–321.

[5] G. Freud, *Orthogonal Polynomials*, Akadémiáí Kiadó, Budapest, 1971.

[6] I. G. Graham and W. R. Mendes, *Nyström-product integration for Wiener-Hopf equations with applications to radiative transfer*, IMA J. Numer. Anal., 9 (1989), pp. 261–284.

[7] G. H. Hardy, J. E. Littlewood, and G. Pólya, *Inequalities*, Cambridge University Press, Cambridge, UK, 1983.

[8] K. G. Ivanov, *On the behaviour of two moduli of functions* II, Serdica, 12 (1986), pp. 196–203.

[9] N. X. Ky, *A contribution to the problem of weighted polynomial approximation of the derivative of a function by the derivative of its approximating polynomial*, Studia Sci. Math. Hungar., 10 (1975), pp. 309–316.

[10] C. Laurita and G. Mastroianni, *Condition numbers in numerical methods for Fredholm integral equations of the second kind*, J. Integral Equations Appl., 14 (2002), pp. 311–341.

[11] G. Mastroianni and G. Monegato, *Convergence of product integration rules over* $(0, \infty)$ *for functions with weak singularities at the origin*, Math. Comp., 64 (1995), pp. 237–249.

[12] G. Mastroianni and G. Monegato, *Nyström interpolants based on zeros of Laguerre polynomials for some Wiener-Hopf equations*, IMA J. Numer. Anal., 17 (1997), pp. 621–642.

[13] G. Mastroianni and G. Monegato, *Truncated Gauss-Laguerre rules*, in Recent Trends in Numerical Analysis, D. Trigiante, ed., Nova Science, Hauppauge, NY, 2000, pp. 213–221.

[14] G. Mastroianni and D. Occorsio, *Lagrange interpolation at Laguerre zeros in some weighted uniform spaces*, Acta Math. Hungar., 91 (2001), pp. 27–52.

[15] G. Mastroianni and M. G. Russo, *Lagrange interpolation in weighted Besov spaces*, Constr. Approx., 14 (1998), pp. 1–33.

[16] G. Mastroianni, *Polynomial inequalities, functional spaces and best approximation on the real semiaxis with Laguerre weights*, Electron. Trans. Numer. Anal., 14 (2002), pp. 125–134.

[17] B. Muckenhoupt, *Mean convergence of Hermite and Laguerre series* II, Trans. Amer. Math. Soc., 147 (1970), pp. 433–460.

[18] S. Prössdorf and B. Silbermann, *Numerical Analysis for Integral and Related Operator Equations*, Birkhäuser, Basel, 1991.

[19] I. H. Sloan and W. E. Smith, *Properties of interpolatory product integration rules*, SIAM J. Numer. Anal., 19 (1982), pp. 427–442.

[20] G. Szegö, *Orthogonal Polynomials*, Amer. Math. Soc. Colloq. Publ. 23, AMS, Providence, RI, 1975.

[21] A. F. Timan, *Theory of Approximation of Functions of a Real Variable*, Dover, New York, 1994.

SIAM J. NUMER. ANAL.
Vol. 41, No. 5, pp. 1893–1925

A NEW LOOK AT PROPER ORTHOGONAL DECOMPOSITION[*]

MURUHAN RATHINAM[†] AND LINDA R. PETZOLD[‡]

Abstract. We investigate some basic properties of the proper orthogonal decomposition (POD) method as it is applied to data compression and model reduction of finite dimensional nonlinear systems. First we provide an analysis of the errors involved in solving a nonlinear ODE initial value problem using a POD reduced order model. Then we study the effects of small perturbations in the ensemble of data from which the POD reduced order model is constructed on the reduced order model. We explain why in some applications this sensitivity is a concern while in others it is not. We also provide an analysis of computational complexity of solving an ODE initial value problem and study the computational savings obtained by using a POD reduced order model. We provide several examples to illustrate our theoretical results.

Key words. proper orthogonal decomposition, model reduction, dynamical systems, numerical methods

AMS subject classification. 37M99

DOI. 10.1137/S0036142901389049

1. Introduction.

1.1. Background on proper orthogonal decomposition.

Proper orthogonal decomposition (POD), also known as Karhunen–Loève decomposition or principal component analysis, provides a technique for analyzing multidimensional data. This method essentially provides an orthonormal basis for representing the given data in a certain least squares optimal sense. The POD method may be applied to infinite dimensional data such as fluid flow patterns as well. Truncation of the optimal basis provides a way to find optimal lower dimensional approximations of the given data.

In addition to being optimal in a least squares sense, POD has the property that it uses a modal decomposition that is completely data dependent and does not assume any prior knowledge of the process that generates the data. This property is advantageous in situations where a priori knowledge of the underlying process is insufficient to warrant a certain choice of basis. It also helps in exploring patterns in data that may reveal some insight into the underlying process that generates it.

Combined with the Galerkin projection procedure, POD provides a powerful method for generating lower dimensional models of dynamical systems that have a very large or even infinite dimensional phase space. The fact that this approach always looks for linear (or affine) subspaces instead of curved submanifolds makes it computationally tractable. However, it must be noted that POD does not neglect the nonlinearities of the original vector-field. This is so because if the original dynamical system is nonlinear, then the resulting POD reduced order model will also typically be nonlinear.

These properties of POD are the reason for its wide application in data analysis, data compression, and model reduction in various fields of engineering and science.

[*] Received by the editors May 7, 2001; accepted for publication (in revised form) March 6, 2003; published electronically November 19, 2003. This research was supported in part by grants EPRI WO-8333-06, NSF/KDI ATM-9873133, and NSF ACI-0086061.

http://www.siam.org/journals/sinum/41-5/38904.html

[†] Mathematics and Statistics, University of Maryland Baltimore County, Baltimore, MD 21250 (muruhan@math.umbc.edu).

[‡] Computational Science and Engineering, University of California Santa Barbara, Santa Barbara, CA 93106 (petzold@engineering.ucsb.edu).

Applications of POD include image processing [22], data compression, signal analysis [2], modeling and control of chemical reaction systems [12, 25, 26], turbulence models [14], coherent structures in fluids [14], control of fluids [10], electrical power grids [21, 20, 18], and wind engineering to name a few.

Extensions and modifications to POD have been proposed by various researchers to accommodate properties of the applications at hand. For instance, instead of time averaging, arclength-based averaging has been found to be useful in capturing dynamics involving “intermittent” attractors in [12]. The predefined POD method has been studied in [8], where modes are selected not only on the basis of energy of the data but also on some prior knowledge of the system. Structure preserving model reduction based on POD for mechanical systems with Lagrangian structure has been developed in [15].

Systems with symmetry deserve special attention. Several authors have made important contributions. Expanding the data set using symmetry was proposed in [27, 28, 29], and later works have shown that it is an essential step in capturing the correct dynamics [3, 4]. Methods for combining reduction theory with POD have been developed in [23].

1.2. Contributions of this work.

In this paper we study some basic questions about POD. We focus on finite dimensional systems and follow a deterministic approach. The contributions of this paper include a study of the errors involved in solving an initial value problem using a POD reduced order model of a dynamical system, the sensitivity of the results of POD to perturbations in the data that is used to form the reduced model, as well as computational efficiency gained in using POD in model reduction applications. Even though these are some fundamental questions relating to POD, we believe that they have not been given sufficient attention in the literature.

1.3. Outline of the paper.

The rest of the paper is organized as follows. In section 2, we review the POD method as it is applied in data representation as well as in model reduction. In section 3, we present some mathematical preliminaries on the manifold of projection matrices and finite time solution norms of linear time invariant systems. The former is relevant in the sensitivity analysis, and the latter will be useful since throughout this paper we derive particular results for linear time invariant systems. In section 4, we provide an error analysis of the POD method of model reduction as applied to a general nonlinear system. An example is provided to illustrate the various factors affecting the errors. In section 5, we study the sensitivity of the POD projection matrix P (Proposition 5.4), the projected data $\tilde{y}$, and the reduced model solution $\hat{y}$ to perturbations in the data x that is used to form the reduced model. We also study the particular case $y = x$, where the particular data/solution y for which the reduced model is applied is the same as the ensemble of data x from which the reduced model is constructed. Two examples are provided to illustrate the sensitivity results, one focusing on the $y = x$ case. In section 6, we present an estimate of the computational complexity involved in integrating a system of ODEs with and without the use of POD reduced order models. We also provide two examples to illustrate the various factors affecting the computational savings. Finally, in section 7, we make concluding remarks.

2. Proper orthogonal decomposition (POD).

POD provides a method for finding the best approximating subspace to a given set of data. Originally POD was used as a data representation technique. For model reduction of dynamical systems, POD may be used on data points obtained from system trajectories obtained via experiments, numerical simulations, or analytical derivations. Additional information may be found in [14, 19, 17, 16].

2.1. POD in data representation.

We shall assume that the *data points* lie in $\mathbb{R}^n$. In the case of a dynamical system this is the phase space. A *data set* is a collection $x^\alpha \in \mathbb{R}^n$, where $\alpha \in \mathcal{I}$. The *index set* $\mathcal{I}$ may be a finite set $\{1, \ldots, N\}$, or a time interval $[0, T]$, or more generally of the form $\mathcal{I} = [0, T] \times \{1, \ldots, N\}$. The latter corresponds to a collection of trajectories. For example, in an image coding problem $\mathcal{I}$ is a finite discrete set. In model reduction of dynamical systems, $\mathcal{I}$ could be of the more general form above. We define an inner product between sets of data x_1 and x_2 with the same index set $\mathcal{I}$ in the obvious way. For example, if x_1 and x_2 are each a collection of N trajectories in the common interval $[0, T]$, i.e., $x_i^\alpha : [0, T] \to \mathbb{R}^n$ for $\alpha = 1, \ldots, N$ and $i = 1, 2$, then

$$(x_1, x_2) = \sum_{\alpha=1}^{N} \int_0^T (x_1^\alpha(t))^T x_2^\alpha(t) dt.$$

The corresponding norm is denoted $\|.\|$.

Remark 2.1. Note that we are using the inner product in our data space ($\mathbb{R}^n$) to induce an inner product in the space of data sets with the same index set.

We shall explain the POD method using the index set $\mathcal{I} = [0, T] \times \{1, \ldots, N\}$. Given a data set x, POD seeks a subspace $S \subset \mathbb{R}^n$ so that the total square distance

$$\|x - \rho_S x\|^2 = \sum_{\alpha=1}^{N} \int_0^T \|x^\alpha(t) - \rho_S x^\alpha(t)\|^2 dt$$

is minimized. Here ρ_S is the orthogonal projection onto the subspace S and $\rho_S x$ is the projected data set. The solution to this problem requires the construction of the *correlation matrix* defined by

$$R = \sum_{\alpha=1}^{N} \int_0^T x^\alpha(t)(x^\alpha(t))^T dt.$$

Note that R is symmetric positive semidefinite. Let $\lambda_1 \geq \lambda_2 \cdots \geq \lambda_N \geq 0$ be the ordered eigenvalues of R. Then the minimum value of $\|x - \rho_S x\|^2$ over all $k (\leq n)$ dimensional subspaces S is given by $\sum_{j=k+1}^{n} \lambda_j$ [14]. In addition the minimizing S is the invariant subspace corresponding to the eigenvalues $\lambda_1, \ldots, \lambda_k$.

Often it may be best to find an affine subspace as opposed to a linear subspace. This requires us first to find the mean value of the data points

$$\bar{x} = \frac{1}{NT} \sum_{\alpha=1}^{N} \int_0^T x^\alpha(t) dt$$

and then construct the *covariance matrix* $\bar{R}$ given by

$$\bar{R} = \sum_{\alpha=1}^{N} \int_0^T (x^\alpha(t) - \bar{x})(x^\alpha(t) - \bar{x})^T dt.$$

Let S_0 be the invariant subspace of the largest k eigenvalues of $\bar{R}$. Then the best approximating affine subspace S passes through $\bar{x}$ and is obtained by shifting S_0 by $\bar{x}$. Algebraically the projection onto the subspace S is given by

$$z = \rho(x - \bar{x}), \tag{2.1}$$

where $z \in \mathbb{R}^k$ are coordinates in the subspace S, $x \in \mathbb{R}^n$ are coordinates in the original coordinate system in $\mathbb{R}^n$, and the matrix ρ of the projection consists of row vectors ϕ_i^T $(i = 1, \ldots, k)$, where ϕ_i are the unit eigenvectors corresponding to the largest k eigenvalues of $\bar{R}$. Note that given any point $p \in S$ with coordinates $z \in \mathbb{R}^k$, the coordinates $x \in \mathbb{R}^n$ of the same point in the original coordinate system are given by

$$x = \rho^T z + \bar{x}.$$

The affine projection $\tilde{x} \subset S$ of a point $x \in \mathbb{R}^n$ in the original coordinates is given by

$$\tilde{x} = P(x - \bar{x}) + \bar{x},$$

where $P = \rho^T \rho \in \mathbb{R}^{n \times n}$ is the matrix of the (linear) projection expressed in the original coordinate system in $\mathbb{R}^n$.

Remark 2.2. Note that the reduced subspace is uniquely characterized by the pair $(\bar{x}, P)$. Different data sets may lead to the same pair $(\bar{x}, P)$, and the detailed information about the data x is lost.

2.2. POD in model reduction. The POD method may also be used in obtaining a lower dimensional model of a dynamical system. In this case, having found the approximating subspace for our system data, the next task is to construct a vector-field on this subspace that represents the reduced order model. The procedure we describe is known as Galerkin projection and has been widely used in reducing PDEs to ODEs by projecting onto appropriate basis functions that describe the spatial variations in the solution. The procedure is applicable to any subspace; the subspace need not be obtained from the POD method. See [14] for more details.

Suppose the original dynamical system in $\mathbb{R}^n$ is given by a vector-field f,

$$\dot{x} = f(x, t).$$

Let $S \subset \mathbb{R}^n$ be the best k dimensional approximating affine subspace with projection given by (2.1). A vector-field f_a in the subspace S is constructed by the following rule: for any point $p \in S$ compute the vector-field $f(p, t)$ and take the projection $\rho f(p, t)$ onto the subspace S to be the value of $f_a(p, t)$. If z are the subspace coordinates of p, then $f_a(z, t) = \rho f(\rho^T z + \bar{x}, t)$. Thus we obtain the following reduced model:

$$\dot{z} = f_a(z, t) = \rho f(\rho^T z + \bar{x}, t). \tag{2.2}$$

If we are solving an initial value problem with $x(0) = x_0$, then in the reduced model one has the initial condition $z(0) = z_0$, where

$$z_0 = \rho(x_0 - \bar{x}).$$

Hence the approximating solution $\hat{x}(t)$ in the original coordinates in $\mathbb{R}^n$ is given by

$$\hat{x}(t) = \rho^T z(t) + \bar{x}.$$

From the above it is easy to see that the approximating solution $\hat{x}(t)$ is the solution to the following initial value problem:

$$\dot{\hat{x}} = Pf(\hat{x}, t); \quad \hat{x}(0) = \hat{x}_0 = P(x_0 - \bar{x}) + \bar{x}. \tag{2.3}$$

Note that $\hat{x}_0$ is just the projection of x_0 onto the affine subspace S.

3. Mathematical preliminaries.

3.1. Manifold of projection matrices.

Let $\mathcal{P} \subset \mathbb{R}^{n\times n}$ be the manifold of all rank $k(< n)$ (orthogonal) projection matrices. ($\mathcal{P}$ is known in the literature as the Grassmannian [6].) For a general introduction to manifolds, tangent spaces, and differentiation on manifolds, see [1, 6]. Since we will be dealing with variations E of projections $P \in \mathcal{P}$ in our POD sensitivity analysis, we need a characterization of the tangent space $T_P\mathcal{P}$ to $\mathcal{P}$ at a given point $P \in \mathcal{P}$. The variation $E \in T_P\mathcal{P}$ cannot be any arbitrary matrix. In fact, the dimension of $\mathcal{P}$ and hence that of $T_P\mathcal{P}$ for any P is $k(n-k)$.

Without loss of generality, we can induce an orthonormal change of coordinates in $\mathbb{R}^n$ such that a given projection P becomes the canonical projection (P_0) in these coordinates, i.e.,

$$P_0 = \begin{bmatrix} I_{k\times k} & 0_{k\times n-k} \\ 0_{n-k\times k} & 0_{n-k\times n-k} \end{bmatrix}.$$

In many places in our analysis we shall assume the use of these canonical coordinates.

Let $V = T_{P_0}\mathcal{P}$, i.e., the tangent space to $\mathcal{P}$ at the canonical projection P_0. Using the relations $P^2 = P$ and $P^T = P$ (symmetric) and letting $P = P_0$, it is easy to see that V consists of matrices of the form

$$\begin{bmatrix} 0_{k\times k} & X_{k\times n-k} \\ X^T_{n-k\times k} & 0_{n-k\times n-k} \end{bmatrix},$$

where X is arbitrary. We will use the Frobenius norm for projection matrices P and their variations E in our analysis. We consider the basis $\{E^{ij} \ : \ i = 1, \dots, k; j = 1, \dots, n-k\}$ for V, where

$$E^{ij} = \begin{bmatrix} 0 & X_{ij} \\ X^T_{ij} & 0 \end{bmatrix}$$

and X_{ij} is the $k \times (n-k)$ matrix with all zeros except for a 1 in the (i, j)th element. Clearly $\|E^{ij}\| = \sqrt{2}$.

Remark 3.1. It may be noted that E^{ij} corresponds to an infinitesimal rotation of the subspace S (onto which P_0 projects) in the plane of the coordinates x_i and x_{j+k}. Consider the family of subspaces $S(\theta)$ which are spanned by

$$\{e_1, \dots, e_{i-1}, \cos\theta e_i + \sin\theta e_{j+k}, e_{i+1}, \dots, e_k\},$$

where $e_1, \dots, e_n$ are the canonical basis vectors in $\mathbb{R}^n$. Note that when $\theta = 0$, S corresponds to the image of P_0. Computing the corresponding family of projection matrices $P(\theta)$, we can see that $E^{ij} = \frac{dP}{d\theta}(\theta = 0)$.

3.2. Finite time response of a linear time invariant system with time varying input.

Some of the analysis in this paper requires estimating the norm of the trajectory of a linear time invariant system in a finite interval in response to a forcing input term.

Consider the system

$$\dot{x} = Ax + u$$

(where $x \in \mathbb{R}^n$) with input $u(t) \in \mathbb{R}^n$ and initial condition $x(0) = x_0$ in the interval $[0, T]$. The solution is

$$x(t) = \int_0^t e^{A(t-\tau)} u(\tau) d\tau + e^{At} x_0.$$

This may be written in the form

$$x = F(T; A)u + G(T; A)x_0, \tag{3.1}$$

where $F(T; A) : \mathcal{L}_2([0, T], \mathbb{R}^n) \to \mathcal{L}_2([0, T], \mathbb{R}^n)$ and $G(T; A) : \mathbb{R}^n \to \mathcal{L}_2([0, T], \mathbb{R}^n)$ are linear operators. It is in general very difficult to obtain sharp estimates for the norms of $F(T; A)$ and $G(T; A)$, and in fact this basically reduces to the problem of estimating the norm of the matrix exponential. As such we shall not provide an estimate, but we remark that these norms grow exponentially with T at a rate that is determined by the largest real part of any eigenvalue of A and in addition depend on the nonnormality of A. See [11] for an estimate of matrix exponential. In our analysis we shall estimate $\|x\|$ as

$$\|x\| \leq \|F(T; A)\| \|u\| + \|G(T; A)\| \|x_0\|, \tag{3.2}$$

expressing the results in terms of $\|F(T; A)\|$ and $\|G(T; A)\|$.

4. Error analysis of the POD method of model reduction. Consider solving the initial value problem $\dot{x} = f(x, t)$, $x(0) = x_0$, using a POD reduced order model in the interval $[0, T]$. Then in effect we are solving the initial value problem (2.3). We shall derive an estimate for the error $e(t) = \hat{x}(t) - x(t)$. Denote the component of $e(t)$ orthogonal to the subspace S by $e_o(t)$ and the component parallel to S by $e_i(t)$. Thus $e_o(t)$ and $e_i(t)$ are orthogonal vectors. Hence by definition $Pe_o(t) = 0$ and $Pe_i(t) = e_i(t)$. It is important to observe that $e_o(t)$ comes from the first part of the method, i.e., the subspace approximation. It is the error between $x(t)$ and its projection onto the subspace S. If one is considering a data compression problem, then $e_o(t) = e(t)$. But since we form a reduced order model by projecting the vector-field onto S, we make further approximations resulting in the additional error $e_i(t)$.

Remark 4.1. Note that for any function $g : [0, T] \to \mathbb{R}^n$, $\|g(t)\|$ is a norm in $\mathbb{R}^n$ which shall be the 2-norm throughout this paper. The function norm will be denoted by $\|g\|$, and unless explicitly stated otherwise it will be assumed to be the 2-norm.

We can derive an error estimate for $e_i(t)$ in terms of $e_o(t)$. Differentiating $e_o(t) + e_i(t) = \hat{x}(t) - x(t)$ and substituting into the ODEs for $\hat{x}$ and x, we get

$$\dot{e}_o + \dot{e}_i = Pf(\hat{x}, t) - f(x, t).$$

Multiplying on the left by P and using $P^2 = P$, we obtain the initial value problem for $e_i(t)$:

$$\dot{e}_i = P(f(x(t) + e_o(t) + e_i, t) - f(x(t), t)); \quad e_i(0) = 0. \tag{4.1}$$

Note that $e_i(0) = 0$ since the starting point $\hat{x}_0$ is the projection of x_0 onto S. Thus the error e_i is governed by (4.1), where we may regard $x(t)$ and $e_o(t)$ as forcing terms. See Figure 4.1, where x is the true solution, $\tilde{x}$ the projected solution, and $\hat{x}$ the solution of the reduced model. The errors e_i and e_o are also shown.

In the case of a linear time invariant system $\dot{x} = Ax$, (4.1) takes a simple form:

$$\dot{e}_i = PAe_i + PAe_o(t); \quad e_i(0) = 0. \tag{4.2}$$

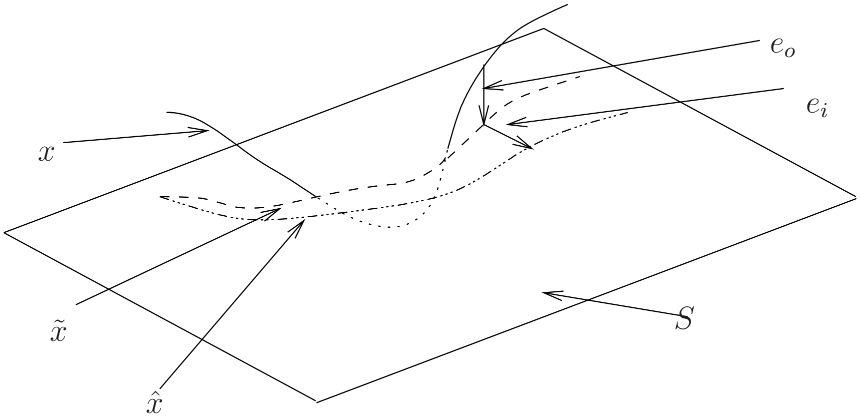


FIG. 4.1. *POD error.*

Applying the notation of (3.1), we get the estimate

$$\|e_i\|_2 \leq \|F(T;\hat{A})\|\|\tilde{A}\|\epsilon.$$

Hence the total error is

$$\|e\|_2 \leq \left(\|F(T;\hat{A})\|\|\tilde{A}\| + 1\right)\epsilon. \tag{4.3}$$

Here $\hat{A} = \rho A \rho^T$ and $\tilde{A} = \rho A \rho_c^T$, where ρ is the projection in subspace coordinates ($P = \rho^T \rho$), and ρ_c is the orthogonal complement to ρ. ϵ is the 2-norm of e_o (i.e., $\|e_o\|_2 = \epsilon$).

Before we state a proposition for the general nonlinear case, recall the definition of a *logarithmic norm related to a* 2*-norm* of a square matrix $A \in \mathbb{R}^{k\times k}$ denoted by $\mu(A)$:

$$\mu(A) = \lim_{h\to 0, h>0} \frac{\|I + hA\|_2 - 1}{h},$$

where I is the identity matrix [13].

PROPOSITION 4.2. *Consider solving the initial value problem $\dot{x} = f(x,t)$, $x(0) = x_0$, using the POD reduced order model in the interval $[0,T]$. Let $\rho \in \mathbb{R}^{k\times n}$ be the relevant projection matrix, and let S denote the affine subspace onto which POD projects. Write the solution (of the full model) $x(t)$ and the solution $\hat{x}(t)$ of the reduced model as*

$$x(t) = \rho^T u(t) + \rho_c^T v(t) + \bar{x}$$

and

$$\hat{x}(t) = \rho^T u(t) + \rho^T w(t) + \bar{x}$$

so that the errors $e_o(t)$ and $e_i(t)$ and the projected solution $\tilde{x}(t)$ are given by

$$e_o(t) = -\rho_c^T v(t),$$

$$e_i(t) = \rho^T w(t),$$

and

$$\tilde{x}(t) = \rho^T u(t) + \bar{x}.$$

Note that $u(t) \in \mathbb{R}^k$, $w(t) \in \mathbb{R}^k$, and $v(t) \in \mathbb{R}^{n-k}$. Let $\gamma \geq 0$ be the Lipschitz constant of $\rho f(x,t)$ in the directions orthogonal to S in a region containing $x(t)$ and $\tilde{x}(t)$. To be precise, suppose

$$\|\rho f(\tilde{x}(t) + \rho_c^T v, t) - \rho f(\tilde{x}(t), t)\| \leq \gamma \|v\|$$

for all $(v,t) \in D \subset \mathbb{R}^{n-k} \times [0,T]$, where the region D is such that the associated region $\tilde{D} = \{(\tilde{x}(t) + \rho_c^T v, t) \ : \ (v,t) \in D\} \subset \mathbb{R}^n \times [0,T]$ contains $(\tilde{x}(t), t)$ and $(x(t), t)$ for all $t \in [0,T]$. Let $\mu(\rho \frac{\partial f}{\partial x}(\bar{x} + \rho^T z, t)\rho^T) \leq \bar{\mu}$ for $(z,t) \in V \subset \mathbb{R}^k \times [0,T]$, where the region V is such that it contains $(u(t), t)$ and $(u(t) + w(t), t)$ for all $t \in [0,T]$ and μ denotes the logarithmic norm related to the 2-norm. Let $\epsilon = \|e_o\|_2$. Then the error e_i in the ∞-norm satisfies

$$\|e_i\|_\infty \leq \epsilon \frac{\gamma}{\sqrt{2\bar{\mu}}} \sqrt{e^{2\bar{\mu}T} - 1}, \tag{4.4}$$

and the 2-norm of the total error satisfies

$$\|e\|_2 \leq \epsilon \sqrt{1 + \frac{\gamma^2}{4\bar{\mu}^2}(e^{2\bar{\mu}T} - 1 - 2\bar{\mu}T)}. \tag{4.5}$$

Proof. We shall closely follow the ideas in [13, pp. 54–60]. Since

$$\dot{w}(t) = \rho f(\bar{x} + \rho^T u(t) + \rho^T w(t), t) - \rho f(\bar{x} + \rho^T u(t) + \rho_c^T v(t), t),$$

for $h > 0$ using Taylor expansion we have

$$\begin{aligned} \|w(t+h)\| &= \|w(t) + h\rho f(\bar{x} + \rho^T u(t) + \rho^T w(t), t) - h\rho f(\bar{x} + \rho^T u(t) + \rho_c^T v(t), t)\| \\ &\quad + O(h^2) \\ &\leq \|w(t) + h\rho f(\bar{x} + \rho^T u(t) + \rho^T w(t), t) - h\rho f(\bar{x} + \rho^T u(t), t)\| \\ &\quad + h\|\rho f(\bar{x} + \rho^T u(t) + \rho_c^T v(t), t) - \rho f(\bar{x} + \rho^T u(t), t)\| + O(h^2). \end{aligned}$$

Applying the mean value theorem to $\eta \mapsto \eta + h\rho f(\bar{x} + \rho^T \eta, t)$, we get

$$\begin{aligned} &\|w(t) + h\rho f(\bar{x} + \rho^T u(t) + \rho^T w(t), t) - h\rho f(\bar{x} + \rho^T u(t), t)\| \\ &\qquad \leq \left(\max_{\eta \in [u(t), u(t)+w(t)]} \left\| I + h\rho \frac{\partial f}{\partial x}(\bar{x} + \rho^T \eta, t)\rho^T \right\| \right) \|w(t)\|, \end{aligned}$$

where for any two vectors η_1, η_2 in $\mathbb{R}^k$, $[\eta_1, \eta_2]$ denotes the line segment joining the two. It follows that

$$\frac{\|w(t+h)\| - \|w(t)\|}{h} \leq \bar{\mu}\|w(t)\| + \gamma\|v(t)\| + O(h),$$

where the $O(h)$ term may be uniformly bounded independent of $w(t)$ [13]. Then it follows from Theorem 10.3 of [13] (also see Theorem 10.6 in [13]) that

$$\|e_i(t)\| = \|w(t)\| \leq \gamma \int_0^t e^{\bar{\mu}(t-\tau)} \|v(t)\| d\tau,$$

since $e_i(t) = \rho^T w(t)$. After applying the Cauchy–Schwarz inequality on the right side, we get

$$\|e_i(t)\| \leq \frac{\gamma}{\sqrt{2\bar{\mu}}} \sqrt{e^{2\bar{\mu}t} - 1} \sqrt{\int_0^t \|e_o(\tau)\|^2 d\tau}. \tag{4.6}$$

From this we readily obtain (4.4). Also bounding $\sqrt{\int_0^t \|e_o(\tau)\|^2 d\tau}$ by ϵ and integrating (4.6), we obtain an upper bound for $\|e_i\|_2$ which can be combined with $\|e_o\|_2$ to get (4.5). □

Remark 4.3. This analysis separates the two different errors and provides a bound for the total error in terms of the projection error ϵ of the true solution $x(t)$. The value of ϵ depends only on the true solution $x(t)$ and on the pair $(P, \bar{x})$ which determines the reduced order model (but not directly on f). If P and $\bar{x}$ were computed from the true solution $x(t)$ in the interval $[0, T]$ (this is somewhat an ideal situation), then $\epsilon = \sqrt{\sum_{j=k+1}^n \lambda_j}$, where λ_i are the eigenvalues of the covariance matrix. However, if the reduced model was computed from some other trajectories as often is the case in applications of model reduction methods, then ϵ would depend on how close $x(t)$ was to the trajectories used as data in addition to the quantity $\sum_{j=k+1}^n \lambda_j$ (typically the fractional error $\frac{\epsilon}{\|x\|_2}$ will be larger than $\frac{\sum_{j=k+1}^n \lambda_j}{\sum_{j=1}^n \lambda_j}$). For instance, in hybrid systems such as power systems where discrete events abruptly change some system parameters, data obtained from trajectories before the event results in a reduced order model with a large ϵ for simulations after the event [7].

Example 1. This example serves to illustrate the various factors that affect e_i given the same projection error e_o. We shall consider a linear time invariant system $\dot{x} = Ax$; $x(0) = x_0$. Assume A has distinct eigenvalues and that it possesses some fast decaying modes (eigenvalues with large negative real parts). Let $S \subset \mathbb{R}^n$ be the invariant subspace corresponding to the rest of the eigenvalues, where S is $k(< n)$ dimensional. If we have sufficiently many trajectories that have initial conditions symmetrically placed with respect to S, then the POD method will pick S as the subspace to project onto. We shall assume this to be the case. Performing an orthonormal change of coordinates if needed, we may assume that S corresponds to the last $n - k$ coordinates being zero. In these coordinates, the A matrix has the form

$$A = \begin{bmatrix} A_1 & A_{12} \\ 0 & A_2 \end{bmatrix},$$

where $A_1 \in \mathbb{R}^{k \times k}$, $A_{12} \in \mathbb{R}^{k \times (n-k)}$, and $A_2 \in \mathbb{R}^{(n-k)\times(n-k)}$. In fact the real Schur decomposition of A will put it in the above form. We shall say A is "block normal" if the off diagonal block $A_{12} = 0$.

Also note that

$$\rho = \begin{bmatrix} I_{k\times k} & 0_{k\times(n-k)} \end{bmatrix}$$

and

$$\rho_c = \begin{bmatrix} 0_{(n-k)\times k} & I_{(n-k)\times(n-k)} \end{bmatrix}.$$

Hence $\bar{\mu} = \mu(A_1)$ and $\gamma = \|A_{12}\|$.

For a given initial condition and time interval, the error e_o relates to the last three components of the solution and does not change if A_2 is unchanged. We can independently change $\bar{\mu}$ and γ by changing A_1 and A_{12}, respectively. We kept A_2 unchanged, thus keeping $\epsilon = \|e_o\|_2$ unchanged, and studied the effects of changing A_1 (and hence $\bar{\mu}$) and A_{12} (and hence γ) independently on the POD error. First we chose

$$A_1 = \begin{bmatrix} -0.1000 & 0 & 0 \\ 0 & -0.1732 & 2.0 \\ 0 & -2.0 & -0.1732 \end{bmatrix},$$

$$A_{12} = \begin{bmatrix} 0.3893 & 0.5179 & -1.543 \\ 1.390 & 1.300 & 0.8841 \\ 0.06293 & -0.9078 & -1.184 \end{bmatrix},$$

and

$$A_2 = \begin{bmatrix} -1.0 & 0 & 0 \\ 0 & -1.226 & -0.7080 \\ 0 & 0.7080 & -1.226 \end{bmatrix}.$$

Note that the eigenvalues of A_2 have large negative real parts compared to the eigenvalues of A_1, and the eigenvalues of A are the union of these two sets. The corresponding $\bar{\mu} = -1$ and $\gamma = 2.4419$. We randomly chose an initial condition and computed $x(t)$, $\tilde{x}(t)$, and $\hat{x}(t)$ in the interval $[0, 5]$. Note that the reduced model has dimension 3 and that the last three components of both $\tilde{x}(t)$ and $\hat{x}(t)$ are zero. Similarly, the first three components of $e_i(t)$ and $e_o(t)$ are zero. See Figure 4.2, where only the nonzero components are plotted. The computed value of the projection error was $\epsilon = \|e_o\|_2 = 1.4575$. The sup-norm and the 2-norm of the error in the subspace S were also computed and found to be $\|e_i\|_\infty = 1.5589$ and $\|e_i\|_2 = 2.5733$. The bounds provided by the theory were $\|e_i\|_\infty \leq 6.3271$ and $\|e_i\|_2 \leq 10.7930$.

The second choice was to keep A_1 and A_2 the same but scale A_{12} down by a factor of 2. We kept the same initial condition and time interval. This results in the same $\bar{\mu}$, but $\gamma = 1.2209$. In fact, according to (4.2), the effect of scaling A_{12} affects the error e_i linearly, and we expect $e_i(t)$ to be scaled down by the same factor of 2. Figure 4.3 shows a plot of $e_i(t)$ for both cases. This highlights how the rate of change of S components of the vector-field in the directions orthogonal to S affect the error. In the extreme case when $A_{12} = 0$ (i.e., A is "block normal"), the components of the vector-field parallel to S are invariant in the directions perpendicular to S, and the error e_i is zero. Thus the error e_i is zero for a matrix that is "block normal" (with respect to a decomposition of the space based on "fast decay" and the rest of the eigenmodes) if the POD indeed captures the attracting subspace S correctly.

The error e_i is more influenced by $\bar{\mu}$ than γ (as long as $\gamma > 0$), and $\bar{\mu}$ is supposed to capture the growth or decay of solutions of the vector-field of the reduced model. Keeping A_2 and A_{12} the same, we changed A_1 so that

$$A_1 = \begin{bmatrix} -0.1000 & 0 & 1.0 \\ 0 & -0.1732 & 2.0 \\ 0 & -2.0 & -0.1732 \end{bmatrix}.$$

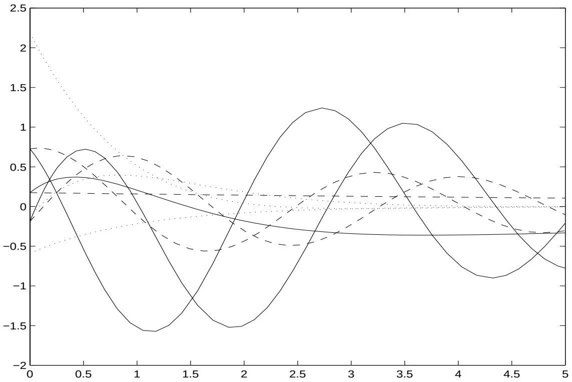


FIG. 4.2. *Example* 1 *on POD error. Solid: Projected solution $\tilde{x}(t)$. Dashed: Reduced model solution $\hat{x}(t)$. Dotted: Projection error $e_o(t)$. Only the three nonzero components $\tilde{x}_1, \tilde{x}_2, \tilde{x}_3$, $\hat{x}_1, \hat{x}_2, \hat{x}_3$ and e_{o4}, e_{o5}, e_{o6} are plotted.*

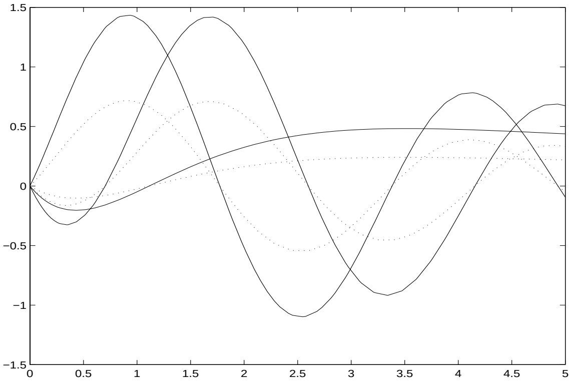


FIG. 4.3. *Example* 1 *on POD error. The effect of scaling A_{12} on the error e_i in the subspace S. Solid: e_i for unscaled A_{12}. Dotted: e_i for scaled down A_{12}. Only the three nonzero components are plotted.*

The corresponding $\bar{\mu} = 0.3647$. Note that this does not change the eigenvalues of A_1, but it does change its normality. This choice of A_1 is no longer normal, and even though the eigenvalues remain the same, the short term behavior of $e_i(t)$ is changed. In fact, $e_i(t)$ does not decay as much as in the normal case. This results in $\|e_i\|_\infty = 2.2088$ and $\|e_i\|_2 = 3.4565$. The bounds provided by the theory are $\|e_i\|_\infty \leq 25.4739$ and $\|e_i\|_2 \leq 28.3330$.

5. Sensitivity of POD to perturbations in data. Given a data set x, POD constructs a projection $P(x)$ onto a subspace which may then be used to approximate some other data set y. If POD is applied to model reduction to compute the approximation $\hat{y}$ to the true solution y of some ODE initial value problem, then the projection $P(x)$ will influence $\hat{y}$. Typically in POD applications the data set x comes from experimental measurement or numerical computations. Hence the data x has some error associated with it. Therefore, it is important to study the effect of these errors on the outcome of the POD model reduction procedure. In this section, we shall theoretically investigate the effect of infinitesimal perturbations of x on $P(x)$, $\tilde{y}$, and $\hat{y}$. We also look at the special case when $y = x$.

5.1. POD sensitivity factor. Let x be a data set, and let $P(x)$ be the corresponding POD projection. In this section, we analyze the sensitivities of the POD projection matrix $P(x)$, with respect to variations in the data x. Our analysis applies to any data set x taking values in $\mathbb{R}^n$, but for simplicity of exposition we assume x to be a single trajectory ($x : [0, T] \to \mathbb{R}^n$) whenever we need to be concrete.

Shifting the origin in data space if necessary, we may assume the mean data values $\bar{x} = 0$. In addition, we can find an orthonormal change of coordinates such that the covariance matrix of x is diagonal. We shall call this a *canonical coordinate system* for data set x. Assuming the use of these canonical coordinates, let

$$x = \sum_{\alpha=1}^{n} x_\alpha e_\alpha, \tag{5.1}$$

where e_α are the standard basis vectors in $\mathbb{R}^n$. Then it can be shown that the scalar data sets x_α are orthogonal. More specifically,

$$(x_\alpha, x_\beta) = \lambda_\alpha \delta_{\alpha,\beta}.$$

Here $\lambda_1, \lambda_2, \dots, \lambda_n$ are the eigenvalues of the covariance matrix of x. If, in addition, we permute the coordinates such that $\lambda_1 \geq \lambda_2 \cdots \geq \lambda_n \geq 0$ are the ordered eigenvalues, then we call this an *ordered canonical coordinate system.* Throughout the rest of the analysis, we shall assume that, after ordering, $\lambda_k > \lambda_{k+1}$ unless stated otherwise.

The POD projection matrix $P \in \mathcal{P} \subset \mathbb{R}^{n \times n}$ is defined as the minimizer of the function

$$e(P, x) = (Px - x, Px - x).$$

Differentiating with respect to P in the direction of E, we obtain

$$\frac{\partial e}{\partial P}(E) = 2(Px - x, Ex).$$

Thus stationary points P of e are given by the condition

$$(Px - x, Ex) = 0, \quad E \in T_P\mathcal{P}.$$

Performing an orthonormal change of coordinates if necessary, we may assume $P = P_0$ (canonical projection) is a stationary point. Then all variations $E \in V = T_{P_0}\mathcal{P}$. Requiring $(P_0 x - x, E^{ij} x) = 0$ for all E^{ij} gives us the conditions that $(x_i, x_{j+k}) = 0$ for all $1 \leq i \leq k$ and $1 \leq j \leq n - k$. This shows that all the stationary points of e are given by P that project onto any of the k dimensional invariant subspaces of the covariance matrix of x. A solution P to the above equation will be a strong local minimum if and only if the second derivative $\frac{\partial^2 e}{\partial P^2}$ is positive definite and this may be shown to be equivalent to $\lambda_k > \lambda_{k+1}$. Under this assumption, P is also a well-defined function of x locally.

Lemma 5.1. *Without loss of generality, let $P = P_0$ be a stationary point in some canonical coordinate system (this may not be ordered). Let $E \in V$ and $\tilde{E} \in V$ be given by*

$$E = \begin{bmatrix} 0_{k \times k} & X_{k \times n-k} \\ X^T_{n-k \times k} & 0_{n-k \times n-k} \end{bmatrix}$$

and

$$\tilde{E} = \begin{bmatrix} 0_{k\times k} & \tilde{X}^T_{k\times n-k} \\ \tilde{X}_{n-k\times k} & 0_{n-k\times n-k} \end{bmatrix}.$$

Then the Hessian satisfies

$$\frac{\partial^2 e}{\partial P^2}(E)(\tilde{E}) = 2(X^T x_1, \tilde{X}^T x_1) - 2(X x_2, \tilde{X} x_2). \tag{5.2}$$

Proof. Since e is a function on the manifold $\mathcal{P}$, one could introduce local coordinates on $\mathcal{P}$ to compute the Hessian of e at a stationary point. However, we shall use a coordinate independent method which allows us to work with matrices and keep the algebra simple. It may be shown that if $P(t,s)$ is a smooth mapping from $\mathbb{R}^2$ into $\mathcal{P}$ such that $P(0,0) = P_0$, $P_t(0,0) = E \in V$ and $P_s(0,0) = \tilde{E} \in V$ and if $\frac{\partial e}{\partial P}(P_0) = 0$, then the Hessian at $P = P_0$ is given by

$$\frac{\partial^2 e}{\partial P^2}(E)(\tilde{E}) = e_{ts}(0,0),$$

where P and e are regarded as functions of t and s and subscripts denote partial derivatives.

Differentiating $e = 2(Px - x, Px - x)$ with respect to t, we get

$$e_t = 2(P_t x, Px - x),$$

and differentiating again with respect to s, we get

$$e_{ts} = 2(P_{ts} x, Px - x) + 2(P_t x, P_s x). \tag{5.3}$$

Suppose

$$P_{ts}(0,0) = \begin{bmatrix} W_1 & W_2 \\ W_3 & W_4 \end{bmatrix}.$$

The matrices W_1, W_2, W_3, and W_4 are not arbitrary but satisfy some relations. These are obtained by differentiating the relation $P^2 = P$ twice. In fact, we get

$$P_{ts}P + P_t P_s + P_s P_t + P P_{ts} = P_{ts},$$

and after substituting expressions for P, P_t, P_s, and P_{ts} (at $(t,s) = (0,0)$) in the above and using the fact that P_{ts} is symmetric, we obtain that $W_1 = -\tilde{X}X^T - X\tilde{X}^T, W_4 = \tilde{X}^T X + X^T \tilde{X}, W_3^T = W_2$, where W_2 is an arbitrary $k \times (n-k)$ matrix. It then follows that $P_{ts}(0,0) = F + W$, where

$$F = \begin{bmatrix} -\tilde{X}X^T - X\tilde{X}^T & 0 \\ 0 & \tilde{X}^T X + X^T \tilde{X} \end{bmatrix}$$

and

$$W = \begin{bmatrix} 0 & W_2 \\ W_2^T & 0 \end{bmatrix},$$

and hence $W \in V$. Hence from (5.3) we obtain

$$e_{ts}(0,0) = 2(Fx, P_0 x - x) + 2(Wx, P_0 x - x) + 2(Ex, \tilde{E}x).$$

Since $\frac{\partial e}{\partial P} = 0$ at $P(0,0) = P_0$ by assumption and $W \in V$, it follows that $(Wx, P_0 x - x) = 0$. Let $x = (x_1, x_2)$, where $x_1(t) \in \mathbb{R}^k$ and $x_2(t) \in \mathbb{R}^{n-k}$. It is easy to see that

$$(Ex, \tilde{E}x) = (Xx_2, \tilde{X}x_2) + (X^T x_1, \tilde{X}^T x_1),$$

where the inner products on the right-hand side are in the appropriate function spaces. Since $(Fx, P_0 x - x) = ((P_0 - 1)Fx, x)$, after computing $(P_0 - 1)F$ it can be shown that

$$(Fx, P_0 x - x) = -(\tilde{X}^T X x_2 + X^T \tilde{X} x_2, x_2) = -2(Xx_2, \tilde{X}x_2).$$

Equation (5.2) follows from this. □

Remark 5.2. From (5.2) it may be shown that the Hessian $\frac{\partial^2 e}{\partial P^2}$ has E^{ij} as its eigenvectors with corresponding eigenvalues $2(\lambda_i - \lambda_{j+k})$ for $1 \le i \le k$ and $1 \le j \le n-k$ (note that we did not order the eigenvalues). Thus the stationary point $P = P_0$ is a strong minimum (maximum) if and only if the first k eigenvalues are strictly greater (smaller) than the rest. It is clear that if, after ordering, $\lambda_k > \lambda_{k+1}$, then there is a unique strong minimum and a unique strong maximum. The rest of the stationary points are saddle points.

The sensitivity of P to variations in data x is given by $\frac{dP}{dx}(\delta x)$, the directional derivative of P with respect to x in the direction δx, where $\delta x : [0,T] \to \mathbb{R}^n$ is assumed to be a unit-norm variation of x ($\|\delta x\| = 1$). It suffices to consider zero mean variations. This is because one may decompose any variation $\delta x \in \mathcal{L}_2([0,T] \to \mathbb{R}^n)$ into a constant function plus a zero mean function, and it is easy to see that the constant function part affects only the mean value $\bar{x}$ of the data while the zero mean function part affects only the projection P.

Remark 5.3. Variations of variables are denoted by prefix δ except for variations of P, which are denoted by E (or $\tilde{E}$, etc.). We will use the 2-norm for functions and the Frobenius norm for matrices P and E.

The norm $\|\frac{dP}{dx}\|$ is defined by

$$\left\| \frac{dP}{dx} \right\| = \sup_{\|\delta x\|=1} \left\| \frac{dP}{dx}(\delta x) \right\|$$

and measures the worst-case sensitivity of P to unit-norm variations of x. However, it makes more sense to consider the nondimensional quantity defined by

$$S_k(x) = \left\| \frac{dP}{dx} \right\| \|x - \bar{x}\|, \tag{5.4}$$

which we shall call the *POD sensitivity factor*. It is the worst-case ratio (in the limit of zero perturbation) of the perturbation of P to the fractional perturbation $\frac{\delta(x-\bar{x})}{\|x-\bar{x}\|}$. We use $x - \bar{x}$ instead of x because P depends only on $x - \bar{x}$. If we scale the data set x by a constant $c \in \mathbb{R}$, then both $\bar{x}$ and $x - \bar{x}$ also scale by c, but P remains unchanged ($P(cx) = P(x)$). The definition of S_k takes care of this scaling symmetry. In fact, we get $S_k(x) = S_k(cx)$. Note that the suffix k stands for the dimension of the reduced subspace $S \subset \mathbb{R}^n$ in which the projected data lives.

Proposition 5.4. *Consider applying POD to a data set x to find the best approximating $k(< n)$ dimensional subspace. Let the ordered eigenvalues of the covariance matrix of the data x be given by $\lambda_1 \ge \cdots \ge \lambda_n$. Suppose $\lambda_k > \lambda_{k+1}$, which ensures* *that $P(x)$ is well defined. Then*

$$S_k(x) = \max_{i\leq k,\, j\leq n-k} \sqrt{2}\frac{\sqrt{\lambda_i+\lambda_{j+k}}}{\lambda_i-\lambda_{j+k}}\sqrt{\lambda_1+\cdots+\lambda_n} \geq \sqrt{2}. \tag{5.5}$$

Furthermore, in the ordered canonical coordinates corresponding to data x, the unit-norm variation δx that causes the maximal variation in P is given by

$$\delta x = \frac{E^{\tilde{i},\tilde{j}}(x-\bar{x})}{\sqrt{\lambda_{\tilde{i}}+\lambda_{\tilde{j}+k}}}, \tag{5.6}$$

where $i=\tilde{i}$ and $j=\tilde{j}$ maximize the right-hand side of (5.5).

Proof. In this proof we will use ordered canonical coordinates. Differentiating $\frac{\partial e}{\partial P}(E)=0$ totally with respect to x in the δx direction, we get

$$\frac{\partial^2 e}{\partial P^2}(E)\left(\frac{dP}{dx}(\delta x)\right) + \frac{\partial^2 e}{\partial P \partial x}(E)(\delta x) = 0 \quad \forall E \in V. \tag{5.7}$$

Hence $\frac{dP}{dx}(\delta x)$ is implicitly defined through the above equation.

The mixed partial $\frac{\partial^2 e}{\partial P\partial x}$ is given by

$$\begin{aligned}\frac{\partial^2 e}{\partial P\partial x}(E)(\delta x) &= 2(Ex,(P-1)\delta x) + 2(E\delta x,(P-1)x)\\ &= 2((PE-E)x,\delta x) + 2((EP-E)x,\delta x)\\ &= -2(Ex,\delta x),\end{aligned} \tag{5.8}$$

where in the first step we used the fact that $E^T=E$ and $P^T=P$, and in the second step we used the fact that $PE+EP=E$ (which comes from $P^2=P$). Note that if $\bar{x}\neq 0$, then $e=(P(x-\bar{x})-(x-\bar{x}),P(x-\bar{x})-(x-\bar{x}))$. Even though we assumed without loss of generality that $\bar{x}=0$, when we take variations of x we need to consider the corresponding variations of $\bar{x}$. However, since we care only about zero mean variations δx, for those the corresponding variation $\delta\bar{x}=0$. Hence we are justified in neglecting the term $\bar{x}$.

It is instructive to examine the finite dimensional space U of $\mathbb{R}^n$-valued functions defined by

$$U = \operatorname{span}\{E'(x-\bar{x}) \ : \ E' \in V\}. \tag{5.9}$$

From (5.8) (note that we assumed $\bar{x}=0$) it can be seen that if δx is orthogonal to U, then $\frac{\partial^2 e}{\partial P\partial x}=0$, and hence by (5.7) $\frac{dP}{dx}(\delta x)=0$. Since we are only interested in variations δx that introduce nonzero variations in P, we shall assume $\delta x\in U$. It can be shown that the map $E'\in V\to E'(x-\bar{x})\in U$ is an isomorphism. This is readily seen by evaluating this map on the basis E^{ij} and showing that $E^{ij}(x-\bar{x})=E^{ij}x$ form an independent set. In fact,

$$E^{ij}x = x_i e_{j+k} + x_{j+k}e_i,$$

and hence

$$\begin{aligned}(E^{ij}x,E^{lm}x) &= \lambda_i+\lambda_{j+k}, \quad l=i, j=m,\\ &= 0 \quad \text{otherwise.}\end{aligned} \tag{5.10}$$

Hence $\{E^{ij}x \ : \ i = 1, \dots, k; j = 1, \dots, n-k\}$ is an orthogonal set (and is clearly independent as well).

Substitute (5.2) and (5.8) into the implicit equation (5.7) for $\frac{dP}{dx}$, and let

$$\frac{dP}{dx}(\delta x) = \begin{bmatrix} 0 & \tilde{X} \\ \tilde{X}^T & 0 \end{bmatrix}$$

and $\delta x = (\delta x_1, \delta x_2)$, where $\delta x_1(t) \in \mathbb{R}^k$ and $\delta x_2(t) \in \mathbb{R}^{n-k}$. Then we obtain an equation for $\tilde{X}$:

$$(5.11)\quad (X^T x_1, \tilde{X}^T x_1) - (X x_2, \tilde{X} x_2) = (X x_2, \delta x_1) + (X^T x_1, \delta x_2) \quad \forall X \in \mathbb{R}^{k \times (n-k)}.$$

Let $\delta x = E' x \in U$, where

$$E' = \sum_{l,m} \alpha_{lm} \frac{E^{lm}}{\sqrt{\lambda_l + \lambda_{m+k}}},$$

and let $\tilde{X} = \sum_{l,m} \beta_{lm} X_{lm}$. Substituting these into (5.11) for $X = X_{ij}$, we get

$$\beta_{ij} = \alpha_{ij} \frac{\sqrt{\lambda_i + \lambda_{j+k}}}{\lambda_i - \lambda_{j+k}}.$$

Hence it follows that

$$(5.12)\qquad \frac{dP}{dx}(\delta x) = \sum_{i,j} \alpha_{ij} \frac{\sqrt{\lambda_i + \lambda_{j+k}}}{\lambda_i - \lambda_{j+k}} E^{ij}.$$

The requirement that $\|\delta x\| = 1$ is equivalent to $\sum_{i,j} \alpha_{ij}^2 = 1$. Since $\|E^{ij}\| = \sqrt{2}$, it follows that

$$(5.13)\qquad \left\| \frac{dP}{dx} \right\| = \max_{i \le k,\ j \le n-k} \sqrt{2} \frac{\sqrt{\lambda_i + \lambda_{j+k}}}{\lambda_i - \lambda_{j+k}},$$

with the maximizing unit-norm variation δx given by (5.6). (Note that we need to replace x by $x - \bar{x}$, since we assumed for simplicity that $\bar{x} = 0$.) The equation in (5.5) follows from this. The inequality in (5.5) follows because

$$\max_{i \le k, j \le n-k} \frac{\sqrt{\lambda_i + \lambda_{j+k}}}{\lambda_i - \lambda_{j+k}} \sqrt{\lambda_1 + \cdots + \lambda_n} \ge \max_{i \le k, j \le n-k} \frac{\lambda_i + \lambda_{j+k}}{\lambda_i - \lambda_{j+k}} \ge 1. \qquad \square$$

The following corollary is obvious from the above proof.

COROLLARY 5.5. *Assuming $\lambda_k > \lambda_{k+1}$ as before and the use of ordered canonical coordinates, the linear map $\delta x \in U \mapsto E(x - \bar{x}) \in U$, where $E = \frac{dP}{dx}(\delta x)$, is self-adjoint and has as its eigenvectors the orthonormal basis of U given by $\{u_{ij}\}$ for $i = 1, \dots, k$ and $j = 1, \dots, n-k$, which are defined by*

$$u_{ij} = E^{ij}(x - \bar{x}) = \frac{x_i e_{j+k} + x_{j+k} e_i}{\sqrt{\lambda_i + \lambda_{j+k}}}.$$

The corresponding eigenvalues are $\frac{\lambda_i + \lambda_{j+k}}{\lambda_i - \lambda_{j+k}}$. Hence the induced 2-norm of this operator is $\frac{\lambda_k + \lambda_{k+1}}{\lambda_k - \lambda_{k+1}}$.

$$\begin{tikzcd}
 & & \begin{array}{c}\tilde{x}\\ \xi = \|\tilde{x} - x\|^2\end{array} \\
x \arrow[r] \arrow[urr, dashed] \arrow[ddrr, dashed] & (\bar{x}, P(x) = P) \arrow[ur, dotted] \arrow[dr, dotted] & \\
 & & \begin{array}{c}\tilde{x} + \delta\tilde{x}_1\\ \xi + \delta\xi_1 = \|\tilde{x} + \delta\tilde{x}_1 - x\|^2\end{array} \\
 & & \begin{array}{c}\tilde{x} + \delta\tilde{x}_2\\ \xi + \delta\xi_2 = \|\tilde{x} + \delta\tilde{x}_2 - x\|^2\end{array} \\
x + \delta x \arrow[r] \arrow[uurr, dashed] \arrow[drr, dashed] & (\bar{x} + \delta\bar{x},\ P(x + \delta x) = P + E) \arrow[ur, dotted] \arrow[dr, dotted] & \\
 & & \begin{array}{c}\tilde{x} + \delta\tilde{x}\\ \xi + \delta\xi = \|\tilde{x} + \delta\tilde{x} - x\|^2\end{array}
\end{tikzcd}$$

FIG. 5.1. *POD sensitivity for y near x: This shows the four different possibilities of using the reduced models computed from x and $y = x + \delta x$ to approximate these data sets. The solid arrows indicate the construction of a POD reduced model from a data set. A pair of dashed and dotted arrows together show the reduced model being applied to a data set to obtain a reduced and approximate data set. The approximate data obtained and its square error with respect to x are shown on the far right.*

Remark 5.6. Proposition 5.4 was concerned with the POD method of finding the best approximating affine subspace using the mean and the covariance matrix of the data x. Instead, if we considered the POD method of finding the best approximating linear subspace using the correlation matrix, then we get the same equations and the same final expression (5.5) for $S_k(x)$ (in the definition of the space U, $(x - \bar{x})$ needs to be replaced by x). However, x, the perturbation δx, and the worst-case perturbation of δx as well as functions in the space U are no longer necessarily zero mean.

5.2. Sensitivity of the projected data $\tilde{y} = P(x)(y - \bar{x}) + \bar{x}$ and the error $\|\tilde{y} - y\|^2$ when $y = x$ and/or $y = x + \delta x$. In some applications the POD reduced model $(\bar{x}, P(x))$ constructed from a data set x may be used to approximate x itself ($y = x$ situation) or some nearby data $y = x + \delta x$. For instance, consider coding a 512×512 grey scale image by dividing it into subimages of size 8×8 to provide an ensemble of $4096 (= 64 \times 64)$ points in the 64 dimensional subimage space. Suppose that by applying POD to this ensemble we find a subspace of dimension 6 that captures 99.9% of the energy. We could then apply the POD projection to the subimages and code the entire image using 4096×6 grey scale values. This is the $y = x$ situation. If we have a sequence of nearby images (such as in video), then we can use the same reduced model $(\bar{x}, P(x))$ for the nearby images $y = x + \delta x$.

From a theoretical point of view, several different sensitivities may be of interest. These are shown in Figure 5.1. The sensitivities $\delta\tilde{x}_1$ and $\delta\xi_1$ correspond to the situation where the same reduced model $(\bar{x}, P(x))$ (obtained from x) is applied to both x and to a nearby $y = x + \delta x$. The quantity $\delta\tilde{x}_1$ is the perturbation of the approximate data, and $\delta\xi_1$ is the perturbation of the square error $\xi = \|\tilde{x} - x\|^2$. Thus

$$\delta\tilde{x}_1 = (P(x)(x + \delta x - \bar{x}) + \bar{x}) - \tilde{x}$$

and

$$\delta\xi_1 = \|(\tilde{x} + \delta\tilde{x}_1) - (x + \delta x)\|^2 - \xi,$$

where $\tilde{x} = P(x)(x - \bar{x}) + \bar{x}$. This is the most common kind of sensitivity one is interested in in practice. Note that the reason for considering the square of the error rather than the error $\|\tilde{x} - x\|$ itself is because the square root is not smooth when its argument is zero.

The sensitivities $\delta\tilde{x}_2$ and $\delta\xi_2$ correspond to the situation where two nearby reduced models $(\bar{x}, P(x))$ and $(\bar{x} + \delta\bar{x}, P(x + \delta x)$ are applied to the same data x. Thus

$$\delta\tilde{x}_2 = (P(x + \delta x)(x - (\bar{x} + \delta\bar{x})) + (\bar{x} + \delta\bar{x})) - \tilde{x}$$

and

$$\delta\xi_2 = \|(\tilde{x} + \delta\tilde{x}_2) - x\|^2 - \xi.$$

The sensitivities $\delta\tilde{x}$ and $\delta\xi$ correspond to the situation where two nearby reduced models $(\bar{x}, P(x))$ and $(\bar{x} + \delta\bar{x}, P(x + \delta x))$ are applied to the respective data sets x and $x + \delta x$ from which they were constructed. Thus

$$\delta\tilde{x} = (P(x + \delta x)((x + \delta x) - (\bar{x} + \delta\bar{x})) + (\bar{x} + \delta\bar{x})) - \tilde{x}$$

and

$$\delta\xi = \|(\tilde{x} + \delta\tilde{x}) - (x + \delta x)\|^2 - \xi.$$

We provide a useful and easy-to-prove lemma stated without proof.

Lemma 5.7. *Let $L : H \to H$ be a linear operator in the Hilbert space H. Let $K \subset H$ be a closed linear subspace of H. Then we can write $H = K \oplus K^\perp$. Furthermore, suppose $L(K) \subset K$ and $L(K^\perp) \subset K^\perp$ and that the restrictions $L|_K$ and $L|_{K^\perp}$ are bounded operators. Then $\|L\|_2 = \max\{\|L|_K\|_2, \|L|_{K^\perp}\|_2\}$.*

Proposition 5.8. *Consider applying POD to a data set x to find the best approximating $k(< n)$ dimensional subspace. Let the ordered eigenvalues of the covariance matrix of the data x be given by $\lambda_1 \geq \cdots \geq \lambda_n$. Suppose $\lambda_k > \lambda_{k+1}$, which ensures that $P(x)$ is well defined. Consider the sensitivities depicted in Figure* 5.1. *For unit-norm (infinitesimal) variations δx of x, the worst-case variations are given by*

$$\|\delta\tilde{x}_1\| = \|\delta x\|, \tag{5.14}$$

$$\|\delta\tilde{x}_2\| = \frac{\lambda_k + \lambda_{k+1}}{\lambda_k - \lambda_{k+1}} > 1, \tag{5.15}$$

$$\|\delta\tilde{x}\| = \frac{\lambda_k + \sqrt{\lambda_k\lambda_{k+1}}}{\lambda_k - \lambda_{k+1}} > 1, \tag{5.16}$$

$$|\delta\xi_1| = |\delta\xi| = 2\|\tilde{x} - x\| = 2\sqrt{\xi}, \tag{5.17}$$

$$\delta\xi_2 = 0. \tag{5.18}$$

(All norms are 2*-norms.)*

Proof. We shall use the ordered canonical coordinate system whenever necessary.

We define some relevant subspaces. As before we shall consider the data x to be a single trajectory $x : [0,T] \to \mathbb{R}^n$ for simplicity of exposition. However, the results will hold for more general types of data. We shall assume $x \in \mathcal{L}_2([0,T],\mathbb{R}^n)$, the space of all square integrable $\mathbb{R}^n$-valued functions in $[0,T]$. Let $Z \subset \mathcal{L}_2([0,T],\mathbb{R}^n)$ denote the (closed) subspace of all zero mean functions:

$$Z = \left\{ x \in \mathcal{L}_2([0,T],\mathbb{R}^n) \ : \ \int_0^T x dt = 0 \right\}.$$

Its orthogonal complement $Z^\perp$ is finite dimensional and consists of functions that are constant-valued (almost everywhere) in $[0,T]$. We shall further decompose Z into the orthogonal sum $Z = W \oplus Y$, where

$$W = \operatorname{span}\left\{ \frac{x_\alpha}{\sqrt{\lambda_\alpha}} e_\beta \ : \ \alpha = 1,\dots,\tilde{n},\ \beta = 1,\dots,n \right\}. \tag{5.19}$$

Here $\tilde{n}$ is the number of nonzero eigenvalues of the covariance matrix associated with trajectory x (thus W depends on x), and x_α are its components in the ordered canonical coordinate system. Since Y is the orthogonal complement of W in Z, it is closed. The $n\tilde{n}$ dimensional W is further decomposed into the orthogonal sum $W = U \oplus U_2 \oplus V_1 \oplus V_2$, where

$$\begin{aligned}
U &= \operatorname{span}\left\{ u_{ij} = \frac{x_i e_{j+k} + x_{j+k} e_i}{\sqrt{\lambda_i + \lambda_{j+k}}} \ : \ i = 1,\dots,k;\ j = 1,\dots,n-k \right\},\\
U_2 &= \operatorname{span}\left\{ u_{ij}^2 = \frac{\lambda_{j+k} x_i e_{j+k} - \lambda_i x_{j+k} e_i}{\sqrt{\lambda_i \lambda_{j+k} (\lambda_i + \lambda_{j+k})}} \ : \ i = 1,\dots,k;\ j = 1,\dots,\tilde{n}-k \right\},\\
V_1 &= \operatorname{span}\left\{ \frac{x_i e_\alpha}{\sqrt{\lambda_i}} \ : \ i = 1,\dots,k;\ \alpha = 1,\dots,k \right\},\\
V_2 &= \operatorname{span}\left\{ \frac{x_{j+k} e_{\beta+k}}{\sqrt{\lambda_{j+k}}} \ : \ j = 1,\dots,\tilde{n}-k;\ \beta = 1,\dots,n-k \right\}.
\end{aligned}$$

It should be noted that the spanning elements above form orthonormal bases for the respective subspaces. Furthermore, define $\widetilde{U} = U \oplus U_2$ and $V = V_1 \oplus V_2$. Also note that U defined above is the same as in (5.9).

The perturbation $\delta\tilde{x}_1$ is given by

$$\delta\tilde{x}_1 = P(x)(x + \delta x - \bar{x}) + \bar{x} - P(x)(x - \bar{x}) - \bar{x}$$

and simplifies to $\delta\tilde{x}_1 = P(x)\delta x$. Hence the worst perturbation δx is in the image of $P(x)$, resulting in $\delta\tilde{x}_1 = \delta x$. Note that this holds for finite as well as infinitesimal perturbations.

The variation $\delta\tilde{x}_2$ is given by

$$\delta\tilde{x}_2 = E(x - \bar{x}) + (1 - P(x))\delta\bar{x},$$

where $E = \frac{dP}{dx}(\delta x)$. Note that $\delta x \in Z^\perp$ implies that $E = 0$ and hence that $\delta\tilde{x}_2 = (1 - P)\delta\bar{x} \in Z^\perp$. Also note that $\delta x \in Z$ implies $\delta\tilde{x}_2 \in Z$. Furthermore, if $\delta x \in Z$ and $\delta x \perp U$, then $\delta\tilde{x}_2 = 0$. If $\delta x \in U$, then $\delta\tilde{x}_2 = E(x - \bar{x}) \in U \subset Z$. From Corollary

5.5 and Lemma 5.7 it is clear that the worst-case variation $\|\delta\tilde{x}_2\| = \frac{\lambda_k+\lambda_{k+1}}{\lambda_k-\lambda_{k+1}} > 1$ and that it corresponds to $\delta x = \frac{x_k e_{k+1} + x_{k+1} e_k}{\sqrt{\lambda_k+\lambda_{k+1}}}$.

The variation $\delta\tilde{x}$ is given by

$$\delta\tilde{x} = E(x-\bar{x}) + P\delta x - P\delta\bar{x} + \delta\bar{x}.$$

Denote by L the operator that maps δx to $\delta\tilde{x}$. The following are easy to establish: $L(Z) \subset Z$, $L(Z^\perp) \subset Z^\perp$, $L(W) \subset W$, $L(Y) \subset Y$, $L(V) \subset V$, and $L(\tilde{U}) \subset \tilde{U}$. If $\delta x \in Z^\perp$, then $\delta\tilde{x} = \delta x = \delta\bar{x}$, so $\|L|_{Z^\perp}\| = 1$. If $\delta x \in Z$ and $\delta x \perp \tilde{U}$, it follows that $\delta\tilde{x} = P\delta x$. Hence by Lemma 5.7

$$\|L\| = \max\{\|L|_{\tilde{U}}\|, 1\}.$$

It can be verified that the following orthonormal basis of $\tilde{U}$ are eigenvectors of the finite dimensional operator $L|_{\tilde{U}} : \tilde{U} \to \tilde{U}$:

$$\left\{ \frac{x_i e_{j+k}}{\sqrt{2\lambda_i}} + \frac{x_{j+k} e_i}{\sqrt{2\lambda_{j+k}}}, \frac{x_i e_{j+k}}{\sqrt{2\lambda_i}} - \frac{x_{j+k} e_i}{\sqrt{2\lambda_{j+k}}}, \frac{x_i e_{\tilde{j}+k}}{\sqrt{2\lambda_i}} \ : i = 1,\dots,k;\ j = 1,\dots,\tilde{n}-k; \right.$$
$$\left. \tilde{j} = \tilde{n}-k+1,\dots,n-k \right\}.$$

The eigenvectors $\{\frac{x_i e_{\tilde{j}+k}}{\sqrt{2\lambda_i}}\}$ all have eigenvalue 1. The eigenvectors $\{\frac{x_i e_{j+k}}{\sqrt{2\lambda_i}} + \frac{x_{j+k} e_i}{\sqrt{2\lambda_{j+k}}}\}$ have corresponding eigenvalues $\frac{\lambda_i + \sqrt{\lambda_i\lambda_{j+k}}}{\lambda_i - \lambda_{j+k}} > 1$. The eigenvectors $\{\frac{x_i e_{j+k}}{\sqrt{2\lambda_i}} - \frac{x_{j+k} e_i}{\sqrt{2\lambda_{j+k}}}\}$ have corresponding (positive) eigenvalues $\frac{\lambda_i - \sqrt{\lambda_i\lambda_{j+k}}}{\lambda_i - \lambda_{j+k}} < 1$. Hence the norm $\|L|_{\tilde{U}}\|$ is given by the largest eigenvalue $\frac{\lambda_k + \sqrt{\lambda_k\lambda_{k+1}}}{\lambda_k - \lambda_{k+1}} > 1$. Hence $\|L\| = \frac{\lambda_k + \sqrt{\lambda_k\lambda_{k+1}}}{\lambda_k - \lambda_{k+1}}$.

The (infinitesimal) variation $\delta\xi_1$ is given by

$$\delta\xi_1 = 2(\tilde{x} - x, \delta\tilde{x}_1 - \delta x) = (\tilde{x} - x, (P-1)\delta x),$$

and hence the worst case is when $\delta x = \pm\frac{\tilde{x}-x}{\|\tilde{x}-x\|}$ and results in $|\delta\xi_1| = 2\|\tilde{x} - x\|$.

The variation $\delta\xi$ is given by

$$\begin{aligned}\delta\xi &= 2(\tilde{x} - x, \delta\tilde{x} - \delta x) \\ &= 2(\tilde{x} - x, E(x-\bar{x})) + 2(\tilde{x} - x, (1-P)\delta\bar{x}) + 2(\tilde{x} - x, (1-P)\delta x) \\ &= 2(\tilde{x} - x, (1-P)\delta x).\end{aligned}$$

As before, the worst-case variation is given by $\delta x = \pm\frac{\tilde{x}-x}{\|\tilde{x}-x\|}$ and results in $|\delta\xi| = 2\|\tilde{x} - x\|$.

The variation $\delta\xi_2$ is given by

$$\begin{aligned}\delta\xi_2 &= 2(\tilde{x} - x, \delta\tilde{x}_2) \\ &= 2(\tilde{x} - x, E(x-\bar{x})) + 2(\tilde{x} - x, (1-P)\delta\bar{x}).\end{aligned}$$

Since $\tilde{x} - x \in V_2$, $E(x-\bar{x}) \in U$, and $(1-P)\delta\bar{x} \in Z^\perp$, it follows that $\delta\xi_2 = 0$. □

Remark 5.9. The above proposition shows that the projected data $\tilde{x}$ may become extremely sensitive to perturbations in the data set when $\lambda_k \approx \lambda_{k+1}$. However, the (square of the) error itself does not show this sensitivity. This is related to the fact that when $\lambda_k = \lambda_{k+1}$ there are infinitely many choices for $P(x)$ and thus for $\tilde{x}$, and these different choices for $\tilde{x}$ may be quite different from each other. However, they all have exactly the same error $\sqrt{\lambda_{k+1} + \cdots + \lambda_n}$. It should also be noted that our sensitivity results hold for infinitesimal variations, and finite perturbations are likely not to be as sensitive as the first derivative may suggest.

Example 2. This example illustrates a situation where the reduced model solution is very sensitive to perturbations of the trajectory x used as POD data. We consider a dissipative ODE example which has a periodic orbit which is a global attractor. Consider the ODE

$$\dot{x} = A(x - f(t)) + f'(t), \quad x \in \mathbb{R}^n,$$

where $f : \mathbb{R} \to \mathbb{R}^n$ is smooth. Observe that for any choice of A and f, $x = f(t)$ is a trajectory of this system. We exploit this fact to independently choose $f(t)$ and A to create an interesting example which highlights some of the potential problems with the POD procedure.

We chose $f(t)$ to be periodic and A to be a constant matrix with all of its eigenvalues in the complex left half plane. Thus $x = f(t)$ will be a global attractor of this system. Specifically we chose $f(t)$ to be of the form

$$f(t) = \left(\sqrt{a_1} \sin\left(\frac{2\pi t}{25}\right),\ \sqrt{a_2} \cos\left(\frac{2\pi t}{25}\right),\ \sqrt{a_3} \sin\left(\frac{4\pi t}{25}\right),\ \sqrt{a_4} \cos\left(\frac{4\pi t}{25}\right) \right)^T,$$

where a_i are real nonnegative constants. This trajectory has period 25. If we use this trajectory in the interval $[0, 50]$ (two periods) as POD data, we will get a reduced model with $\bar{x} = 0$ and a diagonal covariance matrix R with $R_{ii} = 25a_i$. This is because the component functions of $f(t)$ in the interval $[0, 50]$ form an orthogonal set. If we choose $a_4 = 0$ (or very small), then the POD procedure based on this trajectory will give a reduced model ODE by projecting onto the first three components in $\mathbb{R}^4$. This projection will preserve all (or almost all) of the energy of the POD data trajectory. Now consider a matrix A that has all of its eigenvalues in the complex left half plane, but its submatrix consisting of the first three rows and columns (i.e., the projection of A onto the first three components in $\mathbb{R}^4$) has an eigenvalue in the complex right half plane. Such a choice of A will lead to a reduced model which is unstable. If $a_4 = 0$, the global attractor $x = f(t)$ will still be a trajectory of the reduced model but it will not be an attractor. Thus the qualitative behavior of the reduced model will be quite different even though the POD procedure is based on a global attractor of a dissipative system.

In order to find such an A, we first chose A to be diagonalizable with eigenvalues

$$\lambda(A) = \{-0.7 + 0.4i, -0.7 - 0.4i, -0.2, -0.1\}.$$

Then by trial and error, applying random similarity transformations, we found an A with the above canonical form such that its submatrix consisting of the first three rows and columns had an eigenvalue of about 1.8 in the complex right half plane.

If we choose $a_4 = 0$, then with $k = 3$ we do not expect high sensitivity to perturbations in the data. However, we get an interesting example where doing POD on a lower dimensional global attractor still leads to a reduced model which is unstable and qualitatively different. Since we were interested in studying the effects of perturbations in the POD data on the final outcome of a POD reduced model solution, instead of choosing $a_4 = 0$, we chose the following values for a_i:

$$a_1 = 5,\ a_2 = 0.5,\ a_3 = 0.011,\ a_4 = 0.01,$$

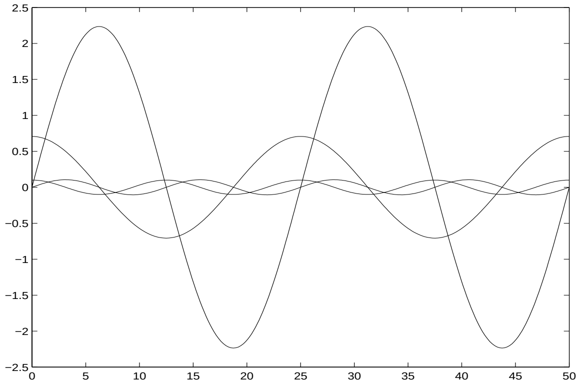


FIG. 5.2. *Example* 2: *True solution* $x(t)$.

where $a_3 \approx a_4$. With this choice, a reduced model of dimension $k = 3$ will capture most of the energy, but we will expect high sensitivity to small perturbations in the POD data. We chose the initial value problem with $x(0) = f(0)$ and time interval $[0, 50]$. Thus the solution trajectory is $x = f(t)$ and consists of two periods. In order to incoorporate the effects of numerical errors, we computed the solution using the MATLAB solver `ode45` and then computed the POD reduced model of dimension $k = 3$ numerically. We also numerically computed the projected trajectory $\tilde{x}$ as well as the solution $\hat{x}$ of the reduced ODE model. We found that the POD procedure preserved 99.8% of the energy and that the sensitivity factor was $S_k = 480$. We found $\|x\| = 11.75$ and $\|\tilde{x}\| = 11.74$. The reduced model solution was highly unstable and $\|\hat{x}\| = 1.25 \times 10^{38}$. The eigenvalues of the reduced model matrix were $\{1.80, -0.281 + 0.217i, -0.281 - 0.217i\}$.

Then we perturbed the trajectory x by δx in the direction given by (5.6) that creates the worst perturbation in P. We chose $\|\delta x\| = 0.1$. We then computed the POD reduced model corresponding to $x + \delta x$ and also computed the perturbed projected trajectory $\tilde{x} + \delta \tilde{x}$ as well as the perturbed reduced model solution $\hat{x} + \delta \hat{x}$. Figure 5.2 shows a plot of the numerically computed true solution $x(t)$ (i.e., the full model solution). Figures 5.3 and 5.4 show how a small perturbation in x leads to a larger perturbation in $\tilde{x}$, and Figure 5.5 shows an even larger perturbation in $\hat{x}$. We also observed that while the unperturbed reduced model projected almost onto the first three components in $\mathbb{R}^4$, the perturbed reduced model was projecting onto a subspace that consisted of the span of $\{e_1, e_2\}$ and a combination of e_3 and e_4, and this subspace was rotated from the span of $\{e_1, e_2, e_3\}$ by an angle of about 41° (e_i being the standard basis vectors in $\mathbb{R}^4$). It was also observed that the eigenvalues of the perturbed reduced model matrix were $\{2.89, -0.337, -0.166\}$, which correspond to a larger instability and a qualitatively different nonoscillatory behavior from that of the unperturbed reduced model.

This example illustrates two potential inadequacies of the POD method. One is that even capturing 100% of the energy of a globally attracting low dimensional trajectory may still lead to a POD reduced model with the wrong dynamics. Second, it also illustrates how POD sensitivity to the data trajectory may lead to qualitatively different reduced models.

The first problem is related to two factors. One is that a single trajectory (even a global attractor) or a set of trajectories alone does not carry all the information

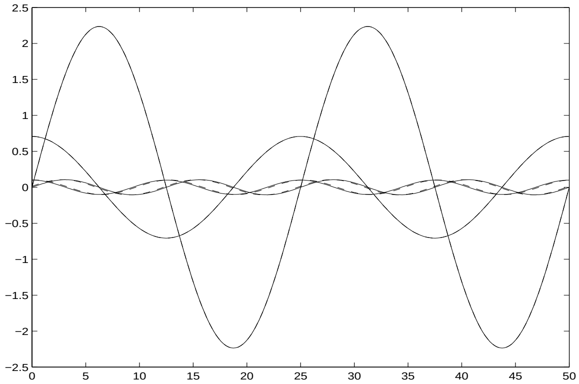


FIG. 5.3. *Example* 2. *Perturbation of x: Solid: x; dashed: $x+\delta x$. The perturbation is so small* ($\frac{\|\delta x\|}{\|x\|} = 0.0085$) *that the two trajectories are barely distinguishable. Note that all four components are plotted for both x and $x+\delta x$. The perturbation is only noticeable in the two smaller components.*

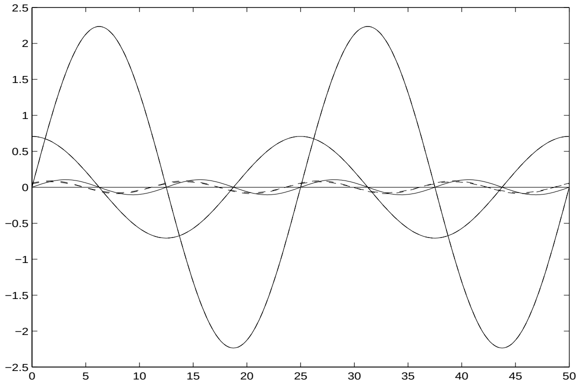


FIG. 5.4. *Example* 2. *Perturbation of $\tilde{x}$: Solid: $\tilde{x}$; dashed: $\tilde{x}+\delta\tilde{x}$. The perturbation is larger than that of x but still not noticeable for the two large components.*

about the dynamics. Second, the projection of a vector-field onto a given subspace does not preserve its stability characteristics. Our example has these characteristics and in addition has a high sensitivity factor.

5.3. Effect of POD sensitivity in data representation and model reduction. Consider the reduced model $(\bar{x}, P(x))$ obtained from a data set x. Suppose we apply this reduced model to represent another data set y and obtain $\tilde{y} = P(x)(y-\bar{x})+\bar{x}$. The previous subsection was concerned with the special situation where $y = x$. In general situations, the data set y is different from x. The variation $\delta\tilde{y}$ due to a variation δx is given by

$$\delta\tilde{y} = E(y-\bar{x}) + \delta\bar{x} - P\delta\bar{x},$$

where E is the corresponding variation of P. Since $\|E\| \leq S_k \frac{\|\delta(x-\bar{x})\|}{\|x-\bar{x}\|}$ (assuming $\|x-\bar{x}\| \neq 0$), we obtain

$$\|\delta\tilde{y}\| \leq S_k \frac{\|y-\bar{x}\|}{\|x-\bar{x}\|} \|\delta(x-\bar{x})\| + \|\delta\bar{x}\|.$$

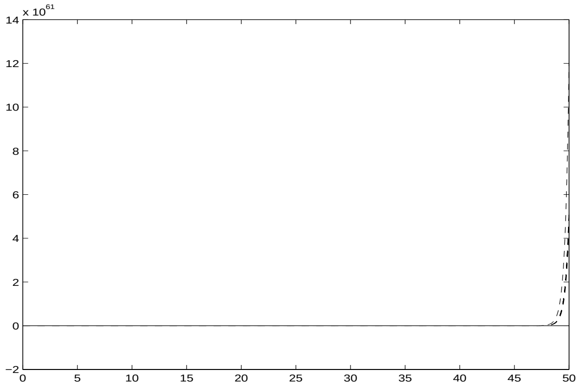


FIG. 5.5. *Example* 2. *Perturbation of $\hat{x}$: Solid: $\hat{x}$; dashed: $\hat{x}+\delta\hat{x}$. Note that $\hat{x}$ appears to be zero since it is of the order* 10^{38}*, while $\hat{x}+\delta\hat{x}$ is of the order* 10^{61}.

Assuming further that $\|y\| \neq 0$, we obtain the following fractional sensitivity relation:

$$\frac{\|\delta\tilde{y}\|}{\|y\|} \leq S_k \frac{\|y-\bar{x}\|}{\|y\|}\frac{\|\delta(x-\bar{x})\|}{\|x-\bar{x}\|} + \frac{\|\delta\bar{x}\|}{\|y\|}. \tag{5.20}$$

Now let us consider the case where we use the reduced model $(\bar{x}, P(x))$ to compute the solution of an ODE initial value problem for a linear time invariant system $\dot{y} = Ay$, with initial condition $y(0) = y_0$ in the interval $[0, T]$. Let y denote the true solution and $\hat{y}$ denote the reduced model solution. Then $\hat{y}$ satisfies the initial value problem $\dot{\hat{y}} = PA\hat{y}$, $\hat{y}(0) = P(y_0 - \bar{x}) + \bar{x}$. Taking variations, we get

$$\delta\dot{\hat{y}} = PA\delta\hat{y} + EA\hat{y},$$

with initial condition $\delta\hat{y}(0) = E(y_0 - \bar{x}) - P\delta\bar{x} + \delta\bar{x}$. Hence applying the estimate (3.2), we get

$$\|\delta\hat{y}\| \leq \|F(T;PA)\|\|E\|\|A\|\|\hat{y}\| + \|G(T;PA)\|\|E\|\|y_0-\bar{x}\| + \|G(T;PA)\|\|\delta\bar{x}\|,$$

where F and G are defined by (3.1). Since $\|E\| \leq S_k \frac{\|\delta(x-\bar{x})\|}{\|x-\bar{x}\|}$, it follows that

$$\begin{aligned}\frac{\|\delta\hat{y}\|}{\|\hat{y}\|} \leq & \left(\|F(T;PA)\|\|A\| + \|G(T;PA)\|\frac{\|y_0-\bar{x}\|}{\|\hat{y}\|}\right) S_k \frac{\|\delta(x-\bar{x})\|}{\|x-\bar{x}\|} \\ & + \|G(T;PA)\|\frac{\|\delta\bar{x}\|}{\|x-\bar{x}\|}.\end{aligned} \tag{5.21}$$

Example 3. We considered the same initial value problem of Example 2 in the same interval. However, instead of using the true solution as POD data, we used the set of eight trajectories x obtained by solving the system in the same interval with the symmetrically placed initial conditions $x(0) = e_i$ and $x(0) = -e_i$ for $i = 1, \ldots, 4$, where $e_i \in \mathbb{R}^4$ are the standard basis vectors as POD data, and computed the rank 3 projection matrix $P(x)$. The sensitivity factor was $S_k = 10.140$. We then perturbed this data set x in the direction given by (5.6) (this gives the worst perturbation in P) by an amount $\|\delta x\| = 0.5$. The norm of the data set was $\|x\| = 54.070$, and $\|x - \bar{x}\| = 52.90$. Thus we had the fractional change $\|\delta x\|/\|x - \bar{x}\| = 0.0095$. We also

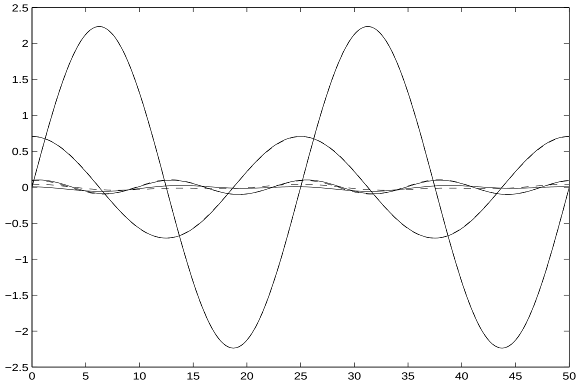


FIG. 5.6. *Example* 3. *Perturbation of $\tilde{y}$: Solid: $\tilde{y}$; dashed: $\tilde{y} + \delta\tilde{y}$.*

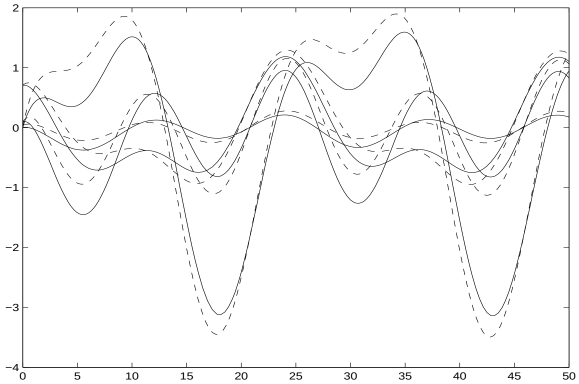


FIG. 5.7. *Example* 3. *Perturbation of $\hat{y}$: Solid: $\hat{y}$; dashed: $\hat{y} + \delta\hat{y}$.*

computed the projections $P(x)$ and $P(x + \delta x)$ corresponding to the data sets x and $x + \delta x$.

Denote by y the true solution of the initial value problem. (This is the same as $x(t)$ of Example 2 in Figure 5.2.) We applied both reduced models $P(x)$ and $P(x+\delta x)$ to compute $\tilde{y} = P(x)(y - \bar{x}) + \bar{x}$ and $\tilde{y} + \delta\tilde{y} = P(x + \delta x)(y - \bar{x} - \delta\bar{x}) + \bar{x} + \delta\bar{x}$, the projected solutions. Figure 5.6 shows the two different projected solutions.

We then computed the reduced model solutions $\hat{y}$ and $\hat{y}+\delta\hat{y}$ corresponding to $P(x)$ and $P(x + \delta x)$, respectively. These are plotted in Figure 5.7. This again illustrates how a small perturbation in the POD data set may cause a large perturbation in the reduced model solution.

Remark 5.10. In this section we basically saw how POD results may be very sensitive to slight perturbations in the data when $\lambda_k \approx \lambda_{k+1}$. However, one needs to be careful in interpreting these results. This raises the question of whether one should consider the sensitivity factor S_k (in addition to the projection error $\sqrt{\lambda_{k+1} + \cdots + \lambda_n}$) as an important factor in choosing an appropriate dimension k for the reduced model. Sometimes the distribution of eigenvalues may be such that seeking higher accuracy may lead to a high sensitivity factor S_k. The importance of S_k depends on the nature of the application. It must also be noted that our sensitivity analysis holds only for infinitesimal perturbations; the sensitivity for finite perturbations is likely to be different. For instance, infinitesimal analysis predicts that in the limit $\lambda_k \to \lambda_{k+1}$ the sensitivity of P with respect to x grows indefinitely. However, since projection matrices live on a compact set ($\|P\| = 1$), the finite perturbations of P cannot grow indefinitely. As mentioned in Remark 5.9, when $y \approx x$, for data compression type problems we have already argued that the sensitivity factor is not a serious issue. However, in reduced order models for ODEs, Examples 2 and 3 show how a small perturbation in the POD data may lead to very large perturbations in the reduced model solutions.

In addition, we would like to note that in iterative methods based on POD such as the DIRM method [20, 21], the convergence of the iterations will depend on the sensitivity factor. This has been theoretically and numerically demonstrated in [20].

6. Computational complexity of POD reduced order models. Since the POD method of model reduction results in a smaller dimensional system of ODEs, one might expect computational savings when integrating the resulting system. However, this may not be so in practice. In this section we shall take a close look at the complexity of computation for integrating a system of ODEs (initial value problems) and evaluate the savings if any in using the POD method.

To be precise, by complexity we shall mean the asymptotic behavior of the number of floating point operations (flops—addition, multiplication, and elementary functions each count as one flop) involved per integration step as the system size n (k for reduced models) becomes very large. Table 6.1 shows the complexity of various basic operations that are used in integration of ODEs. All matrices are $n \times n$ and vectors are n dimensional. Banded matrices are assumed to have $b+1$ nonzero entries symmetrically placed around the diagonal. See [11] for details on complexity of linear algebraic operations. We slightly abuse the notation and denote by $f(n)$ the number of flops involved in computing a nonlinear vector-field $f(x,t)$, where $f : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^n$. Computing the reduced order vector-field $\rho f(\rho^T z + \bar{x}, t)$ could potentially be more expensive than $f(n)$. If this is naively treated as a composition of functions, then the complexity is $f(n) + 4nk$ (two matrix-vector multiplications should be included). Depending on the form of f, one may not be able to improve on this. However, often the analytical formula $\rho f(\rho^T z + \bar{x}, t)$ may be simplified, especially if f is a polynomial in x. We shall denote the complexity of this term by $\hat{f}(k,n)$. This may be bounded as follows:

$$\hat{f}(k,n) \leq f(n) + 4nk.$$

For Jacobian evaluations we have assumed the use of centered finite differences. See [9] for efficient numerical evaluation of banded Jacobians by finite difference approximation. Throughout the rest of this section we will assume that $f(n)$ and $\hat{f}(k,n)$ are of order greater than or equal to n and k, respectively. Under this assumption we can ignore the subtractions and divisions involved in computing the finite differences. If analytical Jacobians are used, the corresponding complexity is likely to be similar [5]. It must be noted that even if the original Jacobian is banded, the reduced model Jacobian is not likely to be.

First we shall consider a linear time invariant system $\dot{x} = Ax$. Table 6.2 shows the asymptotic complexities for various cases. The explicit method considered is forward Euler, and the implicit method considered is backward Euler. The explicit case ($x_n = x_{n-1} + h_n A x_{n-1}$) involves basically a matrix-vector product. The implicit case involves solving the equation $(I - h_n A)x_n = x_{n-1}$ at each time step. We assumed that this is done by Gaussian elimination, first doing an LU decomposition and then two

Table 6.1

Complexity of some basic operations. Dense and banded refer to the full model Jacobians. Jacobian evaluation assumes centered finite differencing.

	Dense	Banded
Matrix-vector product Ab	$2n^2$	$2bn$
LU decomposition	$\frac{2n^3}{3}$	$\frac{b^2n}{2}$
Triangular linear system solve	n^2	bn
Nonlinear function evaluation	$f(n)$	$f(n)$
Nonlinear function evaluation (reduced model)	$\hat{f}(k,n)$	$\hat{f}(k,n)$
Nonlinear Jacobian evaluation	$2nf(n)$	$2(b+1)f(n)$
Nonlinear Jacobian evaluation (reduced model)	$2k\hat{f}(k,n)$	$2k\hat{f}(k,n)$

Table 6.2

Asymptotic complexity for linear systems.

	Full model explicit	Reduced model explicit	Full model implicit	Reduced model implicit
Dense	$2n^2$	$2k^2$	$\frac{n^3}{15}$	$\frac{k^3}{15}$
Banded	$2bn$	$2k^2$	$(\frac{b^2}{20}+3b+2)n$	$\frac{k^3}{15}$

triangular system solves (one forward and one backward). Usually the LU decomposition needs to be computed only whenever the time step h_n changes. Throughout this section we shall assume that on average, the LU decomposition needs to be computed only once every 10 time steps. Thus we obtain a complexity of $\frac{n^3}{15}+2n^2+\frac{n^2}{10}+\frac{n}{10}\sim\frac{n^3}{15}$ for a dense matrix A and $(\frac{b^2}{20}+3b+2)n$ for a banded matrix A (the quantity b is held constant). If a POD reduced order model of dimension $k(<n)$ is used, then we replace n by k in most of the expressions except for that for banded A (the reduced model matrix $\rho A\rho^T$ is not likely to be banded, and we shall assume it to be dense). It must be noted that in several examples when $n\to\infty$, the adequate size k of a reduced model remains constant after an initial growth. This is especially true in discretized PDE systems, since a finite number of empirical modes are adequate to capture any given percentage of the energy. As a result the asymptotic formulae for $k\to\infty$ are often not applicable.

For nonlinear systems $\dot{x}=f(x,t)$, the explicit method (forward Euler) involves evaluating $x_n=x_{n-1}+h_nf(x_{n-1},t_{n-1})$; hence the complexity is $f(n)+2n\sim f(n)$. If the reduced model is used, then one needs to evaluate $\rho f(\rho^Tz_{n-1}+\bar{x},t_{n-1})$, and the corresponding complexity is $\hat{f}(k,n)+2k\sim\hat{f}(k,n)$.

For the implicit case one needs to solve the nonlinear system of equations

$$F(x_n)\triangleq x_n-x_{n-1}-h_nf(x_n,t_n)=0$$

for x_n. This is done typically by Newton iteration

$$\left[I-h_n\frac{\partial f}{\partial x}\right]\delta_n^{(m)}=x_{n-1}-x_n^{(m-1)}+h_nf(x_n^{(m-1)},t_n), \tag{6.1}$$

where $\delta_n^{(m)}=x_n^{(m)}-x_n^{(m-1)}$ is the correction. Ideally

$$\frac{\partial f}{\partial x}=\frac{\partial f}{\partial x}(x_n^{(m-1)},t_n);$$

Table 6.3
Asymptotic complexity for nonlinear systems.

	Full model explicit	Reduced model explicit	Full model implicit	Reduced model implicit
Dense	$f(n)$	$\hat{f}(k,n)$	$\frac{nf(n)}{5}+\frac{n^3}{15}$	$\frac{k\hat{f}(k,n)}{5}+\frac{k^3}{15}$
Banded	$f(n)$	$\hat{f}(k,n)$	$\frac{bf(n)}{5}+\frac{b^2n}{20}$	$\frac{k\hat{f}(k,n)}{5}+\frac{k^3}{15}$

hence the Jacobian should be evaluated at every Newton iteration inside a given time step. In practice it has been observed that one could get away with keeping the Jacobian unchanged not only during the Newton iteration but also for a few time steps, without severely compromising the accuracy. We assumed that the Jacobian update and the LU decomposition typically need to be done only once every 10 time steps. Note that the right-hand side of (6.1), however, needs to be computed at every Newton iteration. The evaluation of the right-hand side alone requires an asymptotic complexity of $f(n)$ at every Newton iteration for the full model. For the model reduced system one needs to evaluate $\rho f(\rho^T z_n + \bar{x}, t_n)$, which involves an asymptotic complexity of $\hat{f}(k,n)$. It is reasonable to assume that the number of Newton iterations per time step is on average a number independent of system size n (or k for the reduced model). Often in practice this could be about 2. Thus asymptotically the complexity of the Jacobian evaluations dominates over the complexity of evaluating the right-hand side of (6.1). Under these assumptions we get the asymptotic complexities shown in Table 6.3.

For nonlinear systems with explicit solver the asymptotic savings depend solely on the complexity of the nonlinear function evaluations $f(n)$ and $\hat{f}(k,n)$. Hence savings can be expected only if $\rho f(\rho^T z + \bar{x}, t)$ can be analytically simplified.

For nonlinear systems with implicit solver the complexity has two components: one from the nonlinear function evaluations and the other from the linear algebra. Depending on the complexity of f (or $\hat{f}$ for reduced models), one of these terms may be dominant. Asymptotic savings achieved depend on several factors including complexity of $f(n)$ and $\hat{f}(k,n)$ as well as the assumptions on asymptotic behavior of k as $n \to \infty$.

In our complexity analysis we have made several assumptions which are not always valid in practice. Even though most of these assumptions are reasonable, the combined error in our estimate of computational savings can sometimes be wrong by more than a factor of 10. For instance, we looked at complexity per time step. This is useful only if both the full model and the reduced model took more or less the same number of time steps. In the two examples here, the number of time steps did not vary by more than a factor of 2, except for the case of the explicit solver applied to the nonlinear PDE example with reduced model dimension $k = 6$. The asymptotic formulae may not be very applicable for the reduced models because of their smaller size. In addition we ignored computations associated with adaptive stepsize control, which seem to be a significant percentage, for the banded Jacobian case.

It must also be noted that we assumed that we care only about the solution value at the final time (or perhaps only at a few different points in the time interval), and this allowed us to ignore the cost of computing $\hat{x} = \rho^T z + \bar{x}$. The next two examples illustrate how various complex factors can affect the computational savings achieved by the use of POD reduced order models.

Example 4 (RC circuit—dense Jacobian). We considered the example of an electric circuit with resistors and capacitors. By connecting each node to every other node, we obtain a dense Jacobian. One of the nodes is considered the ground (has zero voltage). Such a circuit with n nodes other than ground is described by a first order system of n linear ODEs. The current i_{jk} from the jth node to the k node is given by

$$i_{jk} = g_{jk}(v_j - v_k) + c_{jk}(\dot{v}_j - \dot{v}_k),$$

where g_{jk} and c_{jk} are the appropriate conductances and the capacitances, and $v \in \mathbb{R}^n$ is the vector of node voltages. We may add a nonlinearity to the resistors to obtain

$$i_{jk} = g_{jk}(v_j - v_k) + h_{jk}(v_j - v_k)^3 + c_{jk}(\dot{v}_j - \dot{v}_k).$$

These equations when combined with Kirchoff's current law give rise to an equation $\dot{v} = f(v)$.

We chose the parameter values somewhat ad hoc so that the resulting linearized system had a reasonable range of eigenvalues. We chose a system with $n = 500$. A random initial condition was chosen, and the time interval was chosen to be $[0, 2]$. Both linear and nonlinear versions were simulated, with both explicit and implicit solvers `ode45` and `ode15s` from MATLAB. Reduced order models were computed from the resulting trajectories and applied to both the linear and nonlinear systems. A reduced order model of size $k = 50$ was used, even though a size of $k = 2$ would have preserved more than 99% of the energy. The reason for using $k = 50$ is that $k = 2$ is too small for the asymptotic formulae to be valid. The number of floating point operations as counted in MATLAB are shown in Table 6.4.

The asymptotic complexity of function evaluations for this example are given by

$$f(n) = Cn^2,$$

where $C = 13$ for the nonlinear (cubic) case and $C = 2$ for the linear case. The complexity for the nonlinear reduced model is

$$\hat{f}(k, n) = Cn^2 + 4nk,$$

when $\rho f(\rho^T z + \bar{x})$ is not analytically simplified. Since $f(x)$ is cubic for the nonlinear case, it is possible to analytically simplify $\rho f(\rho^T z + \bar{x})$. This will improve the savings achieved by the reduced order model. In fact assuming that we get a cubic polynomial (in z) with all the possible monomials (dense cubic), we can estimate the complexity of $\hat{f}$. Table 6.5 compares the complexities of $\hat{f}$ after analytical simplification (assuming a dense cubic) for different values of k with that of $f(n)$. It is clear that for values of $k = 15$ or $k = 8$ we can expect significant savings.

We can see from Table 6.6 that the savings predicted by our theory is within an order of magnitude of the actual savings. It must be noted that our theory is only valid asymptotically as n and k get arbitrarily large and that our theory was based on forward and backward Euler methods, while the example used a Runge–Kutta method for the explicit case and a numerical differentiation formula (NDF) for the implicit case [24]. The discrepancies are due to several other factors as well. One is that the reduced model size $k = 50$ is not large enough to use the asymptotic formula. This is an important factor in the linear implicit case but not in the explicit case. Another reason is that the number of steps taken were different for the full model and the reduced model. Furthermore, there are computations associated with

Table 6.4

Computational cost in 10^3 flops: RC circuit. Full model versus unsimplified reduced model with $k = 50$.

	Full model	Reduced model (unsimplified, $k = 50$)
Linear explicit	$37,632$	645
Nonlinear explicit	$238,382$	$305,041$
Linear implicit	$646,460$	$2,323$
Nonlinear implicit	$2,609,800$	$400,177$

Table 6.5

Cost (in flops) of f and simplified $\hat{f}$: RC circuit.

Full model $n = 500$	Reduced model $k = 50$	Reduced model $k = 15$	Reduced model $k = 8$
3250000	2342500	24450	2624

Table 6.6

Computational savings ratio: RC circuit, unsimplified reduced model with $k = 50$.

	Observed savings	Asymptotic theoretical savings
Linear explicit	58	100
Nonlinear explicit	0.78	1
Linear implicit	278	1000
Nonlinear implicit	6.5	10

adaptive stepsize control which were not accounted for by our theory. The latter was a significant factor in the explicit case. We found that the computations associated with adaptive stepsize control grew linearly with system dimension. Furthermore, we found that the Jacobian evaluations were done much less often than the LU decomposition, contrary to our initial assumption. In fact there was only one Jacobian evaluation for the whole simulation in all of the implicit solvers.

Example 5 (reaction-diffusion PDE in one dimension—banded Jacobian). We considered the one dimensional reaction diffusion equation

$$x_t = 0.1x_{ss} - cx^3$$

in the spatial interval $s \in [0, 6]$ with zero boundary conditions. We discretized the spatial dimension on a uniform grid of n interior points using centered differences for both first and second derivatives. This yields a system of ODEs: $\dot{x} = f(x)$ with $x \in \mathbb{R}^n$. We chose two values $c = 0$ and $c = 1$. The first gives rise to a linear system and the second to a nonlinear system, both being dissipative. In both cases the Jacobian is a tridiagonal matrix (hence is banded with $b = 2$). We chose a system of size $n = 499$ and used a reduced order model of size $k = 50$. The following smooth initial condition was chosen:

$$x(0, s) = \exp\left(-\frac{(s-3)^2}{9}\right)\sin^2(\frac{\pi s}{2}), \quad s \in [0, 6].$$

The time interval of simulation was $[0, 5]$. We simulated the systems with two different MATLAB solvers: `ode23` (explicit) and `ode15s` (implicit). The cost of the computation is shown in Table 6.7.

The asymptotic complexity of function evaluations for this example is given by

$$f(n) = Cn,$$

Table 6.7

Computational cost in 10^3 flops: One dimensional PDE example, full model and unsimplified reduced model with $k = 50$.

	Full model	Reduced model (unsimplified, $k = 50$)
Linear explicit	149, 150	69, 527
Nonlinear explicit	182, 310	1, 342, 400
Linear implicit	1, 732	3, 508
Nonlinear implicit	2, 089	30, 458

Table 6.8

Computational savings factor: One dimensional PDE example, $k = 50$, unsimplified $\hat{f}$.

	Observed savings	Asymptotic theoretical savings
Linear explicit	2.1	0.6
Nonlinear explicit	0.14	0.048
Linear implicit	0.49	0.49
Nonlinear implicit	0.069	0.002

where the constant $C = 2(b + 1) = 6$ for the linear case and $C = 10$ for the nonlinear case. The complexity of the nonlinear reduced model function evaluations is

$$\hat{f}(k, n) = Cn + 4nk,$$

when $\rho f(\rho^T z + \bar{x})$ is not analytically simplified.

The computational savings achieved and the theoretical asymptotic values are compared in Table 6.8. With the exception of the nonlinear implicit case, the asymptotic theory is within an order of magnitude of the observed values. For the same reasons as in the dense Jacobian case, one cannot expect the theory to be very accurate. In addition, in the banded Jacobian case there are other factors. Since the costs associated with overhead such as stepsize control as well as the rest of the computational costs grow linearly with system size for the unreduced systems, it is no longer valid to neglect the cost of such overhead even asymptotically. Hence the theory underestimates the cost for the unreduced case. This basically explains why the savings were better than predicted by theory. Another reason for the observed discrepancies is that MATLAB (version 5) does not take advantage of the banded structure of the Jacobian. It uses only the sparsity pattern. As a result, the cost of LU decomposition and triangular system solves is greater than that predicted by the theory.

It must be noted that one reason why there are no computational savings is due to the fact that the cost of function evaluation is significantly worse for the reduced model in the nonlinear case: $\hat{f} \approx 10n + 4nk = 210n \gg 10n \approx f(n)$. This was with no expression simplification applied to $\rho f(\rho^T z + \bar{x})$. If we simplify the expression and treat it as a dense cubic, the cost of evaluation (for the $k = 50$ case) is $\hat{f} = 2342500 \gg 104790 = 217n$, which is even worse. This is because the cubic nonlinearity in f is diagonal in the original x coordinates, while the simplified expression for $\hat{f}(z)$ is (typically) a dense cubic. The cost of evaluating a dense cubic $\mathbb{R}^k \to \mathbb{R}^k$ is of order $\frac{k^4}{3}$ and can be prohibitively large if k is not sufficiently small. However, if we use a smaller reduced order model of size $k = 6$, which in this example preserves all of the energy of the solution trajectory up to eight digit accuracy, we indeed get significant computational savings. For $k = 6$, the cost of function evaluation (assuming a dense cubic) is only ≈ 996 flops. Table 6.9 shows the savings factor for $k = 6$ when the simplified expression for nonlinear $\hat{f}$ is used and compares this with the theoretical asymptotic values. The theoretical values are within an order of magnitude

Table 6.9
Computational savings factor: One dimensional PDE example, $k = 6$, simplified nonlinear $\hat{f}$.

	Observed savings	Asymptotic theoretical savings
Nonlinear explicit	860	105
Nonlinear implicit	10.9	1.75

of the observed values. The discrepancies are again due to various factors.

This example illustrates how a multitude of factors affects the computational costs of using a POD reduced model in place of the full original model.

7. Conclusions. We investigated some basic properties of the POD method in finite dimensions. We provided an analysis of the errors involved in computing the solution of a nonlinear ODE initial value problem using a POD reduced order model. In addition to providing quantitatively reasonable error estimates, the analysis also explains the various factors that affect the error.

We also provided a sensitivity analysis of the POD method. We introduced the POD sensitivity factor which was a nondimensional measure of the sensitivity of the resulting projection with respect to perturbations in the data. We studied the effect of data perturbation on the projected data as well as the reduced model solution of the POD method. The POD sensitivity factor is relevant in some applications of POD while it is not in some other applications. We provided a discussion of this issue.

In addition to the error and sensitivity analysis, we also provided an analysis of computational complexity in using the POD reduced model in computing the solution to an ODE initial value problem. Our analysis showed that the computational savings achieved by POD depend on several factors and that the complexity of the nonlinear function evaluations can significantly affect the savings that might be gained by the use of a POD reduced model. Our examples suggest that combining expression simplification with reduced order models (for the class of polynomial vector-fields) may achieve significant savings if the reduced models are small enough.

Acknowledgment. We would like to thank the reviewers for their constructive comments.

REFERENCES

[1] R. Abraham, J. E. Marsden, and T. Ratiu, *Manifolds, Tensor Analysis and Applications*, 2nd ed., Springer-Verlag, New York, 1988.

[2] V. Algazi and D. Sakrison, *On the optimality of Karhunen-Loève expansion*, IEEE Trans. Inform. Theory, 15 (1969), pp. 319–321.

[3] N. Aubry, W. Lian, and E. Titi, *Preserving symmetries in the proper orthogonal decomposition*, SIAM J. Sci. Comput., 14 (1993), pp. 483–505.

[4] G. Berkooz and E. Titi, *Galerkin projections and the proper orthogonal decomposition for equivariant equations*, Phys. Lett. A, 174 (1993), pp. 94–102.

[5] C. Bischof, A. Carle, P. Hovland, P. Khademi, and A. Mauer, *Adifor* 2.0 *Users' Guide (Revision D)*, Technical report CRPC-95516-S, Center for Research on Parallel Computation, Houston, TX, 1995.

[6] W. Boothby, *An Introduction to Differentiable Manifolds and Riemannian Geometry*, 2nd ed., Academic Press, New York, 1994.

[7] Y. Cao and L. R. Petzold, *A Note on Model Reduction for Analysis of Cascading Failures in Power Systems*, manuscript.

[8] E. A. Christensen, M. Brøns, and J. N. Sørensen, *Evaluation of proper orthogonal decomposition–based decomposition techniques applied to parameter-dependent nonturbulent flows*, SIAM J. Sci. Comput., 21 (2000), pp. 1419–1434.

[9] A. Curtis, M. Powell, and J. Reid, *On the estimation of sparse Jacobian matrices*, J. Inst. Math. Appl., 13 (1974), pp. 117–120.

[10] S. Glavaski, J. E. Marsden, and R. Murray, *Model reduction, centering, and the Karhunen-Loève expansion*, in Proceedings of the IEEE Conference on Decision and Control, Tampa, FL, 1998, pp. 2071–2076.

[11] G. Golub and C. Van Loan, *Matrix Computations*, Johns Hopkins University Press, Baltimore, MD, 1996.

[12] M. Graham and I. Kevrekidis, *Alternative approaches to the Karhunen-Loève decomposition for model reduction and data analysis*, Computers and Chemical Engineering, 20 (1996), pp. 495–506.

[13] E. Hairer, S. Norsett, and G. Wanner, *Solving Ordinary Differential Equations* I: *Nonstiff Problems*, Springer-Verlag, New York, 1980.

[14] P. Holmes, J. Lumley, and G. Berkooz, *Turbulence, Coherent Structures, Dynamical Systems and Symmetry*, Cambridge University Press, Cambridge, UK, 1996.

[15] S. Lall, P. Krysl, and J. E. Marsden, *Structure-preserving model reduction for mechanical systems*, Phys. D, 184 (2003), pp. 304–318.

[16] S. Lall, J. E. Marsden, and S. Glavaski, *Empirical model reduction of controlled nonlinear systems*, in Proceedings of the IFAC World Congress, Beijing, 1999, pp. 473–478.

[17] B. Moore, *Principal component analysis in linear systems: Controllability, observability, and model reduction*, IEEE Trans. Automat. Control, 26 (1981), pp. 17–31.

[18] P. Parrilo, F. Paganini, G. Verghese, B. Lesieutre, and J. E. Marsden, *Model reduction for analysis of cascading failures in power systems*, in Proceedings of the American Control Conference, San Diego, CA, 1999, pp. 4028–4212.

[19] K. Pearson, *On lines and planes of closest fit to a system of points in space*, Philosophical Magazine, 2 (1833), pp. 609–629.

[20] M. Rathinam and L. R. Petzold, *Dynamic iteration using reduced order models: A method for simulation of large scale modular systems*, SIAM J. Numer. Anal., 40 (2002), pp. 1446–1474.

[21] M. Rathinam and L. Petzold, *An iterative method for simulation of large scale modular systems using reduced order models*, in Proceedings of the IEEE Conference on Decision and Control, Sydney, Australia, 2000, pp. 4630–4635.

[22] A. Rosenfeld and A. Kak, *Digital Picture Processing*, Academic Press, New York, 1982.

[23] C. Rowley and J. E. Marsden, *Reconstruction equations and the Karhunen-Loève expansion for systems with symmetry*, Phys. D, 142 (2000), pp. 1–19.

[24] L. F. Shampine and M. W. Reichelt, *The MATLAB ODE suite*, SIAM J. Sci. Comput., 18 (1997), pp. 1–22.

[25] S. Shvartsman and I. Kevrekidis, *Low-dimensional approximation and control of periodic solutions in spatially extended systems*, Phys. Rev. E(3), 58 (1998), pp. 361–368.

[26] S. Shvartsman, C. Theodoropoulos, R. Rico-Martinez, I. Kevrekidis, E. Titi, and T. Mountziaris, *Order reduction for nonlinear dynamic models of distributed reacting systems*, J. Process Control, 10 (2000), pp. 177–184.

[27] L. Sirovich, *Turbulence and the dynamics of coherent structures*, I, Quart. Appl. Math., 45 (1987), pp. 561–571.

[28] L. Sirovich, *Turbulence and the dynamics of coherent structures*, II, Quart. Appl. Math., 45 (1987), pp. 573–582.

[29] L. Sirovich, *Turbulence and the dynamics of coherent structures*, III, Quart. Appl. Math., 45 (1987), pp. 583–590.

SIAM J. Numer. Anal.
Vol. 41, No. 5, pp. 1926–1947

CONSERVATIVE FRONT TRACKING WITH IMPROVED ACCURACY*

JAMES GLIMM[†][‡], XIAOLIN LI[†], YINGJIE LIU[†], ZHILIANG XU[†], AND NING ZHAO[§]

Abstract. We propose a fully conservative front tracking algorithm for systems of nonlinear conservation laws. The algorithm improves by one order in its convergence rate over most finite difference schemes. Near tracked discontinuities in the solution, the proposed algorithm has $\mathcal{O}(\Delta x)$ errors, improving over $\mathcal{O}(1)$ errors commonly found near a discontinuity. Numerical experiments which confirm these assertions are presented.

Key words. front tracking, conservation, contact discontinuity

AMS subject classifications. 35L65, 74S10

DOI. 10.1137/S0036142901388627

1. Introduction.

We propose and demonstrate a tracking finite difference algorithm for the problem of nonlinear conservation laws which is (a) fully conservative and (b) improves the local error by one power of Δx near tracked discontinuities. The one dimensional (1D) version of these ideas was presented in [9], and a preliminary (but different) two dimensional (2D) algorithm with the same properties was given in [8], while the results were announced in [10].

Discontinuities in the solution of systems of nonlinear hyperbolic conservation laws are a primary difficulty for numerical simulation. These equations have both linear and nonlinear discontinuities, and (perhaps counterintuitively) the former are more difficult. Nonlinear discontinuities are self-focusing, and their numerical solution does not grow in width with time. The linear discontinuities in contrast do grow and may typically occupy 4 to 10 mesh cells in width.

Front tracking was introduced to give special treatment to discontinuities. A robust validated code has been developed and used in production simulation of fluid instabilities [5, 7, 6, 4]. See also the URL http://www.ams.sunysb.edu/∼shock/FTdoc/FTmain.html.

In this paper, we address an algorithmic issue—formulation of a conservative tracking algorithm. In its original formulation, conservation was enforced only in regular grid cells, those not cut by the tracked front. The missing points of the computational stencil, in the case of a front cutting through the stencil, are filled in as ghost cells, with the state values obtained by extrapolation from nearby front

*Received by the editors May 1, 2001; accepted for publication (in revised form) April 2, 2003; published electronically November 19, 2003. The research of the first author was supported by the Applied Mathematics Subprogram of the U.S. Department of Energy DE-FG02-90ER25084, the Department of Energy Office of Inertial Fusion, the Army Research Office, grant DAAL-03-91-0027, the National Science Foundation, grant DMS-0102480, and Los Alamos National Lab, sponsor ID 26730001014L. The research of the third author was supported by the U.S. Department of Energy, sponsor ID DEFG0398DP. The research of the fourth author was supported by the U.S. Department of Energy, sponsor ID DEFG0290ER and Los Alamos National Lab, sponsor ID 26730001014L.

http://www.siam.org/journals/sinum/41-5/38862.html

[†]Department of Applied Mathematics and Statistics, SUNY at Stony Brook, Stony Brook, NY 11794-3600 (glimm@ams.sunysb.edu, linli@ams.sunysb.edu, yingjie@ams.sunysb.edu).

[‡]Center for Data Intensive Computing, Brookhaven National Laboratory, Upton, NY 11793-5000.

[§]Department of Aerodynamics, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, P.R. China. The research of this author was supported in part by the National Science Foundation of China, grant 10072028.

states of the same component. Thus the state values are double-valued near the front, with the left-component states extending by extrapolation for a small distance into the right component, and vice versa. The use of ghost cell states was introduced into front tracking in 1980 [11]. With the ghost states thus defined, the interior solver follows a conventional finite difference algorithm.

The algorithm proposed in the present study is conservative for all grid cells, including the irregular ones cut by the front. This algorithm presented is related to earlier work of Swartz and Wendroff [18], Harten and Hyman [14], Chern and Colella [2], and Pember et al. [16] but differs from these works in several ways. Chern and Colella and Pember et al. redistribute mass from small cells to nearby large ones to preserve stability and conservation. This issue is addressed here by merging small cells. Swartz and Wendroff discussed only the 1D algorithm. Pember et al. [16] reviews these earlier works in 1995. We emphasize here tracking of a contact rather than the shock tracking of [2].

2. Conservative tracking. Consider the 1D system of conservation laws

$$\frac{\partial u}{\partial t} + \bigtriangledown \cdot f(u) = 0. \tag{1}$$

Weak or discontinuous solutions of this equation are not unique, and the equation must be supplemented by an entropy condition [17]. In the case of discontinuities, the partial derivatives in (1) are not defined, and Rankine–Hugoniot conditions

$$n \cdot ([f] - v[u]) = 0 \tag{2}$$

apply. Here $[A] = A_+ - A_-$ is the jump in the quantity across the interface, v is the velocity of the interface, and n is a unit normal to the interface. In fact, (2) results from (1) if the derivatives in (1) are interpreted in the sense of distributions. Representing (1) in integral form, for a moving discontinuity surface S bounding a time-dependent volume V, we have

$$\frac{\partial}{\partial t}\int_V u dV + \int_S n \cdot (f(u) - vu) dS = 0. \tag{3}$$

Thus $n \cdot (f - vu)$ is the dynamic flux, which replaces the usual flux f for the time-independent surface.

The essence of the new algorithm introduced here is to track the front in space and time, based on the following three steps:

1. Construction of the space-time interface to follow the moving solution discontinuity. This will follow the grid-based construction [7] and extend it to space-time.
2. Construction of space-time finite volume cells, starting as a partition of a regular space-time cell. The cells cut by the space-time interface are defined as irregular. To ensure an adequate Courant–Friedrichs–Lewy (CFL) restriction, portions of such irregular cells with too small a top (at t_{n+1}) or no top at all are merged with neighbor cells.
3. Godunov-type finite volume differencing with limiters to ensure continuity of the dynamic flux (3), so that the algorithm is conservative on a cell by cell basis.

To explain these steps at a more detailed but still simple level, we consider in one dimension an interface whose position at time t is $\sigma_e(t)$, and we assume a linear approximation $\sigma(t)$ to $\sigma_e(t)$ on $[t_n, t_{n+1}]$. The 1D algorithm is divided into two cases. We consider only the first case, in which the cell merger from step 2 above is not required. We assume that the approximate interface does not cross a mesh cell center within the time interval $[t_n, t_{n+1}]$. Thus for some mesh index i, $x_i \leq \sigma(t_n), \sigma(t_{n+1}) \leq x_{i+1}$. We displace the cell boundary located at $x_{1+1/2}$ to the interface location. This change results in a redefinition of the cell average quantity, to yield

$$U_i^m = (\Delta_{i,e}^m)^{-1} \int_{x_{i-1/2}}^{\sigma_e(t_m)} u(x, t_m) dx, \tag{4}$$

$$U_{i+1}^m = (\Delta_{i+1,e}^m)^{-1} \int_{\sigma_e(t_m)}^{x_{i+3/2}} u(x, t_m) dx \tag{5}$$

for $m = n, n+1$, where $\Delta_{j,e}^m$ is the interval over which U_j^m is averaged.

Denote by $\mathcal{U}_i^m$ on $[x_{i-1/2}, \sigma(t_m)]$ and $\mathcal{U}_{i+1}^m$ on $[\sigma(t_m), x_{i+3/2}]$ the numerical approximations of U_i^m and U_{i+1}^m, respectively, and let Δ_j^m be the interval over which $\mathcal{U}_j^m$ is averaged. Integrating the hyperbolic system over the two trapezoidal regions $[x_{i-1/2}, \sigma(t)] \times [t_n, t_{n+1}]$ and $[\sigma(t), x_{i+3/2}] \times [t_n, t_{n+1}]$, the finite difference equation for irregular cells is replaced by

$$\Delta_i^{n+1} \mathcal{U}_i^{n+1} = \Delta_i^n \mathcal{U}_i^n - \Delta t \{\mathcal{F}_{\text{int}}^{n+1/2} - \mathcal{F}_{i-1/2}^{n+1/2}\}, \tag{6}$$

$$\Delta_{i+1}^{n+1} \mathcal{U}_{i+1}^{n+1} = \Delta_{i+1}^n \mathcal{U}_{i+1}^n - \Delta t \{\mathcal{F}_{i+3/2}^{n+1/2} - \mathcal{F}_{\text{int}}^{n+1/2}\}, \tag{7}$$

where $\mathcal{F}_{\text{int}}^{n+1/2}$ is the numerical approximation to the flux

$$F_{\text{int}}^{n+1/2} = \frac{1}{\Delta t} \int_{t_n}^{t_{n+1}} \left(f(u(\sigma_e(t), t)) - s_e u(\sigma_e(t), t) \right) dt \tag{8}$$

across the exact interface. Here $\sigma_e(t)$ and $s_e = d\sigma_e/dt$. The definition (8) gives equal values when evaluated on either side of the interface due to the Rankine–Hugoniot condition (2).

Let $s(t) = d\sigma/dt$ be the speed of the numerically tracked interface $\sigma(t)$. The choice of the numerical shock speed is discussed in [3]. Assume a smooth solution in the interior region excluding the tracked waves. Also we assume that the Riemann solution associated with (1) depends Lipschitz-continuously on the left and right states which define the Riemann problem. Using a second order monotonic upstream-centered scheme for conservation law (MUSCL) reconstruction, we first reconstruct a piecewise linear function on each cell out of the cell averages at $t = t_n$ to yield the approximate left and right states $\mathcal{U}_l^n, \mathcal{U}_r^n$ at $\sigma_e(t_n)$ so that $\mathcal{U}_l^n - u(\sigma_e(t_n)-, t_n) = \mathcal{O}(\Delta x^2)$ and $\mathcal{U}_r^n - u(\sigma_e(t_n)+, t_n) = \mathcal{O}(\Delta x^2)$. Solving the Riemann problem with the above two approximate states, we obtain a shock speed s^n which satisfies $s^n - s_e(t_n) = \mathcal{O}(\Delta x^2)$. Therefore, the approximate tracked interface position at $t = t_n + \frac{1}{2}\Delta t$ is

$$\sigma^{n+1/2} = \sigma(t_n) + \frac{1}{2}\Delta t \cdot s^n = \sigma_e(t_{n+1/2}) + \mathcal{O}(\Delta t^2).$$

Using a Taylor expansion, we reconstruct the approximate left and right states $\mathcal{U}_l^{n+1/2}$, $\mathcal{U}_r^{n+1/2}$ at $(\sigma^{n+1/2}, t_{n+1/2})$ from the MUSCL reconstruction so that

$$\mathcal{U}_l^{n+1/2} - u(\sigma_e(t_{n+1/2})-, t_{n+1/2}) = \mathcal{O}(\Delta x^2) \tag{9}$$

and

$$\mathcal{U}_r^{n+1/2} - u(\sigma_e(t_{n+1/2})+, t_{n+1/2}) = \mathcal{O}(\Delta x^2). \tag{10}$$

Finally, solving a Riemann problem with the left and right states $\mathcal{U}_l^{n+1/2}$ and $\mathcal{U}_r^{n+1/2}$, we obtain the half time step shock speed $s^{n+1/2} = s_e(t_{n+1/2}) + \mathcal{O}(\Delta t^2)$, and the new two sides states $\mathcal{U}_{l1}^{n+1/2}$ and $\mathcal{U}_{r1}^{n+1/2}$ across the interface we want to track. Since the exact solution is smooth near the interface, the new states still satisfy (9) and (10). This construction gives a local error $\mathcal{O}(\Delta x^3)$ for the propagated shock position

$$\sigma_e^{n+1} = \sigma^n + s^{n+1/2}\Delta t + \mathcal{O}(\Delta t^3).$$

In fact,

$$\begin{aligned} \sigma_e^{n+1} - \sigma^n &= \int_{t_n}^{t_{n+1}} s_e(t)dt \\ &= \int_{t_n}^{t_{n+1}} [s_e(t_{n+1/2}) + s_e'(t_{n+1/2})(t - t_{n+1/2}) + \mathcal{O}(\Delta t^2)]dt \\ &= s_e(t_{n+1/2})\Delta t + \mathcal{O}(\Delta t^3) \end{aligned} \tag{11}$$

to give the desired accuracy. Let the numerical flux across the tracked front associated with the Riemann problem defined by these two states be

$$\mathcal{F}_i^{n+1/2} = f(\mathcal{U}_{l1}^{n+1/2}, t_{n+1/2}) - s^{n+1/2}\mathcal{U}_{l1}^{n+1/2}.$$

This flux satisfies

$$\mathcal{F}_i^{n+1/2} = F_i^{n+1/2} + \mathcal{O}(\Delta x^2)$$

and is continuous when evaluated from either side of the discontinuity.

The proof that this algorithm is conservative and (for one dimension only) improves its convergence rate near the tracked discontinuity by $\mathcal{O}(\Delta x^2)$ is given in [9].

3. The 2D algorithm. Consider the two space dimensional system of conservation laws

$$\frac{\partial u}{\partial t} + \frac{\partial f(u)}{\partial x} + \frac{\partial g(u)}{\partial y} = 0, \tag{12}$$

defined in a spatial domain Ω. The discontinuities of u, assumed to lie on curves, are organized to form an *INTERFACE*, which is propagated from one time level to the next.

In the present study, we require at each time level that the *INTERFACE*s are topologically equivalent to a union of nonintersecting line segments or circles [13]. Thus we postulate that triple or multiple *CURVE* intersection points do not occur. Each *CURVE* is assigned an orientation which remains unchanged during the propagation of the *INTERFACE.* The discretized *CURVE* is piecewise linear and connected and composed of *BOND*s. Each *BOND* is a pair of *INTERFACE POINT*s or *POINT*s and (conceptually) the straight line segment joining them. Assume a decomposition of the plane by a rectangular grid with mesh spacing Δx, and assume the boundary $\partial\Omega$ of Ω lies on grid lines. If the *POINT*s are all on the interior of cell edges with at most one *POINT* occurring on the interior of any given grid cell edge, then the *INTERFACE* is called grid-based [7].

The front *POINT*s are propagated through the Riemann solutions in the normal direction followed by a tangential sweep to update the states on the front. Propagation [5, 7, 6, 4] of the *POINT*s of a grid-based *INTERFACE* will yield a general *INTERFACE*, not grid-based, as there is no reason for a propagated *POINT* to lie on a grid cell edge just because it starts on one. The general idea of the grid-based construction is as follows: we consider this propagated *INTERFACE* as a collection of polygonal *CURVE*s in $\Re^2$. Crossing points of the *CURVE* with grid cell edges are inserted as new *POINT*s. The propagated old *POINT*s (named images of propagation in this sense) will be deleted. The *CURVE* is then reconstructed as straight line segments joining these new *POINT*s. In this process, the *CURVE* is displaced by an amount $\mathcal{O}(\Delta x^2)$, assuming that the *CURVE* is smooth, so that all angles between neighboring *BOND*s are $\mathcal{O}(\Delta x)$. See also [15, 7, 6] for detailed discussions of the grid-based INTERFACE construction.

Let B_i^n be a *BOND* on the grid-based *INTERFACE* $\mathcal{I}^n$ at the old time step t^n, and let $\bar{B}_i^{n+1}$ be the image *BOND* after the propagation of the end *POINT*s of B_i^n. A new grid-based *INTERFACE* $\mathcal{I}^{n+1}$ is reconstructed through the new *POINT*s which are produced by the intersection of the image *BOND*s and the gridline segments. Therefore, each new *POINT* P_i^{n+1} corresponds to an old *BOND* B_i^n, but the inverse is not true, because some bonds will not intersect with any gridline segment. On the other hand, since an image *BOND* may have several intersections with different gridline segments, several new *POINT*s may correspond to a single *BOND* B_i^n.

The finite difference method presented here for (12) is an explicit finite volume integration scheme. The spatial domain Ω has two dimensions. The solution of u evolves with respect to time, and we treat the temporal dimension as the third dimension. We join the spatial *INTERFACE*s at two consecutive time steps to form a space-time interface. Assume we have a space-time discretization $\{\mathcal{V}_i\}$ which conforms to the space-time interface as u evolves in one time step from time t_n to t_{n+1}. We solve (12) explicitly in this region. Treating each $\mathcal{V}_i$ as a control volume, we integrate (12) over $\mathcal{V}_i$. By the divergence theorem, we have

$$|\mathcal{V}_i(t_{n+1})|\bar{u}\mid_{t_{n+1}}= |\mathcal{V}_i(t_n)|\bar{u}\mid_{t_n} - \int_{\partial\mathcal{V}_i}(u, f(u), g(u))\cdot ndS, \tag{13}$$

where $\bar{u}\mid_{t_{n+1}}= \frac{1}{|\mathcal{V}_i(t_{n+1})|}\int_{|\mathcal{V}_i(t_{n+1})|} u(x, y, t_{n+1})dxdy$ is defined as a cell average, $|\mathcal{V}_i(t_{n+1})|$ is the face area of $\mathcal{V}_i(t_{n+1})$ at time t_{n+1}, and n is the outward normal to the space-time surfaces of $\mathcal{V}_i$. We wish to calculate $\bar{u}\mid_{t_{n+1}}$, the solution to (12) at time t_{n+1}.

The major issues in designing the conservative algorithm are (1) to obtain the space-time *INTERFACE*, (2) to determine the discretization $\{\mathcal{V}_i\}$, and (3) to calculate the fluxes defined on the space-time surfaces of $\mathcal{V}_i$.

To construct a finite volume decomposition which respects the space-time interface, we identify the crossings of the approximate space-time interface with the space-time hexahedron. We split the space-time hexahedron whose interior is cut by the space-time interface into parts, each of which belongs to only one side of the space time interface. For the purpose of maintaining numerical stability (the CFL time step restriction), we merge those cells with small top area to form a polyhedron with top area bigger than $0.5\Delta x^2$.

3.1. Construction of the space-time interface. In the current section, we solve the following problem: given two piecewise linear spatial grid-based *INTERFACE*s (*CURVE*s) which are separated in time by a step $\triangle t$, construct (triangulate) a surface joining them. We call this joining surface the space-time interface. The space-time interface thus formed is also grid-based, as a three-dimensional (3D) interface (two spatial and one temporal dimensions). The local configurations within a single grid cell for such a 3D grid-based interface have been discussed in [6]. We introduce two hypotheses regarding the old and new spatial interfaces. These hypotheses limit the local complexity of the interface. More complicated topological structures will not be included in the scope of this paper.

Hypothesis 1. *The INTERFACE is assumed to be grid-based. There is no topological change of the INTERFACE during the time interval of computation. Each CURVE is topologically equivalent to a line segment with its two end points on the boundary, or a circle contained in the interior of* Ω. *No CURVE is totally contained within a square of side* $2\Delta x$ *made up of four cells.*

Hypothesis 2. *The CFL number is less than* $\frac{1}{2}$ *so that each POINT of the INTERFACE is propagated a distance less than* $\frac{1}{2}\Delta x$ *within a single time step.*

Assume Hypothesis 1. For a grid-based *INTERFACE* $\mathcal{I}^n$, each *POINT* on $\mathcal{I}^n$ is a crossing *POINT*; there exists at most one crossing *POINT* on each grid cell edge. No crossing is deleted during the reconstruction of the grid-based *INTERFACE*, as such deletion would indicate a change of topology. Propagation of $\mathcal{I}^n$ *POINT*s at one single time step gives a new *INTERFACE* $\mathcal{I}_0^{n+1}$. The new grid-based *INTERFACE* $\mathcal{I}^{n+1}$ is reconstructed from $\mathcal{I}_0^{n+1}$ through the algorithm discussed above.

After the reconstruction of $\mathcal{I}_0^{n+1}$, the order of *POINT*s on the reconstructed *INTERFACE* $\mathcal{I}^{n+1}$ agrees with the natural order of the *POINT*s on $\mathcal{I}^n$ in the following sense: Let B_1^n be an $\mathcal{I}^n$ *BOND* connecting adjacent *POINT*s P_1 and P_2. Let B_2^n be an $\mathcal{I}^n$ *BOND* following B_1^n, connecting adjacent *POINT*s P_2 and P_3. After propagation, B_1^n is mapped onto an $\mathcal{I}_0^{n+1}$ linear segment B_1^{n+1} with a left end point M_1 as the image of P_1 and a right end point M_2 as the image of P_2; similarly, B_2^n is mapped onto B_2^{n+1} of $\mathcal{I}_0^{n+1}$ with a left end point M_2 and a right end point M_3 as the image of P_3. The reconstruction first inserts into $\mathcal{I}_0^{n+1}$ the crossing points of $\mathcal{I}_0^{n+1}$ with grid lines as new *POINT*s. The insertion of new *POINT*s does not change the orientation or order of the polygon $B_1^{n+1}\bigcup B_2^{n+1}$, which preserves its order from the polygon $B_1^n\bigcup B_2^n$. Similarly, the removal of *POINT*s, with the deformation of the polygon to connect with the remaining *POINT*s by linear segment, is order-preserving. Thus grid-based $\mathcal{I}^{n+1}$ reconstructed from $\mathcal{I}_0^{n+1}$ by connecting the new *POINT*s as above described preserves the *POINT* order.

For the grid-based method, every *INTERFACE POINT* lies on a cell edge. A *POINT P* is assigned on index (i,j) if it is located within a half grid size $(0.5\Delta x)$ away from the grid node (i,j). The proximity *Prox P* of *P* includes nine dual grid cells centered at grid node ($i-1$ to $i+1$, $j-1$ to $j+1$).

Assume P_1 and P_2 are the start and end *POINT*s of *BOND* B^{n+1} on $\mathcal{I}^{n+1}$; the (i,j) indices of these two *POINT*s can be identical, adjacent, or diagonally adjacent. The proximity *Prox B* of *BOND* B^{n+1} is defined as *Prox* $P_1 \cap$ *Prox* P_2. Therefore, the following hold.

(1) If (i,j) indices of P_1 and P_2 are identical, say, both are (i,j), *Prox B* is the nine dual grid cells centered at nodes ($i-1$ to $i+1$, $j-1$ to $j+1$).

(2) If (i,j) indices of P_1 and P_2 are adjacent, say, (i,j) for P_1 and $(i+1,j)$ for P_2, then *Prox B* includes the six dual grid cells centered at nodes (i to $i+1$, $j-1$ to $j+1$). *Prox B* for the case in which the second index of P_1 and P_2 differs by 1 is similarly defined.

(3) If (i,j) indices of P_1 and P_2 are diagonally adjacent, say, (i,j) for P_1 and $(i+1,j+1)$ for P_2, *Prox B* includes the four dual grid cells centered at nodes (i to $i+1$, j to $j+1$).

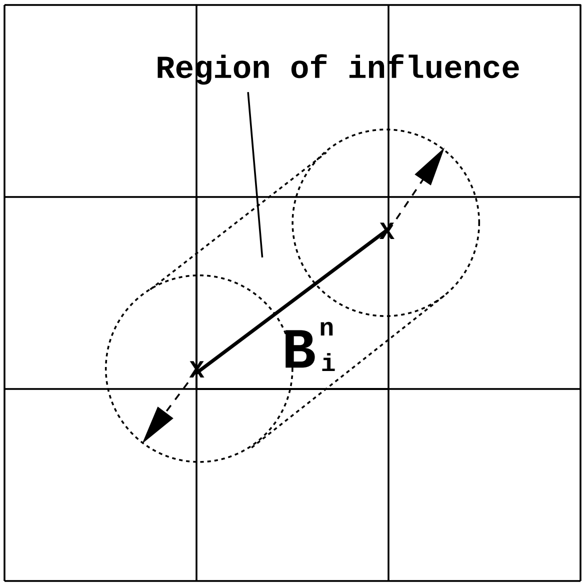


FIG. 1. *Region of influence of a bond B_i^n.*

A *POINT* of $\mathcal{I}^n$ which is located inside the proximity of B^{n+1} is called a *spatially nearest POINT* of B^{n+1} on $\mathcal{I}^n$. If a *BOND* B^{n+1} of $\mathcal{I}^{n+1}$ is the result of propagation followed by the grid-based reconstruction of a single *BOND* B^n on $\mathcal{I}^n$, then B^n is called the *parent BOND* and B^{n+1} is the *child BOND*. In this case, B^{n+1} is formed by the insertion of crossing *POINT*s into the propagated image of B^n. The *region of influence* of any *BOND* in Figure 1 is the region within $0.5\Delta x$ of the points on the *BOND*.

PROPOSITION 1. *Assume Hypotheses* 1, 2*. Let* P_1 *and* P_2 *be two adjacent POINTs connected by a reconstructed grid-based bond* B^{n+1} *of* $\mathcal{I}^{n+1}$*. If* P_1 *is produced (through propagation and intersection of a bond with a gridline segment) by* B_1^n *of* $\mathcal{I}^n$*,* P_2 *is produced by* B_2^n *of* $\mathcal{I}^n$*, and the curve on* $\mathcal{I}^n$ *is oriented so that* B_2^n *follows* B_1^n*, then there exists at least one POINT between the start of* B_1^n *and the end of* B_2^n *which lies within the proximity of* B^{n+1}*.*

Proof. Let C with corner nodes (i, j), $(i+1, j)$, $(i, j+1)$, and $(i+1, j+1)$ be the cell containing B^{n+1}. Assume $B_1^n \neq B_2^n$. In this case, to produce B^{n+1}, all the propagated *POINT*s between (including) the end of B_1^n and the start of B_2^n must lie in the cell C at the new time step t^{n+1}. By Hypothesis 2, all the corresponding old points must be in the proximity defined by P_1 and P_2, because the shortest distance from the boundary of the cell C to the boundary of the proximity is at least $0.5\Delta x$.

Next we assume $B_1^n = B_2^n$. In this case, B_1^n is the *parent BOND* of B^{n+1} and the entire B^{n+1} must be within the *region of influence* of B_1^n. Since the proximity of B^{n+1} is the intersection of the proximities of the two *POINT*s P_1 and P_2, it is the smallest when the indices of P_1 and P_2 are diagonally adjacent, a property we now assume. The proximity is the rectangle *ABCD* in Figure 2. We want to prove that at least one *POINT* of B_1^n is located within the rectangle *ABCD*. To prove this, we show that the parent *BOND* B_1^n cannot have both *POINT*s outside the proximity.

We now draw the boundary of the *region of influence* of all the grid-based bonds with both end *POINT*s outside the proximity of B^{n+1}. The inner boundary of this

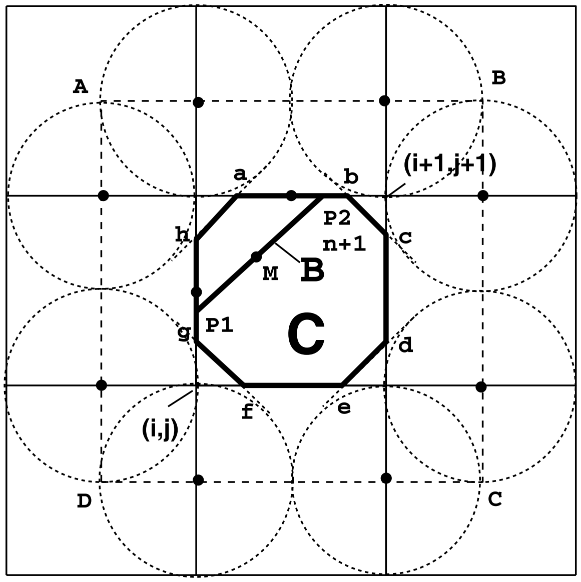


Fig. 2. *Region of influence of all bonds completely outside the proximity of B^{n+1} of which the (i,j) indices of B^{n+1} end points P_1 and P_2 are diagonally adjacent.*

region is the polygon **abcdefgh** as in Figure 2. If the parent *BOND* B_1^n has both end points outside the proximity, then B^{n+1}, lying in its region of influence, should be completely outside the polygon **abcdefgh**. Thus it must lie in one of the four small regions near one corner of the cell C. Since the polygon cuts the edges of the cell C at a distance $0.5(\sqrt{2}-1)\Delta x \approx 0.207\Delta x$ from the four cell corners, the mesh indices of the end *POINT*s of B^{n+1} cannot be diagonally adjacent. Therefore, no bond with both end *POINT*s outside the proximity $ABCD$ can be the *parent* bond of B^{n+1}. Therefore, at least one end point of B^n must be located within the proximity $ABCD$.

The other two cases, when the indices of P_1 and P_2 are identical or adjacent, have a larger proximity for B_1^n. Similar but easier arguments prove Proposition 1 in these cases. This completes the proof.

Proposition 2. *Assume Hypotheses* 1, 2. *Let B_1^{n+1} with end points P_1 and P_2, and B_2^{n+1} with end point P_2 and P_3 be two adjacent BONDs on $\mathcal{I}^{n+1}$ in their natural order. Let B_1^n, B_2^n, and B_3^n be the bonds on $\mathcal{I}^n$ which produce P_1, P_2, and P_3. Denote the spatially nearest POINTs to B_1^{n+1} on $\mathcal{I}^n$ as group* 1 *and the spatially nearest POINTs to B_2^{n+1} on $\mathcal{I}^n$ as group* 2. *There exist a POINT M_1 in group* 1 *and a POINT M_2 in group* 2 *such that* (1) *M_1 is a POINT between (including) the start of B_1^n and the end of B_2^n, and M_2 is a POINT between (including) the start of B_2^n and the end of B_3^n;* (2) *M_1 precedes or equals M_2 in the orientation of $\mathcal{I}^n$.*

Proof. If B_1^n, B_2^n, and B_3^n are three distinct bonds, we take M_1 as any *POINT* between (including) the end of B_1^n and the start of B_2^n and M_2 as any *POINT* between (including) the end of B_2^n and the start of B_3^n. This choice satisfies Proposition 2 in view of Proposition 1.

Next we consider the case $B_1^n = B_2^n \neq B_3^n$. We select M_1 as one of the end *POINT*s of B_1^n which is in group 1. Such a *POINT* exists by Proposition 1. M_2 can be selected between (including) the end of B_2^n to the start of B_3^n. M_1 and M_2 satisfy Proposition 2. The case $B_1^n \neq B_2^n = B_3^n$ is similar.

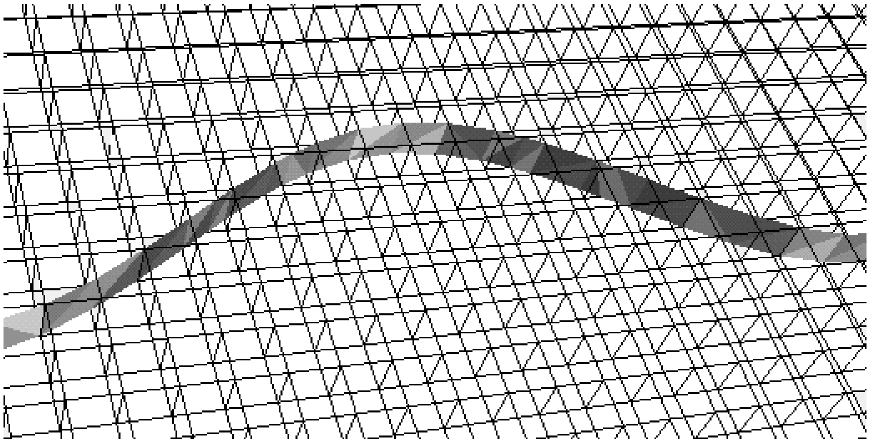


FIG. 3. *The triangulated space-time interface.*

Finally, we consider $B_1^n = B_2^n = B_3^n$. In this case, P_1, P_2, and P_3 lie on a straight line. It is obvious that Proposition 2 holds for the following three cases: (1) the start of B_1^n in both $Prox\ B_1^{n+1}$ and $Prox\ B_2^{n+1}$; (2) the end of B_1^n in both $Prox\ B_1^{n+1}$ and $Prox\ B_2^{n+1}$; and (3) the start of B_1^n in $Prox\ B_1^{n+1}$ and the end of B_1^n in $Prox\ B_2^{n+1}$.

We now prove that it is impossible to have the end of B_1^n in $Prox\ B_1^{n+1} \setminus Prox\ B_2^{n+1}$ and the start of B_1^n in $B_2^{n+1} \setminus Prox\ B_1^{n+1}$. For this to occur, the propagation of both the start and the end points of B_1^n must completely pass through the region of $Prox\ B_1^{n+1} \cap Prox\ B_2^{n+1}$. It is easy to verify that the widths of the intersection in both the x and y directions are at least Δx. However, the maximum propagation distance in one time step is $0.5\Delta x$. The proof is complete.

The surface triangles in space-time are formed by joining the *POINTs* of $\mathcal{I}^{n+1}$ and $\mathcal{I}^n$. Each triangle has a side taken from a single linear segment ($BOND$) of either $\mathcal{I}^{n+1}$ or $\mathcal{I}^n$ and an opposite $POINT$ from the other. We denote a space-time interface triangle which is composed of a $BOND$ at time t_{n+1} and an opposite $POINT$ from $\mathcal{I}^n$ as an upper triangle, and a triangle which is composed of a $BOND$ at time t_n and an opposite $POINT$ from $\mathcal{I}^{n+1}$ as a lower triangle. The space-time interface triangulation is organized into the following two steps:

1. We first form the upper triangles of the space-time interface. For each $\mathcal{I}^{n+1}$ $BOND\ B_m^{n+1}$ whose start and end *POINT*s are P_m and P_{m+1}, we find by Proposition 1 the spatially nearest *POINT*s on $\mathcal{I}^n$. Denote these *POINT*s as group m. Select one $POINT\ M_m$ from each group m to form the list $[M_1, M_2, M_3, \dots]$ with the same orientation as $\mathcal{I}^n$ (due to Proposition 2); M_i and M_{i+1} are not necessarily distinct. Connect each M_m to P_m and P_{m+1} to form upper triangles.
2. The gap on the space-time interface left by step 1 is filled by lower triangles. Each $BOND\ B_k^n$ on the $\mathcal{I}^n$ is located between a pair of distinct *POINT*s M_k and M_{k+1}. M_k and M_{k+1} are connected to a common $POINT$ on $\mathcal{I}^{n+1}$ during the construction of the upper triangles. Connect this common $POINT$ to the start and end *POINT*s of B_k^n to form a lower triangle. This completes the space-time interface triangulation.

Figure 3 shows the triangulated space-time interface. Each triangle is distinguished from its neighbors by contrasting grey shades.

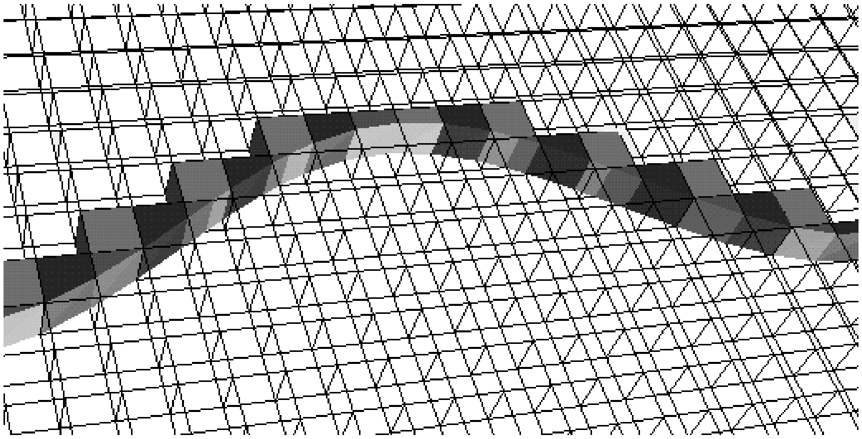


FIG. 4. *Hexahedra and partial polyhedra before volume merging.*

3.2. Construction of the space-time hexahedra. We connect the nodes of a cell D_i^n at time $t = t_n$ to the nodes of the corresponding cell D_i^{n+1} at time $t = t_{n+1}$ to form a space-time hexahedron. We call D_i^{n+1} the top of the hexahedron and D_i^n the bottom. We call a hexahedron *mixed* if the interface passes through its interior, otherwise, it is *pure*. The *mixed* hexahedra are divided into *pure partial* hexahedra, and if necessary, these are combined with neighbors to form a finite volume space-time grid suitable for construction of a conservative difference algorithm in section 3.3. They are adjacent if they share a nontrivial surface which does not meet the space-time interface. Two space-time polyhedra are neighboring if they share a nontrivial vertical line segment which is part of the grid line connecting two corresponding grid nodes at the time levels t_n and t_{n+1} (denoted by a vertical grid line) that does not cross the space-time interface. It is easy to see that two adjacent or neighboring polyhedra must be on the same side of the space-time interface.

The *mixed* hexahedron is separated by the space-time interface into several parts, each of which lies on one side of the space-time interface. These parts are called *pure partial* hexahedra or, in short, *partial* hexahedra. We can similarly define a cell to be *pure*, *mixed*, or *partial*. Any *partial* hexahedron with a trivial or small top will be merged with an adjacent *pure* hexahedron or *partial* hexahedron having a top of minimal size.

Figure 4 shows the control volumes constructed on one side of the space-time interface. Adjacent hexahedra or pure partial polyhedra are represented by contrasting grey shades. Only the volumes near the space-time interface are displayed.

Recalling that two adjacent hexahedra are on the same side of the interface, the following lemma [8] ensures the eventual success of the merging algorithm.

LEMMA 1. *Assume Hypothesis* 1. *If a space-time polyhedron is constructed by merging any number of adjacent partial hexahedra with no top, then the polyhedron will be adjacent to a pure or partial hexahedron on the same side of the space-time interface.*

Proof. At least one nontrivial piece of the side surface of the polyhedron is not on the boundary or the space-time interface; otherwise, the topological structure of the *INTERFACE* changes during this time step and Hypothesis 1 is violated. The proof is complete.

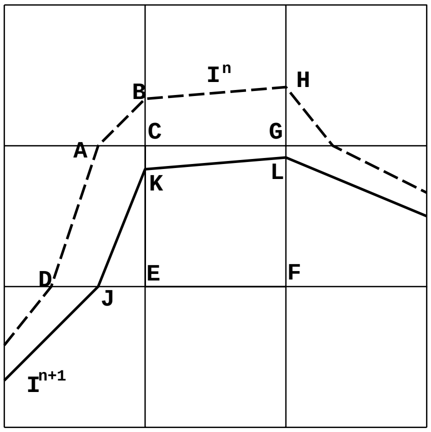


Fig. 5. *A top view of the polyhedra merging process. The solid line represents $\mathcal{I}^{n+1}$, and the dashed line represents $\mathcal{I}^n$. There are four polyhedra in the four upper left mesh blocks: polyhedron* 1 *with bottom ABC and no top, polyhedron* 2 *with bottom face $BHGC$ and no top, polyhedron* 3 *with bottom face $ACED$ and the triangular top KEJ, and polyhedron* 4 *with a square bottom $CGFE$ and the trapezoidal top $KLFE$. They will be merged into one polyhedron with bottom $ACBHGFED$ and top $KLFEJ$.*

Hypotheses 1 and 2 and Lemma 1 ensure that each partial hexahedron with no top and away from the boundary is adjacent to or neighboring one with a nontrivial top. However, for a partial hexahedron with no top and at the boundary, Hypothesis 2 may not be sufficient if the interface intersects the boundary at a small angle. We need to adjust the CFL number so that the intersection point between the *INTERFACE* and the boundary moves a distance less than Δx along the boundary during the time step in order to reach the same property.

We require a hypothesis to limit the local geometric complexity of the *INTERFACE*. To simplify the proof that the merging algorithm converges (rapidly), we state it in a stronger than necessary form. See section 3.4 for a discussion of this issue.

Hypothesis 3. *Each partial hexahedron having top area smaller than $\frac{1}{2}\Delta x^2$ is adjacent to or neighboring one with top area greater than or equal to $\frac{1}{2}\Delta x^2$.*

Because the flux exchange among control volumes is through the shared space-time surfaces, we merge only adjacent *partial* hexahedra on the same side of the space-time interface and not neighboring ones. For this reason, merger is accomplished in stages, i.e., recursively. The merging process then is stated as follows.

Assume Hypothesis 3. *Recursively merge every pure or partial hexahedron having a top area greater than or equal to $\frac{1}{2}\Delta x^2$ with adjacent partial hexahedra having no top or top area smaller than $\frac{1}{2}\Delta x^2$ which have not been merged elsewhere until none of the partial hexahedron having no top or top area smaller than $\frac{1}{2}\Delta x^2$ is left. Denote the resulting space-time polyhedra the big hexahedra. The merging process then is complete. Partial polyhedra generated at each merging stage are called intermediate hexahedra.*

As illustrated in Figure 5, polyhedron 4 with a square bottom face $CGFE$ and top face area $KLFE$ greater than $\frac{1}{2}\Delta x^2$ forms the center of merging. The merged polyhedra include polyhedron 3 with bottom face $ACED$ and a small triangular top KEJ, polyhedron 2 with bottom face $BHGC$ and no top, and polyhedron 1 with bottom face ABC and no top. Polyhedra 2 and 3 are adjacent to 4, while polyhedron

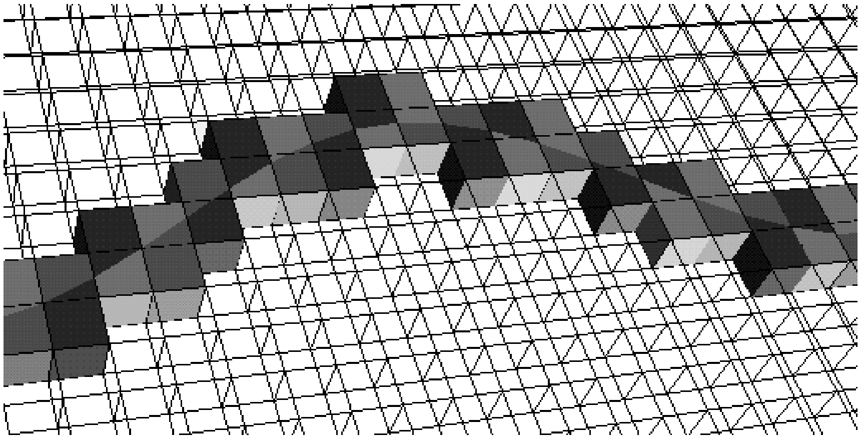


FIG. 6. *Control volumes after merging.*

1 is diagonally adjacent to 4. In the merging process, polyhedra 2 and 3 are absorbed by polyhedron 4 first. A *intermediate* hexahedron with bottom face $ACBHGFED$ is formed. Polyhedron 1 is adjacent to it. Finally, this *intermediate* hexahedron absorbs polyhedron 1, resulting in a *big* hexahedron with bottom face $ABHGFED$ and top face $KLFEJ$.

Determined by Lemma 1 and Hypothesis 3, it is easy to see that a *big* hexahedron contains no more than a fixed number of *pure* or *partial* hexahedra so that the merging process stops rapidly. Actually in most cases the merging process yields *big* hexahedra consisting of two *pure* or *partial* hexahedra. The number of *pure* or *partial* hexahedra in the *big* hexahedron could become larger if the radius of curvature of the moving *CURVE* is small. In fact, we have the following observation.

Assume Hypothesis 3. Let a *pure* or *pure partial* hexahedron $\mathcal{H}$ with top area greater than $\frac{1}{2}\Delta x^2$ be contained inside a space-time cell with cell index (i, j). If $\mathcal{H}$ forms a *big* hexahedron by absorbing *pure partial* hexahedra during the merging process, the bottom faces of these *pure partial* hexahedra which merge with $\mathcal{H}$ are located inside a square, centered at (i, j), with side $3\Delta x$.

Figure 6 shows the control volumes on two sides of the space-time interface after the merging process. Only the volumes near the space-time interface are displayed.

THEOREM 1. *Assume Hypotheses* 1–3*. After the merging algorithm, every partial hexahedron having no top or top area smaller than* $\frac{1}{2}\Delta x^2$ *will be merged into a big hexahedron having top area greater than or equal to* $\frac{1}{2}\Delta x^2$ *on the same side of the interface. The interior of each big hexahedron is connected.*

3.3. The reconstruction, limiter, and numerical scheme. Suppose at the time level $t = t_n$ we know the approximate state averages on each cell, *regular*, *irregular*, or *partial*. We want to reconstruct a piecewise linear state function on these cells with second order accuracy. The reconstruction of the piecewise linear state function on *irregular* cells follows [1], with modifications to the limiter and some simplification. Let D_i^n be a *pure* cell, *regular*, *irregular*, or *partial*, with approximate state average $\mathcal{U}_i$ and cell center (centroid) Y_i, surrounded by any of the types of cells $D_j^n, D_k^n, D_l^n, D_m^n$ with approximate state averages $\mathcal{U}_j^n, \mathcal{U}_k^n, \mathcal{U}_l^n, \mathcal{U}_m^n$ and cell centers Y_j, Y_k, Y_l, Y_m, respectively, on the same side of the *INTERFACE* . Let $\tilde{\mathcal{U}}_i = \mathcal{U}_i + (a, b) \cdot$

$(X - Y_i)$ be the second order accurate linear state function on D_i^n, where a, b are two constants. Choose any two surrounding cells, say, D_j^n, D_k^n so that Y_i, Y_j, Y_k are not colinear. We can determine a, b by solving the following equation:

$$\begin{aligned} \tilde{\mathcal{U}}_i(Y_j) &= \mathcal{U}_j^n, \\ \tilde{\mathcal{U}}_i(Y_k) &= \mathcal{U}_k^n. \end{aligned} \tag{14}$$

Further, for the solution of the above equation to be well conditioned, we require the angle θ formed by line segments $\overline{Y_iY_j}$ and $\overline{Y_iY_k}$ to satisfy $0 < \theta_1 < \theta < \theta_2 < \pi$, where θ_1, θ_2 are two constants. We repeat the above procedure until we find all possible solutions, say, a_i, b_i, for all $0 \leq i \leq I$, where $I \leq 4$. Then we set $a = \text{minmod}\{a_1, \ldots, a_I\}$ and $b = \text{minmod}\{b_1, \ldots, b_I\}$. When there are not enough surrounding cells on the same side of the *INTERFACE*, we choose $a, b = 0$ so that the reconstruction becomes first order.

When D_i^n is a *regular* cell surrounded by *regular* cells, the reconstruction process is simpler. Let the cell center of D_i^n be $(i_1\Delta x, i_2\Delta y)$ with neighboring cell centers $\{((i_1 + k_1)\Delta x, (i_2 + k_2)\Delta y) | k_1, k_2 = -1, 0, 1\}$. Let

$$\begin{aligned} \text{xslope}_\text{i} = \ & \text{minmod}\{[\mathcal{U}((i_1 + k_1)\Delta x, (i_2 + k_2)\Delta y) \\ & -\mathcal{U}((i_1 + k_1 - 1)\Delta x, (i_2 + k_2)\Delta y)]/\Delta x \mid \\ & k_1 = 0, 1; k_2 = -1, 0, 1\}, \end{aligned} \tag{15}$$

and

$$\begin{aligned} \text{yslope}_\text{i} = \ & \text{minmod}\{[\mathcal{U}((i_1 + k_1)\Delta x, (i_2 + k_2)\Delta y) \\ & -\mathcal{U}((i_1 + k_1)\Delta x, (i_2 + k_2 - 1)\Delta y)]/\Delta y \mid \\ & k_1 = -1, 0, 1; k_2 = 0, 1\}, \end{aligned} \tag{16}$$

and define

$$\tilde{\mathcal{U}}_i = \mathcal{U}_i + \text{xslope}_\text{i} \cdot (x - i_1\Delta x) + \text{yslope}_\text{i} \cdot (y - i_2\Delta y).$$

This second order reconstruction is better suited in multiple dimensions than in the operator splitting single line reconstruction (or limiter) for a uniform rectangular grid because, for example, an untracked discontinuity in two dimensions may be in the form of a strip of width between $2\Delta x$ and $3\Delta x$. When the strip is almost parallel to and fully covers the line in which the single line reconstruction occurs, one cannot expect the limiter to choose any smooth solutions nearby.

Next we apply the technique of section 3.2 to generate space-time hexahedra between time levels t^n and t^{n+1}. Let H be a big hexahedron with top D^{n+1}, bottom D^n, and triangle sides $\{S_i\}$ with a unit outer normal n_i and centroid Z_i. Notice that some elements of the $\{S_i\}$ may be on the approximate space-time interface. Integrating (12) over H, we obtain

$$|D^{n+1}| \ U^{n+1} = |D^n| \ U^n - \sum_i \int_{S_i} (u, f, g) \cdot n_i ds. \tag{17}$$

Here $|D^n|$ represents the area of D^n, and similarly $|S_i|$ is the area of S_i.

The numerical scheme can be written as

$$|D^{n+1}|\mathcal{U}^{n+1} = |D^n|\mathcal{U}^n - \sum_i |S_i|(\tilde{\mathcal{U}}_{i,m}, f(\tilde{\mathcal{U}}_{i,m}), g(\tilde{\mathcal{U}}_{i,m})) \cdot n_i. \tag{18}$$

The fluxes through triangle sides $\{S_i\}$ can be calculated by a higher order Godunov-type algorithm.

We first calculate $\tilde{\mathcal{U}}_{i,m}$ as follows: First use a Cauchy–Kowalewski procedure on the reconstructed state function on each side of S_i to get second order approximate states at Z_i on the respective sides of S_i, say, $\mathcal{U}_{i,l}$ and $\mathcal{U}_{i,r}$. If S_i is not on the tracked space-time interface, we can simply use a Riemann solver, say, R, to get the middle state on S_i, i.e.,

$$\tilde{\mathcal{U}}_{i,m} = R(\mathcal{U}_{i,l},\ \mathcal{U}_{i,r}).$$

If S_i is on the tracked space-time interface, we use the Riemann solver to get the left and the right side states $\tilde{\mathcal{U}}_{i,l}$ and $\tilde{\mathcal{U}}_{i,r}$ on the wave we are supposed to track and the wave speed ν_i. Then $\tilde{\mathcal{U}}_{i,m}$ in (18) can be replaced by either $\tilde{\mathcal{U}}_{i,l}$ or $\tilde{\mathcal{U}}_{i,r}$, depending on whether l or r is located within H or not. Also, the n_i in (18) should be replaced by $\tilde{n}_i/|\tilde{n}_i|$, where $\tilde{n}_i = (-\nu_i\sqrt{n_{ix}^2 + n_{iy}^2}, n_{ix}, n_{iy})$, $n_i = (n_{it}, n_{ix}, n_{iy})$. Note that $\tilde{n}_i$ is normal direction of the tracked space-time wave from the Riemann solver; therefore, this modification ensures that the Rankine–Hugoniot condition is satisfied.

The finite volume difference algorithm constitutes a flux through each boundary of the full, *partial*, and *big* hexahedron. Since the flux through a boundary face of the hexahedron is identical when viewed from either side of the face, we have the following theorem.

THEOREM 2. $\sum_{cells} |D^n|\mathcal{U}^n$ *in the finite volume difference scheme is conserved so that its increment over any time interval is equal to the net influx at the boundary.*

Away from the *INTERFACE* the scheme is clearly a second order scheme. For the cells along the *INTERFACE*, its local error is one order lower than in the 1D case since we use a piecewise linear approximation to the smooth *INTERFACE* and the local displacement error of this approximation is $\mathcal{O}(\Delta x^2)$. The scheme is one order better than untracked schemes, which typically have $\mathcal{O}(1)$ local error at the untracked fronts.

THEOREM 3. *Suppose the exact space-time interface and the solution on either side of it are smooth. Then the L_∞ error is $\mathcal{O}(\Delta x)$ for cells adjacent to the INTERFACE.*

Proof. Let the *INTERFACE* at t_n be exact, and let H be a big hexahedron adjacent to the approximate space-time interface. We apply the finite volume scheme to obtain the approximate state average $\mathcal{U}_i^{n+1}$ at the time level t_{n+1}, with top T and bottom B and side boundaries $\{S_i\}$, where each S_i is a triangle. The *INTERFACE* at time t_{n+1} has an $\mathcal{O}(\Delta x^2)$ displacement from the exact interface. The exact space-time interface will cut H into two pieces. Let H_1 be the piece on the same side of the interface as H. Let T_1, B_1, and S^1 be the top, bottom, and side boundaries of H_1, respectively. Let $U_{T_1}^{n+1}, U_{B_1}^n$ be the exact state averages over T_1 and B_1, respectively. Choosing $\mathcal{U}_B^n = U_{B_1}^n$, we want to show that $U_{T_1}^{n+1} - \mathcal{U}_T^{n+1} = \mathcal{O}(\Delta x)$. In fact, from (18),

$$|T|\mathcal{U}_T^{n+1} = |B|\mathcal{U}_B^n - \sum_i |S_i|(\tilde{\mathcal{U}}_{i,m}, f(\tilde{\mathcal{U}}_{i,m}), g(\tilde{\mathcal{U}}_{i,m})) \cdot n_i. \tag{19}$$

The exact solution satisfies

$$|T_1|U_{T_1}^{n+1} = |B_1|U_{B_1}^n - \int_{S^1} (u, f(u), g(u)) \cdot n ds. \tag{20}$$

Note that $|B|\mathcal{U}_B^n = |B_1|U_{B_1}^n$. Also, the numerical flux in (19) approximates the exact flux in (20) to at least $\mathcal{O}(\Delta x^3)$. In fact, when S_i is not on the approximate space-time interface, this is easily seen since $\int_{S_i}(u,f,g)\cdot n_i ds = |S_i|(u,f,g)(Z_i)\cdot n_i + \mathcal{O}(\Delta x^4)$.

Next suppose that S_i is on the approximate space-time interface. Because of the smoothness of the exact space-time interface, it has an $\mathcal{O}(\Delta x^2)$ displacement error to the exact one. The difference between their respective areas is of $\mathcal{O}(\Delta x^3)$. The area of $\bigcup S_i$ is $\mathcal{O}(\Delta x^2)$. Also, the choices of $\tilde{\mathcal{U}}_{i,m}$ and n_i in (19) ensure that $(\tilde{\mathcal{U}}_{i,m}, f(\tilde{\mathcal{U}}_{i,m}), g(\tilde{\mathcal{U}}_{i,m}))\cdot n_i$ in (19) is a first order approximation to the integrand in (20) at any point within an $\mathcal{O}(\Delta x)$ distance from the centroid Z_i of S_i. Thus $\int_{S_i}(u,f,g)\cdot n_i ds = |S_i|(u,f,g)(Z_i)\cdot n_i + \mathcal{O}(\Delta x^3)$ in the case that S_i is on the approximate space-time interface. Therefore, we have

$$\begin{aligned} U_{T_1}^{n+1} - \mathcal{U}_T^{n+1} &= (|T_1|U_{T_1}^{n+1} - |T|\mathcal{U}_T^{n+1})/|T| + U_{T_1}^{n+1}((|T| - |T_1|)/|T|) \\ &= (\mathcal{O}(\Delta x^4) + \mathcal{O}(\Delta x^3))/\mathcal{O}(\Delta x^2) + \mathcal{O}(\Delta x^3)/\mathcal{O}(\Delta x^2) \\ &= \mathcal{O}(\Delta x), \end{aligned} \tag{21}$$

where $\mathcal{O}(\Delta x^4)$ and $\mathcal{O}(\Delta x^3)$ in the first bracket follow from the local error of the numerical approximation of the flux defined on the non space-time interface and space-time interface, respectively. The proof is complete.

3.4. Cell level complexity and interface topological change. Because the dynamic evolution of the $INTERFACE$ often leads to geometrically complex situations, Hypothesis 3 might fail. For example, the Richtmyer–Meshkov (RM) instability develops very long and thin structures at the tips of bubbles and spikes at late time; see Figure 7 for an illustration.

The narrow structures and approximate or actual bifurcations will degrade the algorithm. Excessive cell merging to ensure CFL stability will degrade accuracy, and in any case actual bifurcations are (presently) excluded. We require a robust algorithm to solve problems for which any of the above occurs. We propose that these situations

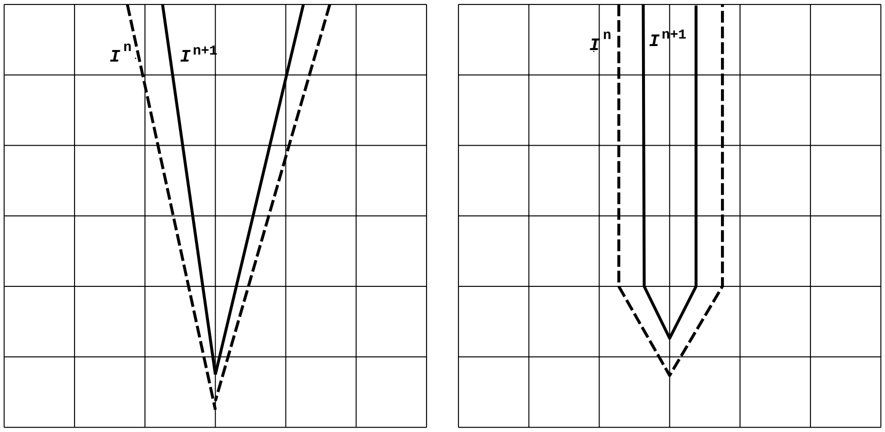


FIG. 7. *Limits on the merging process. $\mathcal{I}^n$ is represented by the dashed line and $\mathcal{I}^{n+1}$ by the solid line. At the time t^{n+1} level, in the first frame, the triangular cell at the tip is adjacent to a triangular cell and quadrilateral only; all of these cells form partial hexahedra having top area smaller than $\frac{1}{2}\Delta x^2$ and thus require further merger. In the second frame, the two branches of the curve near the tip of $\mathcal{I}^n$ and $\mathcal{I}^{n+1}$ are close and parallel to each other (forming a thin wall), thus forming a set of neighboring polyhedra with top area smaller than $\frac{1}{2}\Delta x^2$.*

Table 1
Comparison of error analysis for the test problem (22) *for Burgers' equation.*

Method	N	L_1 error	L_1 order	$\Sigma U_i \Delta x_i$
Untracked	30	6.83e-2	-	1.732
	60	3.49e-2	0.969	1.733
	120	1.63e-2	1.10	1.733
	240	8.24e-3	0.984	1.733
Nonconservatively tracked	30	2.80e-2	-	1.721
	60	6.89e-3	2.02	1.716
	120	4.23e-3	0.704	1.742
	240	2.01e-3	1.07	1.741
Conservatively tracked	30	2.17e-2	-	1.732
	60	7.07e-3	1.62	1.733
	120	2.11e-3	1.74	1.733
	240	6.04e-4	1.80	1.733

should be resolved by locally nonconservative tracking using the ghost cell algorithm of the authors [12]. Since these events will often occur on a lower dimensional space-time manifold, they will not impact the formal order of accuracy of the algorithm.

4. Numerical examples.

In this section we present numerical examples showing the convergence and conservation properties of the conservative front tracking scheme.

4.1. Burgers' equation.

Consider Burgers' equation $\frac{\partial u}{\partial t} + \frac{\partial}{\partial x}(\frac{1}{2}u^2) = 0$ on $[0, 6] \times [0, T]$, with initial conditions

$$u(x,0) = \begin{cases} 0.2 * (x-1)^2 + 0.2, & x \in [1,3], \\ 0.2, & \text{elsewhere.} \end{cases} \tag{22}$$

In Table 1 we present numerical results at ($T = 3.2$) using three different methods: the untracked MUSCL scheme, the nonconservatively (shock) tracked scheme with an MUSCL interior solver, and the conservatively (shock) tracked scheme with an MUSCL interior solver. Here the column labeled L_1 error indicates the L_1 norm of $u - \tilde{U}$, where u is the exact solution and $\tilde{U}$ is the second order approximate solution reconstructed from the piecewise constant numerical solution U at time T. Figure 8 displays the comparison of the numerical results obtained with $N = 30$ cells. For all of section 4, the CFL number is equal to 0.4.

4.2. 1D Euler equations.

Next we conduct a convergence test for the 1D Euler equations for a gamma law gas, $\gamma = 1.4$. We consider a tracked shock wave interacting with C^∞ data (a rarefaction wave with smooth edges). The computational domain is $[0, 4]$ with flow-through boundary conditions. At time $T = 0$ there is a right facing rarefaction wave in $(1, 2)$ and a left moving shock at $x = 3$. The left facing shock interacts with the rarefaction wave by the final time $T = 1$. We first define the initial states V_0 as follows: on $[0, 1]$, the density, pressure, and velocity are $2.0, 0.5$, and -1.0, respectively. On $[1, 2]$, V_0 has a centered rarefaction wave, ending at a pressure 1.5. On $[2, 3]$, the state is constant. On $[3, 4]$, the velocity is -1.5. Since the first

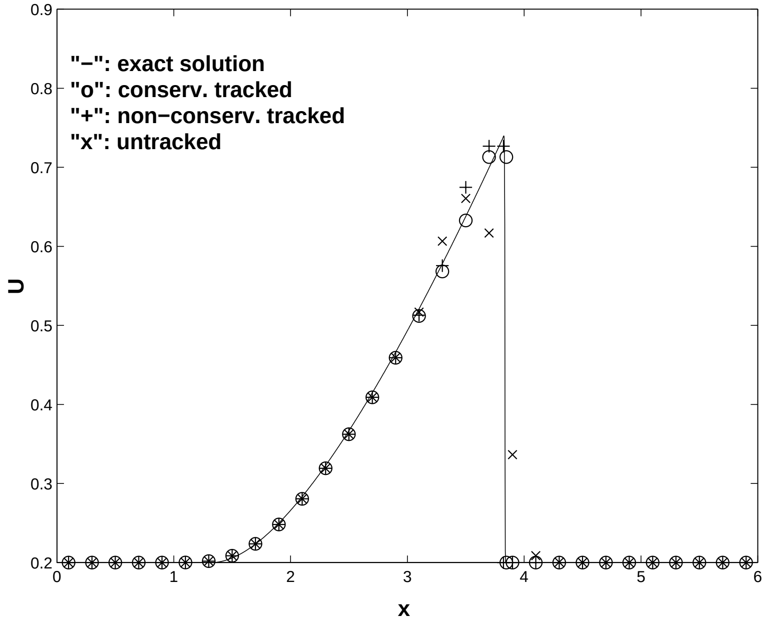


FIG. 8. *Comparison of numerical results for Burgers' equation.*

TABLE 2
Comparison of L_1 errors.

Nonconserv. tracked			Conserv. tracked	
N	L_1 error	L_1 order	L_1 error	L_1 order
100	0.0373	-	0.0395	-
200	0.0135	1.47	0.0106	1.90
400	0.00649	1.06	0.00361	1.55
800	0.00290	1.16	0.000891	2.02
1600	0.00148	0.970	0.000245	1.86
3200	0.000761	0.960	0.0000615	1.99

derivatives of V_0 have jumps at the rarefaction wave edges, we smooth the initial data V_0 so that

$$U_0(x) := \begin{cases} V_0(1)(2-\beta(x)) + V_0(2)(\beta(x)-1), & x \in (1,2), \\ V_0(x), & \text{elsewhere}, \end{cases}$$

where $\beta(x) = \frac{1}{2}(\sin \pi(x - \frac{3}{2})+3)$. We conduct the convergence test with the smoothed initial states U_0. The interior scheme is the second order MUSCL scheme with the shock wave tracked conservatively in one case and nonconservatively in the other. It is compared with a very fine ($N = 12800$, conservatively tracked) numerical solution to calculate the error in the L_1 norm. The comparison of the L_1 errors is shown in Table 2; the shock position errors $\sigma_e - \sigma_n$ are compared in Table 3, where σ_e

Table 3
Comparison of shock position errors.

Nonconserv. tracked			Conserv. tracked	
N	$\sigma_e - \sigma$	order	$\sigma_e - \sigma$	order
100	-4.20e-6	-	2.90e-4	-
200	-4.58e-6	-	9.15e-5	1.66
400	-2.54e-5	-	1.71e-5	2.42
800	-2.29e-5	-	2.85e-6	2.58
1600	-1.12e-5	1.03	9.70e-6	1.55
3200	-5.38e-6	1.06	2.00e-7	2.28

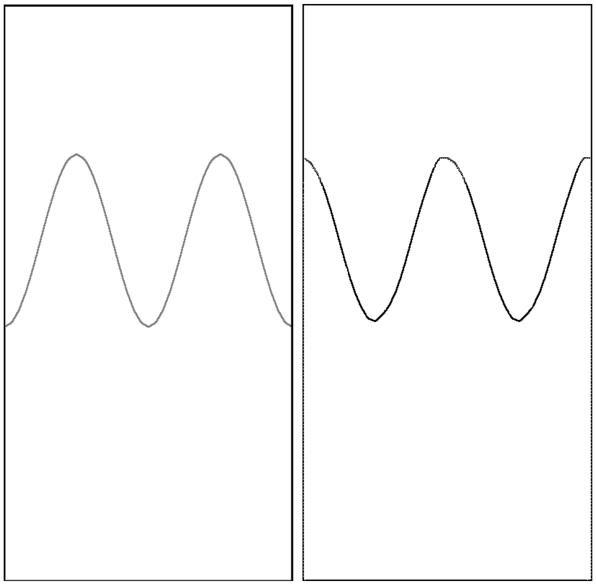


Fig. 9. *Front plot for the simulation of a horizontally moving contact discontinuity. The first frame displays the initial position of the contact; the second displays it after moving horizontally one quarter domain width in* 169 *time steps.*

denotes the exact shock position and σ_n denotes the numerical shock position. The conservatively tracked scheme is second order accurate.

4.3. 2D advection. We conduct a horizontal advection conservation test for the Euler equations to compare the fully conservative tracking scheme to the nonconservative tracking scheme. The numerical experiments were performed on a rectangular 1×2 domain with a 40×80 grid, displacing the interface horizontally half the domain width in 337 time steps. For the lower and upper boundaries of the domain, we use flow-through boundary conditions on which the states are normally extrapolated from the interior, and periodic conditions on the side boundaries. We use a polytropic gas, with polytropic exponent $\gamma = 1.4$. The contact discontinuity separating distinct gas states is tracked. The interface is sinusoidally perturbed with frequency 2.0 and amplitude 0.3. The initial configuration and the one-quarter width displaced configuration are shown in Figure 9. Excellent preservation of the sine wave is evident.

Table 4
Conservation error.

Conservation Error		
	Conservative tracking	Nonconservative tracking
Mass	0.0	0.21%
x-mom	0.0	0.21%
Energy	0.0	0.21%

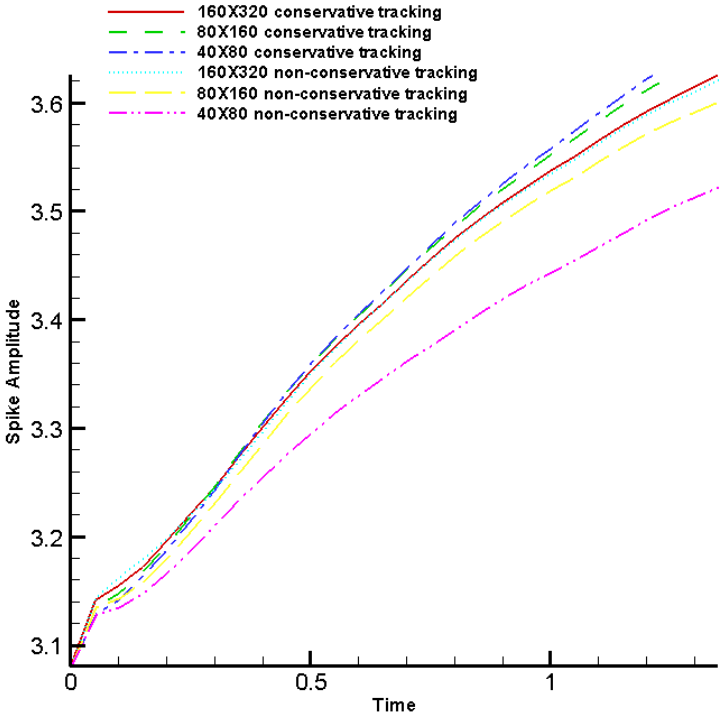


Fig. 10. *Spike amplitude in the RM instability simulations, as functions of time. The conservative tracked amplitude for a coarse grid is in approximate agreement with the nonconservative tracked amplitude for a fine grid.*

In Table 4, we compare the total conservation for the two methods, which is defined for the mass as

$$\text{(final mass} - \text{initial mass} + \text{boundary mass flux)}/\text{(initial mass)}, \tag{23}$$

with similar definitions for other conserved quantities. The conservative quantities refer to the lower gas in Figure 9. The total mass, momentum, and energy in the computational domain for the conservative tracking scheme show essentially perfect conservation, while the nonconserved tracking shows conservation errors of 0.21%.

4.4. Richtmyer–Meshkov instability. A Richtmyer–Meshkov (RM) instability is generated when a shock wave refracts through a perturbed interface which

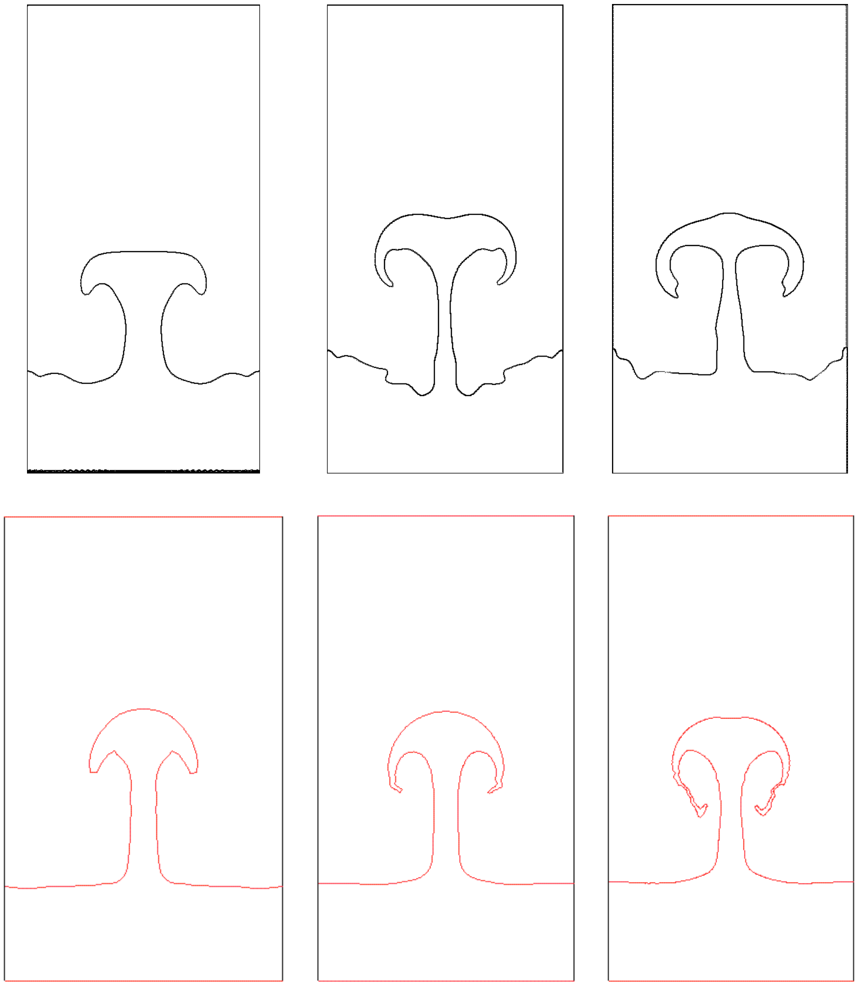


FIG. 11. *Front plot for the RM instability simulations. The upper row shows the plots of nonconservatively tracked interface at time* = 1.38. *The lower row shows the plots of conservatively tracked interface at the same time. For both rows, from left to right, are* 40×80, 80×160 *and* 160×320 *grids, respectively.*

separates fluids of differing densities. We compare simulations produced by the conservative and the nonconservative tracking schemes.

The numerical experiments were performed on a rectangular 1×2 domain, with a 40×80 grid, the lower and upper boundaries with flow-through boundary conditions, and periodic conditions for the side boundaries.

The initial configuration consists of a Mach 5.0 shock in a polytropic gas (with unshocked density 1.0) striking an interface separating two polytropic gases (both have polytropic exponent $\gamma = 1.40$). The preshock contact density ratio is $1 : 5$. The interface is sinusoidally perturbed with wavelength 1.0 and amplitude 0.1. Figure 11 shows the interface evolution of the RM instability; the initial configuration is shown as the left column. We also performed refined nonconservatively tracked simulations with 80×160 and 160×320 grids. The results indicate the convergence of the growth rate with nonconservative simulation to that of the conservative simulation when the computational mesh of the nonconservative simulation is refined. The 40×80 conservatively tracked solution appears to be comparable to both the finest (160×320) nonconservatively and conservatively tracked solutions, while the nonconservatively coarse grid run (40×80) tends to have a smaller growth rate. See Figures 10 and 11.

5. Conclusions.

We have proposed a new fully conservative front tracking algorithm. The algorithm is derived from an integral formulation of the PDEs. The 1D version of the algorithm is fully second order accurate away from the intersection of tracked waves. This has been determined by both the formal derivation and numerical experiments. In two dimensions, we provided the formal derivation that the scheme should be second order in the interior region and first order near the front. The convergence of bubble growth rate in the simulation of the RM instability seems to support this claim. Numerical tests in one and two dimensions demonstrate the improved conservation and convergence properties of the algorithm. The stability of the algorithm is verified by numerical experiments.

Conservative tracking is fundamentally an exercise in computational geometry to define the space-time interface. The finite volume differencing defined by the geometry follows standard algorithms. Further study of the space-time interface construction is called for. A robust algorithm may include nonconservative tracking for regions of greater local complexity than the conservative space-time interface construction will support.

REFERENCES

[1] R. Abgrall, *On essentially non-oscillatory schemes on unstructured meshes: Analysis and implementation*, J. Comput. Phys., 114 (1994), pp. 45–58.

[2] I.-L. Chern and P. Colella, *A Conservative Front Tracking Method for Hyperbolic Conservation Laws*, LLNL report UCRL-97200, Lawrence Livermore National Laboratory, Livermore, CA, 1987.

[3] I.-L. Chern, J. Glimm, O. McBryan, B. Plohr, and S. Yaniv, *Front tracking for gas dynamics*, J. Comput. Phys., 62 (1986), pp. 83–110.

[4] J. Glimm, J. W. Grove, X.-L. Li, W. Oh, and D. H. Sharp, *A critical analysis of Rayleigh-Taylor growth rates*, J. Comput. Phys., 169 (2001), pp. 652–677.

[5] J. Glimm, J. W. Grove, X.-L. Li, K.-M. Shyue, Y. Zeng, and Q. Zhang, *Three-dimensional front tracking*, SIAM J. Sci. Comput., 19 (1998), pp. 703–727.

[6] J. Glimm, J. W. Grove, X.-L. Li, and D. C. Tan, *Robust computational algorithms for dynamic interface tracking in three dimensions*, SIAM J. Sci. Comput., 21 (2000), pp. 2240–2256.

[7] J. Glimm, J. W. Grove, X.-L. Li, and N. Zhao, *Simple front tracking*, in Nonlinear Partial Differential Equations, Contemp. Math. 238, G.-Q. Chen and E. DiBenedetto, eds., AMS, Providence, RI, 1999, pp. 133–149.

[8] J. Glimm, X.-L. Li, and Y.-J. Liu, *Conservative front tracking in higher space dimensions*, in Proceedings of International Workshop on Computational Methods for Continuum Physics and Their Applications (IWCCPA), Nanjing, China, Transactions of Nanjing University of Aeronautics and Astronautics, 18 (2001), suppl. 1–15.

[9] J. Glimm, X.-L. Li, and Y.-J. Liu, *Conservative front tracking in one space dimension*, in Fluid Flow and Transport in Porous Media: Mathematical and Numerical Treatment, Contemp. Math. 295, AMS, Providence, RI, 2002, pp. 253–264.

[10] J. Glimm, X.-L. Li, Y.-J. Liu, and N Zhao, *Conservative front tracking and level set algorithms*, Proc. Natl. Acad. Sci. USA, 98 (2001), pp. 14198–14201.

[11] J. Glimm, D. Marchesin, and O. McBryan, *Subgrid resolution of fluid discontinuities* II, J. Comput. Phys., 37 (1980), pp. 336–354.

[12] J. Glimm, D. Marchesin, and O. McBryan, *A numerical method for two phase flow with an unstable interface*, J. Comput. Phys., 39 (1981), pp. 179–200.

[13] J. Glimm and O. McBryan, *A computational model for interfaces*, Adv. in Appl. Math., 6 (1985), pp. 422–435.

[14] A. Harten and J. Hyman, *Self-adjusting Grid Methods for One-Dimensional Hyperbolic Conservation Laws*, Report LA-9105, Los Alamos National Laboratory, Los Alamos, NM, 1981.

[15] W. E. Lorensen and H. E. Cline, *Marching cubes: A high resolution* 3*D surface construction algorithm*, Computer Graphics, 21 (1987), pp. 163–169.

[16] R. Pember, J. Bell, P. Colella, W. Cruchfield, and M. Welcome, *An adaptive Cartesian grid method for unsteady compressible flow in irregular regions*, J. Comput. Phys., 120 (1995), pp. 278–304.

[17] J. Smoller, *Shock Waves and Reaction-Diffusion Equations*, 2nd ed., Springer-Verlag, New York, 1994.

[18] B. Swartz and B. Wendroff, *Aztec: A front tracking code based on Godunov's method*, Appl. Numer. Math., 2 (1986), pp. 385–397.

SIAM J. NUMER. ANAL.
Vol. 41, No. 5, pp. 1948–1973

RIGOROUS SHADOWING OF NUMERICAL SOLUTIONS OF ORDINARY DIFFERENTIAL EQUATIONS BY CONTAINMENT[*]

WAYNE B. HAYES[†] AND KENNETH R. JACKSON[†]

Abstract. An exact trajectory of a dynamical system lying close to a numerical trajectory is called a *shadow*. We present a general-purpose method for proving the existence of finite-time shadows of numerical ODE integrations of arbitrary dimension in which some measure of hyperbolicity is present and there are either 0 or 1 expanding modes, or 0 or 1 contracting modes. Much of the rigor is provided automatically by interval arithmetic and validated ODE integration software that is freely available. The method is a generalization of a previously published *containment* process that was applicable only to two-dimensional maps. We extend it to handle maps of arbitrary dimension with the above restrictions, and finally to ODEs. The method involves building n-cubes around each point of the discrete numerical trajectory through which the shadow is guaranteed to pass at appropriate times. The proof consists of two steps: first, the rigorous computational verification of a simple geometric property, which we call the *inductive containment property*, and second, a simple geometric argument showing that this property implies the existence of a shadow. The computational step is almost entirely automated and easily adaptable to any ODE problem. The method allows for the rescaling of time, which is a necessary ingredient for successfully shadowing ODEs. Finally, the method is local, in the sense that it builds the shadow inductively, requiring information only from the most recent integration step, rather than more global information typical of several other methods. The method produces shadows of comparable length and distance to all currently published results. Finally, we conjecture that the inductive containment property implies the existence of a shadow without restriction on the number of expanding and contracting modes, although proof currently eludes us.

Key words. error analysis, ordinary differential equations, dynamical systems, chaos, shadowing, computational techniques, interval arithmetic

AMS subject classifications. 37M05, 65C20, 68U20, 70K99, 81T80

DOI. 10.1137/S0036142901399100

1. Introduction.

Consider the *initial value problem* (IVP) for an autonomous *ordinary differential equation* (ODE)

$$\mathbf{y}'(t) = \mathbf{f}(\mathbf{y}(t)), \tag{1.1}$$

$$\mathbf{y}(t_0) = \mathbf{y}_0, \tag{1.2}$$

where the ODE (1.1) is called the *defining equation*, (1.2) is called the *initial condition*, $\mathbf{y}$ is an n-dimensional vector, and f is an n-dimensional vector-valued function. Standard forward error analysis (e.g., Dahlquist and Björck (1974) or Kahaner, Moler, and Nash (1989)) tells us that, for a large class of ODEs, it is impossible in fixed-precision arithmetic to produce a numerical solution to an IVP which remains uniformly close to the exact solution for a long time. As a result, forward error bounds are impractical in such cases. However, one can often guarantee that a numerical solution remains uniformly close not to the solution starting at the initial condition specified, but instead to the exact solution to the ODE (1.1) starting at a nearby initial condition. In other words, if one allows the initial condition to have a nonzero error, just as one is satisfied with a nonzero error at all other times (Murdock (1995)), then it may be possible to guarantee that the numerical solution remains uniformly close to *some* exact

[*] Received by the editors December 3, 2001; accepted for publication (in revised form) March 17, 2003; published electronically November 19, 2003. This work was supported by the NSERC.
http://www.siam.org/journals/sinum/41-5/39910.html

[†] Department of Computer Science, University of Toronto, Toronto, ON, M5S 2E4, Canada (wayne@cs.toronto.edu, krj@cs.toronto.edu).

solution for a long time. Such an exact solution is called a *shadow* of the numerical solution.

Backward error analysis is a general term applied to methods of error analysis that relate a numerical solution to the exact solution of a "nearby" problem (Corless (1994), for example). In the context of IVPs for ODEs, "nearby" has at least two interpretations: we can either perturb the defining equation, or we can perturb the initial condition. Defect-based and other backward error analyses allow a time-dependent perturbation to the defining equation while leaving the initial condition untouched. In contrast, shadowing perturbs only the initial condition. For many physical systems which are modelled using ODEs, the governing equations are well defined, and virtually all error is introduced by imprecise knowledge of initial conditions and/or by numerical error in the computation of the solution. In these contexts, shadowing may a be more appropriate method of error analysis than defect-based methods. On the other hand, rigorous shadowing as presented in this paper and elsewhere is extremely expensive. Whereas nonrigorous defect-controlled methods are of roughly equal expense compared to more traditional integration methods, rigorous shadowing requires validated ODE integration, which at present tends to be several orders of magnitude more expensive in both time and memory than nonvalidated methods, even for low-dimensional problems. Thus, the goal of shadowing should not be to validate every numerical solution computed, but instead to study under what conditions we can expect a numerical solution to have a shadow.

Procedures for finding shadows usually involve some sort of fixed-point method. These include nonrigorous numerical methods akin to Newton's method (Grebogi et al. (1990); Quinlan and Tremaine (1992); Hayes (1995)) and methods that employ a theorem to prove the existence of a shadow, usually relying on Brouwer's fixed-point theorem or the Newton–Kantorovich theorem (Sauer and Yorke 1991; Chow and Palmer 1991, 1992; Chow and Van Vleck 1994).

An important advance has been the realization that ODEs differ fundamentally from maps in that they have errors in *time* as well as in space. By employing a *rescaling of time*, shadow lengths for ODEs can be increased by several orders of magnitude (Coomes, Koçak, and Palmer (1994b), (1995a), (1995b); Van Vleck 1995), even allowing the proof of existence of periodic trajectories near periodic pseudotrajectories (Coomes, Koçak, and Palmer (1994a); Coomes, Koçak, and Palmer (1997)). Shadowing has also been used to demonstrate that conservative integrations that approximately satisfy a first integral can have shadows that exactly satisfy it (Coomes (1997)), and that more explicit control of the numerical error in the stable versus unstable subspaces can lead to better shadowing results (Van Vleck (2000)). An interesting application has been to prove that a chaotic trajectory exists near an apparently chaotic pseudotrajectory (Stoffer and Palmer 1999). Hayes (2001) provides a more detailed survey of ODE shadowing results.

This paper extends the work of Grebogi, Hammel, Yorke, and Sauer (1990)(hereafter GHYS), who introduced an elegant geometrical method called *containment* for proving the existence of shadows. Their proof is valid for iterated maps in two dimensions, and is also practical for two-dimensional ODE problems that do not require a rescaling of time. We extend their results to maps of arbitrary dimension in which some measure of hyperbolicity is present and there are either 0 or 1 expanding modes, or 0 or 1 contracting modes. Although we firmly believe that containment can work with an arbitrary number of expanding and contracting directions, proving the general case is a work in progress. We also introduce a new method complementary to containment that facilitates a rescaling of time. In contrast to the above methods that use a fixed-point result, containment, including our new rescaling of time, uses an entirely geometrical argument. We rigorously verify the conditions of our theorems using validated ODE integration (Nedialkov (1999); Nedialkov, Jackson, and Corliss (1999)) and demonstrate that containment is capable of proving the existence of shadows of IVPs for ODEs that are of comparable quality to any currently in the literature. We also demonstrate how containment can reproduce the proof of chaos given by Stoffer and Palmer (1999).

The outline of the paper is as follows. Section 2 presents the ideas for the proofs of containment in an informal, geometrical setting. We present the actual proofs in section 3. Formally, these proofs break into two steps. First, we must prove that the numerical trajectory satisfies a certain property called the *inductive containment property* (ICP, for short). The ICP can be proven computationally to hold, using a validated ODE integrator; we defer discussion of how this is done until section 4. Second, we must show that a numerical trajectory that satisfies the ICP has a shadow. We prove this for maps in n dimensions for the cases in which there is either one expanding or one contracting direction while all the others do the opposite (3.1), or all directions either expand or contract (3.2). The method to rescale time is presented in section 5. Section 6 presents experimental results and comparisons with previous work, followed in section 7 by our conclusions.

2. Informal description of containment.

Although containment was the first method introduced for proving the existence of finite-time shadows of numerical orbits, it has not, to our knowledge, been pursued beyond its initial conception. In this paper we demonstrate that, at least in the restricted cases discussed, containment is about as strong as any method currently in the literature.

2.1. Definitions.

In this paper, an *orbit* is a discrete sequence of points, a *solution* is a continuous curve, and a trajectory more generally refers to either an orbit or a solution, depending upon the context. The prefix *pseudo-* will be used to denote an approximate orbit, solution, or trajectory, although sometimes it will be omitted if the meaning is clear from the context.

Shadowing of numerical orbits was first applied to iterated maps.

DEFINITION 2.1. *An* orbit *of an* iterated map *consists of a sequence of points* $\mathbf{x}_i$ *generated by the recurrence* $\mathbf{x}_{i+1} = \varphi(\mathbf{x}_i)$ *for some map* φ*.*

DEFINITION 2.2. *A* homeomorphism *is a map which is continuous, one-to-one, and onto.*

For our purposes, we restrict φ to being a homeomorphism.

DEFINITION 2.3. *A* pseudo-orbit*, or* noisy *orbit, for* φ *satisfies* $\mathbf{y}_{i+1} = \varphi(\mathbf{y}_i) + \delta_i$*, where* δ_i *is the noise introduced at step* i*. If* $\|\delta_i\| < \delta$ *for all* i*, then it is called a* δ-pseudo-orbit *for* φ*.*

DEFINITION 2.4. *The exact orbit* $\{\mathbf{x}_i\}_{i=0}^N$ *is an* ε*-shadow of the pseudo-orbit* $\{\mathbf{y}_i\}_{i=0}^N$ *if* $\|\mathbf{y}_i - \mathbf{x}_i\| < \varepsilon$ *for* $i = 0, \ldots, N$*.*

Numerical solutions to ODEs can often be viewed as iterated maps by defining $\mathbf{x}_{i+1} = \varphi_{h_i}(\mathbf{x}_i)$, where φ_{h_i} is the *time-*h_i solution operator for the IVP (1.1), (1.2). The time-h_i solution operator is a homeomorphism as long as $\mathbf{f}$ in (1.1) is bounded and Lipschitz continuous over the domain of interest (Ascher, Mattheij, and Russell (1988)). For small h_i, a *one-step numerical method* approximates φ_{h_i} by $\tilde{\varphi}_{h_i}$ and then computes a sequence of discrete points $\mathbf{y}_{i+1} = \tilde{\varphi}_{h_i}(\mathbf{y}_i)$ representing approximations to $\mathbf{y}(t_{i+1})$, where $t_{i+1} = t_i + h_i$. We will term such a discrete sequence of points a *pseudotrajectory.* If the pseudotrajectory satisfies a *local error tolerance* of δ such that

$\|\mathbf{y}_{i+1} - \varphi_{h_i}(\mathbf{y}_i)\| \leq \delta$, then we call it a *$\delta$-pseudotrajectory.* If h_i is constant, we can drop it as a subscript and treat the pseudotrajectory as a pseudo-orbit of the iterated map $\varphi \equiv \varphi_h$.

2.1.1. Hyperbolicity and pseudohyperbolicity. One of the most important concepts in shadowing is that of *hyperbolicity.* Essentially, a system of ODEs is hyperbolic if the variational equation along a solution $\mathbf{y}(t)$ displays *exponential dichotomy* (Palmer 1988). This means that a perturbation δ to the solution $\mathbf{y}(t)$ at time $t = t_0$, $\mathbf{z}(t_0) = \mathbf{y}(t_0) + \delta$ produces a new solution $\mathbf{z}(t)$ with one of two properties: if δ lies in the *stable subspace* of $\mathbf{y}(t)$, then $\mathbf{z}(t)$ converges exponentially to $\mathbf{y}(t)$ as t increases; if δ lies in the *unstable subspace*, then $\mathbf{z}(t)$ diverges exponentially away from $\mathbf{y}(t)$ as t increases. More details can be found in Hayes (2001) or Palmer (1988). If a system is hyperbolic, then the angle between the stable and unstable subspaces is always bounded away from 0 (see GHYS).

This paper deals not with hyperbolic systems, but with systems whose pseudotrajectories are shadowable for finite but nontrivial lengths of time even though they are not hyperbolic. For this to occur, a system must display pseudohyperbolicity. We say that a system is *pseudohyperbolic* if a small perturbation to a trajectory $\mathbf{y}(t)$ produces a new solution $\mathbf{z}(t)$ which falls into one of two classes: those which *tend to* diverge exponentially away from $\mathbf{y}(t)$ as t increases, and those that *tend to* converge exponentially towards $\mathbf{y}(t)$ as t increases. In addition, $\mathbf{z}(t)$ should behave in this manner over nontrivial periods of time. In short, a pseudohyperbolic system should "mimic" the behavior of a hyperbolic system over finite but nontrivial periods of time. This notion could be quantified by, for example, attempting to find the stable and unstable subspaces using *refinement* (GHYS; Quinlan and Tremaine (1992); Hayes (1995), (2001)), and then performing least-squares fits to exponential curves of the growth and decay of these subspaces.

2.2. Containment in two dimensions. The first studies of shadows of pseudohyperbolic systems with both expanding and contracting directions appear to be Beyn (1987) and Hammel, Yorke, and Grebogi (1987). Hammel, Yorke, and Grebogi (1988) and GHYS provide the first proof of the existence of a shadow for a nonhyperbolic system over a nontrivial length of time. Their method consists of two steps. First, they *refine* a noisy trajectory using an iterative method that produces a nearby trajectory with less noise. When refinement converges to the point that the noise is of the order of the machine precision, they invoke *containment*, which can prove the existence of a nearby exact trajectory. Their containment method, which we now describe, is directly applicable only to two-dimensional maps.

Let $\{\mathbf{y}_i\}_{i=0}^{N} \subset \mathbf{R}^2$ be a two-dimensional δ-pseudo-orbit of φ. As i increases, orbits separated from each other by a small distance along the expanding direction diverge on average away from each other, while orbits separated by a small distance along the contracting direction approach each other on average. The containment process consists of building a parallelogram M_i around each point $\mathbf{y}_i$ of the pseudo-orbit such that two sides $C_i^{\pm 1}$ are approximately normal to, and separated from each other along, the contracting direction, while the other two sides $E_i^{\pm 1}$ are approximately normal to, and separated from each other along, the expanding direction.[1] The diameter of

[1] Note that this naming convention is exactly opposite to that of GHYS, because in two dimensions they emphasized the direction to which the sides of M_i were *parallel.* In higher dimensions, the faces of an n-cube are not parallel to a unique direction, and it is the direction along which a face is separated from the center of the n-cube that matters. We change the naming convention now to avoid confusion later.

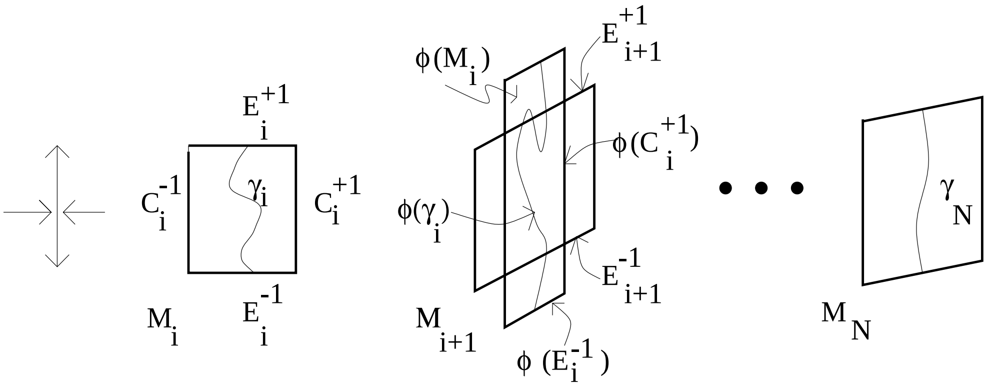


FIG. 2.1. *Containment in two dimensions, reproduced from GHYS. The horizontal direction is contracting, and the vertical direction is expanding.*

M_i will bound the distance from the pseudo-orbit to the shadow. In order to prove the existence of a shadow, the image of M_i under φ must intersect M_{i+1} such that $\varphi(M_i)$ makes a "plus sign" with M_{i+1} (Figure 2.1). To ensure that this property holds, GHYS require a bound on the second derivative of φ, and the amounts of expansion and contraction need to be resolvable to within the machine precision. The proof of the existence of an exact orbit then relies on the following argument. For any $i \in \{0, 1, \ldots, N-1\}$ let γ_i be a continuous curve in M_i connecting the expanding sides E_i^{-1} and E_i^{+1}. Its image $\varphi(\gamma_i)$ is then stretched such that there is a section of $\varphi(\gamma_i)$ lying wholly within M_{i+1}, and in particular $\varphi(\gamma_i)$ leaves M_{i+1} through the expanding sides $E_{i+1}^{\pm 1}$ at both ends. Let γ_{i+1} be a continuous subsection of $\varphi(\gamma_i)$ lying wholly within M_{i+1} connecting the expanding sides $E_{i+1}^{\pm 1}$. Repeating this process along the orbit produces γ_N lying wholly within the final parallelogram M_N. Then any point $\mathbf{x}_N \in \gamma_N$ traced backwards via φ^{-1} yields a point $\mathbf{x}_i \in \gamma_i \subset M_i$, $i = N-1, \ldots, 1, 0$. Note that $\{\mathbf{x}_i\}_{i=0}^N$ is an exact orbit. Moreover, since $\mathbf{x}_i, \mathbf{y}_i \in M_i$, we infer that $\|\mathbf{x}_i - \mathbf{y}_i\| \leq \varepsilon$, where ε bounds the diameter of M_i, $i = 0, \ldots, N$. Thus, $\{\mathbf{x}_i\}_{i=0}^N$ is an ε-shadow of $\{\mathbf{y}_i\}_{i=0}^N$. We make the intuitive argument described here rigorous in section 3.

With this picture in mind, there is a nice geometric interpretation of the requirement that the angle between the stable and unstable directions be bounded away from 0: if the angle gets too small, then the parallelogram essentially loses a dimension, and $\varphi(M_i)$ can not make a "plus sign" with M_{i+1}. Practically speaking, this occurs when the angle becomes comparable to the noise amplitude of the pseudo-orbit. Hence, the more accurate the orbit, the longer it can be shadowed (GHYS, Quinlan and Tremaine (1992)).

2.3. Containment in three dimensions. The process described by GHYS is not directly applicable to systems with more than two dimensions, and GHYS provided no indication of how it could be extended beyond two dimensions. We describe how the method can be extended to three dimensions, in which there are precisely two interesting cases:

(i) One expanding direction and two contracting (Figure 2.2). Assume that the z direction is expanding, while the x and y directions are contracting. (We assume, for simplicity of exposition and for ease of drawing, that these three directions are roughly orthogonal, although in practice they need only

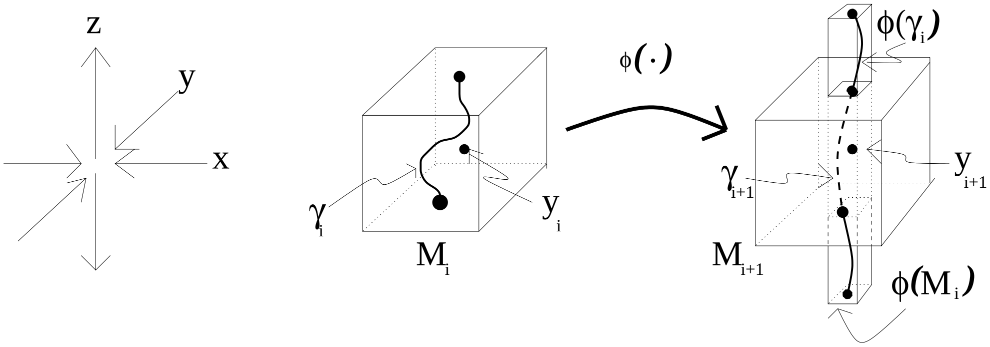


FIG. 2.2. *Containment in three-dimensions, case* (i)*: one expanding direction and two contracting.*

be resolvable from each other.) Then, analogously to the two-dimensional argument, assume we can draw a *cube* M_i of diameter no larger than ε around each noisy point $\mathbf{y}_i$, and, for $i = 0, 1, \ldots, N-1$, assume we can verify that $\varphi(M_i)$ maps over M_{i+1} so that φ stretches M_i into a long, thin tube, a segment of which lies wholly in M_{i+1}. Then, precisely as in the two-dimensional case, we can prove that an ε-shadow of $\{\mathbf{y}_i\}_{i=0}^N$ exists as follows. We introduce a curve γ_i that runs approximately along the expanding (vertical) direction from any point on the top of M_i to its bottom. If $\varphi(M_i)$ maps over M_{i+1} as in Figure 2.2, then we are guaranteed that a contiguous section of $\varphi(\gamma_i)$ lies inside M_{i+1}, connecting its top and bottom along the expanding direction. This segment of $\varphi(\gamma_i)$ becomes γ_{i+1}. Any point $\mathbf{x}_N \in \gamma_N \subset M_N$ can be traced backwards via φ^{-1} to a point $\mathbf{x}_i \in \gamma_i \subset M_i$ for $i = 0, 1, \ldots, N-1$. As in the two-dimensional case, $\{\mathbf{x}_i\}_{i=0}^N$ is an ε-shadow of $\{\mathbf{y}_i\}_{i=0}^N$.

(ii) Two expanding and one contracting direction. We note that if time is reversed in such a system, then expanding and contracting directions reverse their roles. Thus, we simply look at the pseudotrajectory in reverse and apply the above argument. That is, we set $\mathbf{z}_i = \mathbf{y}_{N-i}, i = 0, \ldots, N$, and apply the above argument to the noisy trajectory $\{\mathbf{z}_i\}_{i=0}^N$.

3. Containment theorems and proofs.

3.1. Containment in *n* dimensions with one expanding direction. Here we provide a proof of what we call the $(n, 1)$-inductive containment theorem: the n-dimensional case in which precisely one direction is expanding, while all the others contract. Previous proofs of containment required explicit a priori bounds on spatial derivatives, whereas our proof requires no such bounds.[2]

Let M_i be a parallelepiped in $\mathbf{R}^n$ with faces F_i^j, for $i = 0, \ldots, N$ and $j = \pm 1, \ldots, \pm n$, with opposite signs in the superscript representing opposite faces of a parallelepiped (see Figure 3.1). Without loss of generality, we assume that the first direction is the "expanding" one. We will denote the union of a set of faces by listing all of them in the superscript; for example, $F_i^{\pm 2, \ldots, \pm n}$ represents the set of all the faces of M_i except F_i^{-1} and F_i^{+1}. Let $\partial_E M_i \equiv F_i^{-1} \cup F_i^{+1} \equiv F_i^{\pm 1}$ and $\partial_C M_i \equiv \bigcup_{j=2}^n F_i^{-j} \cup F_i^{+j} \equiv F_i^{\pm 2, \ldots, \pm n}$. Let $\varphi : \mathbf{R}^n \to \mathbf{R}^n$ be a homeomorphism. Let

[2] Of course, our validated ODE integration (Nedialkov (1999)) must compute bounds on derivatives in order to compute enclosures, but these bounds are not a priori; they are computed on-the-fly, and if a bounds check fails, we can always try a smaller timestep to compensate.

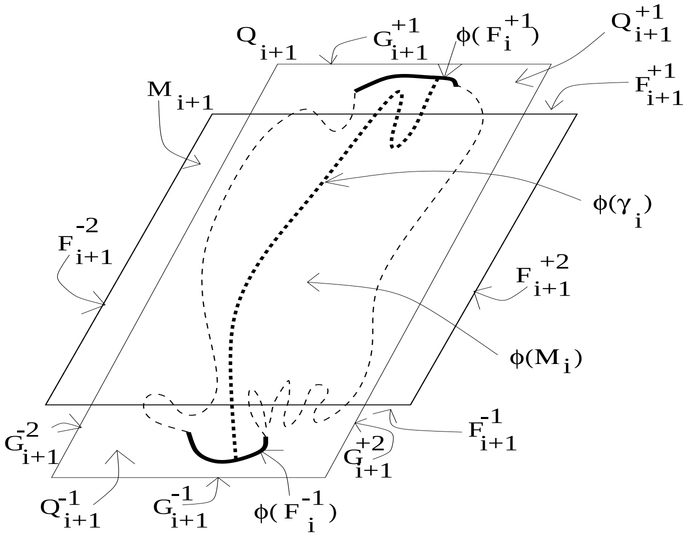


FIG. 3.1. *The image $\varphi(M_i)$ and M_{i+1} for two dimensions. The dark curves at the bottom and top are $\varphi(F_i^{\pm 1})$. The dashed curves at the left and right are $\varphi(F_i^{\pm 2})$.*

int X represent the interior of X. Then M_i and M_{i+1} satisfy the $(n,1)$-ICP if

(1) $\varphi(F_i^{\pm 1}) \cap M_{i+1} = \emptyset$, and $\varphi(F_i^{-1})$ and $\varphi(F_i^{+1})$ are on opposite sides of the infinite slab between the two hyperplanes containing F_{i+1}^{-1} and F_{i+1}^{+1}.

(2) $\exists Q_{i+1}$, a parallelepiped in $\mathbf{R}^n$ with each face G_{i+1}^j parallel to the corresponding face F_{i+1}^j of M_{i+1} for $j = \pm 1, \dots, \pm n$, such that
   (a) $\varphi(M_i) \subset \text{int } Q_{i+1}$,
   (b) $F_{i+1}^{\pm 2,\dots,\pm n} \cap Q_{i+1} = \emptyset$, and $\forall j \in \{2,\dots,n\}$, F_{i+1}^{-j} and F_{i+1}^{+j} are on opposite sides of the infinite slab between the two hyperplanes containing G_{i+1}^{-j} and G_{i+1}^{+j}.

Let $\gamma_0 \subset M_0$ be a simple curve joining F_0^{-1} to F_0^{+1}, but otherwise remaining in the interior of M_0. That is,

$$\gamma_0 \cap F_0^{-1} \neq \emptyset \quad \wedge \quad \gamma_0 \cap F_0^{+1} \neq \emptyset \quad \wedge \quad \text{int } \gamma_0 \subset \text{int } M_0.$$

THEOREM 3.1 ($(n,1)$-inductive containment theorem). *If M_i and M_{i+1} satisfy the $(n,1)$-ICP $\forall i = 0, \dots, N-1$, then $\forall i = 0, \dots, N$*

$$\exists \text{ simple curve } \gamma_i \subseteq \varphi^i(\gamma_0) \text{ s.t. } \gamma_i \cap F_i^{-1} \neq \emptyset \ \wedge \ \gamma_i \cap F_i^{+1} \neq \emptyset \ \wedge \ \text{int } \gamma_i \subset \text{int } M_i. \tag{3.1}$$

That is, γ_i touches the boundary of M_i in precisely two places, connecting F_i^{-1} to F_i^{+1}, but otherwise remains entirely inside M_i.

Proof. We proceed by induction on i. The proof of the base case $i = 0$ is immediate, by the definition of γ_0. For the inductive case, assume $\exists$ a simple curve $\gamma_i \subseteq \varphi^i(\gamma_0)$ such that $\gamma_i \cap F_i^{-1} \neq \emptyset \ \wedge \ \gamma_i \cap F_i^{+1} \neq \emptyset \ \wedge \ \text{int } \gamma_i \subset \text{int } M_i$. From ICP(1), $\varphi(F_i^{\pm 1}) \subset Q_{i+1}$, and the fact that Q_{i+1} is convex, we know that Q_{i+1} intersects both F_{i+1}^{-1} and F_{i+1}^{+1}; and from ICP(2(b)), Q_{i+1} does not intersect $F_{i+1}^{\pm 2,\dots,\pm n}$. Thus, since Q_{i+1} is convex, $Q_{i+1} - M_{i+1}$ is disconnected by the slab defined in ICP(1) into two

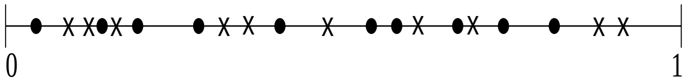


FIG. 3.2. *Schematic representation of the sets $\gamma^{-1}(s^{-1})$ (dots) and $\gamma^{-1}(s^{+1})$ ($\times$'s).*

disjoint components,[3] say Q^{-1}_{i+1} and Q^{+1}_{i+1}, each containing one of $\varphi(F^{\pm 1}_i)$, by ICP(1). Without loss of generality, assume $\varphi(F^j_i) \subset Q^j_{i+1}$, $j = \pm 1$. Now, consider one of the components, say Q^{-1}_{i+1}. It contains one of the two endpoints of $\varphi(\gamma_i)$, since one endpoint is in $\varphi(F^{-1}_i) \subset Q^{-1}_{i+1}$, while the other endpoint of $\varphi(\gamma_i)$ is in $\varphi(F^{+1}_i) \subset Q^{+1}_{i+1}$. Since γ_i is a simple curve and φ is a homeomorphism, $\varphi(\gamma_i)$ is a simple curve. Now, $Q^{-1}_{i+1} \cap Q^{+1}_{i+1} = \emptyset$, and $\varphi(\gamma_i)$ connects the two. Thus, $\varphi(\gamma_i)$ must cross the boundary of Q^{-1}_{i+1}. This boundary consists of exactly two mutually exclusive patches, one of which is a subset of ∂Q_{i+1}, the other a subset of F^{-1}_{i+1}. Since $\varphi(\gamma_i) \subset \varphi(M_i) \subset \text{int } Q_{i+1}$, we infer that $\varphi(\gamma_i) \cap \partial Q_{i+1} = \emptyset$, and so $\varphi(\gamma_i)$ leaves Q^{-1}_{i+1} through F^{-1}_{i+1}. A similar argument shows that $\varphi(\gamma_i)$ leaves Q^{+1}_{i+1} through F^{+1}_{i+1}. Thus, $\varphi(\gamma_i) \cap F^j_{i+1} \neq \emptyset$, $j = \pm 1$. It remains to show that there exists a segment γ_{i+1} of $\varphi(\gamma_i)$ which is a simple curve and maintains the property defined in (3.1).

Since $\varphi(\gamma_i)$ is a simple curve, there exists a parameterization $\gamma(t)$ for $t \in [0,1]$ such that $\gamma([0,1]) = \varphi(\gamma_i)$ and $\gamma(t)$ is a homeomorphism (Munkres (1975)). Let $s^j = \varphi(\gamma_i) \cap F^j_{i+1}$, $j = \pm 1$. Now, s^{-1} and s^{+1} are disjoint sets since $F^{-1}_{i+1} \cap F^{+1}_{i+1} = \emptyset$, and they are compact because (1) F^j_{i+1} for $j = \pm 1$ are compact; (2) γ_i is compact, φ is a homeomorphism, and so $\varphi(\gamma_i)$ is compact; and (3) the intersection of two compact sets in $\mathbf{R}^n$ is compact. Finally, $\gamma^{-1}(s^{\pm 1})$ are compact because γ is a homeomorphism. To prove that there exists a simple curve $\gamma_{i+1} \subset \varphi(\gamma_i)$ such that $\gamma_{i+1} \cap F^{-1}_{i+1} \neq \emptyset$, $\gamma_{i+1} \cap F^{+1}_{i+1} \neq \emptyset$, and $\text{int } \gamma_{i+1} \subset \text{int } M_{i+1}$, we need to show that there exist two points in $[0,1]$, one each from $\gamma^{-1}(s^{-1})$ and $\gamma^{-1}(s^{+1})$, such that no points from either set are between them (see Figure 3.2). This will prove that there exists a simple curve, which is a section of $\varphi(\gamma_i)$, that connects F^{-1}_{i+1} to F^{+1}_{i+1} without otherwise intersecting ∂M_{i+1}. To this end, let $G = \gamma^{-1}(s^{-1})$ and $R = \gamma^{-1}(s^{+1})$, and note that G and R are compact, disjoint, nonempty subsets of $[0,1]$. The following lemma completes the proof. □

LEMMA 3.2. *Let G and R be (possibly infinite) disjoint, compact, nonempty subsets of $[0,1]$. Then $\exists g \in G,\ r \in R$ such that $(g,r) \cap (G \cup R) = \emptyset$, where we have assumed without loss of generality than $g < r$.*

Proof. Consider the function $f(x,y) = |x-y|$ over the subset $G \times R$ of the plane. Since f is continuous and $G \times R$ is compact, f attains its minimum at some point $(g,r) \in G \times R$. That is, $|g-r| \leq |g'-r'|$ for any other $g' \in G$, $r' \in R$. Thus, there is no element of either set G or R between g and r, and so the open interval (g,r) is disjoint from $G \cup R$. □

THEOREM 3.3 (shadowing containment theorem). *Let φ be a homeomorphism. Let $\{M_i\}_{i=0}^N$ be a sequence of parallelepipeds enclosing a pseudotrajectory $\{\mathbf{y}_i\}_{i=0}^N$. Let ε be the maximum diameter of M_i for $i = 0, 1, \ldots, N$. Let $\gamma_i \subset M_i, \gamma_i \neq \emptyset$, $i = 0, \ldots, N$, and let $\gamma_{i+1} \subseteq \varphi(\gamma_i)$, $i = 0, \ldots, N-1$. Then $\exists$ an ε-shadow $\{\mathbf{x}_i\}_{i=0}^N$ of $\{\mathbf{y}_i\}_{i=0}^N$. That is, there is an exact trajectory $\{\mathbf{x}_i\}_{i=0}^N$ of φ such that $\|\mathbf{x}_i - \mathbf{y}_i\| < \varepsilon$, $i = 0, \ldots, N$.*

[3] This is because F^{-1}_{i+1} and F^{+1}_{i+1} are each patches of an $(n-1)$-dimensional hyperplane residing in n dimensions, and so they each disconnect any convex set they intersect, as long as that convex set does not intersect their boundaries ∂F^{-1}_{i+1} and ∂F^{+1}_{i+1}, respectively.

Proof. Since φ is a homeomorphism, φ^{-1} is a well-defined function. Pick any point $\mathbf{x}_N \in \gamma_N$, and recursively define $\mathbf{x}_i = \varphi^{-1}(\mathbf{x}_{i+1}),\ i = N-1, N-2, \ldots, 0$. Since $\gamma_{i+1} \subseteq \varphi(\gamma_i)$, $\varphi^{-1}(\gamma_{i+1}) \subseteq \gamma_i$, and so by induction $\mathbf{x}_i \in \gamma_i$ for $i = N, N-1, \ldots, 0$. Since $\mathbf{y}_i \in M_i$ and $\mathbf{x}_i \in \gamma_i \subset M_i$, $\|\mathbf{y}_i - \mathbf{x}_i\| \leq \operatorname{diam}(M_i) \leq \varepsilon,\ i = 0, \ldots, N$. □

Thus, applying Theorem 3.3 to an orbit satisfying the $(n,1)$-ICP implies the existence of a shadow.

Remark 3.1. Note that Theorem 3.3 is independent of the number of dimensions n, and of the number of expanding and contracting directions, because the only parts of the inductive containment theorem that are used are the conclusions that $\gamma_{i+1} \subseteq \varphi(\gamma_i)$ for $i = 0, 1, \ldots, N-1$ and $\emptyset \neq \gamma_i \subset M_i$ for $i = 0, \ldots, N$. The 0-expanding and 0-contracting directions are handled separately. We conjecture that the general (n,k)-inductive containment theorem (work in progress) will also assert this property, so that the above shadowing containment theorem is applicable to the general (n,k) case, in which k directions are expanding and $n-k$ are contracting.

As mentioned previously, the case with one *contracting* dimension while the other $n-1$ directions expand can be handled simply by reversing the arrow of time and applying the above argument. We call this the $(n, n-1)$ case. Another proof, which is more likely to be generalizable to an arbitrary number of expanding and contracting directions, is presented in Hayes (2001).

3.2. Containment with zero contracting or zero expanding directions. For completeness, we mention the trivial cases in which all directions are contracting, or all directions are expanding. We call these the $(n,0)$ and (n,n) cases, respectively. The former case is entirely trivial, because the problem is stable: if $\varphi(M_i) \subset M_{i+1}$ for all i, then clearly any exact solution starting in M_0 will be in M_i for all $i > 0$. Similarly, if all directions are expanding, then we apply the same argument in the reverse direction: if $\varphi^{-1}(M_{i+1}) \subset M_i$ for all i, then any exact solution *finishing* in M_N, traced backwards, lies in M_i for $i = N-1, N-2, \ldots, 0$.

3.3. Discussion. The four cases $(n,0), (n,1), (n,n-1)$, and (n,n) cover *all* cases for $n = 1, 2, 3$. That is, the theorems in this paper can prove the existence of shadows for any n-dimensional system, $n \leq 3$, in which some measure of pseudohyperbolicity is present. Furthermore, although the proofs, for simplicity, deal only with a single function φ, the induction argument could just as easily use a *different* function φ_i at each step. In particular, φ_i could be the ODE time-h_i solution operator φ_{h_i}. Thus, modulo a rescaling of time (which we discuss below), the above proofs can be used to find shadows of noisy trajectories of ODE systems, as well as maps, with up to three dependent variables. They can also be used in the case of n dependent variables, with the restriction that solutions have either one expanding and $n-1$ contracting directions, or one contracting and $n-1$ expanding directions.

Finally, we believe that a generalized (n,k)-ICP implies the existence of a shadow (work in progress; more discussion in Hayes (2001)).

3.4. Proving the existence of chaotic orbits. Following the analysis of Stoffer and Palmer (1999), we describe how to use containment to prove the existence of chaotic orbits. We quote directly from their introduction.

> The idea is to construct two periodic pseudo-orbits which happen to be close to each other at some point. We call this the branching point. Then it is possible to construct an infinite number of pseudo-orbits as follows. You follow one or the other of the periodic pseudo-orbits. When you reach the branching point, you either stay on your periodic orbit for at least one more loop, or else you switch to the other periodic orbit. Each time you

> arrive at the branching point you can again choose to stay or to switch, ad infinitum. Assume that for each such pseudo-orbit there is a unique *orbit* of the system which is close to the pseudo-orbit. Then the dynamical system indeed behaves chaotically, at least in a certain neighbourhood of the two periodic pseudo-orbits. (Stoffer and Palmer 1999)

To use this approach together with containment to prove the existence of a chaotic orbit, assume that the first orbit has a sequence of N parallelepipeds M_i satisfying the (n, k)-ICP with $M_N = M_0$. Then the (n, k)-inductive containment theorem can be invoked ad infinitum around this periodic pseudo-orbit and proves the existence of infinitely long exact orbits that remain in the vicinity of this pseudo-orbit.[4] Similarly, assume that the second orbit has a sequence of P parallelepipeds Q_i satisfying the (n, k)-ICP with $Q_P = Q_0$. Assume further that $M_i = Q_j$ for some i, j. Then $M_i = Q_j$ is the branching point, the (n, k)-inductive containment theorem can be invoked ad infinitum around *both* of these pseudo-orbits, and each time we pass $M_i = Q_j$ we can choose which pseudo-orbit to follow. The (n, k)-inductive containment theorem proves that a shadow follows us as we go.

4. Verifying the inductive containment property.

We present one method of verifying that the general (n, k)-ICP holds for a given pseudotrajectory derived from the numerical solution of an ODE. (Three more methods for verifying the ICP are presented in Hayes (2001).) We note in passing that this scheme (as well as the other three discussed in Hayes (2001)) could easily be adapted to the simpler problem of maps. We require the use of validated interval arithmetic, or a validated ODE integrator if φ derives from an ODE. The validated ODE integrator that we use is called VNODE (Nedialkov (1999); Nedialkov, Jackson, and Corliss (1999)). VNODE works with n-dimensional parallelepipeds and satisfies the following property: given an n-dimensional parallelepiped A and a timestep h, VNODE will return an n-dimensional parallelepiped B such that $\varphi_h(A) \subset B$, where φ_h is the time-h solution operator. For the purposes of this description, we will denote the output B by $\bar{\varphi}_h(A)$. Thus,

$$\varphi_h(A) \subset \bar{\varphi}_h(A).$$

We will usually omit the timestep parameter h; we will talk only of φ, keeping in mind that, in the induction, φ can be different for each step.

We verify the ICP using an iterative method that we have found empirically to require about 3–4 validated integrations per step on average, independent of n. This method rigorously verifies the ICP in the cases for which we have proven the inductive containment theorem and is the method we actually used to produce our numerical results. Three noniterative, deterministic methods are presented in Hayes (2001); however, we found this method to be the most efficient with the validated ODE solver we used (Nedialkov (1999)).

We first look at the simple two-dimensional case in which one of the directions is expanding, while the other is contracting. To begin, assume that the only information provided by our validated ODE integration is an outer bound $\bar{\varphi}(M_i)$ on $\varphi(M_i)$. Then, it is *not* possible to verify the $(2, 1)$-ICP with only one validated integration, because this information can only prove contraction, not expansion, as shown in Figure 4.1. In both Figures 4.1(a) and 4.1(b), $\bar{\varphi}(M_i)$ is a valid enclosure of $\varphi(M_i)$. In both figures, $\bar{\varphi}(M_i)$ can be used to prove that $\varphi(M_i)$ has contracted in the horizontal

[4] Slightly more is required to prove the existence of periodic orbits or to prove uniqueness.

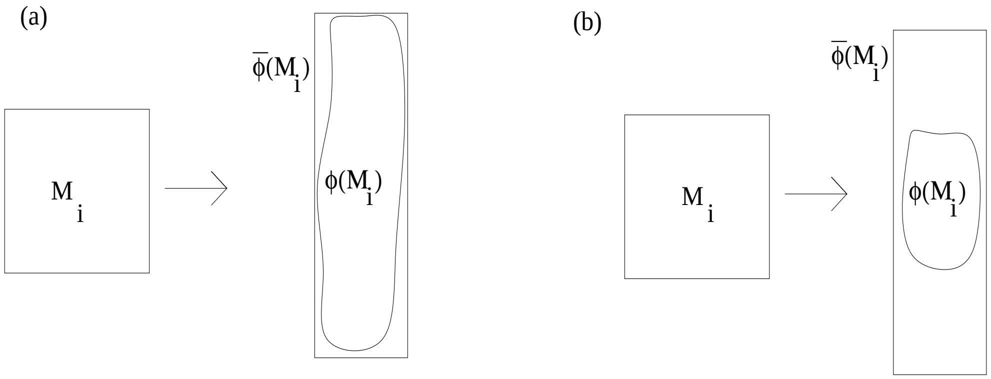


FIG. 4.1. *Enclosure methods can prove contraction but not expansion.*

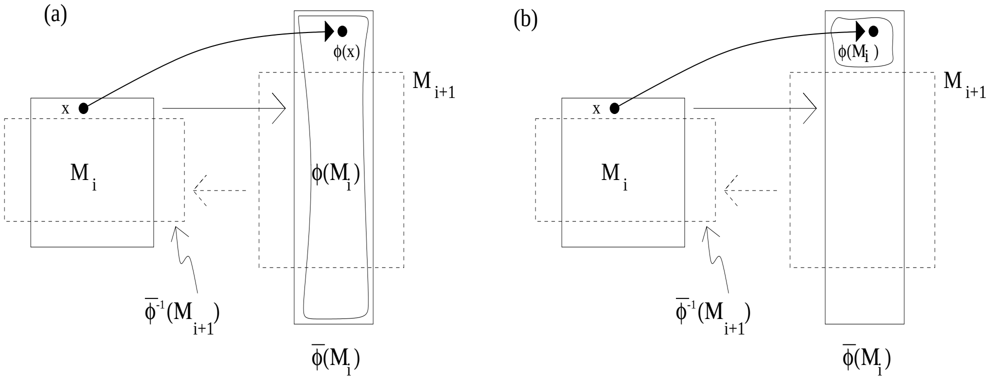


FIG. 4.2. (a) *The two validated integrations required to prove the* $(2,1)$*-ICP.* (b) *A potential problem, which is solved by doing a (cheap) point integration of one point on each expanding face, to verify that there are points of* $\varphi(M_i)$ *on both side of* M_{i+1}*.*

direction. However, enclosure methods cannot directly prove expansion, as Figure 4.1(b) illustrates: although $\bar{\varphi}(M_i)$ is a valid enclosure of $\varphi(M_i)$, it is not a very good one, because the actual image $\varphi(M_i)$ of M_i has not expanded in any direction. To solve this problem, we perform two validated integrations; refer to Figure 4.2(a). The first integration (solid rectangles) is a forward integration that provides $\bar{\varphi}(M_i)$, which in turn gives us a bound on the size of $\varphi(M_i)$ in the contracting directions (depicted as the horizontal direction in the figure). Now, assume we can find an M_{i+1} which satisfies the ICP not with $\varphi(M_i)$, but with $\bar{\varphi}(M_i)$. (If we cannot find such an M_{i+1}, then our method fails and we cannot prove the existence of a shadow beyond step i.) A validated integration *backwards* (dashed rectangles) is then performed on M_{i+1}, giving $\bar{\varphi}^{-1}(M_{i+1})$. If $\bar{\varphi}^{-1}(M_{i+1})$ proves that *contraction* has occurred in the nominally expanding directions when moving back from M_{i+1} to M_i, then we argue that expansion in forward time has occurred, as follows. Choose any $\mathbf{x} \in M_i - \bar{\varphi}^{-1}(M_{i+1})$. Since $\mathbf{x} \notin \bar{\varphi}^{-1}(M_{i+1}) \supset \varphi^{-1}(M_{i+1})$, this implies $\varphi(\mathbf{x}) \in \varphi(M_i) - M_{i+1}$. Since $F_i^{\pm 1} \subset M_i - \bar{\varphi}^{-1}(M_{i+1})$, this tells us that $\varphi(F_i^{\pm 1}) \cap M_{i+1} = \emptyset$. This is insufficient to prove ICP(1), as illustrated in Figure 4.2(b): perhaps $\bar{\varphi}(M_i)$ is a loose enclosure of $\varphi(M_i)$, and all of $\varphi(M_i)$ is actually on one side of M_{i+1}. To verify that this is not the case, we pick one point on each of F_i^{+1} and F_i^{-1} and perform a validated point

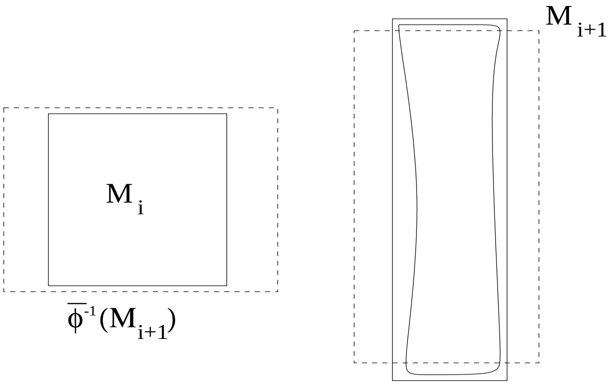


FIG. 4.3. *Shortcomings of the two-integration method: sometimes it can not prove expansion even if the M_{i+1} is valid.*

integration of each (which can be done cheaply) to verify that they land on opposite sides of M_{i+1}.[5] Since there is exactly one expanding direction, M_{i+1} cuts $\bar{\varphi}(M_i)$ into two disjoint sets, and a simple continuity argument shows that the two faces in their entirety land on opposite sides of M_{i+1}, thus verifying ICP(1). A similar argument in reverse time shows that the chosen M_{i+1} also verifies ICP(2(b)).

The argument of the previous paragraph clearly applies just as well in n dimensions when there is one expanding direction and $n-1$ contracting directions, for the same reasons that the two-dimensional proof of containment is easily transformed into Theorem 3.1 (Hayes (2001)). To prove that it also works when there is one contracting direction and $n-1$ expanding directions, note that there is a precise symmetry between the two cases (one expanding vs. one contracting): if we simultaneously reverse the order of $\{M_i\}_{i=0}^N$, giving $L_i = M_{N-i}$, and let $\psi = \varphi^{-1}$, then the above argument applies to the sequence $\{L_i\}_{i=0}^N$ using ψ as the homeomorphism. Thus, by symmetry, this method is also rigorous in the case when there is one contracting direction and $n-1$ expanding ones.

Figure 4.3 illustrates that it is possible to choose an M_{i+1} that satisfies the ICP but for which we cannot *verify* that the ICP holds. This occurs when M_{i+1} is chosen to be "almost as large" as $\bar{\varphi}(M_i)$ in the expanding directions; then, the excess when computing $\bar{\varphi}^{-1}(M_{i+1})$ swamps the contraction that occurs when integrating the expanding direction backwards in time. We solve this problem by iteratively shrinking M_{i+1} in the nominally expanding directions until $\bar{\varphi}^{-1}(M_{i+1})$ fits inside M_i in those directions. If we shrink M_{i+1} to size zero in the expanding direction without being able to integrate it backwards to fit inside M_i, then the method fails, and we cannot prove the existence of a shadow beyond step i. We have found empirically that, when the algorithm is succeeding, no more than 2 to 3 backwards integrations are usually required, independent of n. The number of backwards integrations is occasionally significantly larger, when the system encounters areas of nonhyperbolicity.

If the system were hyperbolic, then the nominally expanding directions would always expand, and the nominally contracting directions would always contract. However, in systems that are only pseudohyperbolic, the nominally expanding directions

[5] We have found empirically that this problem must be very rare, because it has not happened even once during our experiments. We suspect that it may be possible to prove the ICP without this extra point integration, but we have not devoted much thought to this matter.

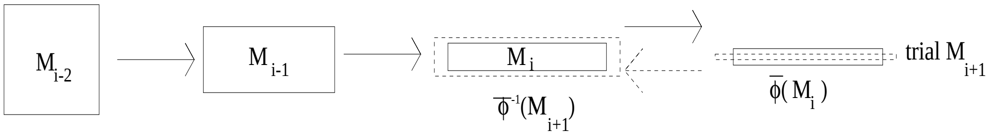


FIG. 4.4. *Example of the nominally expanding direction contracting too much for our integrator to prove contraction in the backwards direction.*

may expand most of the time, but not always, and vice versa for the contracting directions. One of the reasons our shadowing method can fail is if a nominally expanding direction contracts too much or for too long a time (Figure 4.4). Then, the expanding dimensions of M_i can become so small that no backwards integration from M_{i+1} can fit inside M_i in the nominally expanding directions.

4.1. Implementation issues and discussion. In the original paper that described containment, Grebogi et al. (GHYS) appear to have used boxes M_i of fixed size and found that smaller boxes seemed to work better. In contrast, our method dynamically grows and shrinks the M_i as i progresses, in an effort to maintain the ICP. In fact, we find it advantageous to choose the expanding dimension of M_i to be fairly large, to allow us to "absorb" possible future nonexpansion, in an effort to avoid the situation depicted in Figure 4.4. Similarly, we choose the contracting dimensions to be relatively small, to avoid the opposite effect (allowing us to "absorb" noncontraction without the nominal contracting dimensions becoming too large). Practically, we find that our "boxes" can be extremely long and thin: typically, they are of length 10^{-3} to 10^{-6} in the expanding dimensions, and as small as 10^{-12} to 10^{-14} in the contracting dimensions.

Referring once again to Figure 4.4, we note that when containment fails, the "expanding" dimension of M_i has often shrunk to almost the same size as the contracting dimension, and both can be quite small (say, 10^{-12}), whereas when containment is "working," the expanding dimension of M_i can be several orders of magnitude larger than the contracting dimension. It is interesting to note that this implies that the hardest parts of an orbit to shadow are the places where our bounds on the distance between the noisy and shadow orbits are *smallest*, i.e., where we can prove that they are unusually close together. This appears counterintuitive but may be related to the one-dimensional result of Chow and Palmer (1991), in which they proved that shadows must maintain a minimum distance from the noisy orbit.

5. Rescaling time.

5.1. Informal description. Containment as presented thus far has put no restrictions on φ other than that it is a homeomorphism. As has also been mentioned, all of our theorems and proofs have been based on a single application of φ, and there is no explicit connection between the φ used at one step and the one used on the next. Thus, everything said thus far is also applicable if we allow φ to change between steps. In particular, at each step we could use the time-h_i solution operator φ_{h_i}, with h_i being the length of the ODE integration timestep taken at step i. The resulting method for shadowing numerical ODE integrations has been dubbed the *map method* by Coomes, Koçak, and Palmer (1994b), (1995a), (1995b). However, ODE integrations suffer from errors in time. For systems in which the $\mathbf{y}'$ direction lacks even pseudohyperbolicity, errors in time (which manifest themselves in phase space as errors in the $\mathbf{y}'$ direction) can lead to short shadowing times that can be dramatically increased if time is *rescaled.* In this section, we describe how containment can be augmented to allow for the rescaling of time.

Our idea for rescaling time in containment was inspired in part by the rescaling of time developed by Coomes, Koçak, and Palmer (1994b), (1995a) (although our proofs are very different from theirs), and partly by the idea of the *Poincaré section*, also known as a *Poincaré map* or *return map.* There are several variations on this idea, but the one that concerns us is the following. Assume that the solution to an ODE is "almost periodic," in the sense that the solution passes through some fixed neighborhood of a given plane $\mathcal{H}$ approximately every T time units, where $\mathcal{H}$ is approximately perpendicular to the trajectory at the point where it crosses the plane. The Poincaré map generates the sequence of points at which the trajectory intersects $\mathcal{H}$. To accomplish the general rescaling of time, we modify this idea to remove the almost-periodic requirement of the orbit, and simply place a plane $\mathcal{H}_i$ in the vicinity of the solution at time t_i, placed so that $\mathcal{H}_i$ is approximately perpendicular to $\mathbf{y}'(t_i)$. Note that we do not compute $\mathcal{H}_i$; we only prove that it exists.

To facilitate containment, we must extend the idea of the Poincaré section to encompass a small ensemble of solutions. To that effect, we wish to take a set $M_{i-1} \subset \mathcal{H}_{i-1}$, where the diameter of M_{i-1} is small, and place an $(n-1)$-dimensional hyperplane $\mathcal{H}_i$ approximately normal to the flow in the vicinity of $\varphi_{h_{i-1}}(M_{i-1})$. Then we define the Poincaré section of the set $\varphi_{h_{i-1}}(M_{i-1})$ pointwise as follows. Let Δh_{i-1} bound the time interval over which the ensemble $\varphi_{h_{i-1}}(M_{i-1})$ crosses $\mathcal{H}_i$:

$$\forall \mathbf{x} \in M_{i-1} \quad \exists h \in [h_{i-1} - \Delta h_{i-1}/2, h_{i-1} + \Delta h_{i-1}/2] \quad \text{ s.t. } \varphi_h(\mathbf{x}) \in \mathcal{H}_i,$$

where we assume that, for each $\mathbf{x}$, the h chosen is unique. That is, we take the point-by-point Poincaré section of the points in M_{i-1} with respect to the plane $\mathcal{H}_i$. We call this a *splash* operation, because we imagine that the points in M_{i-1}, evolving via φ_h for $h \in [h_{i-1} - \Delta h_{i-1}/2, h_{i-1} + \Delta h_{i-1}/2]$, "splash" through $\mathcal{H}_i$ approximately simultaneously, and we assume that each trajectory intersects $\mathcal{H}_i$ precisely once during that interval; see Figure 5.1.

Our intent is to build $(n-1)$-dimensional parallelepipeds M_i inside $\mathcal{H}_i$ and then show that the point-by-point Poincaré section at $\mathcal{H}_i$—the splash operation—is a homeomorphism. We can then directly apply the previously proven containment theorems to the $(n-1)$-dimensional M_i's, which are each contained in the $(n-1)$-dimensional hyperplane $\mathcal{H}_i$, for an ODE system of n equations.

We note that since rescaling time via the splash operation effectively deletes one dimension from the problem, and since our map containment theorems are rigorous in three dimensions, this means that the methods presented in this paper are capable of rigorously shadowing ODE solutions of up to four dimensions, as long as a rescaling of time is applied.

5.2. Theorem: Splash is a homeomorphism. Refer to Figure 5.2. Let Q_i be an n-dimensional parallelepiped. Let $F_i^{\pm 1}$ be the two opposing faces of Q_i that are approximately normal to $\mathbf{y}'$ inside Q_i, and let $\mathbf{v}_i$ be the unit normal vector to these two faces, with $\mathbf{v}_i$ pointing from F_i^{-1} to F_i^{+1}. That is, $\mathbf{v}_i$ is approximately parallel to $\mathbf{y}'$ inside Q_i. Let D be the distance between F_i^{-1} and F_i^{+1} along $\mathbf{v}_i$. Let the infinite hyperplanes containing F_i^{-1} and F_i^{+1} be H_i^{-1} and H_i^{+1}, respectively, and let Z_i be the closed infinite slab between them. Let B_i be a parallelepiped with faces parallel to Q_i satisfying $Q_i \subset B_i \subset Z_i$, with two of the faces of B_i contained in $H_i^{\pm 1}$. Let $\{\mathbf{f}(\mathbf{x}) \cdot \mathbf{v}_i \mid \mathbf{x} \in B_i\} \subset [v_0, v_1]$, and assume $0 < v_0 \le v_1$.

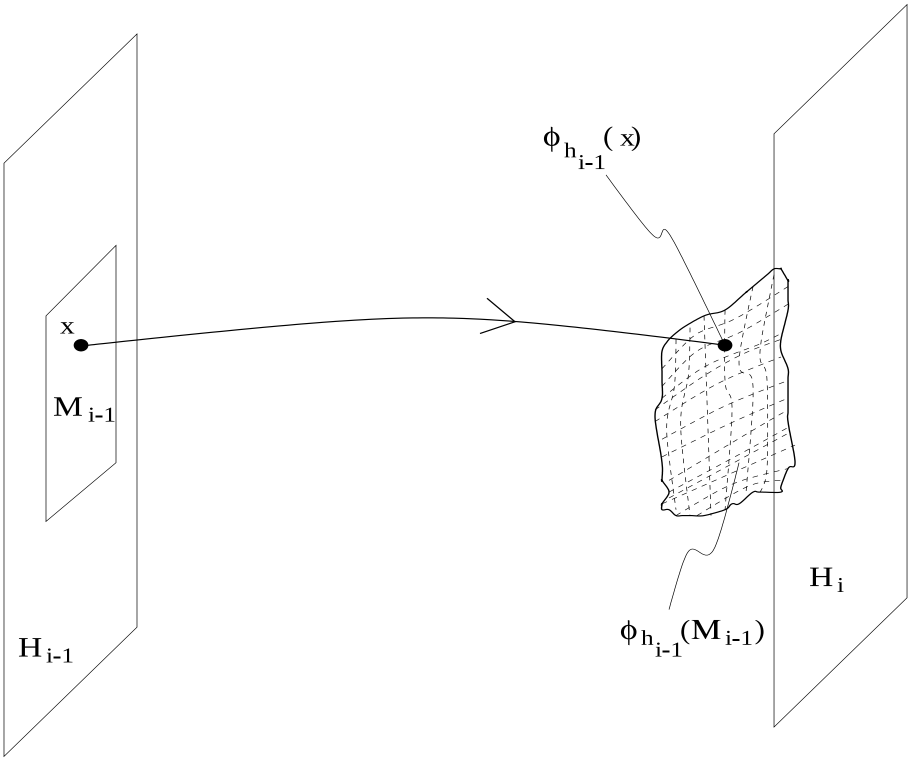


FIG. 5.1. *The "splash" operation depicted for a two-dimensional ensemble evolving in a three-dimensional configuration space. M_{i-1} is embedded in the plane $\mathcal{H}_{i-1}$ and evolves through one timestep to $\varphi_{h_{i-1}}(M_{i-1})$. As depicted, the ensemble is about to splash through $\mathcal{H}_i$.*

LEMMA 5.1. *If a trajectory remains in B_i while it is in Z_i, then it remains in Z_i for at least time $\underline{\varepsilon}_i^t \equiv D/v_1$ and at most $\bar{\varepsilon}_i^t \equiv D/v_0$.*

Proof. Let $\mathbf{y}(t)$ be a trajectory that remains in B_i while it is in Z_i. Let $z(t) = \mathbf{y}(t) \cdot \mathbf{v}_i$. Since $0 < v_0 \leq z'(t) \leq v_1$ and the width of B_i in the $\mathbf{v}_i$ direction is D, the maximum time to cross B_i is D/v_0, while the minimum time to cross is D/v_1. □

Let $\bar{\mathbf{f}}(B_i)$ be an enclosure of $\{\mathbf{f}(\mathbf{x}) \mid \mathbf{x} \in B_i\}$. Let S_i be a parallelepiped enclosure of $\{Z_i \cap (Q_i + h\bar{\mathbf{f}}(B_i)) \mid h \in [-\bar{\varepsilon}_i^t, \bar{\varepsilon}_i^t]\}$, and assume $S_i \subseteq B_i$.

Remark 5.1. S_i is intended to enclose the distance that a trajectory can drift from Q_i along the direction approximately perpendicular to $\mathbf{y}'$ as it travels across Z_i. This is required because a point in Q_i may not remain in Q_i when it is "splashed" onto $\mathcal{H}_i$. The following lemma formalizes this statement.

LEMMA 5.2. *Any trajectory intersecting Q_i remains in S_i while in Z_i, and thus remains in B_i as well.*

Proof. Since $S_i \subseteq B_i$, $\bar{\mathbf{f}}(B_i)$ bounds $\mathbf{y}' \equiv \mathbf{f}$ inside S_i. Since a trajectory remaining in B_i as it crosses Z_i does so in time $\leq \bar{\varepsilon}_i^t$, and since $S_i \subset B_i$, $\{h\bar{\mathbf{f}}(B_i) \mid h \in [-\bar{\varepsilon}_i^t, \bar{\varepsilon}_i^t]\}$ encloses the maximum possible distance from Q_i that a trajectory can travel in time $|\bar{\varepsilon}_i^t|$ while it remains in B_i. Thus, since $Q_i \subset S_i \subseteq B_i$, $\{Q_i + h\bar{\mathbf{f}}(B_i) \mid h \in [-\bar{\varepsilon}_i^t, \bar{\varepsilon}_i^t]\}$ encloses the position of any trajectory $\mathbf{y}(t)$ that is within time $\bar{\varepsilon}_i^t$ of intersecting Q_i, unless $\mathbf{y}(t)$ leaves Z_i during that time. Intersecting with Z_i completes the proof. □

Let $\mathcal{H}_i$ be any plane perpendicular to $\mathbf{v}_i$ which intersects the interior of Q_i. That

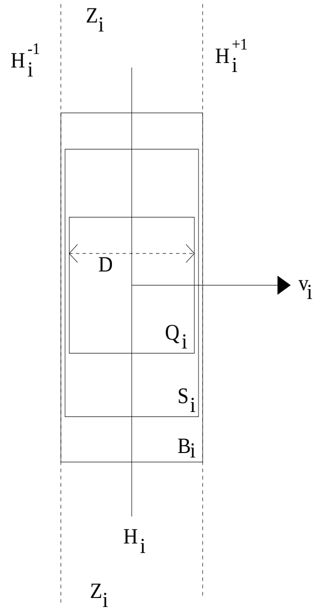


FIG. 5.2. *The objects used in Lemmas* 5.1–5.4. *Note that the left and right sides of* Q_i, S_i, B_i, *and* Z_i *are all in the planes* H_i^{-1}, H_i^{+1}, *respectively; they have been drawn as distinct for illustrative purposes only.*

is, $\mathcal{H}_i$ lies strictly between H_i^{-1} and H_i^{+1}.

LEMMA 5.3. *Every trajectory intersecting* Q_i *intersects* $\mathcal{H}_i$ *at precisely one point while it crosses* Z_i.

Proof. Let $\mathbf{y}(t)$ be a trajectory that intersects Q_i. By Lemma 5.2, $\mathbf{y}(t)$ remains in $S_i \subseteq B_i$ while it crosses Z_i. Let $z(t) = \mathbf{y}(t) \cdot \mathbf{v}_i$. Let the z coordinates of $H_i^{-1}, \mathcal{H}_i, H_i^{+1}$ be z_{-1}, z_0, z_{+1}, respectively. While the trajectory remains in $S_i \subseteq B_i$, $z'(t) \geq v_0 > 0$, and, since $z(t)$ is continuous, it increases monotonically while $\mathbf{y}(t)$ remains in S_i, taking on each value between z_{-1} and z_{+1} precisely once, by the intermediate value theorem. In particular, it takes on the value z_0 precisely once and thus crosses $\mathcal{H}_i$ precisely once. □

Assume that Q_i is an enclosure of $\varphi_{h_{i-1}}(M_{i-1})$. For a point $\mathbf{x} \in M_{i-1}$, let $\varphi_{i-1}(\mathbf{x})$ be the unique point in $\mathcal{H}_i$ defined by Lemma 5.3. Let $\bar{M}_i = S_i \cap \mathcal{H}_i$. Clearly, $\bar{M}_i$ is an enclosure of $\varphi_{i-1}(M_{i-1})$. To show that φ_{i-1} applied to M_{i-1} is a homeomorphism, we need to show that it is continuous and one-to-one. We first prove it is one-to-one.

Let $\varepsilon^t > 0$ be given. Recall $\bar{\varepsilon}_i^t$ as defined in Lemma 5.1.

Assumption 1. Assume $\bar{\varepsilon}_i^t < \varepsilon^t$ and $\nexists$ distinct $\mathbf{x}, \mathbf{y} \in M_{i-1}$ such that $\mathbf{y} = \varphi_t(\mathbf{x})$ for $|t| < \varepsilon^t$.

Each of the assumptions introduced in this section is assumed to hold throughout the remainder of section, once it is introduced.

LEMMA 5.4. *Each point in* $\varphi_{i-1}(M_{i-1})$ *comes from only one point in* M_{i-1}.

Proof. Assume to the contrary that there exist distinct $\mathbf{x}, \mathbf{y} \in M_{i-1}$ such that $\varphi_{i-1}(\mathbf{x}) = \varphi_{i-1}(\mathbf{y}) = \mathbf{z} \in \bar{M}_i$. Since $\varphi_{h_{i-1}}(\mathbf{x}), \varphi_{h_{i-1}}(\mathbf{y})$ both splash to $\mathbf{z}$, they are on the same trajectory, and since they are both in Q_i, the time-shift between them is $\leq \bar{\varepsilon}_i^t$. Thus, $\exists t_1, t_2$ such that $\varphi_{t_1}(\mathbf{x}) = \mathbf{z} = \varphi_{t_2}(\mathbf{y})$ with $|t_1 - t_2| \leq \bar{\varepsilon}_i^t$. Then $\mathbf{y} = \varphi_{t_1 - t_2}(\mathbf{x})$, contradicting Assumption 1. □

THEOREM 5.5. *φ_{i-1} applied to M_{i-1} is one-to-one.*

Proof. Lemma 5.3 proves that $\varphi_{i-1}(M_{i-1})$ is many-to-one, and Lemma 5.4 proves it is one-to-many. Thus, it is one-to-one. □

We now prove that $\varphi_{i-1}(\mathbf{x})$ is continuous for all $\mathbf{x} \in M_{i-1}$.

Assumption 2. $\varphi_t(\mathbf{x})$ exists and is continuous in both t and $\mathbf{x}$ $\forall \mathbf{x} \in M_{i-1}$ and $\forall t$ such that $\varphi_t(\mathbf{x}) \in B_i$. Note that this is true as long as $\mathbf{f}$ is Lipschitz continuous (Stuart and Humphries (1996, Theorem 2.1.12)).

We will need the following theorem.

THEOREM 5.6. *If $\mathbf{y}$ and $\mathbf{z}$ each satisfy the differential equation $\mathbf{y}'(t) = \mathbf{f}(\mathbf{y}(t))$ on the interval $[t_0, t_1]$, and if $\mathbf{f}$ is Lipschitz continuous with constant L, then $\forall t \in [t_0, t_1]$,*

$$\|\mathbf{y}(t) - \mathbf{z}(t)\| \leq \|\mathbf{y}(t_0) - \mathbf{z}(t_0)\| e^{L(t-t_0)}.$$

Proof. See Theorem 112J of Butcher (1987). □

For a point $\mathbf{x} \in M_{i-1}$, let $h_{i-1}(\mathbf{x})$ be that unique timestep defined by $\varphi_{h_{i-1}(\mathbf{x})}(\mathbf{x}) \in \mathcal{H}_i$. That is, $\varphi_{i-1}(\mathbf{x})$ specifies *where* $\mathbf{x}$ goes, and $h_{i-1}(\mathbf{x})$ specifies *how long* it takes to get there.

LEMMA 5.7. *If $\mathbf{f}$ is Lipschitz continuous, then $\forall \mathbf{x}_0 \in M_{i-1}$, $h_{i-1}(\mathbf{x})$ is continuous at $\mathbf{x} = \mathbf{x}_0$.*

Proof. For simplicity, we will drop the subscript from $h_{i-1}(\mathbf{x})$ during this proof.

Let L be the Lipschitz constant for $\mathbf{f}$. Then by Theorem 5.6, for any $\mathbf{x}_0, \mathbf{x}$,

$$\|\varphi_{h(\mathbf{x}_0)}(\mathbf{x}) - \varphi_{h(\mathbf{x}_0)}(\mathbf{x}_0)\| \leq \|\mathbf{x} - \mathbf{x}_0\| e^{Lh(\mathbf{x}_0)} \equiv \delta_3(\mathbf{x}, \mathbf{x}_0).$$

Since we are interested only in the behavior of $h(\mathbf{x})$ in a neighborhood of $\mathbf{x}_0$, choose $\mathbf{x} \in M_{i-1}$ close enough to $\mathbf{x}_0$ so that $\varphi_{h(\mathbf{x}_0)}(\mathbf{x}) \in B_i$. Now, since $\varphi_{h(\mathbf{x}_0)}(\mathbf{x}_0) \in \mathcal{H}_i$, the distance from $\varphi_{h(\mathbf{x}_0)}(\mathbf{x})$ to $\mathcal{H}_i$ is also bounded above by $\delta_3(\mathbf{x}, \mathbf{x}_0)$. Since $\varphi_{h(\mathbf{x}_0)}(\mathbf{x}) \in B_i$, the maximum time to intersect $\mathcal{H}_i$ is $\delta_3(\mathbf{x}, \mathbf{x}_0)/v_0$. Thus, $h(\mathbf{x}) \in [h(\mathbf{x}_0) - \delta_3(\mathbf{x}, \mathbf{x}_0)/v_0, h(\mathbf{x}_0) + \delta_3(\mathbf{x}, \mathbf{x}_0)/v_0]$. The continuity of $h(\mathbf{x})$ at $\mathbf{x}_0$ follows by letting $\mathbf{x} \to \mathbf{x}_0$. □

LEMMA 5.8. *$\varphi_{i-1}(\mathbf{x})$ is continuous $\forall \mathbf{x} \in M_{i-1}$.*

Proof. By definition, $\varphi_{i-1}(\mathbf{x}) = \varphi_{h_{i-1}(\mathbf{x})}(\mathbf{x})$, and by construction, $\varphi_{h_{i-1}(\mathbf{x})}(\mathbf{x}) \in S_i \subseteq B_i$. Since the composition of two continuous functions is continuous and Lemma 5.7 asserts that $h_{i-1}(\mathbf{x})$ is continuous, Assumption 2 directly implies that $\varphi_{i-1}(\mathbf{x})$ is continuous. □

Thus, $\varphi_{i-1}(\mathbf{x}) \equiv \varphi_{h_{i-1}(\mathbf{x})}(\mathbf{x})$ is the unique splash point of $\mathbf{x}$ in $\mathcal{H}_i$.

Finally, the second part of Assumption 1 cannot be taken for granted. The following lemma is applied at step i to give us the second part of Assumption 1 at step $i+1$.

Let W_i be an infinite slab with width $E > D$ in the $\mathbf{v}_i$ direction, parallel to Z_i such that $Z_i \subset W_i$. Let C_i be a parallelepiped with sides parallel to Q_i, also with a width of E in the $\mathbf{v}_i$ direction, satisfying $M_i \subset C_i \subset W_i$, where M_i is built inside $\mathcal{H}_i$ to satisfy the ICP with M_{i-1} under φ_{i-1}. Let $E_{+1} > 0$ be the distance from $\mathcal{H}_i$ in the $\mathbf{v}_i$ direction to the face of W_i, and let $E_{-1} > 0$ be the distance to the opposite face of W_i. Note that $E_{-1} + E_{+1} = E$. Let $\{\mathbf{f}(\mathbf{x}) \cdot \mathbf{v}_i \mid \mathbf{x} \in C_i\} \subset [u_0, u_1]$, and assume $0 < u_0 \leq u_1$. Let $\bar{\mathbf{f}}(C_i)$ be an enclosure of $\{\mathbf{f}(\mathbf{x}) \mid \mathbf{x} \in C_i\}$. Let T_i be a parallelepiped enclosure of $\{W_i \cap (M_i + h\bar{\mathbf{f}}(C_i)) \mid h \in [-\varepsilon^t, \varepsilon^t]\}$, and assume $T_i \subseteq C_i$.

Assumption 3. Assume $E/u_1 > \varepsilon^t$. That is, the minimum crossing time of C_i is greater than ε^t.

LEMMA 5.9. *$\nexists$ distinct $\mathbf{x}, \mathbf{y} \in M_i$ such that $\mathbf{y} = \varphi_t(\mathbf{x})$ for $|t| < \varepsilon^t$.*

Proof. Substituting M_i for Q_i, W_i for Z_i, T_i for S_i, and C_i for B_i in Lemmas 5.1–5.3, we see that

(1) If a trajectory remains in C_i while it is in W_i, then it remains in W_i for at least time E/u_1 and at most E/u_0. By a similar argument, the minimum and maximum times between such a trajectory's entering C_i and intersecting $\mathcal{H}_i$ are E_{-1}/u_1 and E_{-1}/u_0, respectively, and the corresponding times between such a trajectory's intersecting $\mathcal{H}_i$ and exiting C_i are E_{+1}/u_1 and E_{+1}/u_0.

(2) Any trajectory intersecting M_i remains in T_i while it is in W_i, and thus it remains in C_i.

(3) Every trajectory intersecting M_i intersects $\mathcal{H}_i$ at precisely one point while it remains in W_i, where $\mathcal{H}_i \subset W_i$ and $\mathcal{H}_i$ is parallel to the planes enclosing W_i.

Thus, by point (3), to intersect $\mathcal{H}_i$ more than once inside M_i, a trajectory must, at least, first traverse the distance from $\mathcal{H}_i$ to ∂C_i, exit and then reenter C_i, and traverse the distance from ∂C_i back to $\mathcal{H}_i$. By point (1), it takes time at least $E_{-1}/u_1 + E_{+1}/u_1 = E/u_1$ to do so. By Assumption 3, $E/u_1 > \varepsilon^t$. Thus, no trajectory can intersect M_i, exit T_i, and then reenter T_i to again intersect M_i in time less than ε^t. □

Remark 5.2. The base case of the induction is produced by substituting M_0 for M_i in Lemma 5.9, after building suitable W_0, C_0, and T_0.

5.3. Algorithmic details. Algorithmic verification of the requirements for the above theorems and lemmas is fairly straightforward: Q_i is simply the enclosure of $\varphi_{h_{i-1}}(M_{i-1})$ given to us by VNODE; the size of B_i is computed heuristically in an effort to ensure that $S_i \subseteq B_i$, and if our first guess is incorrect, we simply increase its size until $S_i \subseteq B_i$, or fail if increasing the size of B_i results in $0 \in \{\mathbf{f}(\mathbf{x}) \cdot \mathbf{v}_i \mid \mathbf{x} \in B_i\}$; ε^t, which is an upper bound on the time error introduced at each step by the rescaling of time, must currently be prechosen by trial and error, although we believe that good, simple heuristics for choosing it probably exist. The sole complication is in maintaining the property that Q_i has a pair of faces approximately normal to $\mathbf{y}'$ inside Q_i. Note that VNODE maintains a rotation matrix A_i, which represents the orientation of the parallelepiped Q_i. Let the columns of A_i be $\mathbf{a}_i^j, j = 1, \ldots, n$. We simply assign $\mathbf{a}_i^1$ to be parallel to our best estimate of $\mathbf{y}'(t_i)$. VNODE then ensures that $\mathbf{a}_{i+1}^1$ evolves via the variational equation to be approximately parallel to $\mathbf{y}'(t_{i+1})$. To account for the slow buildup of error that would allow $\mathbf{a}_i^1$ to drift away from $\mathbf{y}'(t_i)$, we reset $\mathbf{a}_i^1$ to be parallel to the computed $\mathbf{y}'(t_i)$ at each timestep. This corresponds to rotating Q_i about its center by a small angle θ, computed by solving

$$\cos(\theta) = \frac{\mathbf{a}_i^1 \cdot \mathbf{y}'(t_i)}{\|\mathbf{a}_i^1\| \, \|\mathbf{y}'(t_i)\|},$$

where $\mathbf{a}_i^1$ is the vector computed via evolution of the ODE from the previous timestep, and $\mathbf{y}'(t_i)$ is the value of $\mathbf{y}'$ computed directly from the right-hand side of the ODE at the current timestep. The largest distance a point in Q_i will move as a result of this rotation is $r\theta$, where r is the distance of the furthest corner in Q_i from its center. Thus, after rotating Q_i by θ, we increase its size by $r\theta$ in all directions, thus ensuring that it still encloses $\varphi_{h_{i-1}}(M_{i-1})$.

A simple variable stepsize algorithm was used: whenever containment of a particular step succeeds, we increase the stepsize by a small factor; whenever it fails, we decrease the stepsize by a factor of 2. We do not explicitly fail due to small stepsize, because too small a stepsize results in failures in other parts of the method, for example, as depicted in Figure 4.4.

6. Results and discussion. In this section, we present results of our containment method for ODEs, compare our results to those of others, discuss some of the

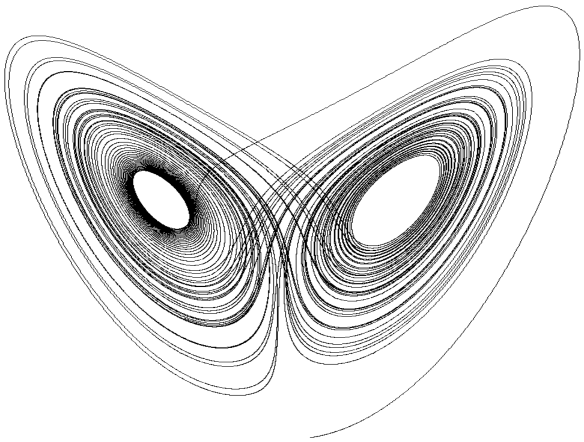


FIG. 6.1. *The "Lorenz butterfly."*

interesting implementation details of our method, and comment on observations of the behavior of our method, including how it sometimes fails.

6.1. Quantitative comparisons with other methods.

6.1.1. The Lorenz system of equations. The Lorenz equations (Lorenz (1963)),

$$\begin{pmatrix} x' \\ y' \\ z' \end{pmatrix} = \begin{pmatrix} \sigma(y-x) \\ \rho x - y - xz \\ xy - \beta z \end{pmatrix}, \tag{6.1}$$

define a dissipative dynamical system (i.e., energy is not conserved), which was originally constructed to be a very simplified weather model. It can be shown (Coomes, Koçak, and Palmer (1995a)) that, under the Lorenz equations, the set

$$U = \{(x,y,z) : \rho x^2 + \sigma y^2 + \sigma(z-2\rho)^2 \le \sigma\rho^2\beta^2/(\beta-1)\}$$

is *forward invariant:* any solution that is in U at time t_0 remains in U for all time $t \ge t_0$. All the methods discussed in this section solve the Lorenz equations with the classical parameter values $\sigma = 10, \rho = 28, \beta = 8/3$ (Lorenz (1963)). It is easy to show that, for these parameter values, the cube $[0,15]^3$ lies in U, and so for our experiments we chose initial conditions randomly inside this cube. A set of initial conditions in this cube will invariably produce a solution whose three-dimensional shape has been dubbed the "Lorenz butterfly" (Figure 6.1). Schematically, the Lorenz butterfly consists of two two-dimensional disks in three-space with a "bridge" between them. The two disks together are termed a "chaotic attractor," because solutions tend to remain in the disks but jump chaotically from one to the other and back again. Solutions lack pseudohyperbolicity in the direction of the flow (Van Vleck (1995); Coomes, Koçak, and Palmer (1994b), (1995a)), and so a rescaling of time is required to shadow them effectively. As should be clear from Figure 6.1 and the above description, in addition to the $\mathbf{y}'$ direction, at any given point a solution has one contracting direction, which is perpendicular to the disk currently housing the solution, and one expanding direction, directed radially from the center of the disk. Provided a rescaling of time is employed, solutions to the Lorenz equations display

Table 6.1

Comparison of shadow lengths for the Lorenz system. VV=Van Vleck (1995)*; CKP = Coomes, Koçak, and Palmer* (1994b), (1995a).

Author	Local error	Global error	Map method	Rescaling time
VV	10^{-6}	10^{-5}	1–2	$10^2 \sim 10^4$
Hayes	10^{-6}	10^{-5}	$10 \sim 50$	$10^3 \sim 10^5$
CKP	10^{-13}	10^{-9}	$10 \sim 100$	$\geq 10^5$
Hayes	10^{-13}	10^{-9}	$10 \sim 1000$	$\geq 7.7 \times 10^5$

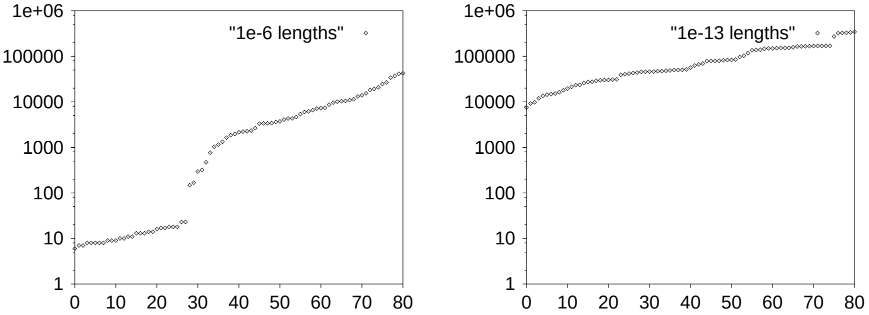


Fig. 6.2. *Distribution of shadow lengths computed by containment with a rescaling of time. Each panel shows a sorted list of shadow lengths for* 80 *simulations of the Lorenz equations. The horizontal axis is simply a label for each shadow; the vertical axis is its length. The magnitude of the noise (i.e., the local error) in the noisy orbits is about* 10^{-6} *in the left graph and* 10^{-13} *in the right.*

remarkable pseudohyperbolicity for extremely long periods of time. Thus, this system is a prime first candidate for testing shadowing methods.

We will compare our results to the only other published results on shadowing the Lorenz equations using a rescaling of time: Van Vleck (1995), whose results could be made rigorous but currently are not; and Coomes, Koçak, and Palmer (1994b), (1995a), whose results are completely rigorous.

First, with *no* rescaling of time (the "map method"), Van Vleck gives two examples of shadows with a local error[6] of about 10^{-5} lasting 1.04 and 1.38 time units; Coomes, Koçak, and Palmer have six examples with local error of about 10^{-13} lasting 9.7, 9.8, 9.9, 9.9, 86, and 126 time units. For this paper, we have simulated hundreds of shadows with various local errors. We have found that with local errors of about 10^{-5}, containment finds shadows that last between 1 and 30 time units, with a median and mean of about 20. With local errors of 10^{-13}, we find shadows lasting between 10 and 1000 time units, again with a mean and median about halfway through that range. Thus, it appears that, without a rescaling of time, the containment method is capable of finding shadows that are about an order of magnitude longer than other existing methods.

With a rescaling of time, Van Vleck gives many examples of shadows (with a local error of about 10^{-6}) ranging from 10^2 to 10^4 time units. Coomes, Koçak, and Palmer (with a local error of 10^{-13}) give six examples of shadows lasting at least 10^5 time units; they do not attempt to find longer shadows, so in fact their method

[6] The local errors used in the current paper were normalized to have comparable size per-unit-step to other methods, even though variable stepsize methods were used both for the validated ODE integration (Nedialkov (1999)) and for choosing the size of shadow steps.

may be capable of finding shadows longer than 10^5. The corresponding numbers for containment are 10^2 to 10^5 for local errors of 10^{-6}, and 10^2 to almost 10^6 for local errors of 10^{-13}. The results are summarized in Table 6.1.[7] It is clear that containment is at least as powerful as the other existing methods. It is worth noting that our results for local errors of 10^{-13} were produced using only a 17th-order Taylor series, whereas Coomes, Koçak, and Palmer used a Taylor series of 31st order.

Figure 6.2 shows two sets of results of shadow lengths, including the rescaling of time. The first is for eighty solutions with local error of approximately 10^{-6}, and the second for eighty solutions with local error of approximately 10^{-13}. The sharp increase in shadow lengths occurring just left of center in the first figure is probably due to the fact that, other than choosing $\mathbf{v}_0$ (cf. Figure 5.2 on page 1963) to be parallel to $\mathbf{y}'(t_0)$, the directions of the faces of M_0 are currently chosen at random. As a result, we sometimes choose nominally expanding and contracting directions that are not sufficiently close to the actual expanding and contracting directions. Thus, many shadows fail early due to this problem. However, if our nominally chosen directions are (by luck) close enough to the actual ones, then we get over this hump to find much longer shadows. There is probably a more clever way to choose the initial M_0, but we have not yet studied this problem closely. This problem becomes less pronounced as the local error decreases and is virtually absent in the right figure, which has local error $\delta = 10^{-13}$.

In addition, our shadowing distances (i.e., the maximum distance between the shadow and the numerical trajectory) are comparable to the methods of the above authors: for orbits with noise 10^{-6} and 10^{-13}, our method and those of Van Vleck and Coomes, Koçak, and Palmer find shadowing distances of approximately 10^{-5} and 10^{-9}, respectively. For containment, these sizes are based on ε^t and the maximum size of M_i over all i, which are at least in part user-controlled. For Van Vleck and Coomes, Koçak, and Palmer the shadowing distances are computed analytically based upon global bounds of various computed quantities.

6.1.2. Other systems of equations. We have reproduced the shadowing experiments of several other authors, usually getting comparable results, as illustrated in Table 6.2. We discussed results for the Lorenz system in the previous section. In this section, we provide results for three other problems.

Forced damped pendulum. We first compare our results for the forced damped pendulum problem,

$$y'' + ay' + \sin y = b \cos t,$$

to those of GHYS, Sauer and Yorke (1991), and Chow and Van Vleck (1994). These authors use the values $a = 0.2, b = 2.4$ and $a = 1, b = 2.4$, with initial conditions $(y, y') = (0, 0)$, and mention that they get similar results with other pairs of values of a, b and initial conditions. We used the above two pairs of values for a, b and various random initial conditions in the unit square $[0, 1]^2$. We convert the second-order equation to two first-order equations by assigning $y_1 = y,\ y_2 = y'$, giving

$$y_1' = y_2,$$

[7] Our attempts to find the longest possible shadows for the latter case have been repeatedly confounded by having either workstation or disk crashes (independent of our code) while our simulations were running. The longest shadow we have observed is thus 7.7×10^5, even though, had our machines not crashed, the shadows might have been longer.

Table 6.2

Comparison of shadow lengths for four systems. For our results, the lengths shown are typical results after attempting many trials with the given local and global errors; the results of others are taken from their respective publications. Legend: δ = local error; ε = global space error; ε_t = global time error (if none is listed for our method, then we did not rescale time); L = shadow length; CKP = Coomes, Koçak, and Palmer (1994b), (1995a)*; SY = Sauer and Yorke* (1991)*; CVV = Chow and Van Vleck* (1994a)*; VV = Van Vleck* (1995)*; NR = not rigorous.*

System	Auth.	δ	ε	ε_t	L	Comment
Lorenz						
	VV	10^{-6}	10^{-5}		10^4	NR
	Hayes	10^{-6}	10^{-5}	2.5×10^{-5}	10^3–10^5	
	CKP	10^{-13}	10^{-9}		$\geq 10^5$	
	Hayes	10^{-13}	10^{-9}	2.5×10^{-9}	$\geq 7.7 \times 10^5$	
Forced damped pendulum						
	SY	10^{-18}	10^{-9}		3×10^4	High machine precision
	Hayes	10^{-15}	10^{-6}	10^{-3}	10^3–3×10^4	
	CVV	10^{-6}	10^{-3}		10^4	NR
	Hayes	10^{-6}	10^{-5}	10^{-3}	10^3	
	CVV	10^{-11}	10^{-8}		10^3	NR
	Hayes	10^{-11}	10^{-8}	10^{-3}	10^3	
Forced van der Pol						Periodic attractor
	VV	10^{-5}	10^{-4}		10^4	NR
	Hayes	10^{-5}	10^{-6}	3×10^{-5}	$\geq 10^5$	
Logistic equation						
	CVV	10^{-7}	5×10^{-6}		9.22	$y_0 = 0.01$, fixed L, NR
	Hayes	10^{-7}	10^{-6}		9.22	
	CVV	10^{-7}	5×10^{-6}		18.46	$y_0 = 10^{-4}$, fixed L, NR
	Hayes	10^{-7}	10^{-6}		18.46	

$$y_2' = b\cos t - \sin y_1 - ay_2.$$

GHYS and Sauer and Yorke (1991) use extended precision arithmetic with a machine epsilon of 10^{-29} to generate a trajectory with local truncation error rigorously bounded by 10^{-18} per step, which allows them to find a shadow of length 3×10^4 and rigorous maximum distance 10^{-9} from their noisy trajectory. In comparison, we use standard IEEE754 floating-point numbers and arithmetic and obtain a local truncation error of about 10^{-15} at best, so our shadow distances are significantly less stringent at 10^{-6}, and tend to be shorter, although in a few instances we successfully found shadows of length $\sim 3\times10^4$. Given that Sauer and Yorke used higher precision, we are not surprised that our shadows tend to be shorter and not as close as theirs. Comparing our results to Chow and Van Vleck (1994), we see our method is capable of rigorously proving the existence of a shadow which is closer, but lasts for a shorter time, than their method does; on the other hand, our result is rigorous, whereas theirs is not, because they do not rigorously bound numerical errors before applying their theorem.

The primary problem with shadowing this system appears to be that it is nonautonomous. We currently handle a nonautonomous system by converting it to an autonomous system with one component of our solution, y_0, representing time: $y_0(0) = t_0,\ \ y_0'(t) = 1$. This has several drawbacks: (1) the new component is decidedly nonhyperbolic; (2) assuming we can solve the linear system $y' = 1$ exactly, the interval representing y_0 then accumulates roundoff error, and as time progresses, the error in

y_0 grows; (3) this is exacerbated by the minimum absolute error in y_0 increasing as $\varepsilon_{mach}t$, where ε_{mach} is the machine precision; (4) finally, the error in the computation of $\cos(y_0)$ adds to the error. These drawbacks, however, do not seem to adequately explain our poor shadowing results for this system. Perhaps the difficulties would vanish if a native procedure for validated integration of nonautonomous systems were used, or if we used higher precision, as did Sauer and Yorke (1991).

Forced van der Pol. The forced van der Pol equation,

$$x'' + \alpha(x^2 - 1)x' + x = \beta\cos(\omega t),$$

is studied by Van Vleck (1995). He defines the parameters implicitly with $\alpha = k = \sigma = 2/5$, where $k = \beta/(2\alpha)$ and $\sigma = (1 - \omega^2)/\alpha$, and uses the initial conditions $(x, x') = (0, 0)$. We try this initial condition, as well as others chosen randomly in the unit square $[0, 1]^2$, and we convert the second-order equation to two first-order equations by assigning $y_1 = x,\ y_2 = x'$, giving

$$y_1' = y_2,$$

$$y_2' = \beta\cos(\omega t) - (y_1^2 - 1)\alpha y_2 - y_1.$$

This equation has a hyperbolic periodic attractor, which all solutions approach asymptotically, and so this system is easy to shadow. With a local truncation error of 10^{-6}, Van Vleck found numerical shadows of length 10^4 and distance 10^{-4}, while we went significantly further, finding rigorous shadows lasting 10^5 and longer with a distance of 10^{-6}. Since solutions asymptotically approach a periodic solution that is hyperbolic, we conjecture that containment could be maintained indefinitely.

Logistic equation. Finally, the logistic equation,

$$y' = y(1 - y), \quad y(0) = \zeta,\ 0 < \zeta \ll 1,$$

was studied by Chow and Van Vleck (1994). In this problem, there is an unstable fixed point at $y = 0$ and a stable fixed point at $y = 1$. Chow and Van Vleck attempt shadowing two solutions, both starting at $y(0) = \zeta$ and integrating until $y(T) \approx 1 - \zeta$. If $\zeta = 10^{-2}$, then $T \approx 9.22$, and if $\zeta = 10^{-4}$, then $T \approx 18.46$. In both cases, we use a local truncation error of $\delta = 10^{-7}$. We find that we easily match their results, noting again that ours are rigorous, while theirs are not. In fact, we find that we can prove the existence of these shadows for $\varepsilon \approx 10\delta$ for δ down to about 10^{-14}.

6.2. Qualitative comparisons with other methods. First and foremost, our method has only been proven to work in a limited number of special cases. Generalizing the (n, k)-ICP to arbitrary (n, k) is straightforward Hayes (2001). Proving that it implies the existence of a shadow is more difficult, and is in progress. See Hayes (2001) for more discussion.

Although containment is rigorous, it appears to be less robust than nonrigorous methods. For example, in two examples out of three, the nonrigorous results of Chow and Van Vleck (1994) produced shadows that were about an order of magnitude longer than we could produce using containment. In addition, Hayes (1995) presented convincing evidence that the gravitational n-body problem is shadowable, and yet containment could prove the existence of shadows lasting only 1% as long as those found nonrigorously in Hayes (1995). Even worse, the VNODE package (Nedialkov (1999)) is capable of providing a validated enclosure of an IVP for the n-body problem which is about ten times as long as the containment-produced shadow! Clearly, if an enclosure of an IVP exists, then a shadow exists for the associated point solution for at least as long. Thus, at least for some problems, our implementation of containment is incapable of finding shadows even though they exist. This does not necessarily imply that the theorems proved in section 2.2 are deficient; it probably means that our implementation for verifying that the ICP holds can be improved, for example by reducing the excess of the validated numerical integrator.

Our method requires some a priori guesses; for example, the maximum and minimum sizes of the M_i, and the maximum time rescaling ε^t, need to be chosen before the algorithm runs. We typically had to choose these numbers by trial and error for each problem; if a certain ε^t did not work, for example, we often found that increasing it or decreasing the maximum size of M_i would allow us to find longer or closer shadows, respectively. Van Vleck's (1995) method also requires some a priori guesswork to make a rescaling of time work. Although Coomes, Koçak, and Palmer do not discuss their choice of parameters, it is likely that they require significant guesswork to find parameters that satisfy their theorems as well. Finally, *all* shadowing methods currently in the literature appear to require guesswork to discover the number of expanding and contracting dimensions and to choose a local error δ which is stringent enough to satisfy their respective theorems.

It is also not trivial to see how containment could be parallelized, since each M_i depends on M_{i-1}. Possibly an iterative method that guesses all the $\{M_i\}_{i=0}^N$ and then iteratively refines them in parallel could be constructed; this may also be related to two-point boundary value problems (Ascher, Mattheij, and Russell (1988)).

On the other hand, containment appears to have several advantages over other methods.

- We use an off-the-shelf validated integrator (Nedialkov (1999)) to verify that ICP holds; this integrator is almost as easy to use as any standard integrator, and thus getting the code "up and running" on a new problem usually takes only a few minutes. Another advantage of this simplicity is that it requires the user to have no deeper understanding of the system than knowing the defining equations.[8]
- Although the success of containment may depend, of course, upon global properties of the system, the method itself is local. By that we mean that it requires information only from the previous step to extend the length of the shadow. Several other methods require computing, storing, and updating global information such as the extent of nonhyperbolicity (cf. Chow and Palmer's p parameter 1991, 1992).

7. Conclusions.

We have extended the simple and elegant *containment* method of producing shadows from two-dimensional maps to maps of arbitrary dimension in which some measure of hyperbolicity is present and there is either 0 or 1 expanding modes, or 0 or 1 contracting modes, and added a rescaling of time to allow containment to work better for ODEs. We have demonstrated that this new method produces shadows of ODE integrations that are of comparable quality and length to any currently in the literature, and noted how it can be used to prove the existence of chaos.

[8] Some may consider this a disadvantage.

REFERENCES

U. M. Ascher, R. M. M. Mattheij, and R. D. Russell (1998), *Numerical Solution of Boundary Value Problems for Ordinary Differential Equations*, Prentice-Hall Series in Comput. Math., Prentice-Hall, Englewood Cliffs, NJ.

W.-J. Beyn (1987), *On invariant closed curves for one-step methods*, Numer. Math., 51, pp. 103–122.

J. C. Butcher (1987), *The Numerical Analysis of Ordinary Differential Equations*, Wiley, New York.

S.-N. Chow and K. J. Palmer (1991), *On the numerical computation of orbits of dynamical systems: The one-dimensional case*, J. Dynam. Differential Equations, 3, pp. 361–380.

S.-N. Chow and K. J. Palmer (1992), *On the numerical computation of orbits of dynamical systems: The higher dimensional case*, J. Complexity, 8, pp. 398–423.

S.-N. Chow and E. S. Van Vleck (1994), *A shadowing lemma approach to global error analysis for initial value ODEs*, SIAM J. Sci. Comput., 15, pp. 959–976.

B. A. Coomes (1997), *Shadowing orbits of ordinary differential equations on invariant submanifolds*, Trans. Amer. Math. Soc., 349, pp. 203–216.

B. A. Coomes, H. Koçak, and K. J. Palmer (1994a), *Periodic shadowing*, in Chaotic Numerics, P. Kloeden and K. Palmer, eds., Contemp. Math. 172, AMS, Providence, RI, pp. 115–130.

B. A. Coomes, H. Koçak, and K. J. Palmer (1994b), *Shadowing orbits of ordinary differential equations*, J. Comput. Appl. Math., 52, pp. 35–43.

B. A. Coomes, H. Koçak, and K. J. Palmer (1995a), *Rigorous computational shadowing of orbits of ordinary differential equations*, Numer. Math., 69, pp. 401–421.

B. A. Coomes, H. Koçak, and K. J. Palmer (1995b), *A shadowing theorem for ordinary differential equations*, Z. Angew. Math. Phys., 46, pp. 85–106.

B. A. Coomes, H. Koçak, and K. J. Palmer (1997), *Long periodic shadowing*, Numer. Algorithms, 14, pp. 55–78.

R. M. Corless (1994), *Error backward*, in Chaotic Numerics, P. Kloeden and K. Palmer, eds., Contemp. Math. 172, pp. 31–62.

G. Dahlquist and Å. Björck (1974), *Numerical Methods*, Prentice-Hall Series in Automatic Computation, Prentice-Hall, Englewood Cliffs, NJ.

C. Grebogi, S. M. Hammel, J. A. Yorke, and T. Sauer (1990), *Shadowing of physical trajectories in chaotic dynamics: Containment and refinement*, Phys. Rev. Lett., 65, pp. 1527–1530.

S. M. Hammel, J. A. Yorke, and C. Grebogi (1987), *Do numerical orbits of chaotic dynamical processes represent true orbits?*, J. Complexity, 3, pp. 136–145.

S. M. Hammel, J. A. Yorke, and C. Grebogi (1988), *Numerical orbits of chaotic dynamical processes represent true orbits*, Bull. Amer. Math. Soc., 19, pp. 465–470.

W. Hayes (1995), *Efficient Shadowing of High Dimensional Chaotic Systems with the Large Astrophysical n-Body Problem as an Example*, Master's thesis, Department of Computer Science, University of Toronto, Toronto.

W. B. Hayes (2001), *Rigorous Shadowing of Numerical Solutions of Ordinary Differential Equations by Containment*, Ph.D. thesis, Department of Computer Science, University of Toronto, Toronto; also available on the web at http://www.cs.toronto.edu/NA/reports.html#hayes-01-phd.

D. Kahaner, C. Moler, and S. Nash (1989), *Numerical Methods and Software*, Prentice-Hall Series in Comput. Math., Prentice-Hall, Englewood Cliff, NJ, 1989.

E. N. Lorenz (1963), *Deterministic nonperiodic flow*, J. Atmospheric Sci., 20, pp. 130–141. (Reprinted in Chaos, by H. Bai-Lin, World Scientific Publishing, Singapore, 1984.)

J. R. Munkres (1975), *Topology: A First Course*, Prentice-Hall, Englewood Cliffs, NJ.

J. Murdock (1995), *Shadowing multiple elbow orbits: An application of dynamical systems to perturbation theory*, J. Differential Equations, 119, pp. 224–247.

N. S. Nedialkov (1999), *Computing Rigorous Bounds on the Solution of an Initial Value Problem for an Ordinary Differential Equation*, Ph.D. thesis, Department of Computer Science, University of Toronto, Toronto.

N. S. Nedialkov, K. R. Jackson, and G. F. Corliss (1999), *Validated solutions of initial value problems for ordinary differential equations*, Appl. Math. Comput., 105, pp. 21–68.

K. J. Palmer (1988), *Exponential dichotomies, The shadowing lemma and transversal homoclinic points*, in Dynamics Reported, U. Kirchgraber and H. O. Walther, eds., Vol. 1, Wiley and Teubner.

G. D. Quinlan and S. Tremaine (1992), *On the reliability of gravitational N-body integrations*, Monthly Notices Roy. Astronom. Soc., 259, pp. 505–518.

T. Sauer and J. A. Yorke (1991), *Rigorous verification of trajectories for the computer simulation* *of dynamical systems*, Nonlinearity, 4, pp. 961–979.

D. Stoffer and K. J. Palmer (1999), *Rigorous verification of chaotic behaviour of maps using validated shadowing*, Nonlinearity, 12, pp. 1683–1698.

A. M. Stuart and A. R. Humphries (1996), *Dynamical Systems and Numerical Analysis*, Cambridge University Press, Cambridge, UK.

E. S. Van Vleck (1995), *Numerical shadowing near hyperbolic trajectories*, SIAM J. Sci. Comput., 16, pp. 1177–1189.

E. S. Van Vleck (2000), *Numerical shadowing using componentwise bounds and a sharper fixed point result*, SIAM J. Sci. Comput., 22, pp. 787–801.

SIAM J. Numer. Anal.
Vol. 41, No. 5, pp. 1974–1995

LEAST SQUARES METHODS FOR THE COUPLING OF FEM AND BEM*

GABRIEL N. GATICA[†], HELMUT HARBRECHT[‡], AND REINHOLD SCHNEIDER[‡]

Abstract. In the present paper we propose least squares formulations for the numerical solution of exterior boundary value problems. The partial differential equation is a first order system in a bounded subdomain, and the unbounded subdomain is treated by means of boundary integral equations. The first order system is derived from a strongly elliptic second order system. The analysis of the present least squares formulations is reduced to the analysis of the Galerkin method for the coupling of finite element and boundary element methods (FEM and BEM) of the second order problem. The least squares approach requires no stability condition. However, it requires the computation of negative as well as of half integer Sobolev norms. The arising linear systems can be preconditioned to have condition numbers ~ 1. The present methods benefit strongly from the use of biorthogonal wavelets on the coupling boundary and the computation of corresponding equivalent norms in Sobolev spaces. In particular, the application of Green's formula leads to an efficient discretization of least squares formulations.

Key words. least squares, coupling of FEM and BEM, wavelets

AMS subject classifications. 65J15, 65N30, 65N38, 65R20

DOI. 10.1137/S0036142902400664

1. Introduction.

The combined use of the finite element method (FEM) and the boundary element method (BEM), also called coupling of FEM and BEM, is already known as a very powerful tool to solve a large class of transmission problems in physics and engineering sciences (see, e.g., [13], [22], [27], [31], [33], [37], and the references therein). In addition, the interest in using mixed FEMs instead of the usual FEM has been increasing during the last few years. Indeed, the combination of mixed finite elements with either boundary integral equations or Dirichlet-to-Neumann mappings has been recently used to solve several interior and exterior boundary value problems appearing in potential theory and elasticity (see, e.g., [2], [8], [20], [23], [25], and [34]).

The reasons for this new interest arise mainly from structural mechanics, where the use of mixed FEMs allows us to compute stresses more accurately than displacements, whereas the utilization of boundary elements or Dirichlet-to-Neumann mappings is more appropriate for linear homogeneous materials in bounded and unbounded domains. In the framework of dual-mixed methods, the recent papers [8] and [34], dealing with an exterior problem from potential theory and the linear elasticity problem, respectively, are the first ones on the subject that consider the $H(\mathrm{div};\Omega)$ spaces in the finite element domain, and the two boundary integral equations approach from [13] and [27] in the boundary element region. Now, the method from [8] and [34]

*Received by the editors January 8, 2002; accepted for publication (in revised form) April 22, 2003; published electronically November 19, 2003. This research was partially supported by Conicyt-Chile through the FONDAP Program in Applied Mathematics, by the Ministry of Education of Chile through the MECESUP project UCO9907, and by the Dirección de Investigación of the Universidad de Concepción through the Advanced Research Groups Program.

http://www.siam.org/journals/sinum/41-5/40066.html

[†] GI^{2}MA, Departamento de Ingeniería Matemática, Universidad de Concepción, Casilla 160-C, Concepción, Chile (ggatica@ing-mat.udec.cl).

[‡] Fakultät für Mathematik, Technische Universität Chemnitz, D-09107 Chemnitz, Germany (helmut.harbrecht@mathematik.tu-chemnitz.de, reinhold.schneider@mathematik.tu-chemnitz.de).

was extended in [20], [25], and [2], where a suitable combination of dual-mixed FEM with either BEM or Dirichlet-to-Neumann mappings was applied to some nonlinear transmission problems. However, this extension has not been completely successful since the derivation of explicit finite element subspaces satisfying the corresponding discrete inf-sup conditions is still an open problem. As a first attempt to overcome this difficulty, we examined in [3] the use of a primal-mixed FEM. More recently, we obtained quite satisfactory results, at both the continuous and discrete levels, by applying what we called a dual-dual mixed variational formulation (see [4], [21], and [24]). This latter approach requires an extension of the usual Babuska–Brezzi theory to a special class of nonlinear variational problems with constraints, which was derived with full details in [19].

On the other hand, a possibility that has not been fully investigated yet is the utilization of least squares methods. As is well known, this approach avoids the necessity of inf-sup conditions, and hence it becomes attractive to use it jointly with mixed finite element formulations. One of the main methods, introduced in [1], uses the general theory of elliptic boundary value problems of Agmon–Douglis–Nirenberg and reduces the system to the minimization of a least squares functional that consists of a weighted sum of the residuals occurring in the equations and the boundary conditions. This is a generalization of both the method of Jespersen [32] and the method of Wendland [39]. Another approach, mostly used for second order elliptic problems written as first order systems, introduces a least squares functional and studies the resulting minimization problem by proving that the hypotheses of the Lax–Milgram lemma are satisfied on appropriate spaces (see, e.g., [9] and [35]). More recently, a least squares functional involving a discrete inner product related to the inner product in the Sobolev space of order -1 was introduced in [5], and an approach more closely coupled to the Galerkin method was studied by the same authors in [6].

Following the approach of [5], [6], the design of the least squares method requires the use of some negative and half integer Sobolev norms, such as the norms of $H^{-1}(\Omega)$ and $H^{-1/2}(\Gamma)$, which seem to be difficult to compute in practice. However, due to recent results in multilevel preconditioning [7] and multiscale methods or wavelet approximations (see [14], [16], [36]), these norms are computable in suitable finite dimensional subspaces. Moreover, in the framework of multiscale methods or biorthogonal wavelets, these computations are fairly simple and can be carried out within optimal complexity. We would like to mention that these approaches give rise to positive definite system matrices that can be easily preconditioned. In particular, using multilevel methods, one can reduce the condition numbers to $\mathcal{O}(1)$.

In the present paper we will discuss various least squares formulations. Most of them follow obviously from the underlying equations. The first approach requires the flux to be in $H(\mathrm{div};\Omega)$ and minimizes the equilibrium equation in the $L^2(\Omega)$-norm (see [9]), whereas the other ones perform the minimization in the $H^{-1}(\Omega)$-norm. Although this and similar approaches lead to some difficulties when developing a functional analytical setting, it turns out from our investigations that one can avoid these problems completely by using the functional $\mathbf{J}_3$ (see section 3 below). This functional is based on partial integration, i.e., Green's formula, and requires no restrictions concerning $H(\mathrm{div};\Omega)$ spaces. In fact, it is easy to implement and needs no further attention, compared to the other schemes, when computing discrete Sobolev norms. The approach of [15] provides an abstract setting considering an invertible operator $\mathcal{L} = (\mathcal{L}_{i,j}) : \prod_i H_i \to \prod_i H_i'$ mapping the product of Sobolev spaces $\prod_i H_i$ into its dual $\prod_i H_i'$. We would like to emphasize that $\mathbf{J}_3$ is the only functional that can be cast into this abstract framework.

For the computation of the discrete Sobolev norms in the bounded subdomain Ω, there are mainly two possibilities, the use of wavelet bases (see [15], [17]) or, alternatively, the utilization of suitable preconditioners (see [5], [6]), which are applicable to standard multigrid finite element discretizations. For the computation of the Sobolev norms along the boundary, we recommend wavelet bases. Obviously, the stability of least squares methods is guaranteed under weak assumptions, e.g., invertibility of the operators. However, it is worth mentioning that the negative and half integer Sobolev norms can be computed only on finite dimensional test spaces. This requires an additional truncation or projection to get a computable discrete formulation. Therefore, stability is not automatically guaranteed by the continuous formulation but must be proven. Our present proofs are completely based on the theory for the Galerkin scheme of the second order problem. From these results, we conclude the stability of the present methods. Only in the case of the functional $\mathbf{J}_4$ (see section 3 below) do we have to enlarge the test spaces slightly. An important feature of the present approaches is that, for the least squares discretizations of the boundary integral operators, we need only the coefficients of the Galerkin matrices of the layer potentials and not of compositions of layer potentials (see section 7). Finally, it is worth mentioning that in the framework of multiscale methods these matrices are sparse and that there are already several techniques available to precondition them (see, e.g., [36], [38]).

Consequently, the purpose of the present work is to examine the use of least squares formulations for the coupling of mixed FEM and BEM, as applied to linear exterior boundary value problems. This must be considered as the first step toward the future extension to nonlinear exterior transmission problems. The rest of the paper is organized as follows. In section 2 we describe the exterior second order model problem and apply the boundary integral equation method to reduce it to an equivalent nonlocal boundary value problem in a bounded annular domain. Then, after setting the flux as a new unknown, the nonlocal problem is rewritten as a first order system, which yields the underlying equations for the discretization. Various continuous least squares formulations, induced by this first order system, are introduced in section 3. Although existence and uniqueness for the least squares minimization problems can be easily deduced from the mapping properties of the underlying operators, we provide explicit proofs by using coercivity estimates of the usual variational formulation for the coupling procedure, since the method of these proofs can be used for the validation of the corresponding results for the discrete least squares formulations in section 5. Next, in section 4 we define the finite dimensional subspaces. The discrete least squares formulations and the corresponding error analysis are studied in section 5. In section 6 we give a brief description of the equivalence of norms based on wavelet bases and indicate the utilization of these functions for the present least squares approach. In addition, we remark how to use the wavelet bases provided by [17] for the treatment of three dimensional problems. In the last section we consider a numerical example and demonstrate how to set up the discrete matrices related to the minimization of $\mathbf{J}_3$.

2. The exterior boundary value problem.

Let G be a bounded and simply connected domain in $\mathbb{R}^2$ with Lipschitz-continuous boundary ∂G, and let Γ_D and Γ_N be two disjoint subsets of ∂G such that $|\Gamma_D| \neq 0$ and $\partial G = \overline{\Gamma}_D \cup \overline{\Gamma}_N$. In addition, let Ω be the annular domain bounded by ∂G and a second Lipschitz-continuous curve Γ whose interior region contains G. We denote $\Omega_e := \mathbb{R}^2 - (\overline{G} \cup \overline{\Omega})$. Then, given $f \in L^2(\Omega)$ and a matrix valued function $\mathbf{a}(\cdot) := (a_{ij}(\cdot))_{2\times 2}$, we consider the following exterior boundary value problem: *Find* $u \in H^1_{loc}(\mathbb{R}^2 - \overline{G})$ *such that*

$$
\begin{gathered}
u \,=\, 0 \quad on \quad \Gamma_D \quad and \quad (\mathbf{a}\,\nabla u)\cdot\mathbf{n} \,=\, 0 \quad on \quad \Gamma_N\,,\\
-\operatorname{div}(\mathbf{a}\,\nabla u) \,=\, f \quad in \quad \Omega\,,\\
\lim_{\substack{x\to x_0\\ x\in\Omega}} u(x) = \lim_{\substack{x\to x_0\\ x\in\Omega_e}} u(x) \quad \forall\, x_0\in\Gamma\,,\\
\lim_{\substack{x\to x_0\\ x\in\Omega}} \mathbf{a}(x)\nabla u(x)\cdot\mathbf{n}(x_0) \,=\, \lim_{\substack{x\to x_0\\ x\in\Omega_e}} \nabla u(x)\cdot\mathbf{n}(x_0) \quad \forall\, x_0\in\Gamma\,,\\
-\,\Delta\, u \,=\, 0 \quad in \quad \Omega_e \quad , \quad u(x) \,=\, O(1) \quad as \quad \|x\|\to+\infty\,,
\end{gathered}
\tag{2.1}
$$

where $\mathbf{n}$ *(resp.,* $\mathbf{n}(x_0)$*) denotes the unit outward normal to* $\partial\Omega$ *(to* $x_0\in\partial\Omega$*).* Here, we assume that $a_{ij}\in L^\infty(\Omega)$ and that there exists $\alpha>0$ such that

$$
\alpha\,\|z\|^2 \,\le\, z^T\,\mathbf{a}(x)\,z \qquad \forall\, z\in\mathbb{R}^2 \quad \text{and for almost all } x\in\Omega\,. \tag{2.2}
$$

We observe that the fourth and fifth equations of (2.1) constitute the usual transmission conditions along the interface Γ.

In what follows, we use the boundary integral equation method in the region Ω_e and reduce the problem (2.1) to a nonlocal boundary value problem on the bounded domain Ω. To this end, we let

$$
E(x,y) \,:=\, -\frac{1}{2\pi}\,\log\,\|x-y\|
$$

be the fundamental solution of the Laplacian and recall that the Green representation formula in Ω_e becomes

$$
u(x) \,=\, \int_\Gamma \left\{ \frac{\partial}{\partial\mathbf{n}(y)}\,E(x,y)\,u(y) - E(x,y)\,\frac{\partial u}{\partial\mathbf{n}}(y)\right\} ds_y - \lambda \quad \forall\, x\in\Omega_e,
$$

where λ is an unknown constant.

Then, according to the well-known jump conditions of the layer potentials, and using the transmission conditions from (2.1), we obtain the integral equations

$$
\begin{aligned}
0 \;&=\; \left(\frac{1}{2}\mathbf{I}\,-\,\mathbf{K}\right)u + \mathbf{V}\sigma + \lambda \quad && \text{on } \Gamma\,,\\
\sigma \;&=\; -\,\mathbf{W}u \,+\, \left(\frac{1}{2}\mathbf{I} - \mathbf{K}'\right)\sigma && \text{on } \Gamma\,,
\end{aligned}
\tag{2.3}
$$

where we have introduced the new unknown $\sigma := (\mathbf{a}\,\nabla u)\cdot\mathbf{n}$ on Γ, and $\mathbf{V}$, $\mathbf{K}$, $\mathbf{K}'$, and $\mathbf{W}$ are the boundary integral operators of the simple, double, adjoint of the double, and hypersingular layer potentials, respectively.

Now, the condition at infinity of u implies that σ satisfies

$$
\int_\Gamma \sigma\,ds = \int_\Gamma (\mathbf{a}\,\nabla u)\cdot\mathbf{n}\,ds = 0\,,
$$

which means that $\sigma\in H_0^{-1/2}(\Gamma)$, where $H_0^{-1/2}(\Gamma) := \{\tau\in H^{-1/2}(\Gamma) : \langle\tau,1\rangle = 0\}$, and, hereafter, $\langle\cdot,\cdot\rangle$ denotes the duality pairing between $H^{-1/2}(\Gamma)$ and $H^{1/2}(\Gamma)$ with respect to the $L^2(\Gamma)$-inner product.

In this way, the original exterior boundary value problem (2.1) reduces to the following nonlocal boundary value problem in Ω: *Find* $(u,\sigma,\lambda) \in H^1(\Omega)\times H_0^{-1/2}(\Gamma)\times \mathbb{R}$ *such that*

$$(2.4)\qquad \begin{aligned} u \,=\, 0 \quad \textit{on} \quad \Gamma_D \quad \textit{and} \quad (\mathbf{a}\,\nabla u)\cdot \mathbf{n} \,=\, 0 \quad \textit{on} \quad \Gamma_N\,,\\ -\operatorname{div}(\mathbf{a}\,\nabla u) \,=\, f \quad \textit{in} \quad \Omega\,,\\ \sigma \,=\, (\mathbf{a}\nabla u)\cdot \mathbf{n} \quad \textit{on} \quad \Gamma\,,\\ \sigma = -\,\mathbf{W}u \,+\, \left(\frac{1}{2}\mathbf{I} - \mathbf{K}'\right)\sigma \quad \textit{on} \quad \Gamma\,,\\ 0 = \left(\frac{1}{2}\mathbf{I} \,-\, \mathbf{K}\right)u + \mathbf{V}\sigma + \lambda \quad \textit{on} \quad \Gamma\,. \end{aligned}$$

We now introduce the flux $\boldsymbol{\theta} := \mathbf{a}\,\nabla u$. Since $\sigma \in H_0^{-1/2}(\Gamma)$ and $\boldsymbol{\theta}\cdot\mathbf{n} = \sigma$ on Γ, we note that the unknown $\boldsymbol{\theta}$ must belong to $H_0(\mathrm{div};\Omega)$, where

$$H_0(\mathrm{div};\Omega) := \left\{\,\boldsymbol{\zeta} \in H(\mathrm{div};\Omega):\quad \boldsymbol{\zeta}\cdot\mathbf{n} = 0 \quad \text{on} \quad \Gamma_N \quad \text{and} \quad \langle \boldsymbol{\zeta}\cdot\mathbf{n}, 1\rangle = 0\,\right\}.$$

As usual, $H(\mathrm{div};\Omega)$ is the space of functions $\boldsymbol{\zeta} \in [L^2(\Omega)]^2$ such that $\operatorname{div}\boldsymbol{\zeta} \in L^2(\Omega)$. Provided with the inner product

$$(\boldsymbol{\theta},\boldsymbol{\zeta})_{H(\mathrm{div};\Omega)} := (\boldsymbol{\theta},\boldsymbol{\zeta})_{[L^2(\Omega)]^2} \,+\, (\operatorname{div}\boldsymbol{\theta}, \operatorname{div}\boldsymbol{\zeta})_{L^2(\Omega)}\,,$$

$H(\mathrm{div};\Omega)$ is a Hilbert space. Here, $(\cdot,\cdot)_{[L^2(\Omega)]^2}$ and $(\cdot,\cdot)_{L^2(\Omega)}$ denote the inner products of the spaces indicated. Moreover, for all $\boldsymbol{\zeta} \in H(\mathrm{div};\Omega)$, $\boldsymbol{\zeta}\cdot\mathbf{n} \in H^{-1/2}(\Gamma)$ and there holds $\|\boldsymbol{\zeta}\cdot\mathbf{n}\|_{H^{-1/2}(\Gamma)} \le \|\boldsymbol{\zeta}\|_{H(\mathrm{div};\Omega)}$ (see [26] for the proof of these results).

Consequently, our problem (2.4) can be rewritten as the following equivalent first order system: *Find* $(\boldsymbol{\theta},u,\sigma,\lambda) \in H_0(\mathrm{div};\Omega)\times H^1(\Omega)\times H_0^{-1/2}(\Gamma)\times\mathbb{R}$ *such that*

$$(2.5)\qquad \begin{aligned} u \,=\, 0 \quad \textit{on} \quad \Gamma_D\,,\\ \boldsymbol{\theta} \,-\, \mathbf{a}\,\nabla\,u \,=\, 0 \quad \textit{and} \quad -\operatorname{div}\boldsymbol{\theta} \,=\, f \quad \textit{in} \quad \Omega\,,\\ \sigma \,=\, \boldsymbol{\theta}\cdot\mathbf{n} \quad \textit{on} \quad \Gamma\,,\\ \sigma = -\,\mathbf{W}u \,+\, \left(\frac{1}{2}\mathbf{I} - \mathbf{K}'\right)\sigma \quad \textit{on} \quad \Gamma\,,\\ 0 = \left(\frac{1}{2}\mathbf{I} \,-\, \mathbf{K}\right)u + \mathbf{V}\sigma + \lambda \quad \textit{on} \quad \Gamma\,. \end{aligned}$$

This system is the starting point for the least squares formulations that we propose below in section 3.

Before ending the present section, we recall that the boundary integral operators used above are formally defined by

$$(\mathbf{V}\tau)(x) := \int_\Gamma E(x,y)\,\tau(y)\,ds_y \quad \forall\,\tau \in H^{-1/2}(\Gamma)\,, \quad \forall\,x\in\Gamma\,,$$

$$(\mathbf{K}\mu)(x) := \int_\Gamma \frac{\partial}{\partial\mathbf{n}(y)}E(x,y)\,\mu(y)\,ds_y \quad \forall\,\mu \in H^{1/2}(\Gamma)\,, \quad \forall\,x\in\Gamma\,,$$

$$(\mathbf{K}'\tau)(x) := \int_\Gamma \frac{\partial}{\partial \mathbf{n}(x)} E(x,y)\,\tau(y)\,ds_y \quad \forall\,\tau \in H^{-1/2}(\Gamma)\,, \quad \forall\, x \in \Gamma\,,$$

$$(\mathbf{W}\mu)(x) := -\,\frac{\partial}{\partial \mathbf{n}(x)} \int_\Gamma \frac{\partial}{\partial \mathbf{n}(y)} E(x,y)\,\mu(y)\,ds_y \quad \forall\,\mu \in H^{1/2}(\Gamma)\,, \quad \forall\, x \in \Gamma\,.$$

Moreover, their main mapping properties are collected in the following lemma.

Lemma 2.1. *Let Γ be a Lipschitz boundary. The operators*

$$\mathbf{V} : H^{-1/2+s}(\Gamma) \longrightarrow H^{1/2+s}(\Gamma), \qquad \mathbf{K} : H^{1/2+s}(\Gamma) \longrightarrow H^{1/2+s}(\Gamma),$$

$$\mathbf{K}' : H^{-1/2+s}(\Gamma) \longrightarrow H^{-1/2+s}(\Gamma), \quad \mathbf{W} : H^{1/2+s}(\Gamma) \longrightarrow H^{-1/2+s}(\Gamma)$$

are continuous for all $s \in [-1/2,\, 1/2]$. Furthermore, there exist positive constants α_1, α_2 such that

$$\langle \tau, \mathbf{V}\tau \rangle \;\geq\; \alpha_1\, \|\tau\|^2_{H^{-1/2}(\Gamma)} \quad \forall\,\tau \in H_0^{-1/2}(\Gamma)$$

and

$$\langle \mathbf{W}\mu, \mu \rangle \;\geq\; \alpha_2\, \|\mu\|^2_{H^{1/2}(\Gamma)} \quad \forall\,\mu \in H_0^{1/2}(\Gamma)\,,$$

where

$$H_0^{1/2}(\Gamma) := \{\mu \in H^{1/2}(\Gamma) : \quad \langle 1, \mu \rangle = 0\}\,.$$

Proof. See [12]. □

3. The continuous least squares formulations. According to the system (2.5), and taking into account the least squares formulations already described in section 1, we consider here four different approaches.

First, we introduce the operator $\mathbf{P}_0 : H^{1/2}(\Gamma) \to H_0^{1/2}(\Gamma)$, where

$$\mathbf{P}_0\,\mu \;:=\; \mu \,-\, \frac{1}{|\Gamma|}\,\langle 1, \mu\rangle \quad \forall\,\mu \in H^{1/2}(\Gamma)\,. \tag{3.1}$$

Note that $\mathbf{P}_0\,\mu \equiv 0$ for all constant μ on Γ and that there exists $C > 0$, depending only on Γ, such that

$$\|\mathbf{P}_0\,\mu\|_{H^{1/2}(\Gamma)} \;\leq\; C\,\|\mu\|_{H^{1/2}(\Gamma)} \quad \forall\,\mu \in H^{1/2}(\Gamma)\,. \tag{3.2}$$

Then, we define the space

$$H^1_{\Gamma_D}(\Omega) := \{\, v \in H^1(\Omega) : \quad v = 0 \quad \text{on} \quad \Gamma_D\}$$

and consider the following minimization problem: *Find $(\boldsymbol{\theta}, u, \sigma) \in \mathbf{X}_1 := H_0(\mathrm{div};\Omega) \times H^1_{\Gamma_D}(\Omega) \times H_0^{-1/2}(\Gamma)$ such that*

$$\mathbf{J}_1(\boldsymbol{\theta}, u, \sigma) \;=\; \min_{(\boldsymbol{\zeta}, v, \tau) \in \mathbf{X}_1} \mathbf{J}_1(\boldsymbol{\zeta}, v, \tau)\,, \tag{3.3}$$

where $\mathbf{J}_1$ is the quadratic functional defined by

$$\begin{aligned} \mathbf{J}_1(\boldsymbol{\zeta}, v, \tau) \;:=\;& \|\mathbf{a}\nabla v - \boldsymbol{\zeta}\|^2_{[L^2(\Omega)]^2} \;+\; \|\mathrm{div}\,\boldsymbol{\zeta} + f\|^2_{L^2(\Omega)} \\ &+\; \left\|\mathbf{W}v + \boldsymbol{\zeta}\cdot\mathbf{n} - \left(\tfrac{1}{2}\mathbf{I} - \mathbf{K}'\right)\tau\right\|^2_{H^{-1/2}(\Gamma)} \;+\; \left\|\mathbf{P}_0\left[\left(\tfrac{1}{2}\mathbf{I} - \mathbf{K}\right)v + \mathbf{V}\tau\right]\right\|^2_{H^{1/2}(\Gamma)}\,. \end{aligned} \tag{3.4}$$

In what follows, let $H^{-1}(\Omega)$ denote the dual of $H^1_{\Gamma_D}(\Omega)$. Then, since the fourth equation from (2.5) must be understood at least in the distributional sense, it suffices to assume that the data f belongs to $H^{-1}(\Omega)$ and that the unknown $\boldsymbol{\theta}$ is sought in the space

$$H \;:=\; \left\{ \boldsymbol{\zeta} \in [L^2(\Omega)]^2 : \quad \boldsymbol{\zeta}\cdot\mathbf{n} = 0 \quad \text{on} \quad \Gamma_N \quad \text{and} \quad \boldsymbol{\zeta}\cdot\mathbf{n} \in H_0^{-1/2}(\Gamma) \right\},$$

which is endowed with the norm of $[L^2(\Omega)]^2$.

The above remark leads us to the following minimization problem: *Find* $(\boldsymbol{\theta}, u, \sigma) \in \mathbf{X}_2 := H \times H^1_{\Gamma_D}(\Omega) \times H_0^{-1/2}(\Gamma)$ *such that*

$$\mathbf{J}_2(\boldsymbol{\theta}, u, \sigma) \;=\; \min_{(\boldsymbol{\zeta}, v, \tau)\in\mathbf{X}_2} \mathbf{J}_2(\boldsymbol{\zeta}, v, \tau)\,, \tag{3.5}$$

where $\mathbf{J}_2$ *is the quadratic functional defined by*

$$\begin{aligned} \mathbf{J}_2(\boldsymbol{\zeta}, v, \tau) \;&:=\; \|\mathbf{a}\nabla v - \boldsymbol{\zeta}\|^2_{[L^2(\Omega)]^2} \;+\; \|\operatorname{div}\boldsymbol{\zeta} + f\|^2_{H^{-1}(\Omega)} \\ &+\; \left\|\mathbf{W}v + \boldsymbol{\zeta}\cdot\mathbf{n} - \left(\tfrac{1}{2}\mathbf{I} - \mathbf{K}'\right)\tau\right\|^2_{H^{-1/2}(\Gamma)} \;+\; \left\|\mathbf{P}_0\left[\left(\tfrac{1}{2}\mathbf{I} - \mathbf{K}\right)v + \mathbf{V}\tau\right]\right\|^2_{H^{1/2}(\Gamma)}\,. \end{aligned} \tag{3.6}$$

We remark that the only differences between (3.3)–(3.4) and (3.5)–(3.6) lie in the norm that measures the error arising from the equilibrium equation $(\operatorname{div}\boldsymbol{\zeta} + f) = 0$, and on the space in which the unknown $\boldsymbol{\theta}$ lives. In any case, it is easy to see that the minimum of both $\mathbf{J}_1$ and $\mathbf{J}_2$ is attained for any solution $(\boldsymbol{\theta}, u, \sigma)$ of problem (2.5). Also, it is important to mention that, instead of the first term in the definitions of $\mathbf{J}_1$ and $\mathbf{J}_2$, one may use the weighted norm $\|\mathbf{a}^{-1/2}\,(\mathbf{a}\nabla v - \boldsymbol{\zeta})\|^2_{[L^2(\Omega)]^2}$, which leads to a better conditioning of the corresponding discrete problems (see [15] for details).

The use of norms is motivated by the proper functional analytical setting $\mathcal{L} : \mathbf{X} \to \mathbf{X}'$. The paper [15] provides a general framework for least squares methods based on variational formulations. In contrast to Galerkin methods, the least square methods are stable if and only if $\mathcal{L}$ is normally solvable, i.e., if $\operatorname{Im}\mathcal{L} \subset \mathbf{X}'$ is a closed subset of $\mathbf{X}'$. However, the previous formulations do not fit exactly into the framework of [15]. Nevertheless, there is a slight modification of the functional $\mathbf{J}_2$ fitting into this setting which can be derived from the variational formulation of the second order problem. This realization facilitates the implementation; see section 7. Taking the equation $-\operatorname{div}\boldsymbol{\zeta} = f$ in its weak form, we can apply Green's theorem

$$\big((\operatorname{div} - \delta_\Gamma \otimes \;\cdot\mathbf{n})\boldsymbol{\zeta}, v\big)_{L^2(\Omega)} \;:=\; (\operatorname{div}\boldsymbol{\zeta}, v)_{L^2(\Omega)} \;-\; \langle\boldsymbol{\zeta}\cdot\mathbf{n}, v\rangle \;=\; -\,(\boldsymbol{\zeta}, \nabla v)_{[L^2(\Omega)]^2}$$

for all $v \in H^1_{\Gamma_D}(\Omega)$. Here $\delta_\Gamma \otimes \tau$ is a distribution in $H^{-1}(\Omega)$ which is supported on the interface boundary Γ. We remark that by duality and the trace theorem we find

$$\|\delta_\Gamma \otimes \tau\|_{H^{-1}(\Omega)} \;\lesssim\; \|\tau\|_{H^{-1/2}(\Gamma)} \quad \forall\,\tau \in H^{-1/2}(\Gamma).$$

Hence we derive

$$\begin{aligned} |(\operatorname{div}\boldsymbol{\zeta} - \delta_\Gamma \otimes \boldsymbol{\zeta}\cdot\mathbf{n}, v)_{L^2(\Omega)}| \;&=\; |(\operatorname{div}\boldsymbol{\zeta}, v)_{L^2(\Omega)} \;-\; (\boldsymbol{\zeta}\cdot\mathbf{n}, v)_{L^2(\Gamma)}| \;=\; |(\boldsymbol{\zeta}, \nabla v)_{L^2(\Omega)}| \\ &\le\; \|\boldsymbol{\zeta}\|_{[L^2(\Omega)]^2}\|v\|_{H^1_{\Gamma_D}(\Omega)}. \end{aligned}$$

According to the abstract setting in [15], we consider the operator

$$\mathcal{L} \;=\; \begin{bmatrix} \mathcal{L}_{1,1} & \mathcal{L}_{1,2} & \mathcal{L}_{1,3} \\ \mathcal{L}_{2,1} & \mathcal{L}_{2,2} & \mathcal{L}_{2,3} \\ \mathcal{L}_{3,1} & \mathcal{L}_{3,2} & \mathcal{L}_{3,3} \end{bmatrix} : \begin{bmatrix} [L^2(\Omega)]^2 \\ H^1_{\Gamma_D}(\Omega) \\ H_0^{-1/2}(\Gamma) \end{bmatrix} \;\to\; \begin{bmatrix} [L^2(\Omega)]^2 \\ H^{-1}(\Omega) \\ H_0^{1/2}(\Gamma) \end{bmatrix}$$

given by

$$\mathcal{L} \;=\; \begin{bmatrix} -\mathbf{I} & a\nabla & 0 \\ \operatorname{div} - \delta_\Gamma \otimes \;\cdot\mathbf{n} & -\delta_\Gamma \otimes \mathbf{W}\gamma & \delta_\Gamma \otimes \left(\frac{1}{2}\mathbf{I} - \mathbf{K}'\right) \\ 0 & \mathbf{P}_0\left(\frac{1}{2}\mathbf{I} - \mathbf{K}\right)\gamma & \mathbf{P}_0\mathbf{V} \end{bmatrix}\,,$$

where γ denotes the trace operator $\gamma : H^1_{\Gamma_D}(\Omega) \to H^{1/2}(\Gamma)$. For the sake of brevity, we use the notation $\mathbf{W}v = \mathbf{W}\gamma v$. Consequently, following [15], we introduce the Hilbert space $\mathbf{X}_3 := [L^2(\Omega)]^2 \times H^1_{\Gamma_D}(\Omega) \times H_0^{-1/2}(\Gamma)$ and $\mathbf{X}_3'$ as the dual space of $\mathbf{X}_3$ with respect to the canonical L^2-inner product. This yields the following least squares minimization problem: *Find* $(\boldsymbol{\theta}, u, \sigma) \in \mathbf{X}_3$ *such that*

$$\mathbf{J}_3(\boldsymbol{\theta}, u, \sigma) \;=\; \min_{(\boldsymbol{\zeta}, v, \tau)\in\mathbf{X}_3} \mathbf{J}_3(\boldsymbol{\zeta}, v, \tau)\,, \tag{3.7}$$

where $\mathbf{J}_3$ *is the quadratic functional defined by*

$$\begin{aligned} \mathbf{J}_3(\boldsymbol{\zeta}, v, \tau) \;:=\; & \|\mathbf{a}\nabla v - \boldsymbol{\zeta}\|^2_{[L^2(\Omega)]^2} \;+\; \left\|\mathbf{P}_0\left[\left(\tfrac{1}{2}\mathbf{I} - \mathbf{K}\right)v + \mathbf{V}\tau\right]\right\|^2_{H^{1/2}(\Gamma)} \\ & + \left\|\operatorname{div}\boldsymbol{\zeta} + f - \delta_\Gamma \otimes \left(\mathbf{W}v + \boldsymbol{\zeta}\cdot\mathbf{n} - \left(\tfrac{1}{2}\mathbf{I} - \mathbf{K}'\right)\tau\right)\right\|^2_{H^{-1}(\Omega)}\,. \end{aligned} \tag{3.8}$$

Though this minimization problem looks unusual, it is relatively simple to implement. One advantage of this formulation is that the flux can be chosen simply in $[L^2(\Omega)]^2$. Let us notice that this approach is the only one which does not assume that $\boldsymbol{\zeta} \in H$ or $\boldsymbol{\zeta}\cdot\mathbf{n} \in H_0^{-1/2}(\Gamma)$, which in turn requires $\boldsymbol{\zeta} \in H(\operatorname{div};\Omega)$. Therefore, it is the only least squares formulation which can be cast in the abstract setting of [15].

For the sake of completeness, we mention that there is another more simplified version which is obtained by inserting the transmission condition $\boldsymbol{\theta}\cdot\mathbf{n} = \sigma$ directly into the above formulation. Then, the trace norms $\|\cdot\|_{H^{1/2}(\Gamma)}$ and $\|\cdot\|_{H^{-1/2}(\Gamma)}$ in (3.6) are redundant, and we can derive a simpler minimization problem: *Find* $(\boldsymbol{\theta}, u) \in \mathbf{X}_4 := H \times H^1_{\Gamma_D}(\Omega)$ *such that*

$$\mathbf{J}_4(\boldsymbol{\theta}, u) \;=\; \min_{(\boldsymbol{\zeta}, v)\in\mathbf{X}_4} \mathbf{J}_4(\boldsymbol{\zeta}, v)\,, \tag{3.9}$$

where $\mathbf{J}_4$ *is the quadratic functional defined by*

$$\begin{aligned} \mathbf{J}_4(\boldsymbol{\zeta}, v) \;:=\; & \|\mathbf{a}\nabla v - \boldsymbol{\zeta}\|^2_{[L^2(\Omega)]^2} \;+\; \|\operatorname{div}\boldsymbol{\zeta} + f\|^2_{H^{-1}(\Omega)} \\ & + \left\|\mathbf{P}_0\left[\left(\tfrac{1}{2}\mathbf{I} - \mathbf{K}\right)v + \mathbf{V}(\boldsymbol{\zeta}\cdot\mathbf{n})\right]\right\|^2_{H^{1/2}(\Gamma)}\,. \end{aligned} \tag{3.10}$$

Since $H^{-1}(\Omega)$, $H^{1/2}(\Gamma)$, and $H^{-1/2}(\Gamma)$ are Hilbert spaces, the norms are defined by the corresponding inner products $(\cdot,\cdot)_{H^{-1}(\Omega)}$, $\langle\cdot,\cdot\rangle_{H^{-1/2}(\Gamma)}$, and $\langle\cdot,\cdot\rangle_{H^{1/2}(\Gamma)}$. For example, the quadratic functional $\mathbf{J}_4$ can be rewritten by

$$\begin{aligned} \mathbf{J}_4(\boldsymbol{\zeta}, v) \;:=\; & (\mathbf{a}\nabla v - \boldsymbol{\zeta}, \mathbf{a}\nabla v - \boldsymbol{\zeta})_{[L^2(\Omega)]^2} \;+\; (\operatorname{div}\boldsymbol{\zeta} + f, \operatorname{div}\boldsymbol{\zeta} + f)_{H^{-1}(\Omega)} \\ & + \left\langle \mathbf{P}_0\left[\left(\tfrac{1}{2}\mathbf{I} - \mathbf{K}\right)v + \mathbf{V}(\boldsymbol{\zeta}\cdot\mathbf{n})\right], \mathbf{P}_0\left[\left(\tfrac{1}{2}\mathbf{I} - \mathbf{K}\right)v + \mathbf{V}(\boldsymbol{\zeta}\cdot\mathbf{n})\right]\right\rangle_{H^{1/2}(\Gamma)}\,. \end{aligned}$$

In what follows, we develop the necessary tools to study the solvability and discrete approximations of our least squares formulations. However, the second, third, and fourth formulation are sharper than the first one. In fact, the resulting convergence rate is higher, and the system matrices can be preconditioned quite well. In the third and fourth formulations, the $H^{-1/2}$-norm is avoided. Moreover, in the third formulation, the flux is computed by means of Green's theorem; see section 7. This means that we need neither $\boldsymbol{\zeta}\cdot\mathbf{n}$ nor the assumption $\boldsymbol{\zeta}\cdot\mathbf{n}\in H^{-1/2}(\Gamma)$ explicitly. In our opinion, it is the most favorable approach. The fourth formulation looks most simple, and it avoids the computation of the hypersingular operator $\mathbf{W}$. However, its discretization requires some kind of stabilization, which will be discussed below in section 5. Therefore, throughout the rest of the paper, we will just concentrate on the problems (3.5)–(3.6), (3.7)–(3.8), and (3.9)–(3.10). Since the computation of the $L^2(\Omega)$-inner product offers no difficulties, the corresponding extension to (3.3)–(3.4) will be straightforward.

Now, following the general setting from [15], we find that (3.3), (3.5), and (3.7) are equivalent to

$$\mathbf{J}_i(\boldsymbol{\theta},u,\sigma) \;=\; \min_{(\boldsymbol{\zeta},v,\tau)\in\mathbf{X}_i}\mathbf{J}_i(\boldsymbol{\zeta},v,\tau)\,, \tag{3.11}$$

with

$$\mathbf{J}_i(\boldsymbol{\zeta},v,\tau) \;=\; \frac{1}{2}\,B_i\big((\boldsymbol{\zeta},v,\tau),(\boldsymbol{\zeta},v,\tau)\big) \;-\; \mathbf{G}_i(\boldsymbol{\zeta},v,\tau) \;+\; \text{const}, \tag{3.12}$$

$i=1,2,3$, with corresponding bilinear forms $B_i:\mathbf{X}_i\times\mathbf{X}_i\to\mathbb{R}$, and linear functionals $\mathbf{G}_i:\mathbf{X}_i\to\mathbb{R}$. An analogous setting holds for (3.10). Then, the minimization problems are equivalent to the following linear equations: *Find* $(\boldsymbol{\theta},u,\sigma)\in\mathbf{X}_i$ *such that*

$$B_i\big((\boldsymbol{\theta},u,\sigma),(\boldsymbol{\zeta},v,\tau)\big) \;=\; \mathbf{G}_i(\boldsymbol{\zeta},v,\tau) \tag{3.13}$$

for all $(\boldsymbol{\zeta},v,\tau)\in\mathbf{X}_i$. This equation is solved approximatively on a finite dimensional subspace in $\mathbf{X}_i$. Therein, the major difficulty is the computation of the underlying bilinear and linear forms. However, this can be done only approximatively, which means B_i is replaced by some discrete bilinear form B_i^h.

4. Coercivity estimates. It is easy to prove, using the mapping properties of the boundary integral operators (cf. Lemma 2.1) and (3.2), that B_2, B_3, and B_4 are symmetric and bounded in the corresponding energy norms. In addition, $\mathbf{G}_2$, $\mathbf{G}_3$, and $\mathbf{G}_4$ are also bounded. Therefore, in order to conclude the unique solvability of our least squares formulations (3.5)–(3.6), (3.7)–(3.8), and (3.9)–(3.10), it remains to show that B_2, B_3, and B_4 are strongly coercive in $\mathbf{X}_2$, $\mathbf{X}_3$, and $\mathbf{X}_4$, respectively. Usually, coercivity estimates for least squares formulations are valid under much weaker conditions than for Galerkin formulations since only the normal solvability of the operator is required. Since the Sobolev norms cannot be computed exactly (see below), we need to apply a more sophisticated tool for the investigation of the present discrete least squares methods. For this purpose, we have to state some previous results concerning the Galerkin scheme of the original second order nonlocal boundary value problem (2.4).

First, proceeding in the usual way (see, e.g. [13], [22], [27]), we find that the weak formulation of (2.4) reduces to the following: *Find* $(u,\sigma)\in\mathbf{H}:=H^1_{\Gamma_D}(\Omega)\times H_0^{-1/2}(\Gamma)$ *such that*

$$A\big((u,\sigma),(v,\tau)\big) \;=\; \mathbf{F}(v,\tau) \quad \forall\,(v,\tau)\in\mathbf{H}\,, \tag{4.1}$$

where $A:\mathbf{H}\times\mathbf{H}\to\mathbb{R}$ *is the bounded bilinear form defined by*

$$\begin{aligned} A\big((u,\sigma),(v,\tau)\big) \;:=\;& (\mathbf{a}\,\nabla u,\nabla v)_{[L^2(\Omega)]^2} \;+\; \langle\mathbf{W}u,v\rangle \;-\; \langle\big(\tfrac{1}{2}\mathbf{I}-\mathbf{K}'\big)\,\sigma,v\rangle \\ &+\; \langle\tau,\mathbf{V}\sigma\rangle \;+\; \langle\tau,\big(\tfrac{1}{2}\mathbf{I}-\mathbf{K}\big)\,u\rangle \end{aligned} \tag{4.2}$$

for all (u,σ), $(v,\tau) \in \mathbf{H}$, *and* $\mathbf{F} \in \mathbf{H}'$ *is given by*

$$\mathbf{F}(v,\tau) \,:=\, \int_\Omega f\,v\,dx \quad \forall\,(v,\tau) \in \mathbf{H}\,. \tag{4.3}$$

The product space $\mathbf{H}$ is endowed with the corresponding norm, that is,

$$\|(v,\tau)\|_{\mathbf{H}} \,:=\, \left\{\, \|v\|^2_{H^1(\Omega)} \,+\, \|\tau\|^2_{H^{-1/2}(\Gamma)} \,\right\}^{1/2}.$$

In what follows, given two expressions a and b, the relation $a \lesssim b$ means that a is bounded by some constant times b uniformly in all parameters upon which a and b may depend. An analogous definition holds for the relation $a \gtrsim b$. Also, $a \sim b$ means that $a \lesssim b$ and $a \gtrsim b$.

LEMMA 4.1. *The bilinear form A is strongly coercive in* $\mathbf{H}$, *that is,*

$$A\big((v,\tau),(v,\tau)\big) \,\gtrsim\, \|(v,\tau)\|^2_{\mathbf{H}} \quad \forall\,(v,\tau) \in \mathbf{H}\,.$$

Proof. Using that $\mathbf{K}'$ is the adjoint of $\mathbf{K}$, we obtain from (4.2) that

$$A\big((v,\tau),(v,\tau)\big) \,=\, (\mathbf{a}\,\nabla v, \nabla v)_{[L^2(\Omega)]^2} \,+\, \langle \mathbf{W}v, v\rangle \,+\, \langle \tau, \mathbf{V}\tau\rangle\,.$$

Since $|\Gamma_D| \neq 0$, Poincaré's inequality yields the equivalence between the norm and the seminorm of $H^1(\Omega)$ in the subspace $H^1_{\Gamma_D}(\Omega)$, which, together with (2.2), implies that

$$(\mathbf{a}\,\nabla v, \nabla v)_{[L^2(\Omega)]^2} \,\gtrsim\, \|v\|^2_{H^1(\Omega)} \quad \forall\, v \in H^1_{\Gamma_D}(\Omega)\,.$$

Then, the above inequality and the coerciveness properties of $\mathbf{V}$ and $\mathbf{W}$ given in Lemma 2.1 complete the proof. □

For the sake of completeness, we also provide the following consequence of the previous lemma.

THEOREM 4.2. *There exists a unique solution* $(u,\sigma) \in \mathbf{H}$ *of the variational formulation* (4.1). *Moreover, there holds the a priori estimate* $\|(u,\sigma)\|_{\mathbf{H}} \lesssim \|\mathbf{F}\|_{\mathbf{H}'}$.

Proof. The proof is a straightforward application of the Lax–Milgram lemma. □

The following lemma reveals a well-known fact about boundary integral operators.

LEMMA 4.3. *For* $u \in H_0^{1/2}(\Gamma)$ *and* $\sigma \in H_0^{-1/2}(\Gamma)$, *the following holds:*

$$\left\|\mathbf{W}u + (\tfrac{1}{2}\mathbf{I} + \mathbf{K}')\sigma\right\|_{H^{-1/2}(\Gamma)} \,\sim\, \left\|\mathbf{P}_0\left[(\tfrac{1}{2}\mathbf{I} - \mathbf{K})u + \mathbf{V}\sigma\right]\right\|_{H^{1/2}(\Gamma)}\,. \tag{4.4}$$

Proof. The proof follows easily from the equality $\mathbf{W} = -(\frac{1}{2}\mathbf{I} + \mathbf{K}')\mathbf{V}^{-1}(\frac{1}{2}\mathbf{I} - \mathbf{K})$ together with the mapping properties of the double layer potential operators in the spaces $H^{\pm 1/2}(\Gamma)$. □

THEOREM 4.4. *For all functions* $(\boldsymbol{\theta}, u, \sigma) \in \mathbf{X}_2 := H \times H^1_{\Gamma_D}(\Omega) \times H_0^{-1/2}(\Gamma)$, *the following a priori estimate is valid:*

$$\begin{aligned} &\|u\|_{H^1(\Omega)} + \|\boldsymbol{\theta}\|_{[L^2(\Omega)]^2} + \|\sigma\|_{H^{-1/2}(\Gamma)} \,\lesssim\, \|\mathbf{a}\,\nabla u - \boldsymbol{\theta}\|_{[L^2(\Omega)]^2} + \|\mathrm{div}\,\boldsymbol{\theta}\|_{H^{-1}(\Omega)} \\ &\quad + \left\|\mathbf{W}u + \boldsymbol{\theta}\cdot\mathbf{n} - (\tfrac{1}{2}\mathbf{I} - \mathbf{K}')\sigma\right\|_{H^{-1/2}(\Gamma)} + \left\|\mathbf{P}_0\left[(\tfrac{1}{2}\mathbf{I} - \mathbf{K})u + \mathbf{V}\sigma\right]\right\|_{H^{1/2}(\Gamma)}\,. \end{aligned} \tag{4.5}$$

Moreover, for all $(\boldsymbol{\theta}, u, \sigma) \in \mathbf{X}_3 := [L^2(\Omega)]^2 \times H^1_{\Gamma_D}(\Omega) \times H_0^{-1/2}(\Gamma)$ *there holds*

$$\begin{aligned}\|u\|_{H^1(\Omega)} + \|\boldsymbol{\theta}\|_{[L^2(\Omega)]^2} + \|\sigma\|_{H^{-1/2}(\Gamma)} \;&\lesssim\; \|\mathbf{a}\,\nabla u - \boldsymbol{\theta}\|_{[L^2(\Omega)]^2} \\ &+ \left\|\operatorname{div}\boldsymbol{\theta} - \delta_\Gamma \otimes \left[\mathbf{W}u + \boldsymbol{\theta}\cdot\mathbf{n} - (\tfrac{1}{2}\mathbf{I} - \mathbf{K}')\sigma\right]\right\|_{H^{-1}(\Omega)} \\ &+ \left\|\mathbf{P}_0\left[(\tfrac{1}{2}\mathbf{I} - \mathbf{K})u + \mathbf{V}\sigma\right]\right\|_{H^{1/2}(\Gamma)}\,. \end{aligned} \tag{4.6}$$

In addition, for any $(\boldsymbol{\theta}, u) \in \mathbf{X}_4 := H \times H^1_{\Gamma_D}(\Omega)$ *there holds the a priori estimate*

$$\begin{aligned}\|u\|_{H^1(\Omega)} + \|\boldsymbol{\theta}\|_{[L^2(\Omega)]^2} + \|\boldsymbol{\theta}\cdot\mathbf{n}\|_{H^{-1/2}(\Gamma)} \;&\lesssim\; \|\mathbf{a}\,\nabla u - \boldsymbol{\theta}\|_{[L^2(\Omega)]^2} \\ &+ \|\operatorname{div}\boldsymbol{\theta}\|_{H^{-1}(\Omega)} + \left\|\mathbf{P}_0\left[(\tfrac{1}{2}\mathbf{I} - \mathbf{K})u + \mathbf{V}(\boldsymbol{\theta}\cdot\mathbf{n})\right]\right\|_{H^{1/2}(\Gamma)}\,. \end{aligned} \tag{4.7}$$

Proof. We provide a particular proof of this result because we need this reasoning below to prove the main theorem, Theorem 5.1. By virtue of Theorem 4.2 we estimate

$$\begin{aligned}\|u\|_{H^1(\Omega)} + \|\sigma\|_{H^{-1/2}(\Gamma)} \;&\lesssim\; \sup_{\delta\in H_0^{-1/2}(\Gamma)} \frac{1}{\|\delta\|_{H^{-1/2}(\Gamma)}} \left\{\langle \delta, (\tfrac{1}{2}\mathbf{I} - \mathbf{K})u + \mathbf{V}\sigma\rangle\right\} \\ &+ \sup_{v\in H^1_{\Gamma_D}(\Omega)} \frac{1}{\|v\|_{H^1(\Omega)}} \left\{(\mathbf{a}\nabla u, \nabla v)_{[L^2(\Omega)]^2} + \langle \mathbf{W}u - (\tfrac{1}{2}\mathbf{I} - \mathbf{K}')\sigma, v\rangle\right\} \\ &\lesssim \sup_{\delta\in H_0^{-1/2}(\Gamma)} \frac{1}{\|\delta\|_{H^{-1/2}(\Gamma)}} \left\{\langle \delta, (\tfrac{1}{2}\mathbf{I} - \mathbf{K})u + \mathbf{V}\sigma\rangle\right\} \\ &+ \sup_{v\in H^1_{\Gamma_D}(\Omega)} \frac{1}{\|v\|_{H^1(\Omega)}} \Big\{ (\mathbf{a}\nabla u - \boldsymbol{\theta}, \nabla v)_{[L^2(\Omega)]^2} + (\boldsymbol{\theta}, \nabla v)_{[L^2(\Omega)]^2} \\ &\qquad + \langle \mathbf{W}u - (\tfrac{1}{2}\mathbf{I} - \mathbf{K}')\sigma, v\rangle \Big\}\,. \end{aligned}$$

Next, we apply the divergence theorem and use that $\boldsymbol{\theta}\cdot\mathbf{n} = 0$ on Γ_N, whence

$$\begin{aligned}\|u\|_{H^1(\Omega)} + \|\sigma\|_{H^{-1/2}(\Gamma)} \;&\lesssim\; \sup_{\delta\in H_0^{-1/2}(\Gamma)} \frac{1}{\|\delta\|_{H^{-1/2}(\Gamma)}} \left\{\langle \delta, (\tfrac{1}{2}\mathbf{I} - \mathbf{K})u + \mathbf{V}\sigma\rangle\right\} \\ &+ \sup_{v\in H^1_{\Gamma_D}(\Omega)} \frac{1}{\|v\|_{H^1(\Omega)}} \Big\{(\mathbf{a}\nabla u - \boldsymbol{\theta}, \nabla v)_{[L^2(\Omega)]^2} - (\operatorname{div}\boldsymbol{\theta}, v)_{[L^2(\Omega)]^2} \\ &\qquad + \langle \mathbf{W}u + \boldsymbol{\theta}\cdot\mathbf{n} - (\tfrac{1}{2}\mathbf{I} - \mathbf{K}')\sigma, v\rangle\Big\}\,, \end{aligned}$$

which implies both estimates, (4.5) and (4.6), immediately. We remark that we have used the trace theorem $\|v\|_{H^{1/2}(\Gamma)} \lesssim \|v\|_{H^1(\Omega)}$ and the fact that $v|_{\Gamma_D} = 0$ for all $v \in H^1_{\Gamma_D}(\Omega)$. To prove (4.7) we choose $\sigma = \boldsymbol{\theta}\cdot\mathbf{n}$ in (4.5) and apply the result of Lemma 4.3. □

5. Finite element approximations. For the definition of the Ritz and Galerkin methods for (3.11)–(3.12), we consider finite dimensional subspaces $\mathbf{X}_i^h := X_h^i \times V_h \times S_h$ of $\mathbf{X}_i$ and assume the following.

- *Approximation property of* V_h. There exists $d := d_V > 1$ such that for all $s < \min\{\frac{3}{2}, d\}$ and for all $u \in H^d(\Omega)$

$$\inf_{v_h\in V_h} \|u - v_h\|_{H^s(\Omega)} \;\lesssim\; h^{d-s}\|u\|_{H^d(\Omega)}\,.$$

- *Inverse property* of V_h. For all $v_h \in V_h$ and for all $t < s < \frac{3}{2}$ there holds

$$\|v_h\|_{H^s(\Omega)} \;\lesssim\; h^{t-s}\|v_h\|_{H^t(\Omega)}\,.$$

Similar properties are also assumed for X_h^i and S_h with constants d_X and d_S, respectively.

Typical candidates for these spaces are finite element spaces V_h, subordinated to a triangulation $\mathcal{T}_h = \{\tau_k\}$ of Ω consisting of triangles or quadrilaterals τ_k with diameter h_k. The above properties are valid for shape regular quasi-uniform triangulations. The results for (3.11)–(3.12) remain valid also for nonuniform triangulations. The results with respect to $\mathbf{J}_4$ seem to be also true on nonuniform grids. (Perhaps the proof becomes rather technical.)

Here d denotes polynomial degree on each triangle. Since $V_h \subset H^1_{\Gamma_D}(\Omega)$, the functions $v_h \in V_h$ are assumed to be continuous on Ω. For a consistent discretization it is sufficient to choose $d_X = d_V - 1 = d - 1$. The spaces S_h are defined analogously on the boundary, and they should be exact at least of order $d_S = d - 1$.

The $H^{-1}(\Omega)$-norm on V_h can be computed by introducing the operator $\mathbf{T}_h : H^{-1}(\Omega) \to V_h$, where for each $f \in H^{-1}(\Omega)$ the function $w_h := \mathbf{T}_h f$ is the unique function in V_h satisfying

$$(\nabla w_h, \nabla v_h)_{[L^2(\Omega)]^2} = (f, v_h)_{L^2(\Omega)} \quad \forall\, v_h \in V_h\,. \tag{5.1}$$

The computation of the operator $\mathbf{T}_h$ requires the solution of a Neumann problem which is relatively expensive. For an efficient computation it is much more feasible to use a symmetric preconditioner $\mathbf{B}_h : V_h^* \to V_h$ instead of $\mathbf{T}_h$ satisfying

$$\|\mathbf{B}_h f_h\|_{H^1_{\Gamma_D}(\Omega)} \sim \|f_h\|_{H^{-1}(\Omega)} \qquad \forall\, f_h \in V_h^* \tag{5.2}$$

or, equivalently,

$$(\mathbf{T}_h f_h, f_h)_{L^2(\Omega)} \sim (\mathbf{B}_h f_h, f_h)_{L^2(\Omega)} \quad \forall\, f_h \in V_h^*\,, \tag{5.3}$$

where $V_h^* \subset H^{-1}(\Omega)$ is a suitable finite dimensional subspace. Such preconditioners are available from multigrid or multilevel algorithms [5], [6], in which case one can choose $V_h^* = V_h$, as well as from wavelet bases [15], where the space V_h^* is generated by the dual wavelet basis.

Now, in order to compute the inner products $\langle\cdot,\cdot\rangle_{H^{\pm 1/2}(\Gamma)}$, one can use, according to Lemma 2.1, that $\langle\lambda,\lambda\rangle_{H^{1/2}(\Gamma)} \sim \langle\mathbf{W}\lambda,\lambda\rangle$ for all $\lambda \in H_0^{1/2}(\Gamma)$ and that $\langle\sigma,\sigma\rangle_{H^{-1/2}(\Gamma)} \sim \langle\sigma,\mathbf{V}\sigma\rangle$ for all $\sigma \in H_0^{-1/2}(\Gamma)$, which, however, are not accessible for numerical computations. Again we have to consider only $H^{\pm 1/2}(\Gamma)$-norms and $\langle\cdot,\cdot\rangle_{H^{\pm 1/2}(\Gamma)}$-inner products on finite dimensional subspaces. However, one can apply a preconditioner $\mathbf{D}_h$ for $\mathbf{W}$ in the same way as described above; see, e.g., [30] and [36]. It is computable on a finite dimensional subspace $\tilde{V}_h(\Gamma)$ of $H_0^{1/2}(\Gamma)$ and satisfies

$$\langle\mathbf{W}\lambda_h,\lambda_h\rangle \sim \langle\mathbf{D}_h\lambda_h,\lambda_h\rangle \quad \forall\, \lambda_h \in \tilde{V}_h(\Gamma)\,. \tag{5.4}$$

Similarly, we introduce an operator $\mathbf{C}_h$ as a preconditioner for $\mathbf{V}$ satisfying

$$\langle\sigma_h,\mathbf{V}\sigma_h\rangle \sim \langle\sigma_h,\mathbf{C}_h\sigma_h\rangle \quad \forall\, \sigma_h \in \tilde{S}_h\,, \tag{5.5}$$

where $\tilde{S}_h$ is a finite dimensional subspace of $H_0^{-1/2}(\Gamma)$. In the case in which one is dealing only with traditional boundary elements, we simply have $\tilde{V}_h = V_h|_\Gamma$ and $\tilde{S}_h = S_h$. For a wavelet preconditioner we refer to the subsequent section.

Since these operators are symmetric and coercive, we define for notation's convenience the square roots $\mathbf{B}_h^{1/2}$ by $(\mathbf{B}_h^{1/2})^*\mathbf{B}_h^{1/2} = \mathbf{B}_h$, and we set up $\mathbf{C}_h^{1/2}$ and $\mathbf{D}_h^{1/2}$

similarly. In addition, we define $V_h(\Gamma) := V_h|_\Gamma \cap H_0^{1/2}(\Gamma)$ and let $P_h : H^1_{\Gamma_D}(\Omega) \to V_h$, $\mathbb{Q}_h : H^{1/2}(\Gamma) \to S_h$, and $Q_h : H^{-1/2}(\Gamma) \to V_h(\Gamma)$ be bounded projectors with *adjoint* operators $P_h^* : H^{-1}(\Omega) \to V_h^*$, $\mathbb{Q}_h^* : H^{-1/2}(\Gamma) \to \tilde{S}_h$ and $Q_h^* : H^{1/2}(\Gamma) \to \tilde{V}_h(\Gamma)$, respectively. Then, according to (5.2), (5.4), and (5.5), we deduce that

$$\begin{aligned} (\mathbf{B}_h P_h^* f, P_h^* f)_{L^2(\Omega)} &\sim \|P_h^* f\|^2_{V_h^*} && \forall\, f \in H^{-1}(\Omega)\,, \\ \langle \mathbf{D}_h Q_h^* \lambda, Q_h^* \lambda\rangle &\sim \|Q_h^* \lambda\|^2_{H^{1/2}(\Gamma)} && \forall\, \lambda \in H_0^{1/2}(\Gamma)\,, \\ \langle\, \mathbb{Q}_h^* \sigma, \mathbf{C}_h\, \mathbb{Q}_h^* \sigma\rangle &\sim \|\,\mathbb{Q}_h^* \sigma\|^2_{H^{-1/2}(\Gamma)} && \forall\, \sigma \in H_0^{-1/2}(\Gamma)\,. \end{aligned} \tag{5.6}$$

The above means that we will use truncated bilinear forms instead of the original ones for the computation of the Galerkin solutions. Certainly, this truncation may influence the stability of the methods. Hence we prove next that stability is not violated by this procedure.

THEOREM 5.1. *For arbitrary functions $(\boldsymbol{\theta}_h, u_h, \sigma_h) \in \mathbf{X}_2^h$, the following a priori estimate holds:*

$$\begin{aligned} \|u_h\|_{H^1(\Omega)} + \|\boldsymbol{\theta}_h\|_{[L^2(\Omega)]^2} + \|\sigma_h\|_{H^{-1/2}(\Gamma)} &\lesssim \|\mathbf{a}\nabla u_h - \boldsymbol{\theta}_h\|_{[L^2(\Omega)]^2} \\ + \|\mathbf{B}_h^{1/2} P_h^* \operatorname{div} \boldsymbol{\theta}_h\|_{L^2(\Omega)} &+ \big\|\mathbf{C}_h^{1/2}\, \mathbb{Q}_h^* \big[\mathbf{W} u_h + \boldsymbol{\theta}_h \cdot \mathbf{n} - (\tfrac{1}{2}\mathbf{I} - \mathbf{K}')\sigma_h\big]\big\|_{L^2(\Gamma)} \\ &+ \big\|\mathbf{D}_h^{1/2} Q_h^* \,\mathbf{P}_0 \big[(\tfrac{1}{2}\mathbf{I} - \mathbf{K}) u_h + \mathbf{V}\sigma_h\big]\big\|_{L^2(\Gamma)}\,, \end{aligned} \tag{5.7}$$

and for $(\boldsymbol{\theta}_h, u_h, \sigma_h) \in \mathbf{X}_3^h$ we find

$$\begin{aligned} \|u_h\|_{H^1(\Omega)} + \|\boldsymbol{\theta}_h\|_{[L^2(\Omega)]^2} + \|\sigma_h\|_{H^{-1/2}(\Gamma)} &\lesssim \|\mathbf{a}\nabla u_h - \boldsymbol{\theta}_h\|_{[L^2(\Omega)]^2} \\ + \big\|\mathbf{B}_h^{1/2} P_h^* \big(\operatorname{div} \boldsymbol{\theta}_h - \delta_\Gamma \otimes \big[\mathbf{W} u_h &+ \boldsymbol{\theta}_h \cdot \mathbf{n} - (\tfrac{1}{2}\mathbf{I} - \mathbf{K}')\sigma_h\big]\big)\big\|_{L^2(\Omega)} \\ &+ \big\|\mathbf{D}_h^{1/2} Q_h^* \,\mathbf{P}_0 \big[(\tfrac{1}{2}\mathbf{I} - \mathbf{K}) u_h + \mathbf{V}\sigma_h\big]\big\|_{L^2(\Gamma)}\,. \end{aligned} \tag{5.8}$$

Proof. We estimate the expression in the same fashion as in the proof of Theorem 4.4. First we observe that the stability of the Galerkin scheme implies the estimate

$$\begin{aligned} \|u_h\|_{H^1(\Omega)} + \|\sigma_h\|_{H^{-1/2}(\Gamma)} &\lesssim \sup_{\delta_h \in S_h} \frac{1}{\|\delta_h\|_{H^{-1/2}(\Gamma)}} \big\{\langle \delta_h, (\tfrac{1}{2}\mathbf{I} - \mathbf{K}) u_h + \mathbf{V}\sigma_h\rangle\big\} \\ &+ \sup_{v_h \in V_h} \frac{1}{\|v_h\|_{H^1(\Omega)}} \big\{(\mathbf{a}\nabla u_h, \nabla v_h)_{[L^2(\Omega)]^2} + \langle \mathbf{W} u_h - (\tfrac{1}{2}\mathbf{I} - \mathbf{K}')\sigma_h, v_h\rangle\big\} \\ &\lesssim \big\|Q_h^* \,\mathbf{P}_0 \big[(\tfrac{1}{2}\mathbf{I} - \mathbf{K}) u_h + \mathbf{V}\sigma_h\big]\big\|_{H^{1/2}(\Gamma)} + \|\mathbf{a}\nabla u_h - \boldsymbol{\theta}_h\|_{[L^2(\Omega)]^2} \\ &+ \big\|P_h^* \big(\operatorname{div} \boldsymbol{\theta}_h - \delta_\Gamma \otimes \big[\mathbf{W} u_h + \boldsymbol{\theta}_h \cdot \mathbf{n} - (\tfrac{1}{2}\mathbf{I} - \mathbf{K}')\sigma_h\big]\big)\big\|_{H^{-1}(\Omega)} \\ &\lesssim \big\|\mathbf{D}_h^{1/2} Q_h^* \,\mathbf{P}_0 \big[(\tfrac{1}{2}\mathbf{I} - \mathbf{K}) u_h + \mathbf{V}\sigma_h\big]\big\|_{L^2(\Gamma)} + \|\mathbf{a}\nabla u_h - \boldsymbol{\theta}_h\|_{[L^2(\Omega)]^2} \\ &+ \big\|\mathbf{B}_h^{1/2} P_h^* \big(\operatorname{div} \boldsymbol{\theta}_h - \delta_\Gamma \otimes \big[\mathbf{W} u_h + \boldsymbol{\theta}_h \cdot \mathbf{n} - (\tfrac{1}{2}\mathbf{I} - \mathbf{K}')\sigma_h\big]\big)\big\|_{L^2(\Omega)}\,, \end{aligned}$$

where we have used the properties of the operators $\mathbf{B}_h$ and $\mathbf{D}_h$. From this estimate the assertion (5.8) follows immediately. One can prove the estimate (5.7) similarly. □

This suggests that one has to solve the linear problem

$$B_i^h\big((\boldsymbol{\theta}_h, u_h, \sigma_h), (\boldsymbol{\zeta}_h, v_h, \tau_h)\big) \;=\; \mathbf{G}_i^h(\boldsymbol{\zeta}_h, v_h, \tau_h)\,, \quad i = 2, 3\,, \tag{5.9}$$

with the truncated bilinear forms $B_i^h : \mathbf{X}_i^h \times \mathbf{X}_i^h \to \mathbb{R}$ and the functionals $\mathbf{G}_i^h : \mathbf{X}_i^h \to \mathbb{R}$. The computation of the bilinear forms B_1^h, B_2^h, and B_4^h requires the computation of $\operatorname{div} \boldsymbol{\zeta}$, which is possible, e.g., if $\boldsymbol{\zeta} \in H(\operatorname{div}; \Omega)$ or $X_h \subset H$, despite the fact that the energy space using B_2^h or B_4^h is $[L^2(\Omega)]^2$. The differentiation of $\boldsymbol{\zeta}$ can be avoided with the aid of Green's theorem. This is used in the third formulation using B_3^h, which requires only that $\boldsymbol{\zeta} \in [L^2(\Omega)]^2$, i.e., $X_h \subset [L^2(\Omega)]^2$.

It turns out that, for the fourth formulation, the truncation must be performed on a finer grid to preserve the stability. Here, the bilinear form $B_4^h : \mathbf{X}_4^h \times \mathbf{X}_4^h \to \mathbb{R}$ is defined by

$$\begin{aligned} B_4^h\big((\boldsymbol{\theta}_h, u_h), (\boldsymbol{\zeta}_h, v_h)\big) \;:=\; & (\mathbf{a}\nabla u_h - \boldsymbol{\theta}_h, \mathbf{a}\nabla v_h - \boldsymbol{\zeta}_h)_{[L^2(\Omega)]^2} \\ & + (\mathbf{B}_h P_h^* \operatorname{div} \boldsymbol{\theta}_h, P_h^* \operatorname{div} \boldsymbol{\zeta}_h)_{L^2(\Omega)} \\ + \langle \mathbf{D}_{h'} Q_{h'}^*\, \mathbf{P}_0 \left[\left(\tfrac{1}{2}\mathbf{I} - \mathbf{K}\right) u_h + \mathbf{V}(\boldsymbol{\theta}_h \cdot \mathbf{n})\right], & \, Q_{h'}^*\, \mathbf{P}_0 \left[\left(\tfrac{1}{2}\mathbf{I} - \mathbf{K}\right) v_h + \mathbf{V}(\boldsymbol{\zeta}_h \cdot \mathbf{n})\right] \rangle_{L^2(\Gamma)}\,, \end{aligned}$$

where the positive parameter h' has to be chosen such that $h' \lesssim h$, that is, $h' \le c\,h$ for a sufficiently small constant c. In fact, we have the following result.

Theorem 5.2. *Assume that $\boldsymbol{\theta}_h \cdot \mathbf{n} \in S_h$ for all $\boldsymbol{\theta}_h \in X_h$. Then there exists a mesh size $h' \lesssim h$ such that for $u_h \in V_h$ and $\boldsymbol{\theta}_h \in X_h$ there holds the a priori estimate*

$$\begin{aligned} \|u_h\|_{H^1(\Omega)} \;+\; \|\boldsymbol{\theta}_h\|_{[L^2(\Omega)]^2} \;+\; \|\boldsymbol{\theta}_h \cdot \mathbf{n}\|_{H^{-1/2}(\Gamma)} \;\lesssim\; \|\mathbf{a}\nabla u_h - \boldsymbol{\theta}_h\|_{[L^2(\Omega)]^2} \\ + \big\|\mathbf{B}_h^{1/2} P_h^* \operatorname{div} \boldsymbol{\theta}_h\big\|_{L^2(\Omega)} \;+\; \big\|\mathbf{D}_{h'}^{1/2} Q_{h'}^*\, \mathbf{P}_0\big[(\tfrac{1}{2}\mathbf{I} - \mathbf{K})u_h + \mathbf{V}(\boldsymbol{\theta}_h \cdot \mathbf{n})\big]\big\|_{L^2(\Gamma)}\,. \end{aligned}$$

Proof. Given $(\boldsymbol{\theta}_h, u_h) \in X_h \times V_h$, we take $\sigma_h := \boldsymbol{\theta}_h \cdot \mathbf{n}$ in the estimate (5.7) and then apply Lemma 4.3 to obtain

$$\begin{aligned} \|u_h\|_{H^1(\Omega)} \;+\; \|\boldsymbol{\theta}_h\|_{[L^2(\Omega]^2} \;+\; \|\boldsymbol{\theta}_h \cdot \mathbf{n}\|_{H^{-1/2}(\Gamma)} \;&\lesssim\; \|\mathbf{a}\nabla u_h - \boldsymbol{\theta}_h\|_{[L^2(\Omega]^2} \\ + \big\|\mathbf{B}_h^{1/2} P_h^* \operatorname{div} \boldsymbol{\theta}_h\big\|_{L^2(\Omega)} + \big\|\mathbf{P}_0\big[(\tfrac{1}{2}\mathbf{I} - \mathbf{K})u_h + &\mathbf{V}(\boldsymbol{\theta}_h \cdot \mathbf{n})\big]\big\|_{H^{1/2}(\Gamma)} \\ \lesssim\; \|\mathbf{a}\nabla u_h - \boldsymbol{\theta}_h\|_{[L^2(\Omega]^2} \;+\; \|\mathbf{B}_h^{1/2} P_h^* \operatorname{div} \boldsymbol{\theta}_h\|_{L^2(\Omega)} & \\ + \big\|Q_{h'}^*\, \mathbf{P}_0\big[(\tfrac{1}{2}\mathbf{I} - \mathbf{K})u_h + \mathbf{V}(\boldsymbol{\theta}_h \cdot \mathbf{n})\big]\big\|_{H^{1/2}(\Gamma)} & \\ + \big\|(I - Q_{h'}^*)\mathbf{P}_0\big[(\tfrac{1}{2}\mathbf{I} - \mathbf{K})u_h + \mathbf{V}(\boldsymbol{\theta}_h \cdot \mathbf{n})\big]\big\|_{H^{1/2}(\Gamma)}\,. & \end{aligned} \tag{5.10}$$

Next, using the approximation and inverse properties of the subspaces involved, we get

$$\begin{aligned} & \big\|(I - Q_{h'}^*)\mathbf{P}_0\big[(\tfrac{1}{2}\mathbf{I} - \mathbf{K})u_h + \mathbf{V}(\boldsymbol{\theta}_h \cdot \mathbf{n})\big]\big\|_{H^{1/2}(\Gamma)} \\ & \lesssim\; (h')^{\alpha} \big\|\mathbf{P}_0\big[(\tfrac{1}{2}\mathbf{I} - \mathbf{K})u_h + \mathbf{V}(\boldsymbol{\theta}_h \cdot \mathbf{n})\big]\big\|_{H^{1/2+\alpha}(\Gamma)} \\ & \lesssim\; (h')^{\alpha} \left\{\|u_h\|_{H^{1/2+\alpha}(\Gamma)} + \|\boldsymbol{\theta}_h \cdot \mathbf{n}\|_{H^{-1/2+\alpha}(\Gamma)}\right\} \\ & \lesssim\; (h')^{\alpha} h^{-\alpha} \left\{\|u_h\|_{H^1(\Omega)} + \|\boldsymbol{\theta}_h \cdot \mathbf{n}\|_{H^{-1/2}(\Gamma)}\right\}\,. \end{aligned} \tag{5.11}$$

Therefore, replacing (5.11) back into (5.10), choosing $h' \lesssim h$, and using (5.6), we conclude the proof. □

The error analysis of both methods then is a standard application of the well-known second Strang lemma.

Theorem 5.3. *The bilinear forms B_i^h, $i = 2, 3$, satisfy*

$$B_i^h\big((\boldsymbol{\theta}_h, u_h, \sigma_h), (\boldsymbol{\theta}_h, u_h, \sigma_h)\big) \ \sim \ \|(\boldsymbol{\theta}_h, u_h, \sigma_h)\|_{\mathbf{X}_i}^2 \,,$$

and the following convergence estimate holds in both cases:

$$\|u - u_h\|_{H^1(\Omega)} \ + \ \|\boldsymbol{\theta} - \boldsymbol{\theta}_h\|_{L^2(\Omega)} \ + \ \|\sigma - \sigma_h\|_{H^{-1/2}(\Gamma)} \ \lesssim \ h^{d-1}\|u\|_{H^d(\Omega)} \,.$$

In addition, there exists $h' \lesssim h$ such that the bilinear form B_4^h satisfies

$$B_4^h\big((\boldsymbol{\theta}_h, u_h), (\boldsymbol{\theta}_h, u_h)\big) \ \gtrsim \ \|u_h\|_{H^1(\Omega)}^2 \ + \ \|\boldsymbol{\theta}_h\|_{L^2(\Omega)}^2 \ + \ \|\boldsymbol{\theta}_h \cdot \mathbf{n}\|_{H^{-1/2}(\Gamma)}^2 \,,$$

and the following convergence estimate holds:

$$\|u - u_h\|_{H^1(\Omega)} \ + \ \|\boldsymbol{\theta} - \boldsymbol{\theta}_h\|_{L^2(\Omega)} \ + \ \|\boldsymbol{\theta} \cdot \mathbf{n} - \boldsymbol{\theta}_h \cdot \mathbf{n}\|_{H^{-1/2}(\Gamma)} \ \lesssim \ h^{d-1}\|u\|_{H^d(\Omega)} \,.$$

6. Wavelet bases and related matrices. In the framework of the present least squares methods, we would like to recommend the use of wavelet bases at least for the discretization of the boundary integral operators. Wavelet bases facilitate the computation of the Sobolev norms. In fact, one can exploit several features simultaneously, namely, the computation of the half integer Sobolev norms [15], the preconditioning [14], [36], together with a sparse discretization by matrix compression [16], [36], [38], and the use of wavelet bases for an adaptive approximation [11]. The matrix compression accelerates computation with the boundary element matrices enormously. In fact, it reduces the quadratic complexity dealing with full matrices of size N to order N or $N \log^a N$; cf. [36]. This might be not a major concern for two dimensional problems, since the finite element part already has N^2 unknowns. However, for three dimensional problems the complexity of the boundary element part would dominate that of the finite element part. Therefore, fast methods for boundary integral equations become necessary when dealing with very large systems of integral equations [28].

Wavelet bases, and, in particular, wavelet bases for boundary integral equations, is by now a well-studied subject. There are many excellent accounts about wavelets in general, and for boundary integral equations we refer the reader to the survey paper [14] and the references therein. Here we focus only on those aspects which are important for the present purpose. Particularly, more information about wavelet least squares methods is contained in [15].

In general, a multiresolution analysis consists of a nested family of finite dimensional subspaces

$$S_0 \subset \cdots \subset S_j \subset S_{j+1} \subset \cdots \,,$$

where, e.g., $\bigcup_{j \geq 0} S_j$ is supposed dense in $L^2(\Gamma)$. For example, we may consider $S_j = S_h$ with $h \sim 2^{-j}$ and where $\bigcup_{j \geq 0} S_j$ is dense in $H^{-1/2}(\Gamma)$.

Each space S_j is defined by a single-scale basis, i.e., $S_j = \operatorname{span}\{\varphi_k^j : k \in \Delta_j\}$, where Δ_j denotes a suitable index set with cardinality $\#\Delta_j \sim 2^{nj}$. These basis functions might be classical piecewise constant or piecewise linear basis functions for boundary element methods. The wavelets $\Psi^j = \{\psi_k^j : k \in \nabla_j = \Delta_{j+1} \setminus \Delta_j\}$ are the bases of complementary spaces $W_j = \operatorname{span}\{\psi_k^j : k \in \nabla_j\}$ of S_j in S_{j+1}, i.e.,

$$S_{j+1} \,=\, S_j \,\oplus\, W_j, \qquad S_j \,\cap\, W_j \,=\, \{0\}\,.$$

In what follows, we adhere to the following shorthand notation. We write $\psi_k^{-1} := \varphi_k^0$ and $\nabla_{-1} := \Delta_0$. By Ψ_j we denote the (column-) vector $\Psi_j = (\psi_k^l)_{k\in\nabla_l,-1\le l<j}$. For a given vector $\mathbf{v} \in \mathbb{R}^{\#\Delta_j}$, we write simply

$$\Psi_j^T\mathbf{v} \,=\, \mathbf{v}^T\Psi_j \,=\, \sum_{l=-1}^{j-1}\sum_{k\in\nabla_l} v_{l,k}\psi_k^l\,.$$

It is supposed that the collection Ψ_j builds a uniformly stable basis of S_{j+1} and a Riesz-basis in L^2. This property is guaranteed if there exists a biorthogonal, or dual, collection $\tilde\Psi = \{\tilde\psi_k^l : k \in \nabla_l,\ l \ge -1\}$ generating spaces $\tilde S_0 \subset \cdots \subset \tilde S_j \subset \cdots$ such that $\langle\tilde\psi_k^j, \psi_l^i\rangle = \delta_{k,l}\delta_{i,j}$. In this case, every $v \in L^2(\Gamma)$ has the representations

$$v \,=\, \langle v, \tilde\Psi\rangle^T\Psi\,, \qquad v \,=\, \langle v, \Psi\rangle^T\tilde\Psi\,. \tag{6.1}$$

Then the projectors Q_j and Q_j^* are given by

$$Q_jv \,=\, \langle v, \tilde\Psi_j\rangle^T\Psi_j, \qquad Q_j^*v \,=\, \langle v, \Psi_j\rangle^T\tilde\Psi_j\,.$$

In addition, the wavelets are supposed to be local on the corresponding scale. We refer to [14], [17], and [36] for further details.

Let $\gamma := \sup\{s \in \mathbb{R} : S_j \subset H^s(\Gamma)\}$ and $\tilde\gamma$ be defined analogously. Then, for a given function v, the following norm equivalences hold:

$$\|v\|_{H^s(\Gamma)}^2 \sim \sum_{l\ge-1} 2^{-2ls}\,\|\langle v, \tilde\Psi^l\rangle\|^2, \qquad \|v\|_{H^{-s}(\Gamma)}^2 \sim \sum_{l\ge-1} 2^{2ls}\,\|\langle v, \Psi^l\rangle\|^2, \tag{6.2}$$

where $-\tilde\gamma < s < \gamma$. It is important to remark that one does not need the dual basis for the computation of the norm.

To describe the application of these norm equivalences, let us take a single operator and consider $h' \le h$, for example. Let $\Phi_{j'}$ be a wavelet basis for the traces of $V_{h'}$ on the boundary Γ. Then, we define the matrix $\mathbf{V}_{h',h} := \langle\mathbf{P}_0(\mathbf{V}\Psi_j), \Phi_{j'}\rangle$, where $h' = 2^{-j'}$ and $h = 2^{-j}$, which is nothing but a part of the Galerkin matrix for the operator $\mathbf{V}$ together with the diagonal matrix $\mathbf{D}_{h',h'}^{-2s} = \operatorname{diag}(2^{-2ls})$. For instance, we compute the $H^{1/2}(\Gamma)$-norm by setting $s = 1/2$ and obtain

$$\|Q_j\mathbf{P}_0(\mathbf{V}\Psi_j)\|_{H^{1/2}(\Gamma)}^2 \sim \mathbf{c}_j^T\mathbf{V}_{h',h}^T\mathbf{D}_{h',h'}^{-1}\mathbf{V}_{h',h}\mathbf{c}_j\,.$$

This means that the preconditioner defined in the previous section is of the form $\mathbf{D}_hu = \Psi_j\mathbf{D}_{h,h}^{-1}\langle\Psi_j, u\rangle$. The other parts of the system matrices are derived similarly. For the combination of finite element spaces and the use of the Bramble–Pasciak–Xu (BPX) preconditioner for the computation of the $H^{-1}(\Omega)$-inner products and wavelet bases on the boundary we need to apply the wavelet transform (we refer to [28] and [29] for further details). We would like to remark that the size of the matrix $\mathbf{V}_{h',h}^T\mathbf{D}_{h',h'}^{-1}\mathbf{V}_{h',h}$ is already $\sim 2^{jn} \times 2^{jn}$ and can be sparsified by wavelet matrix compression.

Remark. Wavelets on surfaces are defined in, e.g., [17] and [18]. The first construction in [17] seems to be simpler than the final one in [18]. Since in [17] the duality is based on a modified inner product $\langle\cdot,\cdot\rangle$ defined via the local parametrizations, a comment about the use of this construction is required for the correct utilization of these bases computing $H^{1/2}(\Gamma)$-inner products according to (6.2). Instead of using the inner products $\langle f,\psi_k^j\rangle$, one has to use the modified inner product $\langle\cdot,\cdot\rangle$, whereas for the computation of the $H^{-1/2}(\Gamma)$-inner products one has to use the canonical inner product $\langle\cdot,\cdot\rangle$.

Remark. A major restriction of the present approach is that the traces along the boundary Γ of the spaces V_h must also admit a multiresolution analysis. This restriction can be removed by introducing an additional unknown $\mu\in H^{1/2}(\Gamma)$ for the traces of u along Γ like in [10]. Here μ will be discretized by wavelet bases. This means that the coupling is defined by a slightly weaker condition; see [10]. The generalization of the present method to this case is rather straightforward.

7. Numerical results. In this section, we show how to compute the corresponding system matrices and right-hand sides for the minimization of the functional $\mathbf{J}_3$ and present some numerical results. The energy space of $\mathbf{J}_3$ is $\mathbf{X}_3=[L^2(\Omega)]^2\times H^1_{\Gamma_D}(\Omega)\times H_0^{-1/2}(\Gamma)$. For a conforming discretization this requires only $\boldsymbol{\zeta}_h\in X_h\subset[L^2(\Omega)]^2$, which, for instance, allows the functions in X_h to be discontinuous. In our tests we use both piecewise constant functions and continuous piecewise linear functions subordinated to the triangulation $\mathcal{T}_h$. The trial functions $u_h\in V_h\subset H^1_{\Gamma_D}(\Omega)$ are chosen piecewise linear and continuous, and $\sigma_h\in S_h\subset H^{-1/2}(\Gamma)$ consists of piecewise constant functions.

It is worthwhile describing in more detail the present realization of that least squares method. Abbreviating

$$g(\boldsymbol{\theta}_h,u_h,\sigma_h)\;:=\;\operatorname{div}\boldsymbol{\theta}_h-\delta_\Gamma\otimes(\boldsymbol{\theta}_h\cdot\mathbf{n})-\left(\delta_\Gamma\otimes\left[\mathbf{W}u_h-(\tfrac12\mathbf{I}-\mathbf{K}')\sigma_h\right]\right),$$

the discrete bilinear form $B_3^h:\mathbf{X}_3^h\times\mathbf{X}_3^h\to\mathbb{R}$ is defined by

$$\begin{aligned}B_3^h\big((\boldsymbol{\theta}_h,u_h,\sigma_h),(\boldsymbol{\zeta}_h,v_h,\tau_h)\big)\;&:=\;(\mathbf{a}\nabla u_h-\boldsymbol{\theta}_h,\mathbf{a}\nabla v_h-\boldsymbol{\zeta}_h)_{[L^2(\Omega)]^2}\\&+\big(\mathbf{B}_hP_h^*\,g(\boldsymbol{\theta}_h,u_h,\sigma_h),P_h^*\,g(\boldsymbol{\zeta}_h,v_h,\tau_h)\big)_{L^2(\Omega)}\\+\langle\mathbf{D}_hQ_h^*\,\mathbf{P}_0\left[\left(\tfrac12\mathbf{I}-\mathbf{K}\right)u_h+\mathbf{V}(\boldsymbol{\theta}_h\cdot\mathbf{n})\right],&\,Q_h^*\,\mathbf{P}_0\left[\left(\tfrac12\mathbf{I}-\mathbf{K}\right)v_h+\mathbf{V}(\boldsymbol{\zeta}_h\cdot\mathbf{n})\right]\rangle_{L^2(\Gamma)},\end{aligned}$$

and the linear functional $\mathbf{G}_3^h:\mathbf{X}_3^h\to\mathbb{R}$ is given by

$$\mathbf{G}_3^h(\boldsymbol{\zeta}_h,v_h,\tau_h)\;:=\;\big(P_h^*\,f,P_h^*\,g(\boldsymbol{\zeta}_h,v_h,\tau_h)\big)_{L^2(\Omega)}.$$

Let us denote by Φ_h the vector of basis functions $\phi_k^h\in V_h$, Θ_h consists of the basis functions in X_h, and Λ_h indicates the vector of basis functions in S_h. To build up the system matrix and the right-hand side for the corresponding least squares method, we require the matrices and vectors

$$\begin{aligned}\mathbf{A}_h\;&:=\;(\mathbf{a}\nabla\Phi_h,\nabla\Phi_h)_{[L^2(\Omega)]^2}, &\mathbf{F}_h\;&:=\;(\nabla\Phi_h,\Theta_h)_{[L^2(\Omega)]^2},\\\mathbf{G}_h\;&:=\;(\Theta_h,\Theta_h)_{[L^2(\Omega)]^2}, &\mathbf{f}_h\;&:=\;(f,\Phi_h)_{[L^2(\Omega)]^2},\end{aligned}$$

together with the matrices of basis functions belonging to the interface boundary

$$\begin{aligned}\mathbf{V}_h\;&:=\;\langle\mathbf{V}\Lambda_h,\Lambda_h\rangle, &\mathbf{K}_h\;&:=\;\langle\mathbf{K}\Phi_h,\Lambda_h\rangle,\\\mathbf{W}_h\;&:=\;\langle\mathbf{W}\Phi_h,\Phi_h\rangle, &\mathbf{I}_h\;&:=\;\langle\Phi_h,\Lambda_h\rangle.\end{aligned}$$

We use the matrices $\mathbf{C}_h$ to define the inner product in $H^{-1/2}(\Gamma)$ and $\mathbf{B}_h$ for the computation of the inner product in $H^{-1}(\Omega)$. We choose $\mathbf{B}_h$ as a BPX preconditioner [7] and $\mathbf{C}_h := (\operatorname{diag}\mathbf{V}_h)^{-1}$, where $\mathbf{V}_h = \langle \mathbf{V}\Psi_h, \Psi_h\rangle$, is given with respect to a wavelet basis Ψ_h of S_h. Then, the corresponding linear system for the present least squares method can be written in the following form:

$$\begin{aligned}&\left\{\begin{bmatrix}\mathbf{G}_h & -\mathbf{F}_h^T & \mathbf{0}\\ -\mathbf{F}_h & \mathbf{A}_h & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \mathbf{0}\end{bmatrix} + \begin{bmatrix}\mathbf{F}_h^T\\ \mathbf{W}_h\\ (\mathbf{K}_h - \frac{1}{2}\mathbf{I}_h)\end{bmatrix}\mathbf{B}_h\begin{bmatrix}\mathbf{F}_h & \mathbf{W}_h & (\mathbf{K}_h - \frac{1}{2}\mathbf{I}_h)^T\end{bmatrix}\right.\\ &\quad\left. + \begin{bmatrix}\mathbf{0}\\ (\frac{1}{2}\mathbf{I}_h - \mathbf{K}_h)^T\\ \mathbf{V}_h\end{bmatrix}\mathbf{C}_h\begin{bmatrix}\mathbf{0} & (\frac{1}{2}\mathbf{I}_h - \mathbf{K}_h) & \mathbf{V}_h\end{bmatrix}\right\}\begin{bmatrix}\boldsymbol{\theta}_h\\ \mathbf{u}_h\\ \boldsymbol{\sigma}_h\end{bmatrix} = -\begin{bmatrix}\mathbf{F}_h\\ \mathbf{W}_h\\ (\mathbf{K}_h - \frac{1}{2}\mathbf{I}_h)\end{bmatrix}\mathbf{B}_h\mathbf{f}_h.\end{aligned}$$

This system is preconditioned by the operator $\operatorname{diag}(\mathbf{Id}, \mathbf{B}_h, \mathbf{C}_h)$. We remark that $P_h^*\big(\operatorname{div}\boldsymbol{\zeta}_h - \delta_\Gamma \otimes (\boldsymbol{\zeta}_h \cdot \mathbf{n})\big)$ is computed from the inner products

$$\big(\operatorname{div}\boldsymbol{\zeta}_h - \delta_\Gamma \otimes (\boldsymbol{\zeta}_h \cdot \mathbf{n}), v_h\big)_{L^2(\Omega)} = -\,(\boldsymbol{\zeta}_h, \nabla v_h)_{[L^2(\Omega)]^2} \quad \forall\, v_h \in V_h\,.$$

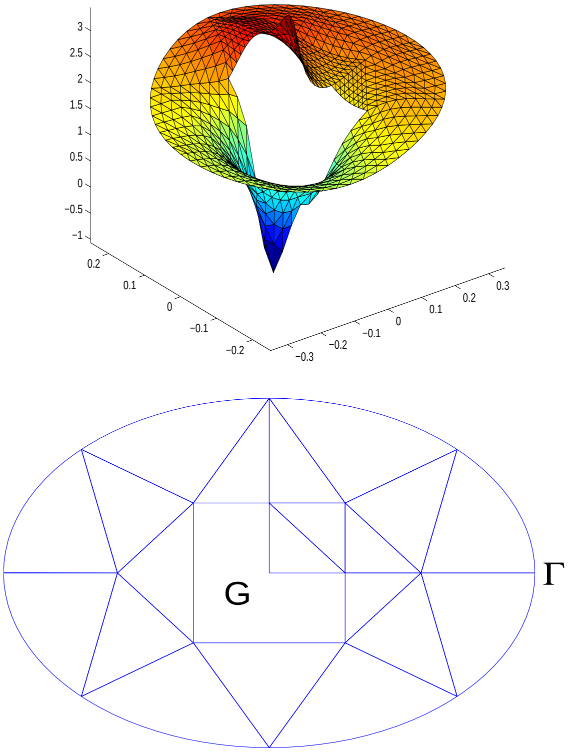


FIG. 7.1. *The solution u and the initial triangulation of Ω.*

For the numerical tests we choose G as the annulus outside the two dimensional L-shape $\left[-\frac{1}{10}, \frac{1}{10}\right]^2 \setminus \left[0, \frac{1}{10}\right]^2$ and inside an ellipse. Similarly to [29], we consider a

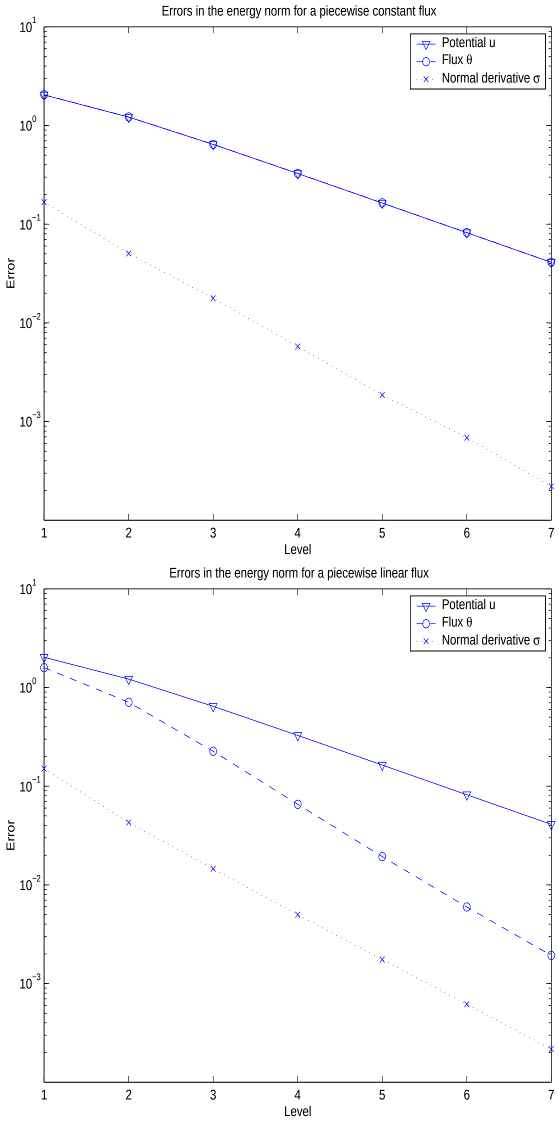


FIG. 7.2. *Error in the energy norm.*

problem for which an analytical solution is known. We split

$$u(x,y) \;=\; u_1(x,y) \;+\; u_2(x,y) \;\in\; C^2\big(\mathbb{R}^2 \setminus \left[\begin{smallmatrix}-1/20\\ 0\end{smallmatrix}\right]\big)$$

with the harmonic function

$$u_1(x,y) \;=\; \frac{1}{100}\cdot\frac{(x+\frac{1}{20})+y}{(x+\frac{1}{20})^2+y^2} \;\in\; C^\infty\big(\mathbb{R}^2 \setminus \left[\begin{smallmatrix}-1/20\\ 0\end{smallmatrix}\right]\big)$$

and the nonharmonic function $u_2 \in C^2(\mathbb{R}^2)$ defined by

$$u_2(x,y) \;=\; 2 + \begin{cases} \left(\frac{x^2}{0.3^2}+\frac{y^2}{0.2^2}-1\right)^3 & \text{if } \frac{x^2}{0.3^2}+\frac{y^2}{0.2^2} \le 1,\\ 0 & \text{if } \frac{x^2}{0.3^2}+\frac{y^2}{0.2^2} > 1. \end{cases}$$

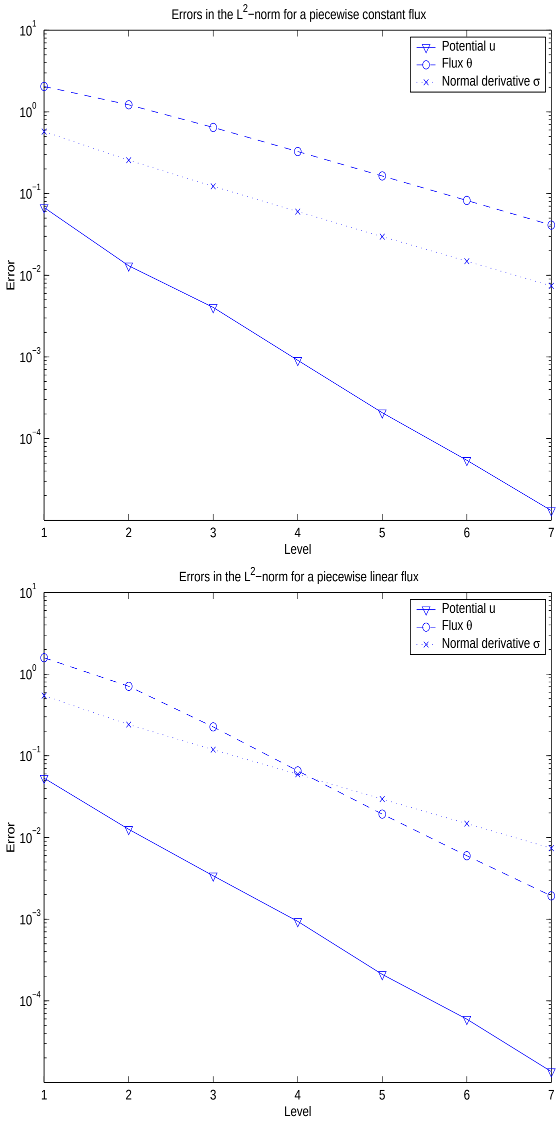


FIG. 7.3. *Error in L^2-norms.*

The function $f := -\triangle u_2 \in C^1(\mathbb{R})$ is supported in the ellipse with semiaxes 0.3 and 0.2. Thus, setting $g := u|_{\partial G}$, we obtain a boundary value problem with nonhomogeneous Dirichlet data at the boundary $\Gamma_D = \partial G$. The interface boundary Γ is chosen as the boundary of the ellipse with semiaxes 0.35 and 0.25. The solution u and the initial triangulation using curved triangles is shown in Figure 7.1.

We depict in Figure 7.2 the errors with respect to the energy norm using piecewise constant and continuous piecewise linear functions, respectively, for the approximation of the flux $\boldsymbol{\theta}_h$. In Figure 7.3 one finds the corresponding errors with respect to the L^2-norms. Note that we use double logarithmic scales.

In the previous sections, we have already proved convergence estimates with respect to the energy norm. The numerical experiments confirm the claimed convergence rate $\mathcal{O}(h)$. This does not include L^2-estimates for the potential u. However, it can be observed that the potential u converges in L^2 with the order h^2, which is optimal for piecewise linear functions. The measured convergence rate for flux in L^2 is h^1 for the piecewise constant approximation. With respect to continuous piecewise linear functions, it seems to be between $h^{3/2}$ and h^2. However, we have not proved these types of convergence rates. But we mention that the application of the Aubin–Nitzsche trick is limited due to the concave vertices of Ω. Obviously, we observe a better approximation of the flux when using piecewise linear functions.

It is confirmed by our experience that the expenses for the boundary integral part is small compared to the finite element part if the integral equations are treated by fast methods. Therefore, the efficiency of the present algorithm is comparable to the efficiency of corresponding finite element least squares methods for interior boundary value problems. Wavelet methods are proved to be an efficient tool for the treatment of boundary integral operators in the coupling. Our approach requires the same boundary element matrices as the FEM-BEM coupling for the second order system.

REFERENCES

[1] A. K. Aziz, R. B. Kellog, and A. B. Stephens, *Least-squares methods for elliptic systems*, Math. Comp., 44 (1985), pp. 53–70.

[2] G. R. Barrenechea, G. N. Gatica, and G. C. Hsiao, *Weak solvability of interior transmission problems via mixed finite elements and Dirichlet-to-Neumann mappings*, J. Comput. Appl. Math., 100 (1998), pp. 145–160.

[3] G. R. Barrenechea, G. N. Gatica, and J.-M. Thomas, *Primal mixed formulations for the coupling of FEM and BEM* I: *Linear problems*, Numer. Funct. Anal. Optim., 19 (1998), pp. 7–32.

[4] M. A. Barrientos, G. N. Gatica, and E. P. Stephan, *A mixed finite element method for nonlinear elasticity: Two-fold saddle point approach and a-posteriori error estimate*, Numer. Math., 91 (2002), pp. 197–222.

[5] J. Bramble, R. D. Lazarov, and J. E. Pasciak, *A least-squares approach based on a discrete minus one inner product for first order systems*, Math. Comp., 66 (1997), pp. 935–955.

[6] J. Bramble, R. D. Lazarov, and J. E. Pasciak, *Least-squares for second order elliptic problems*, Comput. Methods Appl. Mech. Engrg., 152 (1998), pp. 195–210.

[7] J. Bramble, J. E. Pasciak, and J. Xu, *Parallel multilevel preconditioners*, Math. Comp., 55 (1990), pp. 1–22.

[8] U. Brink, C. Carstensen, and E. Stein, *Symmetric coupling of boundary elements and Raviart-Thomas type mixed finite elements in elastostatics*, Numer. Math., 75 (1996), pp. 153–174.

[9] Z. Cai, R. Lazarov, T. A. Manteuffel, and S. F. McCormick, *First-order system least squares for second-order partial differential equations* I, SIAM J. Numer. Anal., 31 (1994), pp. 1785–1799.

[10] C. Carstensen and S. A. Funken, *Coupling of nonconforming finite elements and boundary elements* I: *A-priori estimates*, Computing, 62 (1999), pp. 229–241.

[11] A. Cohen, W. Dahmen, and R. de Vore, *Adaptive wavelet schemes for elliptic operator equations-convergence rates*, Math. Comp., 70 (2001), pp. 27–75.

[12] M. Costabel, *Boundary integral operators on Lipschitz domains: Elementary results*, SIAM J. Math. Anal., 19 (1988), pp. 613–626.

[13] M. Costabel and E. P. Stephan, *Coupling of finite and boundary element methods for an elastoplastic interface problem*, SIAM J. Numer. Anal., 27 (1990), pp. 1212–1226.

[14] W. Dahmen, *Stability of multiscale transformations*, J. Fourier Anal. Appl., 4 (1996), pp. 341–361.

[15] W. Dahmen, A. Kunoth, and R. Schneider, *Wavelet least squares methods for boundary value problems*, SIAM J. Numer. Anal., 39 (2002), pp. 1985–2013.

[16] W. Dahmen, S. Prössdorf, and R. Schneider, *Multiscale methods for pseudo-differential equations on smooth manifolds*, in Proceedings of the International Conference on Wavelets: Theory, Algorithms, and Applications, C. K. Chui, L. Montefusco, and L. Puccio, eds., Academic Press, New York, 1994, pp. 385–424.

[17] W. Dahmen and R. Schneider, *Composite wavelet bases for operator equations*, Math. Comp., 68 (1999), pp. 1533–1567.

[18] W. Dahmen and R. Schneider, *Wavelets on manifolds* I: *Construction and domain decomposition*, SIAM J. Math. Anal., 31 (1999), pp. 184–230.

[19] G. N. Gatica, *Solvability and Galerkin approximations of a class of nonlinear operator equations*, Z. Anal. Anwendungen, 21 (2002), pp. 761–781.

[20] G. N. Gatica, *Combination of mixed finite element and Dirichlet to Neumann methods in nonlinear plane elasticity*, Appl. Math. Lett., 10 (1997), pp. 29–35.

[21] G. N. Gatica and N. Heuer, *A dual-dual formulation for the coupling of mixed-FEM and BEM in hyperelasticity*, SIAM J. Numer. Anal., 38 (2000), pp. 380–400.

[22] G. N. Gatica and G. C. Hsiao, *Boundary-Field Equation Methods for a Class of Nonlinear Problems*, Pitman Res. Notes Math. Ser. 331, Longman, Harlow, UK, 1995.

[23] G. N. Gatica and S. Meddahi, *An a-posteriori error estimate for the coupling of BEM and mixed-FEM*, Numer. Funct. Anal. Optim., 20 (1999), pp. 449–472.

[24] G. N. Gatica and S. Meddahi, *A dual-dual mixed formulation for nonlinear exterior transmission problems*, Math. Comp., 70 (2001), pp. 1461–1480.

[25] G. N. Gatica and W. L. Wendland, *Coupling of mixed finite elements and boundary elements for a hyperelastic interface problem*, SIAM J. Numer. Anal., 34 (1997), pp. 2335–2356.

[26] V. Girault and P.-A. Raviart, *Finite Element Methods for Navier-Stokes Equations. Theory and Algorithms*, Springer-Verlag, New York, 1986.

[27] H. Han, *A new class of variational formulations for the coupling of finite and boundary element methods*, J. Comput. Math., 8 (1990), pp. 223–232.

[28] H. Harbrecht, F. Paiva, C. Perez, and R. Schneider, *Biorthogonal wavelet approximation for the coupling of FEM-BEM*, Numer. Math., 92 (2002), pp. 325–356.

[29] H. Harbrecht, F. Paiva, C. Perez, and R. Schneider, *Multiscale preconditioning for the coupling of FEM-BEM*, Numer. Linear Algebra Appl., 10 (2003), pp. 197–222.

[30] N. Heuer, E. P. Stephan, and T. Tran, *Multilevel additive Schwarz method for the h-p version of the Galerkin boundary element method*, Math. Comp., 67 (1998), pp. 501–518.

[31] G. C. Hsiao, *The coupling of boundary element and finite element methods*, ZAMM. Z. Angew. Math. Mech., 70 (1990), pp. 493–503.

[32] D. C. Jespersen, *A least-square decomposition method for solving elliptic systems*, Math. Comp., 31 (1977), pp. 873–880.

[33] R. Kress, *Linear Integral Equations*, Springer-Verlag, New York, 1989.

[34] S. Meddahi, J. Valdés, O. Menéndez, and P. Pérez, *On the coupling of boundary integral and mixed finite element methods*, J. Comput. Appl. Math., 69 (1996), pp. 113–124.

[35] A. I. Pehlivanov, G. F. Carey, and R. D. Lazarov, *Least-squares mixed finite elements for second-order elliptic problems*, SIAM J. Numer. Anal., 31 (1994), pp. 1368–1377.

[36] R. Schneider, *Multiskalen und Wavelet-Matrixkompression: Analysisbasierte Methoden zur Lösung Großer Vollbesetzter Gleichungssysteme*, Adv. Numer. Math., Teubner, Stuttgart, 1998.

[37] E. P. Stephan, *Coupling of finite elements and boundary elements for some nonlinear interface problems*, Comput. Methods Appl. Mech. Engrg., 101 (1992), pp. 61–72.

[38] T. von Petersdorff, C. Schwab, and R. Schneider, *Multiwavelets for second-kind integral equations*, SIAM J. Numer. Anal., 34 (1997), pp. 2212–2227.

[39] W. L. Wendland, *Elliptic Systems in the Plane*, Pitman, London, 1979.

SIAM J. NUMER. ANAL.
Vol. 41, No. 6, pp. 1997–2031

A FINITE VOLUME SCHEME FOR A NONCOERCIVE ELLIPTIC EQUATION WITH MEASURE DATA*

JÉRÔME DRONIOU[†], THIERRY GALLOUËT[‡], AND RAPHAÈLE HERBIN[‡]

Abstract. We show here the convergence of the finite volume approximate solutions of a convection-diffusion equation to a weak solution, without the usual coercitivity assumption on the elliptic operator and with weak regularity assumptions on the data. Numerical experiments are performed to obtain some rates of convergence in two and three space dimensions.

Key words. noncoercive elliptic equations, measure data, finite volume

AMS subject classifications. 65N12, 35G15

DOI. 10.1137/S0036142902405205

1. Introduction.

The scope of this work is the discretization by the cell-centered finite volume method of convection-diffusion problems on general structured or nonstructured grids. Let Ω be a polygonal (or polyhedral) open subset of $\mathbb{R}^d$ ($d = 2$ or 3); the problem under study can be written as

$$\left\{\begin{array}{ll} -\Delta u + \operatorname{div}(\mathbf{v}u) + bu = \mu & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega, \end{array}\right. \tag{1}$$

with the following hypotheses on the data:

$$\begin{gathered} \mathbf{v} \in (C(\overline{\Omega}))^d, \\ b \in L^2(\Omega),\ b \geq 0 \quad \text{a.e. on } \Omega, \\ \mu \in M(\overline{\Omega}), \end{gathered} \tag{2}$$

where $M(\overline{\Omega}) = (C(\overline{\Omega}))'$ is the dual space of $C(\overline{\Omega})$, which may also be identified with the set of bounded measures on $\overline{\Omega}$. In what follows, we shall consider the usual infinity norm on $C(\overline{\Omega})$, and we shall denote by $||\cdot||_{M(\overline{\Omega})}$ its dual norm on $M(\overline{\Omega})$.

Our purpose is to prove the convergence of the cell-centered finite volume scheme for the discretization of problem (1). Cell-centered schemes for convection-diffusion equations using rectangular, triangular, or Voronoï grids have been analyzed in a number of papers, including [27], [18], [23], [26], [29], and [8]. The analysis which we develop here uses some of the tools which were developed in [14], [20], [15], and [19]. In [15], a convergence result without any assumption of regularity of the solution is proved. An approximate gradient is constructed in [7]. Noncoercive elliptic equations with regular H^{-1} right-hand sides were also recently studied [12]. Finally, a thorough study of finite volume schemes for linear or nonlinear elliptic, parabolic, and hyperbolic equations may be found in [14], to which we refer for further details. The discretization grids which are considered here and in these latter works consist of

*Received by the editors April 9, 2002; accepted for publication (in revised form) May 23, 2003; published electronically November 25, 2003.
http://www.siam.org/journals/sinum/41-6/40520.html

[†]Département de Mathématiques, Université de Montpellier II, CC 051, place Eugène Bataillon, 34095 Montpellier cedex 5, France (droniou@math.univ-montp2.fr).

[‡]Laboratoire d'Analyse, Topologie et Probabilités, Unité Mixte de Recherche 6632, Université de Provence, 30 rue Joliot Curie, 13453 Marseille Cedex 13, France (gallouet@cmi.univ-mrs.fr, herbin@cmi.univ-mrs.fr).

polygonal (or polyhedral) control volumes satisfying adequate geometrical conditions (which are stated in what follows) and not necessarily ordered in a Cartesian grid.

Let us remark that the analysis which is developed here still holds for equations of the type

$$-\mathrm{div}\big(k(x)\nabla u(x)\big)+\mathrm{div}\big(\mathbf{v}(x)\,u(x)\big)+b(x)u(x)=f(x),\quad x\in\Omega, \tag{3}$$

with the following hypotheses on k:

$$\begin{aligned}&k \text{ is a piecewise } C^1 \text{ function from } \overline{\Omega} \text{ to } \mathbb{R};\\&\text{there exists } k_0\in\mathbb{R}_+^* \text{ such that } k(x)\geq k_0 \text{ for almost every } x\in\Omega.\end{aligned}\tag{4}$$

For the sake of the simplicity of notation, we prefer to deal with the Laplace operator here, but we shall point out the modifications which are required if the operator div$(k\nabla.)$ is considered instead: see Remarks 2.2, 2.4, and 2.6. If now k is a tensor satisfying the hypotheses

$$\begin{aligned}&k \text{ is a piecewise } C^1 \text{ function from } \overline{\Omega} \text{ to } \mathbb{R}^{d\times d},\\&\text{for all } x\in\overline{\Omega},\ k(x) \text{ is a symmetric matrix},\\&\text{there exists } k_0\in\mathbb{R}_+^* \text{ such that } k(x)\xi\cdot\xi\geq k_0\\&\qquad\text{for almost every } x\in\Omega \text{ and for all } \xi\in\mathbb{R}^d,\end{aligned}\tag{5}$$

then one may still write the finite volume scheme and obtain some error estimates in the regular case, but the assumptions on the mesh have to be modified; see [20], [24], [25], and [8]. However, if the mesh is Cartesian and if for all $x\in\overline{\Omega}$ the matrix $k(x)$ is diagonal, then the mesh is "aligned" with the grid, and the analysis is similar to the (nonconstant) scalar case of (3).

The originality of the present work with respect to the above-cited works is threefold: first, the elliptic operator associated with the convection-diffusion equation is not assumed to be coercive; second, the convection velocity $\mathbf{v}$ is assumed only to be continuous (it was assumed to be C^1 in previous works); third, the right-hand side μ is supposed only to be a Radon measure.

In the next section, the finite volume scheme for the discretization of (1) is presented, along with the admissible meshes. We then state the main convergence theorem of this paper (Theorem 2.1), along with some preliminary technical results similar to those used in [14], [20], [15], the proof of which is given in an appendix. Section 3 is devoted to a priori estimates on the approximate solutions (existence is not proved at this stage), which will be needed in order to obtain compactness results, and which also yield the existence and uniqueness of the approximate solution. The proof of Theorem 2.1, that is, the proof of the convergence of the approximate solutions to the weak solution of (1), is then given in section 4. Section 5 presents a modified finite volume scheme, where the measure data whose support is on the edges of the mesh are taken into account through a jump of the flux between two neighboring cells; comparing this scheme to the scheme of section 2, the convergence result is easy to obtain. Finally, in section 6 we present some numerical results in two and three space dimensions, using Cartesian or unstructured triangular meshes (in two dimensions), and a spherical mesh in the case of a spherical geometry. These results allow us to derive some rates of convergence of the method, even though no error estimate is known theoretically.

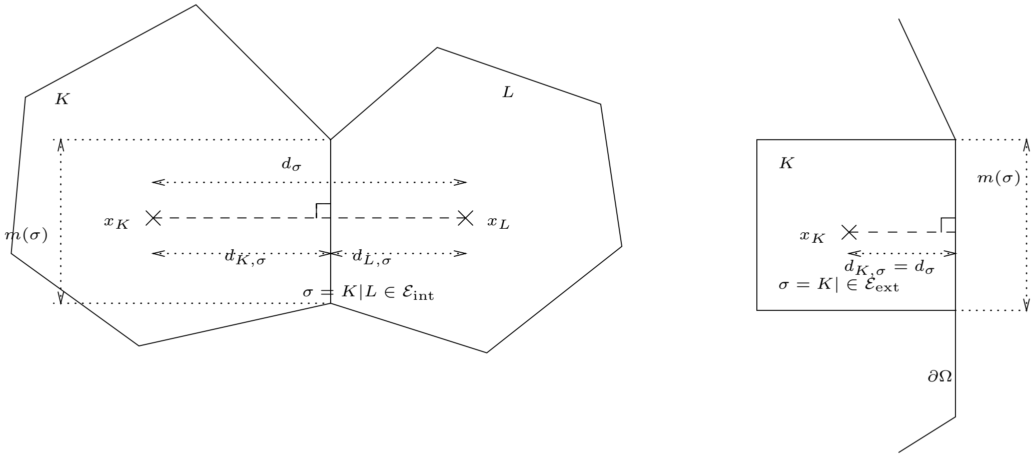

K
L
d_σ
x_K
x_L
$m(\sigma)$
$d_{K,\sigma}$
$d_{L,\sigma}$
$\sigma = K|L \in \mathcal{E}_{\rm int}$
K
$m(\sigma)$
x_K
$d_{K,\sigma} = d_\sigma$
$\sigma = K| \in \mathcal{E}_{\rm ext}$
$\partial\Omega$

FIG. 1. *Notations for an admissible mesh.*

2. Conservative finite volume discretization and convergence result.

DEFINITION 2.1. *An admissible mesh of Ω, denoted by $\mathcal{M}$, is given by a finite partition $\mathcal{T}$ of Ω in polygonal (or polyhedral) convex sets (the "control volumes"), by a finite family $\mathcal{E}$ of disjoint subsets of $\overline{\Omega}$ contained in affine hyperplanes (the "edges"), and by a family $\mathcal{P} = (x_K)_{K\in\mathcal{T}}$ of points in Ω such that*

(i) *each $\sigma \in \mathcal{E}$ is a nonempty open subset of ∂K for some $K \in \mathcal{T}$;*

(ii) *by denoting $\mathcal{E}_K = \{\sigma \in \mathcal{E} \mid \sigma \subset \partial K\}$, one has $\partial K = \cup_{\sigma\in\mathcal{E}_K} \overline{\sigma}$ for all $K \in \mathcal{T}$;*

(iii) *for all $K \neq L$ in $\mathcal{T}$, either the $(d-1)$-dimensional measure of $\overline{K} \cap \overline{L}$ is null, or $\overline{K} \cap \overline{L} = \overline{\sigma}$ for some $\sigma \in \mathcal{E}$, which we denote then by $\sigma = K|L$;*

(iv) *for all $K \in \mathcal{T}$, x_K is in the interior of K;*

(v) *for all $\sigma = K|L \in \mathcal{E}$, the line (x_K, x_L) intersects and is orthogonal to σ;*

(vi) *for all $\sigma \in \mathcal{E}$, $\sigma \subset \partial\Omega \cap \partial K$, the line which is orthogonal to σ and going through x_K, intersects σ.*

Remark 2.1 (other admissible meshes). Note that property (v) in the above definition is required in order to obtain a consistent discretization of the normal fluxes over the boundary of the control domains when using the two-point finite difference scheme to discretize the normal flux. In fact, the above definition of an admissible mesh may be extended to geometries of Ω other than a polygon or a polyhedron. For instance, if $\Omega = \{x \in \mathbb{R}^d;\ |x| \leq r\}$ is a spherical ball of radius r, then a natural mesh is defined by the control volumes $K_0 = \{x \in \mathbb{R}^d;\ |x| \leq r_{1/2}\}$ and, for $i = 1, N$, $K_i = \{x \in \mathbb{R}^d;\ r_{i-1/2} \leq |x| \leq r_{i+1/2}\}$, where $(r_{i+1/2})_{i=1,N} \subset (0, r]$ is a given increasing sequence such that $r_{N+1/2} = r$. Let $x_0 = 0$ and, for $i = 1, \ldots, N$, $r_i \in (r_{i-1/2}, r_{i+1/2})$; then a discretization of the normal diffusive flux $\nabla u \cdot \mathbf{n}$ (where $\mathbf{n}$ is the outward normal unit vector) over the sphere $\{x \in \mathbb{R}^d;\ |x| = r_{i+1/2}\}$ by the two-point scheme $\frac{u_{i+1}-u_i}{r_{i+1}-r_i}$ is clearly consistent if the solution u to (1) depends only on r. Moreover, if $r_{i+1/2} = \frac{1}{2}(r_{i+1} - r_i)$, it is consistent of order 2. Hence this class of spherical discretizations is clearly admissible for the analysis that follows.

The size of the mesh is then defined by $\mathrm{size}(\mathcal{M}) = \sup_{K\in\mathcal{T}} \mathrm{diam}(K)$. We denote by $\mathrm{meas}(K)$ the Lebesgue measure of $K \in \mathcal{T}$. The unit normal to $\sigma \in \mathcal{E}_K$ outward to K is denoted by $\mathbf{n}_{K,\sigma}$. An example of interior neighboring cells and of a boundary cell is given in Figure 1, along with notation.

We define $\mathcal{E}_{\rm int} = \{\sigma \in \mathcal{E} \mid \sigma \not\subset \partial\Omega\}$ and $\mathcal{E}_{\rm ext} = \mathcal{E} \backslash \mathcal{E}_{\rm int}$. If $\sigma \in \mathcal{E}$, $\mathrm{meas}(\sigma)$ is the

$(d-1)$-dimensional measure of σ; if $\sigma = K|L \in \mathcal{E}_{\text{int}}$, d_σ is the distance between the points (x_K, x_L), and $d_{K,\sigma}$ denotes the distance between x_K and σ; if $\sigma \in \mathcal{E}_{\text{ext}} \cap \mathcal{E}_K$, $d_\sigma = d_{K,\sigma}$ is the distance between x_K and σ. The transmissivity through an edge σ is

$$\tau_\sigma = \frac{\text{meas}(\sigma)}{d_\sigma}.$$

Within the integrals, the letter λ (resp., γ) stands for the d- (resp., $(d-1)$)-dimensional measure on the domain Ω (resp., on the edges of the mesh). Note that both measures are denoted by "meas" when applied to a control volume or an edge.

We shall naturally identify the set $\mathbb{R}^{\text{Card}(\mathcal{T})}$ with the set $X(\mathcal{T})$ of functions defined a.e. on Ω and constant on each control volume $K \in \mathcal{T}$.

Remark 2.2. In the case of the operator $\text{div}(k\nabla.)$, which is considered in (3), where k is a function from $\overline{\Omega}$ to $\mathbb{R}$ or $\mathbb{R}^{d\times d}$ which satisfies (4) or (5), admissible meshes must satisfy the following additional condition:

(vii) For any $K \in \mathcal{T}$, the restriction $k|_K$ of the function k to any given control volume K belongs to $C^1(\overline{K})$.

Furthermore if k is a piecewise C^1 function from $\overline{\Omega}$ to $\mathbb{R}^{d\times d}$, the orthogonality conditions (iv) and (v) have to be modified into the following:

(iv)$'$ For any $K \in \mathcal{T}$, let k_K denote the mean value of k on K, that is,

$$k_K = \frac{1}{\text{meas}(K)} \int_K k d\lambda. \tag{6}$$

The set $\mathcal{T}$ is such that there exists a family of points

$$\mathcal{P} = (x_K)_{K\in\mathcal{T}} \quad \text{such that} \quad x_K = \cap_{\sigma\in\mathcal{E}_K} \mathcal{D}_{K,\sigma,k} \in \overline{K},$$

where $\mathcal{D}_{K,\sigma,k}$ is a straight line perpendicular to σ with respect to the scalar product induced by k_K^{-1} such that $\mathcal{D}_{K,\sigma,k} \cap \sigma = \mathcal{D}_{L,\sigma,k} \cap \sigma \neq \emptyset$ if $\sigma = K|L$. Furthermore, if $\sigma = K|L$, let $y_\sigma = \mathcal{D}_{K,\sigma,k} \cap \sigma$ $(= \mathcal{D}_{L,\sigma,k} \cap \sigma)$ and assume that $x_K \neq x_L$.

(v)$'$ For any $\sigma \in \mathcal{E}_{\text{ext}}$, let K be the control volume such that $\sigma \in \mathcal{E}_K$, and let $\mathcal{D}_{K,\sigma,k}$ be the straight line going through x_K and orthogonal to σ with respect to the scalar product induced by k_K^{-1}; then there exists $y_\sigma \in \sigma \cap \mathcal{D}_{K,\sigma,k}$.

If $\mathcal{M}$ is an admissible mesh, and under hypothesis (2), we can define the finite volume discretization of (1).

By denoting, for $K \in \mathcal{T}$ and $\sigma \in \mathcal{E}_K$,

$$b_K = \frac{1}{\text{meas}(K)} \int_K b d\lambda \quad \text{and} \quad v_{K,\sigma} = \int_\sigma \mathbf{v}\cdot\mathbf{n}_{K,\sigma}\, d\gamma, \tag{7}$$

the scheme is defined by

$$\text{for all } \ K \in \mathcal{T}, \ \sum_{\sigma\in\mathcal{E}_K} F_{K,\sigma} + \sum_{\sigma\in\mathcal{E}_K} v_{K,\sigma} u_{\sigma,+} + \text{meas}(K) b_K u_K = \mu(K), \tag{8}$$

$$\begin{aligned} &\text{for all } \ \sigma = K|L \in \mathcal{E}_{\text{int}}, \quad F_{K,\sigma} = -\tau_\sigma(u_L - u_K), \\ &\text{for all } \ \sigma \in \mathcal{E}_{\text{ext}} \cap \mathcal{E}_K, \qquad F_{K,\sigma} = \tau_\sigma u_K, \end{aligned} \tag{9}$$

$$\begin{aligned} &\text{for all } \ \sigma = K|L \in \mathcal{E}_{\text{int}}\,, \quad u_{\sigma,+} = u_K \ \text{ if } \ v_{K,\sigma} \geq 0, \quad u_{\sigma,+} = u_L \ \text{ otherwise}, \\ &\text{for all } \ \sigma \in \mathcal{E}_{\text{ext}} \cap \mathcal{E}_K\,, \quad\ u_{\sigma,+} = u_K \ \text{ if } \ v_{K,\sigma} \geq 0, \quad u_{\sigma,+} = 0 \ \text{ otherwise}. \end{aligned} \tag{10}$$

Equations (8)–(10) form a linear system in $(u_K)_{K\in\mathcal{T}}$ of size $\mathrm{Card}(\mathcal{T})$. Notice that this scheme is conservative in the sense that if $\sigma = K|L$, then $F_{K,\sigma} = -F_{L,\sigma}$ and $v_{K,\sigma} = -v_{L,\sigma}$.

Remark 2.3. The approximation (10) of the convective flux is the classical upwind scheme, which we choose here because it ensures both the existence of a solution to the scheme and the maximum principle without any condition on the size of the mesh. If, instead of the upwind scheme, we used the central difference scheme, then we would need a condition on the size of the mesh in order to have the existence of a solution to the scheme and in order for the maximum principle to hold. However, when the size of the mesh tends to 0, the centered scheme may also be shown to converge. The upwind scheme is often preferred in applications because of its robustness on coarse meshes.

Also note that if $v_{K,\sigma} = 0$, for some $\sigma = K|L$, for example, then (10) does not determine $u_{\sigma,+}$ uniquely since one may take either $u_{\sigma,+} = u_K$ (since $v_{K,\sigma} \geq 0$) or $u_{\sigma,+} = u_L$ (since $v_{L,\sigma} = -v_{K,\sigma} = 0 \geq 0$). However, this is no real problem since $u_{\sigma,+}$ always appears multiplied by $v_{K,\sigma}$ or $v_{L,\sigma}$ and thus, if $v_{K,\sigma} = 0$, the value of $u_{\sigma,+}$ does not matter. (One can, for example, reduce the second sum of (8) to the $\sigma \in \mathcal{E}_K$ such that $v_{K,\sigma} \neq 0$.)

Remark 2.4. In the case of a nonconstant diffusion coefficient as in (3), where k is a function from Ω to $\mathbb{R}$ satisfying (4) or from Ω to $\mathbb{R}^{d\times d}$ satisfying (5), one considers admissible meshes satisfying (vii) of Remark 2.2 and, in the tensor case, also (iv)′ and (v)′ instead of (iv) and (v). For $K \in \mathcal{T}$ and $\sigma \in \mathcal{E}_K$, let

$$k_{K,\sigma} = \left| \frac{1}{\mathrm{meas}(K)} \int_K k\, d\lambda\, \mathbf{n}_{K,\sigma} \right| \tag{11}$$

(where $|\cdot|$ denotes the Euclidean norm). Note that, in the scalar case, this yields in fact $k_{K,\sigma} = \frac{1}{\mathrm{meas}(K)} \int_K k d\lambda$. The exact diffusion fluxes $k(x)\nabla u \cdot \mathbf{n}_{K,\sigma}$ on an edge σ of the mesh may then be approximated in a consistent way (see [14] and [24]) by replacing the formulae in (9) by

- internal edges:

$$F_{K,\sigma} = -\tau_\sigma(u_L - u_K) \quad \text{if } \sigma \in \mathcal{E}_{\mathrm{int}},\ \sigma = K|L, \tag{12}$$

where

$$\tau_\sigma = \mathrm{meas}(\sigma) \frac{k_{K,\sigma} k_{L,\sigma}}{k_{K,\sigma} d_{L,\sigma} + k_{L,\sigma} d_{K,\sigma}};$$

- boundary edges:

$$F_{K,\sigma} = -\tau_\sigma(u_\sigma - u_K) \quad \text{if } \sigma \in \mathcal{E}_{\mathrm{ext}} \text{ and } x_K \notin \sigma, \tag{13}$$

where

$$\tau_\sigma = \mathrm{meas}(\sigma) \frac{k_{K,\sigma}}{d_{K,\sigma}}\cdot$$

Let us now state our main result, which we shall prove in the following sections.

THEOREM 2.1. *If $\mathcal{M}$ is an admissible mesh, then there exists a unique solution to* (8)–(10). *Moreover, if $(\mathcal{M}_n)_{n\geq 1}$ is a sequence of admissible meshes such that there exists $\zeta > 0$ satisfying*

$$\textit{for all } n \geq 1, \textit{ for all } K \in \mathcal{T}_n, \textit{ for all } \sigma \in \mathcal{E}_K, \quad d_{K,\sigma} \geq \zeta d_\sigma,$$

and such that $\mathrm{size}(\mathcal{M}_n) \to 0$, *then, defining* $u_n \in X(\mathcal{T}_n)$ *as the solution of* (8)–(10) *for* $\mathcal{M} = \mathcal{M}_n$, $(u_n)_{n\geq 1}$ *converges to* u *in* $L^p(\Omega)$ *for all* $p \in [1, \frac{d}{d-2})$, *where* u *is the unique solution to* (1) *in the sense*

$$(14)\quad \left\{ \begin{array}{l} u \in \displaystyle\bigcap_{q<\frac{d}{d-1}} W_0^{1,q}(\Omega), \\ \displaystyle\int_\Omega \nabla u \cdot \nabla \varphi d\lambda - \int_\Omega u\mathbf{v} \cdot \nabla \varphi d\lambda + \int_\Omega bu\varphi d\lambda = \int_\Omega \varphi\, d\mu,\ \forall \varphi \in \bigcup_{s>d} W_0^{1,s}(\Omega), \end{array} \right.$$

where $\int_\Omega \varphi\, d\mu = \langle \mu, \varphi \rangle_{(C(\bar\Omega))', C(\bar\Omega)}$. *(We recall that* $W^{1,q}(\Omega)$ *is the set of functions which belong to* $L^q(\Omega)$ *and such that their derivatives are also in* $L^q(\Omega)$ *and* $W_0^{1,q}(\Omega) = \overline{C_c^\infty(\Omega)}^{W^{1,q}(\Omega)}$. *We also recall that* $W_0^{1,s}(\Omega) \subset C_0(\bar\Omega)$ *for* $s > d$.*)*

Remark 2.5. Notice that we do not suppose the existence and uniqueness of a solution to (14); we will prove both.

Remark 2.6. A convergence result still holds if a nonconstant piecewise C^1 diffusion scalar coefficient is considered, i.e., if k satisfies (4) and if (3) is discretized by the scheme (7), (8), (11)–(13). In fact, in the two-dimensional case, the proof follows the one given below in the case $k = Id$. In the three-dimensional case, however, the regularity of the solution to the dual problem (47), which is used in the proof of the uniqueness of a solution to (14) (see section 4) is not so clear. Hence in the 3D case, uniqueness of a solution to (14) is not known, and the convergence result of Theorem 2.1 still holds, but only up to a subsequence.

If one now considers the general tensor case, then some more restrictive assumptions are needed on the mesh in order to obtain consistency of the fluxes; see [14] and [24].

The proof of the existence and uniqueness of a solution to (8)–(10) is based on a priori estimates on the solutions to this problem, which are obtained with the discrete $W_0^{1,q}$ norm defined as follows.

Definition 2.2 (discrete $W^{1,q}$ norm). *If* $\mathcal{M}$ *is an admissible mesh,* $v_\mathcal{T} = (v_K)_{K\in\mathcal{T}} \in \mathbb{R}^{\mathrm{Card}(\mathcal{T})}$, *and* $1 \leq q < \infty$, *we define*

$$||v_\mathcal{T}||_{1,q,\mathcal{M}} = \left(\sum_{\sigma\in\mathcal{E}} \mathrm{meas}(\sigma) d_\sigma \left(\frac{D_\sigma v_\mathcal{T}}{d_\sigma} \right)^q \right)^{\frac{1}{q}},$$

where $D_\sigma v_\mathcal{T} = |v_K - v_L|$ *if* $\sigma = K|L \in \mathcal{E}_{\mathrm{int}}$, *and* $D_\sigma v_\mathcal{T} = |v_K|$ *if* $\sigma \in \mathcal{E}_{\mathrm{ext}} \cap \mathcal{E}_K$.

Let us now state the main a priori estimate, which will be proved in section 3. This estimate is crucial to proving the existence of a solution to (8)–(10), and also to obtaining the compactness properties on approximate solutions which will eventually yield the convergence result.

Theorem 2.2. *Let* $\mathcal{M}$ *be an admissible mesh and* $\zeta > 0$ *satisfy*

$$(15)\qquad \text{for all } K \in \mathcal{T} \text{ and all } \sigma \in \mathcal{E}_K, \quad d_{K,\sigma} \geq \zeta d_\sigma.$$

Then for all $q \in [1, \frac{d}{d-1})$ *there exists* $C > 0$ *depending only on* $(\Omega, \mathbf{v}, q, \zeta)$ *such that, if* $u_\mathcal{T} \in X(\mathcal{T})$ *is a solution to* (8)–(10), *then* $||u_\mathcal{T}||_{1,q,\mathcal{M}} \leq C||\mu||_{M(\overline{\Omega})}$.

In what follows, we shall use the following properties of the discrete $W_0^{1,q}$ norm.

Proposition 2.1 (discrete Poincaré inequality). *If* $1 \leq q \leq 2$, $\mathcal{M}$ *is an admissible mesh, and* $v_\mathcal{T} \in X(\mathcal{T})$, *then*

$$(16)\qquad ||v_\mathcal{T}||_{L^q(\Omega)} \leq \mathrm{diam}(\Omega)||v_\mathcal{T}||_{1,q,\mathcal{M}}.$$

PROPOSITION 2.2 (discrete Sobolev inequality). *Let $1 \le q \le 2$, $\mathcal{M}$ be an admissible mesh, and $\zeta > 0$ satisfy* (15). *Then, with $q^* = \frac{dq}{d-q}$ if $q < d$ and $q^* < \infty$ if $q = d = 2$, there exists $C > 0$ depending only on (Ω, q, q^*, ζ) such that, for all $v_{\mathcal{T}} \in X(\mathcal{T})$,*

$$||v_{\mathcal{T}}||_{L^{q^*}(\Omega)} \le C||v_{\mathcal{T}}||_{1,q,\mathcal{M}}.$$

In fact, it is easily seen that the above inequality also holds for any $r \le q^*$, that is,

$$||v_{\mathcal{T}}||_{L^r(\Omega)} \le C||v_{\mathcal{T}}||_{1,q,\mathcal{M}} \quad \text{for any } r \le q^*.$$

PROPOSITION 2.3 (discrete Rellich theorem). *Let $1 \le q \le 2$ and $\mathcal{M}$ be an admissible mesh. Then there exists $C > 0$ depending only on (Ω, q) such that, for all $h \in \mathbb{R}^d$ and all $v_{\mathcal{T}} \in X(\mathcal{T})$, denoting as $w_{\mathcal{T}}$ the extension of $v_{\mathcal{T}}$ to $\mathbb{R}^d$ by 0 outside Ω, we have*

$$\int_{\mathbb{R}^d} |w_{\mathcal{T}}(x+h) - w_{\mathcal{T}}(x)|^q \, d\lambda(x) \le |h|(|h| + C\text{size}(\mathcal{M}))^{q-1}||v_{\mathcal{T}}||^q_{1,q,\mathcal{M}}. \tag{17}$$

In particular, if $(\mathcal{M}_n)_{n\ge1}$ is a sequence of admissible meshes and $v_n \in X(\mathcal{T}_n)$ is such that $(||v_n||_{1,q,\mathcal{M}_n})_{n\ge1}$ is bounded, then $(v_n)_{n\ge1}$ is relatively compact in $L^q(\Omega)$.

PROPOSITION 2.4 (regularity of the limit). *Let $q \in (1, 2]$ and $(\mathcal{M}_n)_{n\ge1}$ be a sequence of admissible meshes such that* size$(\mathcal{M}_n) \to 0$. *If $v_n \in X(\mathcal{T}_n)$, $(||v_n||_{1,q,\mathcal{M}_n})_{n\ge1}$ is bounded, and $v_n \to v$ in $L^q(\Omega)$, then $v \in W_0^{1,q}(\Omega)$.*

These propositions are easy adaptations of similar results in [14] for the case $q = 2$ (see also [6] for Proposition 2.2 and [19] for Proposition 2.3). We sketch the proofs of these propositions in the appendix for the sake of completeness.

3. A priori estimates.

The aim of this section is to prove the discrete $W^{1,q}$ a priori estimate of Theorem 2.2, which is crucial in the proof of existence of the scheme and also in obtaining a compactness result which will allow us to prove the convergence of a sequence of approximate solutions (Theorem 2.1 and its proof in section 4).

Such a priori estimates were already used for the study of the finite volume approximation of nonlinear elliptic or parabolic equations; see, e.g., [15], [16]. However, in those previous works, the estimates were obtained in a discrete H^1 norm, in accord with the regularity of the solution of the continuous problem.

We prove here some a priori estimates on the solution to (8)–(10) in a discrete $W^{1,q}$ norm, since the solution to the continuous problem is in $W^{1,q}$. As in the continuous case, it is difficult to obtain an estimate on $u_{\mathcal{T}}$ itself. (Note that, in the continuous case, u is not allowed as a test function in (14).) Hence, as in [19], we shall obtain estimates on truncations of the approximate solutions, that is, the functions $T_k(u_{\mathcal{T}})$, where T_k is defined in Figure 2. However, in [19], we dealt only with the Laplace operator, whereas here we allow noncoercive convection-diffusion operators. Because of this noncoercivity, we shall need to start with some weaker estimates, namely, an estimate on $\ln(1 + |u_{\mathcal{T}}|)$, as was done in [9] in the continuous case (see also [13]). In order to obtain this estimate, we shall obtain some estimates on $S_k(u_{\mathcal{T}})$, where $S_k = Id - T_k$ is also defined in Figure 2 and section 3.2. Note that, in the diffusion dominated case, the operator becomes coercive, and the discrete $W^{1,q}$ estimate may be directly obtained from the estimates on $T_k(u_{\mathcal{T}})$, as in [19].

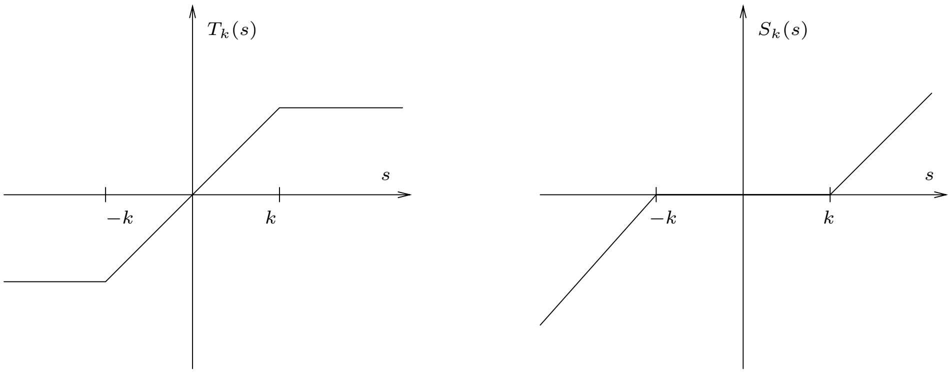


FIG. 2. *The functions T_k and S_k.*

Since the function T_k is bounded, the estimate on $T_k(u_{\mathcal{T}})$ is easy to obtain. The estimate on $S_k(u_{\mathcal{T}})$ is more tricky. The convective term is controlled through a bound of meas(E_k), where $E_k = \{|u_{\mathcal{T}}| > k\}$ (see Corollary 3.1), which is a consequence of an estimate on $\ln(1 + |u_{\mathcal{T}}|)$ (see Proposition 3.1).

Each of the estimates we present here has a continuous counterpart; see, for example, [2], [3] for estimates on nonlinear elliptic equations with measure data and [9], [10] for estimates on linear and nonlinear noncoercive variational elliptic problems. Mixing the techniques of [3] and [9] (or [10]), we can prove estimates (and an existence result) on solutions to linear or nonlinear noncoercive elliptic equations with measure data.

To obtain the estimates on the solutions to (8)–(10), we adapt to the discrete setting this mix of techniques of [3] and [9]. Thus, to make the following proofs easier to understand, we sketch, for each of the discrete estimates, the proof of the corresponding continuous estimate.

3.1. Estimate on $\ln(1 + |u_{\mathcal{T}}|)$.

PROPOSITION 3.1. *Let $\mathcal{M}$ be an admissible mesh. If $u_{\mathcal{T}} = (u_K)_{K\in\mathcal{T}}$ is a solution to* (8)–(10)*, then*

$$||\ln(1 + |u_{\mathcal{T}}|)||_{1,2,\mathcal{T}}^2 \leq 2||\mu||_{M(\overline{\Omega})} + d\text{meas}(\Omega)\,||\,|\mathbf{v}|\,||_{L^\infty(\Omega)}^2 \tag{18}$$

(where $|\mathbf{v}|$ denotes the Euclidean norm of $\mathbf{v}$ in $\mathbb{R}^d$).

Before we prove Proposition 3.1, let us state an easy corollary, which is used in the proof of the estimate of Proposition 3.2.

COROLLARY 3.1. *Let $\mathcal{M}$ be an admissible mesh. If $u_{\mathcal{T}} = (u_K)_{K\in\mathcal{T}}$ is a solution to* (8)–(10) *and, for $k > 0$, $E_k = \{|u_{\mathcal{T}}| > k\}$, then there exists $C \in \mathbb{R}_+^*$ depending only on $(\Omega, \mathbf{v})$ such that*

$$\text{meas}(E_k) \leq \frac{C(1 + ||\mu||_{M(\overline{\Omega})})}{(\ln(1 + k))^2}.$$

Proof of Corollary 3.1. By Proposition 3.1, we get that

$$||\ln(1 + |u_{\mathcal{T}}|)||_{1,2,\mathcal{T}}^2 \leq (2 + d\text{meas}(\Omega)\,||\,|\mathbf{v}|\,||_{L^\infty(\Omega)}^2)(1 + ||\mu||_{M(\overline{\Omega})}).$$

Therefore, using the discrete Poincaré inequality (Proposition 2.1), we get that there exists $C \in \mathbb{R}_+^*$ depending only on $(\Omega, \mathbf{v})$ such that

$$||\ln(1 + |u_{\mathcal{T}}|)||_{L^2(\Omega)}^2 \leq C(1 + ||\mu||_{M(\overline{\Omega})}).$$

Finally, since $\text{meas}(E_k) = \text{meas}(\{\ln(1+|u_{\mathcal{T}}|) \geq \ln(1+k)\})$, the Chebyshev inequality yields that $\text{meas}(E_k) \leq \frac{C(1+||\mu||_{M(\overline{\Omega})})}{(\ln(1+k))^2}$. □

Proof of Proposition 3.1. *Step* 0: Sketch of the proof in the continuous case. Let $\varphi(s) = \int_0^s \frac{dt}{(1+|t|)^2}$. Suppose that $\mu \in H^{-1}(\Omega) \cap L^1(\Omega)$, and let $u \in H_0^1(\Omega)$ be a variational solution of (1). Using $\varphi(u)$ as a test function in the equation satisfied by u, and since φ is bounded by 1, we find

$$\int_\Omega \nabla u \cdot \frac{\nabla u}{(1+|u|)^2}\,d\lambda + \int_\Omega bu\varphi(u)\,d\lambda \leq ||\mu||_{L^1(\Omega)} + ||\,|\mathbf{v}|\,||_{L^\infty(\Omega)} \int_\Omega |u|\frac{|\nabla u|}{(1+|u|)^2}\,d\lambda$$
$$\leq C + C\int_\Omega \frac{|\nabla u|}{(1+|u|)}\,d\lambda,$$

where C depends only on $||\mu||_{L^1(\Omega)}$ and $\mathbf{v}$. Since $\nabla(\ln(1+|u|)) = \text{sgn}(u)\frac{\nabla u}{(1+|u|)}$ and $bu\varphi(u) \geq 0$ (b is nonnegative and $\varphi(s)$ has the same sign as s), we deduce that

$$||\,|\nabla(\ln(1+|u|))|\,||^2_{L^2(\Omega)} \leq C + C\text{meas}(\Omega)^{1/2}||\,|\nabla(\ln(1+|u|))|\,||_{L^2(\Omega)},$$

which gives an estimate on $||\,|\nabla(\ln(1+|u|))|\,||_{L^2(\Omega)}$ (and thus, by the Poincaré inequality, also on $||\ln(1+|u|)||_{L^2(\Omega)}$).

Step 1: Proof of a first discrete estimate. Let $\varphi(s) = \int_0^s \frac{dt}{(1+|t|)^2}$. Multiplying each equality of (8) by $\varphi(u_K)$ and summing on $K \in \mathcal{T}$, we have

$$\sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{E}_K} F_{K,\sigma}\varphi(u_K) + \sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{E}_K} v_{K,\sigma}u_{\sigma,+}\varphi(u_K) + \sum_{K\in\mathcal{T}}\text{meas}(K)b_Ku_K\varphi(u_K)$$
$$= \sum_{K\in\mathcal{T}}\mu(K)\varphi(u_K). \tag{19}$$

Gathering by edges and using (9), we can write

$$\sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{E}_K} F_{K,\sigma}\varphi(u_K) = \sum_{\sigma\in\mathcal{E}}\tau_\sigma(u_K-u_L)(\varphi(u_K)-\varphi(u_L)), \tag{20}$$

where we let $\sigma = K|L$ if $\sigma \in \mathcal{E}_{\text{int}}$, and $u_L = 0$ if $\sigma \in \mathcal{E}_{\text{ext}} \cap \mathcal{E}_K$.

By the conservativity of the fluxes, still gathering by edges, we find

$$\sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{E}_K} v_{K,\sigma}u_{\sigma,+}\varphi(u_K) = \sum_{\sigma\in\mathcal{E}} u_{\sigma,+}v_{K,\sigma}(\varphi(u_K)-\varphi(u_L))$$

(recall that $u_L = 0$—so that $\varphi(u_L) = 0$—if $\sigma \in \mathcal{E}_{\text{ext}} \cap \mathcal{E}_K$). If $\sigma \in \mathcal{E}$, we denote $v_\sigma = |v_{K,\sigma}|$ for a $K \in \mathcal{T}$ such that $\sigma \in \mathcal{E}_K$ (the definition of v_σ does not depend on the choice of such a K) and $u_{\sigma,-}$ the downstream choice of u; i.e., $u_{\sigma,-}$ is such that $\{u_{\sigma,+}, u_{\sigma,-}\} = \{u_K, u_L\}$ (where $\sigma = K|L$ if $\sigma \in \mathcal{E}_{\text{int}}$ and $u_L = 0$ if $\sigma \in \mathcal{E}_{\text{ext}} \cap \mathcal{E}_K$).

Let $\sigma \in \mathcal{E}$; if $v_{K,\sigma} \geq 0$, then $u_{\sigma,+} = u_K$ and $u_{\sigma,-} = u_L$ so that $v_{K,\sigma}(\varphi(u_K) - \varphi(u_L)) = v_\sigma(\varphi(u_{\sigma,+}) - \varphi(u_{\sigma,-}))$; if $v_{K,\sigma} < 0$, then $u_{\sigma,+} = u_L$ and $u_{\sigma,-} = u_K$, which gives $v_{K,\sigma}(\varphi(u_K)-\varphi(u_L)) = -v_\sigma(\varphi(u_{\sigma,-})-\varphi(u_{\sigma,+})) = v_\sigma(\varphi(u_{\sigma,+})-\varphi(u_{\sigma,-}))$. Thus,

$$\sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{E}_K} v_{K,\sigma}u_{\sigma,+}\varphi(u_K) = \sum_{\sigma\in\mathcal{E}} v_\sigma u_{\sigma,+}(\varphi(u_{\sigma,+})-\varphi(u_{\sigma,-})). \tag{21}$$

Since b is nonnegative and $\varphi(s)$ has the same sign as s,

$$\sum_{K\in\mathcal{T}}\text{meas}(K)b_Ku_K\varphi(u_K) \geq 0. \tag{22}$$

Since φ is bounded by 1 and $\mathcal{T}$ is a partition of Ω,

$$\left|\sum_{K\in\mathcal{T}} \mu(K)\varphi(u_K)\right| \le \sum_{K\in\mathcal{T}} |\mu(K)| \le |\mu|(\Omega) = ||\mu||_{M(\overline{\Omega})}. \tag{23}$$

Using (20), (21), (22), and (23) in (19), we get

$$\sum_{\sigma\in\mathcal{E}} \tau_\sigma(u_K - u_L)(\varphi(u_K) - \varphi(u_L)) \le ||\mu||_{M(\overline{\Omega})} + \sum_{\sigma\in\mathcal{E}} v_\sigma u_{\sigma,+}(\varphi(u_{\sigma,-}) - \varphi(u_{\sigma,+})). \tag{24}$$

We now study each term of the last sum a little more precisely. We use the fact that φ is nondecreasing.

- If $u_{\sigma,+} \ge u_{\sigma,-}$ and $u_{\sigma,+} \ge 0$, then $\varphi(u_{\sigma,-}) - \varphi(u_{\sigma,+}) \le 0$ and $u_{\sigma,+}(\varphi(u_{\sigma,-}) - \varphi(u_{\sigma,+})) \le 0$.
- If $u_{\sigma,+} \ge u_{\sigma,-}$ and $u_{\sigma,+} < 0$, then $0 > u_{\sigma,+} \ge u_{\sigma,-}$, so that $(u_{\sigma,+}, u_{\sigma,-})$ have the same sign and $|u_{\sigma,+}| \le |u_{\sigma,-}|$.
- If $u_{\sigma,+} < u_{\sigma,-}$ and $u_{\sigma,+} \ge 0$, then $0 \le u_{\sigma,+} < u_{\sigma,-}$, so that $(u_{\sigma,+}, u_{\sigma,-})$ have the same sign and $|u_{\sigma,+}| \le |u_{\sigma,-}|$.
- If $u_{\sigma,+} < u_{\sigma,-}$ and $u_{\sigma,+} < 0$, then $\varphi(u_{\sigma,-}) - \varphi(u_{\sigma,+}) \ge 0$ and $u_{\sigma,+}(\varphi(u_{\sigma,-}) - \varphi(u_{\sigma,+})) \le 0$.

By defining $\mathcal{A} = \{\sigma \in \mathcal{E} \mid u_{\sigma,+} \ge u_{\sigma,-},\ u_{\sigma,+} < 0\} \cup \{\sigma \in \mathcal{E} \mid u_{\sigma,+} < u_{\sigma,-},\ u_{\sigma,+} \ge 0\}$, we notice that, for all $\sigma \in \mathcal{E}\backslash\mathcal{A}$, $v_\sigma u_{\sigma,+}(\varphi(u_{\sigma,-}) - \varphi(u_{\sigma,+})) \le 0$. This gives

$$\sum_{\sigma\in\mathcal{E}} v_\sigma u_{\sigma,+}(\varphi(u_{\sigma,-}) - \varphi(u_{\sigma,+})) \le \sum_{\sigma\in\mathcal{A}} v_\sigma u_{\sigma,+}(\varphi(u_{\sigma,-}) - \varphi(u_{\sigma,+})).$$

As $v_\sigma \le \mathrm{meas}(\sigma)||\,|\mathbf{v}|\,||_{L^\infty(\Omega)}$, we deduce, using the Cauchy–Schwarz inequality, that

$$\begin{aligned}
&\sum_{\sigma\in\mathcal{E}} v_\sigma u_{\sigma,+}(\varphi(u_{\sigma,-}) - \varphi(u_{\sigma,+}))\\
&\le ||\,|\mathbf{v}|\,||_{L^\infty(\Omega)} \sum_{\sigma\in\mathcal{A}} \mathrm{meas}(\sigma)|u_{\sigma,+}||\varphi(u_{\sigma,-}) - \varphi(u_{\sigma,+})|\\
&\le ||\,|\mathbf{v}|\,||_{L^\infty(\Omega)} \left(\sum_{\sigma\in\mathcal{A}} \mathrm{meas}(\sigma) d_\sigma\right)^{\frac{1}{2}} \left(\sum_{\sigma\in\mathcal{A}} \tau_\sigma u_{\sigma,+}^2(\varphi(u_{\sigma,-}) - \varphi(u_{\sigma,+}))^2\right)^{\frac{1}{2}}.
\end{aligned}$$

However, $\sum_{\sigma\in\mathcal{A}} \mathrm{meas}(\sigma)d_\sigma \le \sum_{\sigma\in\mathcal{E}} \mathrm{meas}(\sigma)d_\sigma = d\mathrm{meas}(\Omega)$ and, if $\sigma \in \mathcal{A}$, $(u_{\sigma,+}, u_{\sigma,-})$ have the same sign and $|u_{\sigma,+}| \le |u_{\sigma,-}|$; thus, by Lemma 3.1 below and Young's inequality,

$$\begin{aligned}
&\sum_{\sigma\in\mathcal{E}} v_\sigma u_{\sigma,+}(\varphi(u_{\sigma,-}) - \varphi(u_{\sigma,+}))\\
&\le (d\mathrm{meas}(\Omega))^{\frac{1}{2}}||\,|\mathbf{v}|\,||_{L^\infty(\Omega)} \left(\sum_{\sigma\in\mathcal{A}} \tau_\sigma(u_{\sigma,-} - u_{\sigma,+})(\varphi(u_{\sigma,-}) - \varphi(u_{\sigma,+}))\right)^{\frac{1}{2}}\\
&\le \frac{1}{2} d\mathrm{meas}(\Omega)\,||\,|\mathbf{v}|\,||_{L^\infty(\Omega)}^2 + \frac{1}{2}\sum_{\sigma\in\mathcal{E}} \tau_\sigma(u_{\sigma,-} - u_{\sigma,+})(\varphi(u_{\sigma,-}) - \varphi(u_{\sigma,+})).
\end{aligned}$$

For all $\sigma \in \mathcal{E}$, we have $\{u_{\sigma,+}, u_{\sigma,-}\} = \{u_K, u_L\}$, so that $(u_{\sigma,-} - u_{\sigma,+})(\varphi(u_{\sigma,-}) - \varphi(u_{\sigma,+})) = (u_K - u_L)(\varphi(u_K) - \varphi(u_L))$. Coming back to (24), we obtain

$$\sum_{\sigma\in\mathcal{E}} \tau_\sigma (u_K - u_L)(\varphi(u_K) - \varphi(u_L)) \le 2||\mu||_{M(\overline{\Omega})} + d\mathrm{meas}(\Omega)\, ||\, |\mathbf{v}|\, ||^2_{L^\infty(\Omega)}\,, \tag{25}$$

which concludes this step.

Step 2: Estimate on $\ln(1 + |u_\mathcal{T}|)$. We notice that, for all $s \in \mathbb{R}$, $\ln(1 + |s|) = \int_0^s \frac{\mathrm{sgn}(t)\,dt}{1+|t|}$. Thus, for all $(x, y) \in \mathbb{R}^2$, by the Cauchy–Schwarz inequality and since φ is nondecreasing,

$$\begin{aligned}(\ln(1 + |x|) - \ln(1 + |y|))^2 = \left(\int_y^x \frac{\mathrm{sgn}(t)\,dt}{1 + |t|}\right)^2 &\le |x - y| \left|\int_y^x \frac{dt}{(1 + |t|)^2}\right| \\ &= |x - y||\varphi(x) - \varphi(y)| = (x - y)(\varphi(x) - \varphi(y)).\end{aligned}$$

Using this upper bound and (25), we deduce the result of the proposition. □

Let us now state and prove the technical result that was used in Step 1 of the above proof.

Lemma 3.1. *Let* $\varphi(s) = \int_0^s \frac{dt}{(1+|t|)^2}$. *If* $(x, y) \in \mathbb{R}^2$ *have the same sign and* $|x| \le |y|$, *then*

$$x^2(\varphi(y) - \varphi(x))^2 \le (y - x)(\varphi(y) - \varphi(x)). \tag{26}$$

Proof of Lemma 3.1. Since φ is C^1-continuous on $\mathbb{R}$, there exists $\theta \in [x, y]$ such that $\varphi(y) - \varphi(x) = \varphi'(\theta)(y - x)$, so that, since φ is nondecreasing,

$$\begin{aligned}x^2(\varphi(y) - \varphi(x))^2 &\le \frac{x^2}{(1 + |\theta|)^2}|y - x|\,|\varphi(y) - \varphi(x)| \\ &\le \frac{x^2}{(1 + |\theta|)^2}(y - x)(\varphi(y) - \varphi(x)).\end{aligned}$$

But $|x| \le |y|$ and x and y have the same sign, so that, since $\theta \in [x, y]$, we have $|\theta| \ge |x|$, and (26) is thus a consequence of the previous inequality. □

3.2. Estimate on $||u_\mathcal{T}||_{1,q,\mathcal{M}}$.

We denote, for $k > 0$, $T_k(s) = \max(-k, \min(s, k))$ and $S_k(s) = s - T_k(s)$ (see Figure 2).

Proposition 3.2. *Let* $\mathcal{M}$ *be an admissible mesh and let* $\zeta > 0$ *be defined by* (15). *We suppose that* μ *satisfies* $||\mu||_{M(\overline{\Omega})} \le 1$. *Then there exists* $k_0 > 0$ *depending only on* $(\Omega, \mathbf{v}, \zeta)$ *and, for all* $m \in (1, 2)$, $C > 0$ *depending only on* $(\Omega, \mathbf{v}, m, \zeta)$ *such that, if* $u_\mathcal{T}$ *is a solution to* (8)–(10) *and* $\varphi_m(s) = \int_0^s \frac{dt}{(1+|t|)^m}$, *we have*

$$\sum_{\sigma\in\mathcal{E}} \tau_\sigma (S_{k_0}(u_K) - S_{k_0}(u_L))(\varphi_m(S_{k_0}(u_K)) - \varphi_m(S_{k_0}(u_L))) \le C \tag{27}$$

and

$$\sum_{\sigma\in\mathcal{E}} \tau_\sigma (T_{k_0}(u_K) - T_{k_0}(u_L))(\varphi_m(T_{k_0}(u_K)) - \varphi_m(T_{k_0}(u_L))) \le C, \tag{28}$$

where we let $\sigma = K|L$ *if* $\sigma \in \mathcal{E}_{\mathrm{int}}$ *and* $u_L = 0$ *if* $\sigma \in \mathcal{E}_{\mathrm{ext}} \cap \mathcal{E}_K$.

Remark 3.1. Problem (8)–(10) being linear, there is no loss of generality in the estimate if we consider measures of norm less than 1, as we will see in Theorem 2.2.

Proof of Proposition 3.2. *Step* 0: Sketch of the estimate in the continuous case. Suppose that $u \in H_0^1(\Omega)$ is a variational solution of (1) with $\mu \in H^{-1}(\Omega) \cap L^1(\Omega)$ satisfying $||\mu||_{L^1(\Omega)} \leq 1$, and take $\varphi_m(S_k(u))$ as a test function in (14). Using the fact that $bu\varphi_m(S_k(u)) \geq 0$ (b is nonnegative and $\varphi_m(s)$ and $S_k(s)$ have the same sign as s), that $\nabla(S_k(u)) = \nabla u$, where $\nabla(S_k(u)) \neq 0$, and that φ_m is bounded by $1/(m-1)$, we have

$$\int_\Omega \frac{|\nabla(S_k(u))|^2}{(1+|S_k(u)|)^m}\,d\lambda \leq \frac{1}{m-1} + ||\,|\mathbf{v}|\,||_{L^\infty(\Omega)} \int_\Omega |u| \frac{|\nabla(S_k(u))|}{(1+|S_k(u)|)^m}\,d\lambda.$$

However, $|u| \leq k + |S_k(u)|$ and $(1+|S_k(u)|)^{2m} \geq (1+|S_k(u)|)^m$, so that, by the Cauchy–Schwarz inequality,

$$\begin{aligned}\int_\Omega \frac{|\nabla(S_k(u))|^2}{(1+|S_k(u)|)^m}\,d\lambda &\leq \frac{1}{m-1} + C_1 k \left(\int_\Omega \frac{|\nabla(S_k(u))|^2}{(1+|S_k(u)|)^{2m}}\,d\lambda\right)^{\frac{1}{2}}\\ &\quad + C_1 \int_\Omega \frac{|S_k(u)|}{(1+|S_k(u)|)^{\frac{m}{2}}} \frac{|\nabla(S_k(u))|}{(1+|S_k(u)|)^{\frac{m}{2}}}\,d\lambda\\ &\leq \frac{1}{m-1} + C_1 k \left(\int_\Omega \frac{|\nabla(S_k(u))|^2}{(1+|S_k(u)|)^{m}}\,d\lambda\right)^{\frac{1}{2}}\\ &\quad + C_1 ||\psi(S_k(u))||_{L^2(\Omega)} \left(\int_\Omega \frac{|\nabla(S_k(u))|^2}{(1+|S_k(u)|)^{m}}\,d\lambda\right)^{\frac{1}{2}},\end{aligned}\tag{29}$$

where C_1 depends only on $(\Omega, \mathbf{v})$ and $\psi(s) = \frac{|s|}{(1+|s|)^{\frac{m}{2}}}$.

Now, by the Hölder inequality and the Sobolev injection, and since $\psi(S_k(u)) = 0$ outside $E_k = \{|u| > k\}$, there exists $r > 2$ depending only on d, and C_2 depending only on (Ω, r) (notice that a dependence on Ω takes into account a dependence on d), such that

$$\begin{aligned}||\psi(S_k(u))||_{L^2(\Omega)} &\leq \mathrm{meas}(E_k)^{\frac{1}{2}-\frac{1}{r}} ||\psi(S_k(u))||_{L^r(\Omega)}\\ &\leq C_2 \mathrm{meas}(E_k)^{\frac{1}{2}-\frac{1}{r}} ||\,|\nabla(\psi(S_k(u)))|\,||_{L^2(\Omega)}.\end{aligned}\tag{30}$$

Since $|\psi'(s)| \leq \frac{1+\frac{m}{2}}{(1+|s|)^{\frac{m}{2}}} \leq \frac{2}{(1+|s|)^{\frac{m}{2}}}$, one has

$$||\,|\nabla(\psi(S_k(u)))|\,||_{L^2(\Omega)} \leq 2 \left(\int_\Omega \frac{|\nabla(S_k(u))|^2}{(1+|S_k(u)|)^m}\,d\lambda\right)^{\frac{1}{2}}.\tag{31}$$

Gathering (29), (30), and (31), we find C_3 depending only on $(\Omega, \mathbf{v})$ such that

$$\begin{aligned}\int_\Omega \frac{|\nabla(S_k(u))|^2}{(1+|S_k(u)|)^m}\,d\lambda &\leq \frac{C_1}{m-1} + C_1 k \left(\int_\Omega \frac{|\nabla(S_k(u))|^2}{(1+|S_k(u)|)^m}\,d\lambda\right)^{\frac{1}{2}}\\ &\quad + C_3 \mathrm{meas}(E_k)^{\frac{1}{2}-\frac{1}{r}} \int_\Omega \frac{|\nabla(S_k(u))|^2}{(1+|S_k(u)|)^m}\,d\lambda.\end{aligned}\tag{32}$$

Thanks to a continuous equivalent of Corollary 3.1, there exists C_4 depending only on $(\Omega, \mathbf{v})$ such that $\mathrm{meas}(E_k) \leq \frac{C_4}{(\ln(1+k))^2}$. Thus, there exists $k_0 > 0$ depending only on

(C_4, C_3, r) (i.e., on $(\Omega, \mathbf{v})$) such that $C_3\text{meas}(E_{k_0})^{\frac{1}{2}-\frac{1}{r}} \le \frac{1}{2}$. Applying (32) to this k_0 gives

$$\int_\Omega \frac{|\nabla(S_{k_0}(u))|^2}{(1+|S_{k_0}(u)|)^m}\,d\lambda \le C_5,$$

where C_5 depends only on $(\Omega, \mathbf{v}, m)$, which is the continuous equivalent of (27).

The estimate on $T_{k_0}(u)$ is quite simple and well known (see [2]). Take $\varphi_m(T_{k_0}(u))$ as a test function in the equation satisfied by u; since $\nabla(T_{k_0}(u)) = 0$ outside $\{|u| \le k_0\}$ and $(1+|T_{k_0}(u)|)^{2m} \ge (1+|T_{k_0}(u)|)^m$, we find

$$\begin{aligned}\int_\Omega \frac{|\nabla(T_{k_0}(u))|^2}{(1+|T_{k_0}(u)|)^m}\,d\lambda &\le \frac{1}{m-1} + ||\,|\mathbf{v}|\,||_{L^\infty(\Omega)} \int_{\{|u|\le k_0\}} |u| \frac{|\nabla(T_{k_0}(u))|}{(1+|T_{k_0}(u)|)^m}\,d\lambda \\ &\le \frac{1}{m-1} + ||\,|\mathbf{v}|\,||_{L^\infty(\Omega)} k_0 \text{meas}(\Omega)^{\frac{1}{2}} \left(\int_\Omega \frac{|\nabla(T_{k_0}(u))|^2}{(1+|T_{k_0}(u)|)^m}\,d\lambda\right)^{\frac{1}{2}}.\end{aligned}$$

This gives an estimate on $T_{k_0}(u)$ which is the continuous equivalent of (28).

Step 1: Estimate on $S_k(u_\mathcal{T})$. Let $\mathcal{M}$ be an admissible mesh, and take $u_\mathcal{T}$ as a solution of (8)–(10). Multiplying each equation of (8) by $\varphi_m(S_k(u_K))$, summing on $K \in \mathcal{T}$, and gathering by edges, we find

$$\begin{aligned}&\sum_{\sigma\in\mathcal{E}} \tau_\sigma(u_K - u_L)(\varphi_m(S_k(u_K)) - \varphi_m(S_k(u_L))) + \sum_{K\in\mathcal{T}} \text{meas}(K) b_K u_K \varphi_m(S_k(u_K)) \\ &\quad = \sum_{K\in\mathcal{T}} \mu(K)\varphi_m(S_k(u_K)) - \sum_{\sigma\in\mathcal{E}} v_{K,\sigma} u_{\sigma,+}(\varphi_m(S_k(u_K)) - \varphi_m(S_k(u_L))).\end{aligned} \tag{33}$$

(Recall that if $\sigma \in \mathcal{E}_{\text{int}}$, we use the notation $\sigma = K|L$, and if $\sigma \in \mathcal{E}_{\text{ext}} \cap \mathcal{E}_K$, we set $u_L = 0$.)

The function φ_m is bounded by $\frac{1}{m-1}$, and $\mathcal{T}$ is a partition of Ω, so that

$$\left|\sum_{K\in\mathcal{T}} \mu(K)\varphi_m(S_k(u_K))\right| \le \frac{1}{m-1}\sum_{K\in\mathcal{T}} |\mu(K)| \le \frac{||\mu||_{M(\overline{\Omega})}}{m-1} \le \frac{1}{m-1}. \tag{34}$$

We again denote by $u_{\sigma,-}$ the downstream choice of u_σ (i.e., $u_{\sigma,-} = u_L$ if $v_{K,\sigma} \ge 0$ and $u_{\sigma,-} = u_K$ otherwise), and $v_\sigma = |v_{K,\sigma}|$ (for a $K \in \mathcal{T}$ such that $\sigma \in \mathcal{E}_K$); we then have

$$\begin{aligned}&-\sum_{\sigma\in\mathcal{E}} v_{K,\sigma} u_{\sigma,+}(\varphi_m(S_k(u_K)) - \varphi_m(S_k(u_L))) \\ &\qquad = \sum_{\sigma\in\mathcal{E}} v_\sigma u_{\sigma,+}(\varphi_m(S_k(u_{\sigma,-})) - \varphi_m(S_k(u_{\sigma,+}))).\end{aligned}$$

However, as in the proof of Proposition 3.1 (because $\varphi_m \circ S_k$ is nondecreasing), we have $u_{\sigma,+}(\varphi_m(S_k(u_{\sigma,-})) - \varphi_m(S_k(u_{\sigma,+}))) \le 0$ if $\sigma \notin \mathcal{A}$, where $\mathcal{A} = \{\sigma \in \mathcal{E} \mid u_{\sigma,+} \ge u_{\sigma,-}, u_{\sigma,+} < 0\} \cup \{\sigma \in \mathcal{E} \mid u_{\sigma,+} < u_{\sigma,-}, u_{\sigma,+} \ge 0\}$. Thus,

$$\begin{aligned}&-\sum_{\sigma\in\mathcal{E}} v_{K,\sigma} u_{\sigma,+}(\varphi_m(S_k(u_K)) - \varphi_m(S_k(u_L))) \\ &\le \sum_{\sigma\in\mathcal{A}} v_\sigma u_{\sigma,+}(\varphi_m(S_k(u_{\sigma,-})) - \varphi_m(S_k(u_{\sigma,+}))) \\ &\le ||\,|\mathbf{v}|\,||_{L^\infty(\Omega)} \sum_{\sigma\in\mathcal{A}} \text{meas}(\sigma)|u_{\sigma,+}|\,|\varphi_m(S_k(u_{\sigma,-})) - \varphi_m(S_k(u_{\sigma,+}))|.\end{aligned} \tag{35}$$

Let $a_{k,\sigma} = \int_0^1 \varphi_m'(S_k(u_{\sigma,+}) + t(S_k(u_{\sigma,-}) - S_k(u_{\sigma,+})))\,dt \geq 0$, so that

$$\varphi_m(S_k(u_{\sigma,-})) - \varphi_m(S_k(u_{\sigma,+})) = a_{k,\sigma}(S_k(u_{\sigma,-}) - S_k(u_{\sigma,+})). \tag{36}$$

We can write

$$\begin{aligned}
&\sum_{\sigma\in\mathcal{A}} \mathrm{meas}(\sigma)|u_{\sigma,+}|\,|\varphi_m(S_k(u_{\sigma,-})) - \varphi_m(S_k(u_{\sigma,+}))|\\
&= \sum_{\sigma\in\mathcal{A}} \mathrm{meas}(\sigma)a_{k,\sigma}^{\frac12}|u_{\sigma,+}|a_{k,\sigma}^{\frac12}|S_k(u_{\sigma,-}) - S_k(u_{\sigma,+})|\\
&\leq \left(\sum_{\sigma\in\mathcal{A}} \mathrm{meas}(\sigma)d_\sigma a_{k,\sigma}u_{\sigma,+}^2\right)^{\frac12}\left(\sum_{\sigma\in\mathcal{A}} \tau_\sigma a_{k,\sigma}(S_k(u_{\sigma,-}) - S_k(u_{\sigma,+}))^2\right)^{\frac12}.
\end{aligned}$$

But, by (36), $a_{k,\sigma}(S_k(u_{\sigma,-}) - S_k(u_{\sigma,+}))^2 = (S_k(u_{\sigma,-}) - S_k(u_{\sigma,+}))(\varphi_m(S_k(u_{\sigma,-})) - \varphi_m(S_k(u_{\sigma,+})))$, so that

$$\begin{aligned}
&\sum_{\sigma\in\mathcal{A}} \mathrm{meas}(\sigma)|u_{\sigma,+}|\,|\varphi_m(S_k(u_{\sigma,-})) - \varphi_m(S_k(u_{\sigma,+}))|\\
&\leq \left(\sum_{\sigma\in\mathcal{A}} \mathrm{meas}(\sigma)d_\sigma a_{k,\sigma}u_{\sigma,+}^2\right)^{\frac12}\\
&\times \left(\sum_{\sigma\in\mathcal{A}} \tau_\sigma (S_k(u_{\sigma,-}) - S_k(u_{\sigma,+}))(\varphi_m(S_k(u_{\sigma,-})) - \varphi_m(S_k(u_{\sigma,+})))\right)^{\frac12}.
\end{aligned}\tag{37}$$

Moreover, for all $\sigma \in \mathcal{A}$, $u_{\sigma,+}$ and $u_{\sigma,-}$ have the same sign and $|u_{\sigma,+}| \leq |u_{\sigma,-}|$. Thus, for such σ, $(S_k(u_{\sigma,+}), S_k(u_{\sigma,-}))$ have the same sign and $|S_k(u_{\sigma,+})| \leq |S_k(u_{\sigma,-})|$ and, by Lemma 3.2 stated after this proof, we deduce that

$$a_{k,\sigma} \leq \frac{1}{(1+|S_k(u_{\sigma,+})|)^m} \leq 1.$$

Since $|u_{\sigma,+}| \leq k + |S_k(u_{\sigma,+})|$, we deduce that

$$a_{k,\sigma}u_{\sigma,+}^2 \leq 2k^2 + 2\frac{|S_k(u_{\sigma,+})|^2}{(1+|S_k(u_{\sigma,+})|)^m},$$

which gives, in (37), using $\sum_{\sigma\in\mathcal{A}} \mathrm{meas}(\sigma)d_\sigma \leq \sum_{\sigma\in\mathcal{E}} \mathrm{meas}(\sigma)d_\sigma = d\mathrm{meas}(\Omega)$ and $(\alpha+\beta)^{1/2} \leq \alpha^{1/2} + \beta^{1/2}$ for all nonnegative (α, β),

$$\begin{aligned}
&\sum_{\sigma\in\mathcal{A}} \mathrm{meas}(\sigma)|u_{\sigma,+}|\,|\varphi_m(S_k(u_{\sigma,-})) - \varphi_m(S_k(u_{\sigma,+}))|\\
&\leq \sqrt{2d\mathrm{meas}(\Omega)}kA_k + \sqrt{2}A_k\left(\sum_{\sigma\in\mathcal{A}} \mathrm{meas}(\sigma)d_\sigma\psi(S_k(u_{\sigma,+}))^2\right)^{\frac12},
\end{aligned}\tag{38}$$

where $\psi(s) = \frac{|s|}{(1+|s|)^{\frac{m}{2}}}$ and

$$\begin{aligned}
A_k &= \left(\sum_{\sigma\in\mathcal{E}} \tau_\sigma (S_k(u_{\sigma,-}) - S_k(u_{\sigma,+}))(\varphi_m(S_k(u_{\sigma,-})) - \varphi_m(S_k(u_{\sigma,+})))\right)^{\frac12}\\
&= \left(\sum_{\sigma\in\mathcal{E}} \tau_\sigma (S_k(u_K) - S_k(u_L))(\varphi_m(S_k(u_K)) - \varphi_m(S_k(u_L)))\right)^{\frac12}.
\end{aligned}$$

(Recall that $\sigma = K|L$ if $\sigma \in \mathcal{E}_{\rm int}$, that $u_L = 0$ if $\sigma \in \mathcal{E}_{\rm ext} \cap \mathcal{E}_K$, and that $\{u_{\sigma,+}, u_{\sigma,-}\} = \{u_K, u_L\}$ for all $\sigma \in \mathcal{E}$.)

We have, since $d_{K,\sigma} \geq \zeta d_\sigma$ for all $K \in \mathcal{T}$ and all $\sigma \in \mathcal{E}_K$,

$$\begin{aligned}\sum_{\sigma\in\mathcal{A}} {\rm meas}(\sigma) d_\sigma \psi(S_k(u_{\sigma,+}))^2 &\leq \sum_{K\in\mathcal{T}} \psi(S_k(u_K))^2 \left(\sum_{\sigma\in\mathcal{A}\cap\mathcal{E}_K \ | \ v_{K,\sigma}\geq 0} {\rm meas}(\sigma) d_\sigma\right)\\ &\leq \frac{1}{\zeta}\sum_{K\in\mathcal{T}} \psi(S_k(u_K))^2 \left(\sum_{\sigma\in\mathcal{E}_K} {\rm meas}(\sigma) d_{K,\sigma}\right)\\ &= \frac{1}{\zeta}\sum_{K\in\mathcal{T}} \psi(S_k(u_K))^2 \times d{\rm meas}(K) = \frac{d}{\zeta}||\psi(S_k(u_\mathcal{T}))||^2_{L^2(\Omega)}.\end{aligned}$$

By Proposition 2.2, and since $\psi(S_k(u_\mathcal{T})) = 0$ outside $E_k = \{|u_\mathcal{T}| > k\}$, we can thus find $r > 2$ and $C_1 > 0$ depending only on (Ω, ζ) such that

$$\left(\sum_{\sigma\in\mathcal{A}} {\rm meas}(\sigma) d_\sigma \psi(S_k(u_{\sigma,+}))^2\right)^{\frac{1}{2}} \leq C_1 {\rm meas}(E_k)^{\frac{1}{2}-\frac{1}{r}} ||\psi(S_k(u_\mathcal{T}))||_{1,2,\mathcal{M}}.$$

But, by Lemma 3.3 below and the definition of A_k,

$$||\psi(S_k(u_\mathcal{T}))||^2_{1,2,\mathcal{M}} = \sum_{\sigma\in\mathcal{E}} \tau_\sigma \psi(S_k(u_K)) - \psi(S_k(u_L))^2 \leq 4A_k^2,$$

so that

$$\left(\sum_{\sigma\in\mathcal{A}} {\rm meas}(\sigma) d_\sigma \psi(S_k(u_{\sigma,+}))^2\right)^{\frac{1}{2}} \leq 2C_1 A_k {\rm meas}(E_k)^{\frac{1}{2}-\frac{1}{r}}.$$

Returning to (38), we thus find

$$\begin{aligned}&\sum_{\sigma\in\mathcal{A}} {\rm meas}(\sigma)|u_{\sigma,+}|\,|\varphi_m(S_k(u_{\sigma,-})) - \varphi_m(S_k(u_{\sigma,+}))|\\ &\qquad \leq \sqrt{2d{\rm meas}(\Omega)}kA_k + 2\sqrt{2}C_1 {\rm meas}(E_k)^{\frac{1}{2}-\frac{1}{r}} A_k^2.\end{aligned}\tag{39}$$

Then (33), (34), (35), (39), and the fact that $b_K u_K \varphi_m(S_k(u_K)) \geq 0$ give

$$\begin{aligned}&\sum_{\sigma\in\mathcal{E}} \tau_\sigma (u_K - u_L)(\varphi_m(S_k(u_K)) - \varphi_m(S_k(u_L)))\\ &\leq \frac{1}{m-1} + ||\,|\mathbf{v}|\,||_{L^\infty(\Omega)}\sqrt{2d{\rm meas}(\Omega)}kA_k + 2\sqrt{2}||\,|\mathbf{v}|\,||_{L^\infty(\Omega)} C_1 {\rm meas}(E_k)^{\frac{1}{2}-\frac{1}{r}} A_k^2\\ &\leq \frac{1}{m-1} + C_2k^2 + \frac{1}{2}A_k^2 + C_2{\rm meas}(E_k)^{\frac{1}{2}-\frac{1}{r}} A_k^2, \qquad (40)\end{aligned}$$

where C_2 depends only on $(\Omega, \mathbf{v}, \zeta)$. However, φ_m and S_k are nondecreasing and S_k is Lipschitz-continuous with Lipschitz constant 1, and thus

$$\begin{aligned}&(S_k(u_K) - S_k(u_L))(\varphi_m(S_k(u_K)) - \varphi_m(S_k(u_L)))\\ &\qquad \leq (u_K - u_L)(\varphi_m(S_k(u_K)) - \varphi_m(S_k(u_L))),\end{aligned}$$

and (40) gives

$$\sum_{\sigma\in\mathcal{E}}\tau_\sigma(S_k(u_K)-S_k(u_L))(\varphi_m(S_k(u_K))-\varphi_m(S_k(u_L)))\le\frac{2}{m-1}+2C_2k^2$$
$$+\,2C_2\mathrm{meas}(E_k)^{\frac12-\frac1r}\sum_{\sigma\in\mathcal{E}}\tau_\sigma(S_k(u_K)-S_k(u_L))(\varphi_m(S_k(u_K))-\varphi_m(S_k(u_L))).\tag{41}$$

By Corollary 3.1, there exists $k_0>0$ depending only on $(\Omega,\mathbf{v},C_2,r)$ (i.e., depending only on $(\Omega,\mathbf{v},\zeta)$) such that $2C_2\mathrm{meas}(E_k)^{\frac12-\frac1r}\le\frac12$. We deduce from (41) that

$$\sum_{\sigma\in\mathcal{E}}\tau_\sigma(S_{k_0}(u_K)-S_{k_0}(u_L))(\varphi_m(S_{k_0}(u_K))-\varphi_m(S_{k_0}(u_L)))\le\frac{4}{m-1}+4C_2k_0^2,$$

which gives (27).

Step 2: Estimate on $T_{k_0}(u_\mathcal{T})$. Multiplying each equation of (8) by $\varphi_m(T_{k_0}(u_K))$, summing on $K\in\mathcal{T}$, and reordering the sums on the edges, we find

$$\sum_{\sigma\in\mathcal{E}}\tau_\sigma(u_K-u_L)(\varphi_m(T_{k_0}(u_K))-\varphi_m(T_{k_0}(u_L)))+\sum_{K\in\mathcal{T}}\mathrm{meas}(K)b_Ku_K\varphi_m(T_{k_0}(u_K))$$
$$=\sum_{K\in\mathcal{T}}\mu(K)\varphi_m(T_{k_0}(u_K))-\sum_{\sigma\in\mathcal{E}}v_{K,\sigma}u_{\sigma,+}(\varphi_m(T_{k_0}(u_K))-\varphi_m(T_{k_0}(u_L))).\tag{42}$$

As before, we have

$$\left|\sum_{K\in\mathcal{T}}\mu(K)\varphi_m(T_{k_0}(u_K))\right|\le\frac{1}{m-1},\tag{43}$$

and, with the previous notations,

$$\begin{aligned}-\sum_{\sigma\in\mathcal{E}}&v_{K,\sigma}u_{\sigma,+}(\varphi_m(T_{k_0}(u_K))-\varphi_m(T_{k_0}(u_L)))\\&=\sum_{\sigma\in\mathcal{E}}v_\sigma u_{\sigma,+}(\varphi_m(T_{k_0}(u_{\sigma,-}))-\varphi_m(T_{k_0}(u_{\sigma,+})))\\&\le\sum_{\sigma\in\mathcal{A}}v_\sigma u_{\sigma,+}(\varphi_m(T_{k_0}(u_{\sigma,-}))-\varphi_m(T_{k_0}(u_{\sigma,+}))).\end{aligned}$$

If $\sigma\in\mathcal{A}$, then $0\le u_{\sigma,+}\le u_{\sigma,-}$ or $u_{\sigma,-}\le u_{\sigma,+}\le 0$. In either case, if $|u_{\sigma,+}|\ge k_0$, then $T_{k_0}(u_{\sigma,+})=T_{k_0}(u_{\sigma,-})$, so that $\varphi_m(T_{k_0}(u_{\sigma,-}))-\varphi_m(T_{k_0}(u_{\sigma,+}))=0$. Thus, in the previous sum, we can suppress the terms $\sigma\in\mathcal{A}$ such that $|u_{\sigma,+}|\ge k_0$ and we have

$$\begin{aligned}&-\sum_{\sigma\in\mathcal{E}}v_{K,\sigma}u_{\sigma,+}(\varphi_m(T_{k_0}(u_K))-\varphi_m(T_{k_0}(u_L)))\\&\le k_0\sum_{\sigma\in\mathcal{A}}v_\sigma|\varphi_m(T_{k_0}(u_{\sigma,-}))-\varphi_m(T_{k_0}(u_{\sigma,+}))|\\&\le k_0||\,|\mathbf{v}|\,||_{L^\infty(\Omega)}\left(\sum_{\sigma\in\mathcal{E}}\mathrm{meas}(\sigma)d_\sigma\right)^{\frac12}\left(\sum_{\sigma\in\mathcal{E}}\tau_\sigma(\varphi_m(T_{k_0}(u_{\sigma,-}))-\varphi_m(T_{k_0}(u_{\sigma,+})))^2\right)^{\frac12}.\end{aligned}$$

Here φ_m and T_{k_0} are nondecreasing and φ_m is Lipschitz-continuous with Lipschitz constant 1; thus, for all $\sigma \in \mathcal{E}$,

$$\begin{aligned}(\varphi_m(T_{k_0}(u_{\sigma,-})) &- \varphi_m(T_{k_0}(u_{\sigma,+})))^2\\ &\le (T_{k_0}(u_{\sigma,-}) - T_{k_0}(u_{\sigma,+}))(\varphi_m(T_{k_0}(u_{\sigma,-})) - \varphi_m(T_{k_0}(u_{\sigma,+})))\\ &= (T_{k_0}(u_K) - T_{k_0}(u_L))(\varphi_m(T_{k_0}(u_K)) - \varphi_m(T_{k_0}(u_L))).\end{aligned}$$

Using this inequality and the fact that $\sum_{\sigma\in\mathcal{E}} \mathrm{meas}(\sigma) d_\sigma = d\mathrm{meas}(\Omega)$, we find

$$\begin{aligned}&-\sum_{\sigma\in\mathcal{E}} v_{K,\sigma} u_{\sigma,+}(\varphi_m(T_{k_0}(u_K)) - \varphi_m(T_{k_0}(u_L)))\\ &\le k_0 ||\,|\mathbf{v}|\,||_{L^\infty(\Omega)} \sqrt{d\mathrm{meas}(\Omega)}\\ &\quad\times \left(\sum_{\sigma\in\mathcal{E}} \tau_\sigma (T_{k_0}(u_K) - T_{k_0}(u_L))(\varphi_m(T_{k_0}(u_K)) - \varphi_m(T_{k_0}(u_L)))\right)^{\frac{1}{2}}.\end{aligned} \tag{44}$$

Since φ_m and T_{k_0} are nondecreasing and T_{k_0} is Lipschitz-continuous with Lipschitz constant 1, we have

$$\begin{aligned}(T_{k_0}(u_K) &- T_{k_0}(u_L))(\varphi_m(T_{k_0}(u_K)) - \varphi_m(T_{k_0}(u_L)))\\ &\le (u_K - u_L)(\varphi_m(T_{k_0}(u_K)) - \varphi_m(T_{k_0}(u_L))).\end{aligned}$$

Combined with (42), (43), (44), and the fact that $b_K u_K \varphi_m(T_{k_0}(u_K)) \ge 0$, this inequality gives

$$\begin{aligned}&\sum_{\sigma\in\mathcal{E}} \tau_\sigma (T_{k_0}(u_K) - T_{k_0}(u_L))(\varphi_m(T_{k_0}(u_K)) - \varphi_m(T_{k_0}(u_L)))\\ &\le \frac{1}{m-1} + k_0 ||\,|\mathbf{v}|\,||_{L^\infty(\Omega)} \sqrt{d\mathrm{meas}(\Omega)}\\ &\quad\times \left(\sum_{\sigma\in\mathcal{E}} \tau_\sigma (T_{k_0}(u_K) - T_{k_0}(u_L))(\varphi_m(T_{k_0}(u_K)) - \varphi_m(T_{k_0}(u_L)))\right)^{\frac{1}{2}},\end{aligned}$$

from which we deduce (28). □

What remains is to state and prove the two technical lemmas which were used in Step 1 of the above proof.

LEMMA 3.2. *Let $m \in (1,2)$ and $\varphi_m(s) = \int_0^s \frac{dt}{(1+|t|)^m}$. If (x,y) have the same sign and $|x| \le |y|$, then*

$$\int_0^1 \varphi_m'(x + t(y-x))\,dt \le \frac{1}{(1+|x|)^m}.$$

Proof of Lemma 3.2. Suppose that $0 \le x \le y$. Then, for all $t \in [0,1]$, $0 \le x \le x + t(y-x)$, so that $\varphi_m'(x+t(y-x)) = \frac{1}{(1+(x+t(y-x)))^m} \le \frac{1}{(1+|x|)^m}$. Integrating this relation on $[0,1]$ gives the desired inequality. If $y \le x \le 0$, we use the fact that φ_m' is even and apply the previous result to $(-x,-y)$. □

LEMMA 3.3. *Let $m \in (1,2)$, $\varphi_m(s) = \int_0^s \frac{dt}{(1+|t|)^m}$, and $\psi(s) = \frac{|s|}{(1+|s|)^{\frac{m}{2}}}$. Then for all $(x,y) \in \mathbb{R}^2$, one has*

$$(\psi(x) - \psi(y))^2 \le 4(x-y)(\varphi_m(x) - \varphi_m(y)).$$

Proof of Lemma 3.3. The function ψ is Lipschitz-continuous and, for all $s \in \mathbb{R}$,

$$|\psi'(s)| = \left| \frac{\mathrm{sgn}(s)}{(1+|s|)^{\frac{m}{2}}} - \frac{\frac{m}{2}\mathrm{sgn}(s)|s|}{(1+|s|)^{1+\frac{m}{2}}} \right| \leq \frac{1+\frac{m}{2}}{(1+|s|)^{\frac{m}{2}}} \leq \frac{2}{(1+|s|)^{\frac{m}{2}}},$$

so that, for all $(x,y) \in \mathbb{R}^2$, by the Cauchy–Schwarz inequality,

$$\begin{aligned}|\psi(x)-\psi(y)| = \left| \int_y^x \psi'(s)\,ds \right| &\leq \left| \int_y^x \frac{4\,ds}{(1+|s|)^m} \right|^{\frac{1}{2}} |x-y|^{\frac{1}{2}} \\ &\leq 2|\varphi_m(x)-\varphi_m(y)|^{\frac{1}{2}}|x-y|^{\frac{1}{2}}.\end{aligned}$$

Taking the power 2 of this inequality and using the fact that φ_m is nondecreasing, we deduce the desired inequality. ◻

We shall now deduce the key estimate on $u_{\mathcal{T}}$ (Theorem 2.2) from Proposition 3.2 and the following lemma.

LEMMA 3.4. *Let $\mathcal{M}$ be an admissible mesh and let $\zeta > 0$ be defined by* (15)*. Let $F : (1,2) \to \mathbb{R}^+$ be a function. For $m \in (1,2)$, we define $\varphi_m(s) = \int_0^s \frac{dt}{(1+|t|)^m}$. If $v_{\mathcal{T}} = (v_K)_{K\in\mathcal{T}} \in X(\mathcal{T})$ satisfies, for all $m \in (1,2)$,*

$$\sum_{\sigma\in\mathcal{E}} \tau_\sigma (v_K - v_L)(\varphi_m(v_K) - \varphi_m(v_L)) \leq F(m)$$

(where we have denoted, as usual, $\sigma = K|L$ if $\sigma \in \mathcal{E}_{\mathrm{int}}$ and $u_L = 0$ if $\sigma \in \mathcal{E}_{\mathrm{ext}} \cap \mathcal{E}_K$), then, for all $q \in [1, \frac{d}{d-1})$, there exists $C > 0$ depending only on (Ω, ζ, F, q) such that $||v_{\mathcal{T}}||_{1,q,\mathcal{M}} \leq C$.

Proof of Lemma 3.4. Let $q \in [1, \frac{d}{d-1})$.

Take $m \in (1,2)$ (fixed later as a function of d and q) and define $a_{m,\sigma} = \int_0^1 \varphi_m'(v_K + t(v_L - v_K))\,dt$. We have $\varphi_m(v_K) - \varphi_m(v_L) = (v_K - v_L)a_{m,\sigma}$, so that

$$\sum_{\sigma\in\mathcal{E}} \tau_\sigma a_{m,\sigma} (D_\sigma v_{\mathcal{T}})^2 \leq F(m).$$

By Hölder's inequality, we have, since $1 \leq q < 2$,

$$\begin{aligned}&\sum_{\sigma\in\mathcal{E}} \mathrm{meas}(\sigma) d_\sigma \left(\frac{D_\sigma v_{\mathcal{T}}}{d_\sigma}\right)^q \\ &\qquad\leq \left(\sum_{\sigma\in\mathcal{E}} \mathrm{meas}(\sigma) d_\sigma a_{m,\sigma} \left(\frac{D_\sigma v_{\mathcal{T}}}{d_\sigma}\right)^2\right)^{\frac{q}{2}} \left(\sum_{\sigma\in\mathcal{E}} \mathrm{meas}(\sigma) d_\sigma a_{m,\sigma}^{-\frac{q}{2-q}}\right)^{\frac{2-q}{2}} \\ &\qquad\leq F(m)^{\frac{q}{2}} \left(\sum_{\sigma\in\mathcal{E}} \mathrm{meas}(\sigma) d_\sigma a_{m,\sigma}^{-\frac{q}{2-q}}\right)^{\frac{2-q}{2}}. \qquad (45)\end{aligned}$$

For all $(x,y) \in \mathbb{R}^2$ and all $t \in [0,1]$, one has $|x + t(y-x)| \leq \sup(|x|,|y|)$, so that

$$\begin{aligned}\varphi_m'(x + t(y-x)) = \frac{1}{(1+|x+t(y-x)|)^m} &\geq \frac{1}{(1+\sup(|x|,|y|))^m} \\ &\geq \inf\left(\frac{1}{(1+|x|)^m}, \frac{1}{(1+|y|)^m}\right).\end{aligned}$$

Taking $x = v_K$, $y = v_L$ and integrating the previous inequality on $t \in [0,1]$, we find

$$a_{m,\sigma} \geq \inf\left(\frac{1}{(1+|v_K|)^m}, \frac{1}{(1+|v_L|)^m}\right),$$

which implies

$$a_{m,\sigma}^{-\frac{q}{2-q}} \leq \sup\left((1+|v_K|)^{\frac{mq}{2-q}}, (1+|v_L|)^{\frac{mq}{2-q}}\right) \leq 2^{\frac{mq}{2-q}}\left(1+|v_K|^{\frac{mq}{2-q}} + |v_L|^{\frac{mq}{2-q}}\right).$$

We deduce from (45), using the fact that $\sum_{\sigma\in\mathcal{E}} \mathrm{meas}(\sigma)d_\sigma = d\mathrm{meas}(\Omega)$ and reordering the sum on the control volumes,

$$||v_{\mathcal{T}}||_{1,q,\mathcal{M}}^q \leq C_1\left(1 + \sum_{K\in\mathcal{T}} |v_K|^{\frac{mq}{2-q}}\left(\sum_{\sigma\in\mathcal{E}_K} \mathrm{meas}(\sigma)d_\sigma\right)\right)^{\frac{2-q}{2}},$$

where C_1 depends only on (F, m, q, Ω). But since $d_{K,\sigma} \geq \zeta d_\sigma$ for all $K \in \mathcal{T}$ and all $\sigma \in \mathcal{E}_K$, we have $\sum_{\sigma\in\mathcal{E}_K} \mathrm{meas}(\sigma)d_\sigma \leq \frac{1}{\zeta}\sum_{\sigma\in\mathcal{E}_K} \mathrm{meas}(\sigma)d_{K,\sigma} = \frac{d}{\zeta}\mathrm{meas}(K)$, and we thus obtain

$$||v_{\mathcal{T}}||_{1,q,\mathcal{M}}^q \leq C_2\left(1 + ||v_{\mathcal{T}}||_{L^{\frac{mq}{2-q}}(\Omega)}^{\frac{mq}{2}}\right), \tag{46}$$

where C_2 depends only on (F, m, q, Ω). (Notice that, since $m > 1$, we always have $\frac{mq}{2-q} \geq 1$.)

By Proposition 2.2, there exists C_3 depending only on (Ω, q, ζ) such that, if $q^* = \frac{dq}{d-q}$ (note that $q < \frac{d}{d-1} \leq d$),

$$||v_{\mathcal{T}}||_{L^{q^*}(\Omega)} \leq C_3||v_{\mathcal{T}}||_{1,q,\mathcal{M}}.$$

Using this in (46), we obtain

$$||v_{\mathcal{T}}||_{L^{q^*}(\Omega)}^q \leq C_3^q C_2\left(1 + ||v_{\mathcal{T}}||_{L^{\frac{mq}{2-q}}(\Omega)}^{\frac{mq}{2}}\right).$$

If $q < \frac{d}{d-1}$, one has $\frac{q}{2-q} < q^*$, so that we can choose $m \in (1,2)$ (depending only on (q,d)) such that $\frac{mq}{2-q} \leq q^*$. We thus obtain, with such a choice of m and Hölder's inequality,

$$||v_{\mathcal{T}}||_{L^{q^*}(\Omega)}^q \leq C_4\left(1 + ||v_{\mathcal{T}}||_{L^{q^*}(\Omega)}^{\frac{mq}{2}}\right),$$

where C_4 depends only on (Ω, ζ, q, F). Since $\frac{mq}{2} < q$ (recall that $m < 2$), this inequality gives us C_5 depending only on (Ω, ζ, q, F) such that $||v_{\mathcal{T}}||_{L^{q^*}(\Omega)} \leq C_5$, and, returning to (46), we deduce the desired estimate on $||v_{\mathcal{T}}||_{1,q,\mathcal{M}}$. ∎

3.3. Proof of Theorem 2.2. Here we give the proof of the key estimate on $||u_{\mathcal{T}}||_{1,q,\mathcal{M}}$, which was stated in Theorem 2.2 and which is crucial to showing the existence and convergence of the solution to the finite volume scheme.

Proof of Theorem 2.2. Let $\Lambda > ||\mu||_{M(\overline{\Omega})}$ (to avoid dividing by 0). Since (8)–(10) make up a linear problem, we see that $u_{\mathcal{T}}/\Lambda$ is a solution to (8)–(10) with μ/Λ instead of μ.

Since $||\mu/\Lambda||_{M(\overline{\Omega})} \le 1$, we can apply Proposition 3.2 to $u_{\mathcal{T}}/\Lambda$; let $k_0 > 0$ depending only on $(\Omega, \mathbf{v}, \zeta)$ given by this proposition. $S_{k_0}(u_{\mathcal{T}}/\Lambda)$ and $T_{k_0}(u_{\mathcal{T}}/\Lambda)$ then satisfy the hypotheses of Lemma 3.4 with a function F depending only on $(\Omega, \mathbf{v}, \zeta)$. We deduce from this lemma that, for all $q \in [1, \frac{d}{d-1})$, there exists $C > 0$ depending only on $(\Omega, \mathbf{v}, \zeta, q)$ such that

$$||S_{k_0}(u_{\mathcal{T}}/\Lambda)||_{1,q,\mathcal{M}} \le C \quad \text{and} \quad ||T_{k_0}(u_{\mathcal{T}}/\Lambda)||_{1,q,\mathcal{M}} \le C.$$

Since $u_{\mathcal{T}}/\Lambda = S_{k_0}(u_{\mathcal{T}}/\Lambda) + T_{k_0}(u_{\mathcal{T}}/\Lambda)$ and $||\cdot||_{1,q,\mathcal{M}}$ is a norm, this gives $||u_{\mathcal{T}}/\Lambda||_{1,q,\mathcal{M}} \le C$, that is to say, $||u_{\mathcal{T}}||_{1,q,\mathcal{M}} \le C\Lambda$. Letting Λ tend to $||\mu||_{M(\overline{\Omega})}$, we obtain the desired estimate on $u_{\mathcal{T}}$. □

4. Proof of Theorem 2.1. We first prove the uniqueness of the solution to (14), which does not involve numerical analysis methods, and then the existence and convergence of the approximate solutions (which yields the existence of a solution to (14)).

Proof of the uniqueness of the solution to (14). This proof uses the regularity results of [22] on the variational solution to $-\Delta v = f \in L^2(\Omega)$, $v_{|\partial\Omega} = 0$, for Ω polygonal (or polyhedral) open set in $\mathbb{R}^d$, $d = 2$ or 3.

Problem (14) being linear, it is sufficient to prove that, if u is a solution to (14) with $\mu = 0$, then $u = 0$.

Let $\theta \in L^\infty(\Omega)$, and take $\varphi \in H_0^1(\Omega) \cap L^\infty(\Omega)$ as the solution to

$$(47) \quad \int_\Omega \nabla\varphi \cdot \nabla\psi \, d\lambda - \int_\Omega \psi \mathbf{v} \cdot \nabla\varphi \, d\lambda + \int_\Omega b\varphi\psi \, d\lambda = \int_\Omega \theta\psi \, d\lambda \quad \text{for all } \psi \in H_0^1(\Omega).$$

The existence of such a φ is ensured by the results of [9]. Letting $\Theta = \theta + \mathbf{v}\cdot\nabla\varphi - b\varphi \in L^2(\Omega)$, we see that $\varphi \in H_0^1(\Omega)$ satisfies $-\Delta\varphi = \Theta$ on Ω.

Since Ω is a polygonal (or polyhedral) open set in $\mathbb{R}^2$ or $\mathbb{R}^3$, the results of [22] give us $\eta > 0$ such that $\varphi \in H^{\frac{3}{2}+\eta}(\Omega)$. Thus, by the Sobolev injections (see [1]), there exists $s > d$ such that $\varphi \in W_0^{1,s}(\Omega)$. (In the case $d = 2$, to obtain such an $s > 2$ we could also have used the result of [28], which is stated for regular open sets but is also true for open sets with Lipschitz-continuous boundary; see [21].)

Thanks to this additional regularity, a density argument allows us to see that (47) is also true for $\psi \in W_0^{1,s'}(\Omega)$, where s' is the conjugate exponent to s, that is, such that $\frac{1}{s} + \frac{1}{s'} = 1$.

We can thus use φ in the equation satisfied by u and u in the equation satisfied by φ to obtain

$$0 = \int_\Omega \nabla u \cdot \nabla\varphi \, d\lambda - \int_\Omega u\mathbf{v} \cdot \nabla\varphi \, d\lambda + \int_\Omega bu\varphi \, d\lambda = \int_\Omega \theta u \, d\lambda.$$

We deduce from this equality, satisfied for all $\theta \in L^\infty(\Omega)$, that $u = 0$, i.e., the uniqueness of the solution to (14). □

Proof of the existence and convergence results. The existence of a unique solution to (8)–(10) is an immediate consequence of the estimate of Theorem 2.2: Indeed, if $\mu = 0$, then this theorem shows that any solution to (8)–(10) is null, that is to say, the square matrix defining this linear system is invertible.

Let us now prove the convergence result. The techniques used here are easy adaptations of the convergence proof of [14].

Let $(u_n)_{n\in\mathbb{N}}$ be a sequence of functions of $L^2(\Omega)$ such that u_n is a solution to (8)–(10) for $\mathcal{M} = \mathcal{M}_n$, where $(\mathcal{M}_n)_{n\in\mathbb{N}}$ is a family of admissible meshes $\mathcal{M}$ satisfying (15) (for some fixed $\zeta > 0$), and such that size$(\mathcal{M}_n)$ tends to 0 as n tends to $+\infty$.

We first prove (Steps 0 to 5) that if $(u_n)_{n\in\mathbb{N}}$ tends to u in $L^p(\Omega)$ for all $p<\frac{d}{d-2}$ as n tends to $+\infty$ (and $\mathrm{size}(\mathcal{M}_n)\to 0$), with $u\in\cap_{q<\frac{d}{d-1}}W_0^{1,q}(\Omega)$, then u is a solution to (14).

We then prove (Step 6), thanks to the a priori estimates of section 3, the compactness of the sequence $(u_n)_{n\in\mathbb{N}}$ and conclude, thanks to the uniqueness result which was proved above, the convergence of $(u_n)_{n\in\mathbb{N}}$ to the solution u to (14).

Step 0: Density argument. By the density of $C_c^\infty(\Omega)$ in $W_0^{1,s}(\Omega)$ for all $s\in(d,\infty)$ and by the regularity results on u, it is clearly sufficient to prove that u satisfies the equations of (14) for all $\varphi\in C_c^\infty(\Omega)$. Take such a φ. Multiplying (8) by $\varphi(x_K)$ and summing over $K\in\mathcal{T}$, we have, by the conservativity of the fluxes and by dropping the index n,

$$\sum_{\sigma\in\mathcal{E}}\tau_\sigma(u_K-u_L)(\varphi(x_K)-\varphi(x_L))+\sum_{\sigma\in\mathcal{E}}v_{K,\sigma}u_{\sigma,+}(\varphi(x_K)-\varphi(x_L)) \\ +\sum_{K\in\mathcal{T}}\mathrm{meas}(K)b_Ku_K\varphi(x_K)=\sum_{K\in\mathcal{T}}\varphi(x_K)\mu(K). \tag{48}$$

We shall now pass to the limit as $\mathrm{size}(\mathcal{M})$ tends to 0 in (48) and prove the convergence of each of the terms to the corresponding term in (14). In fact, the proof of convergence of the first and third terms of the left-hand side can be found in [14] or [15], as can the proof of the second term under a stronger regularity condition. The proof of convergence of the right-hand side may be found in [19], so that the only new part in this proof is Step 4, which shows the convergence of the convective term with a continuous convection velocity (rather than C^1 in previous works). However, we give a quick proof for all terms for the sake of completeness.

Step 1: Convergence of the lower order terms. Denote by $\varphi_\mathcal{T}\in X(\mathcal{T})$ the function defined by $\varphi_K=\varphi(x_K)$ for all $K\in\mathcal{T}$. By the regularity of φ, we have $\varphi_\mathcal{T}\to\varphi$ uniformly on Ω as $\mathrm{size}(\mathcal{M})\to 0$, and thus

$$\sum_{K\in\mathcal{T}}\varphi(x_K)\mu(K)=\int_\Omega\varphi_\mathcal{T}\,d\mu\to\int_\Omega\varphi\,d\mu \tag{49}$$

as $\mathrm{size}(\mathcal{M})\to 0$. (Notice that $\varphi_\mathcal{T}=0$ near $\partial\Omega$ for $\mathrm{size}(\mathcal{M})$ small enough.)

By regularity of b, $b_\mathcal{T}=(b_K)_{K\in\mathcal{T}}$ tends to b in $L^2(\Omega)$ as $\mathrm{size}(\mathcal{M})\to 0$; thus, since $\varphi_\mathcal{T}\to\varphi$ in $L^\infty(\Omega)$ and $u_\mathcal{T}\to u$ in $L^2(\Omega)$ (because $2<d/(d-2)$) as $\mathrm{size}(\mathcal{M})\to 0$, we have

$$\sum_{K\in\mathcal{T}}\mathrm{meas}(K)b_Ku_K\varphi(x_K)=\int_\Omega b_\mathcal{T}u_\mathcal{T}\varphi_\mathcal{T}\,d\lambda\to\int_\Omega bu\varphi\,d\lambda \tag{50}$$

as $\mathrm{size}(\mathcal{M})\to 0$.

Step 2: Convergence of the diffusion term. Gathering by control volumes, we have

$$\sum_{\sigma\in\mathcal{E}}\tau_\sigma(u_K-u_L)(\varphi(x_K)-\varphi(x_L))=\sum_{K\in\mathcal{T}}u_K\sum_{\sigma\in\mathcal{E}_K}\tau_\sigma(\varphi(x_K)-\varphi(x_L)).$$

But, by regularity of φ,

$$\tau_\sigma(\varphi(x_K)-\varphi(x_L))=-\int_\sigma\nabla\varphi\cdot\mathbf{n}_{K,\sigma}\,d\gamma+\mathrm{meas}(\sigma)R_{K,\sigma},$$

where $|R_{K,\sigma}| \le C_1 \mathrm{size}(\mathcal{M})$ (C_1 does not depend on the mesh) and $R_{K,\sigma} = -R_{L,\sigma}$ whenever $\sigma = K|L \in \mathcal{E}_{\mathrm{int}}$. Thus, gathering by edges,

$$\begin{aligned}\sum_{\sigma\in\mathcal{E}} \tau_\sigma(u_K - u_L)(\varphi(x_K) - \varphi(x_L)) + \sum_{K\in\mathcal{T}} u_K \int_{\partial K} \nabla\varphi\cdot\mathbf{n}_K\,d\gamma \\ = \sum_{K\in\mathcal{T}} u_K \sum_{\sigma\in\mathcal{E}_K} \mathrm{meas}(\sigma) R_{K,\sigma} = \sum_{\sigma\in\mathcal{E}} \mathrm{meas}(\sigma) R_{K,\sigma}(u_K - u_L).\end{aligned}$$

However,

$$\begin{aligned}\left|\sum_{\sigma\in\mathcal{E}} \mathrm{meas}(\sigma) R_{K,\sigma}(u_K - u_L)\right| &\le C_1 \mathrm{size}(\mathcal{M}) \sum_{\sigma\in\mathcal{E}} \mathrm{meas}(\sigma) d_\sigma \frac{D_\sigma u_{\mathcal{T}}}{d_\sigma} \\ &= C_1 \mathrm{size}(\mathcal{M}) ||u_{\mathcal{T}}||_{1,1,\mathcal{M}},\end{aligned}$$

and this last quantity tends to 0 as $\mathrm{size}(\mathcal{M}) \to 0$ (because, by Theorem 2.2, $||u_{\mathcal{T}}||_{1,1,\mathcal{M}}$ stays bounded). By noticing that

$$\sum_{K\in\mathcal{T}} u_K \int_{\partial K} \nabla\varphi\cdot\mathbf{n}_K\,d\gamma = \sum_{K\in\mathcal{T}} u_K \int_K \Delta\varphi\,d\lambda = \int_\Omega u_{\mathcal{T}} \Delta\varphi\,d\lambda,$$

and since $u_{\mathcal{T}} \to u$ in $L^1(\Omega)$ as $\mathrm{size}(\mathcal{M}) \to 0$, we deduce that

$$\sum_{\sigma\in\mathcal{E}} \tau_\sigma(u_K - u_L)(\varphi(x_K) - \varphi(x_L)) \to -\int_\Omega u\Delta\varphi\,d\lambda = \int_\Omega \nabla u\cdot\nabla\varphi\,d\lambda \tag{51}$$

as $\mathrm{size}(\mathcal{M}) \to 0$.

Step 3: Preliminary to the convergence of the convection term. (In fact, we prove here the convergence of the convection term if $\mathbf{v}$ is regular.)

Let $\mathbf{w} \in (C^1(\overline{\Omega}))^d$, and define, for $K \in \mathcal{T}$ and $\sigma \in \mathcal{E}_K$, $w_{K,\sigma} = \int_\sigma \mathbf{w}\cdot\mathbf{n}_{K,\sigma}\,d\gamma$. (Notice that, if $\sigma = K|L \in \mathcal{E}_{\mathrm{int}}$, then $w_{K,\sigma} = -w_{L,\sigma}$.) We want to study the limit, as $\mathrm{size}(\mathcal{M}) \to 0$, of $\sum_{\sigma\in\mathcal{E}} w_{K,\sigma} u_{\sigma,+}(\varphi(x_K) - \varphi(x_L))$ (that is to say, the convection term of (48) with $\mathbf{w}$ instead of $\mathbf{v}$).

We have

$$\begin{aligned}&\sum_{\sigma\in\mathcal{E}} w_{K,\sigma} u_{\sigma,+}(\varphi(x_K) - \varphi(x_L)) \\ &= \sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}_K} w_{K,\sigma} u_{\sigma,+}\varphi(x_K) \\ &= \sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}_K} w_{K,\sigma}(u_{\sigma,+} - u_K)\varphi(x_K) + \sum_{K\in\mathcal{T}} \varphi(x_K) u_K \sum_{\sigma\in\mathcal{E}_K} w_{K,\sigma}. \qquad (52)\end{aligned}$$

Since $\sum_{\sigma\in\mathcal{E}_K} w_{K,\sigma} = \int_{\partial K} \mathbf{w}\cdot\mathbf{n}_K\,d\gamma = \int_K \mathrm{div}(\mathbf{w})\,d\lambda$, we have

$$\sum_{K\in\mathcal{T}} \varphi(x_K) u_K \sum_{\sigma\in\mathcal{E}_K} w_{K,\sigma} = \int_\Omega u_{\mathcal{T}} \varphi_{\mathcal{T}} \mathrm{div}(\mathbf{w})\,d\lambda \to \int_\Omega u\varphi \mathrm{div}(\mathbf{w})\,d\lambda \tag{53}$$

as $\mathrm{size}(\mathcal{M}) \to 0$.

Moreover,

$$\sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{E}_K} w_{K,\sigma}(u_{\sigma,+}-u_K)\varphi(x_K) = \sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{E}_K}(u_{\sigma,+}-u_K)\int_\sigma \varphi\mathbf{w}\cdot\mathbf{n}_{K,\sigma}\,d\gamma \\ + \sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{E}_K}(u_{\sigma,+}-u_K)\int_\sigma (\varphi(x_K)-\varphi)\mathbf{w}\cdot\mathbf{n}_{K,\sigma}\,d\gamma.$$

Since, for size($\mathcal{M}$) small enough, the support of φ does not intersect the cells K such that $\partial K\cap\partial\Omega\neq\emptyset$, we have

$$\sum_{K\in\mathcal{T}}\sum_{\sigma=K|L\in\mathcal{E}_K} u_{\sigma,+}\int_\sigma \varphi\mathbf{w}\cdot\mathbf{n}_{K,\sigma}\,d\gamma \\ = \sum_{\sigma\in\mathcal{E}_{\rm int}} u_{\sigma,+}\left(\int_\sigma \varphi\mathbf{w}\cdot\mathbf{n}_{K,\sigma}\,d\gamma + \int_\sigma \varphi\mathbf{w}\cdot\mathbf{n}_{L,\sigma}\,d\gamma\right) = 0,$$

because $\mathbf{n}_{K,\sigma}=-\mathbf{n}_{L,\sigma}$ if $\sigma=K|L\in\mathcal{E}_{\rm int}$. On the other hand,

$$-\sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{E}_K} u_K\int_\sigma \varphi\mathbf{w}\cdot\mathbf{n}_{K,\sigma}\,d\gamma = -\sum_{K\in\mathcal{T}} u_K\int_K {\rm div}(\varphi\mathbf{w})\,d\lambda \\ = -\int_\Omega u_{\mathcal{T}}{\rm div}(\varphi\mathbf{w})\,d\lambda \to -\int_\Omega u{\rm div}(\varphi\mathbf{w})\,d\lambda$$

as size($\mathcal{M}$) → 0. By regularity of φ, we have C_5 depending only on φ such that

$$\left|\sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{E}_K}(u_{\sigma,+}-u_K)\int_\sigma(\varphi(x_K)-\varphi)\mathbf{w}\cdot\mathbf{n}_{K,\sigma}\,d\gamma\right| \\ \le C_5||\,|\mathbf{w}|\,||_{C(\overline{\Omega})}{\rm size}(\mathcal{M})\sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{E}_K}{\rm meas}(\sigma)|u_{\sigma,+}-u_K| \\ \le C_5||\,|\mathbf{w}|\,||_{C(\overline{\Omega})}{\rm size}(\mathcal{M})\sum_{\sigma\in\mathcal{E}}{\rm meas}(\sigma)D_\sigma u_{\mathcal{T}} \\ = C_5||\,|\mathbf{w}|\,||_{C(\overline{\Omega})}{\rm size}(\mathcal{M})||u_{\mathcal{T}}||_{1,1,\mathcal{M}}.$$

The last quantity tending to 0 as size($\mathcal{M}$) → 0, we deduce from the preceding that

$$\sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{E}_K} w_{K,\sigma}(u_{\sigma,+}-u_K)\varphi(x_K)\to -\int_\Omega u{\rm div}(\varphi\mathbf{w})\,d\lambda \tag{54}$$

as size($\mathcal{M}$) → 0.

Using (53) and (54) in (52), we obtain

$$\sum_{\sigma\in\mathcal{E}} w_{K,\sigma}u_{\sigma,+}(\varphi(x_K)-\varphi(x_L)) \to \int_\Omega u\varphi{\rm div}(\mathbf{w})\,d\lambda - \int_\Omega u{\rm div}(\varphi\mathbf{w})\,d\lambda \\ = -\int_\Omega u\mathbf{w}\cdot\nabla\varphi\,d\lambda \tag{55}$$

as size($\mathcal{M}$) → 0.

Step 4. Convergence of the convection term. Let $\varepsilon > 0$ and take $\mathbf{w} \in (C^1(\overline{\Omega}))^d$ such that $|| \, |\mathbf{v} - \mathbf{w}| \, ||_{C(\overline{\Omega})} \leq \varepsilon$. By the regularity of φ,

$$\left|\sum_{\sigma\in\mathcal{E}} v_{K,\sigma} u_{\sigma,+}(\varphi(x_K) - \varphi(x_L)) - \sum_{\sigma\in\mathcal{E}} w_{K,\sigma} u_{\sigma,+}(\varphi(x_K) - \varphi(x_L))\right| \\ \leq C_2\varepsilon \sum_{\sigma\in\mathcal{E}} \mathrm{meas}(\sigma) d_\sigma |u_{\sigma,+}|,$$

where C_2 depends only on φ. Gathering by control volumes, we deduce that

$$\left|\sum_{\sigma\in\mathcal{E}} v_{K,\sigma} u_{\sigma,+}(\varphi(x_K) - \varphi(x_L)) - \sum_{\sigma\in\mathcal{E}} w_{K,\sigma} u_{\sigma,+}(\varphi(x_K) - \varphi(x_L))\right| \\ \leq C_2\varepsilon \sum_{K\in\mathcal{T}} |u_K| \sum_{\sigma\in\mathcal{E}_K \,|\, v_{K,\sigma}\geq 0} \mathrm{meas}(\sigma) d_\sigma.$$

However, by our hypothesis on the mesh, $\sum_{\sigma\in\mathcal{E}_K \,|\, v_{K,\sigma}\geq 0} \mathrm{meas}(\sigma) d_\sigma \leq \zeta^{-1} \sum_{\sigma\in\mathcal{E}_K} \mathrm{meas}(\sigma) d_{K,\sigma} = \zeta^{-1} d\mathrm{meas}(K)$, so that

$$\left|\sum_{\sigma\in\mathcal{E}} v_{K,\sigma} u_{\sigma,+}(\varphi(x_K) - \varphi(x_L)) - \sum_{\sigma\in\mathcal{E}} w_{K,\sigma} u_{\sigma,+}(\varphi(x_K) - \varphi(x_L))\right| \\ \leq C_3\varepsilon \sum_{K\in\mathcal{T}} \mathrm{meas}(K)|u_K| \leq C_4\varepsilon, \tag{56}$$

where C_3 and C_4 depend neither on the mesh $\mathcal{M}$ nor on ε. ($\sum_{K\in\mathcal{T}} \mathrm{meas}(K)|u_K| = ||u_\mathcal{T}||_{L^1(\Omega)}$ is bounded.)

We also notice that

$$\left|\int_\Omega u\mathbf{v} \cdot \nabla\varphi \, d\lambda - \int_\Omega u\mathbf{w} \cdot \nabla\varphi \, d\lambda\right| \leq C_6\varepsilon, \tag{57}$$

where C_6 does not depend on ε.

Next using (55) and (57) in (57), we obtain

$$\limsup_{\mathrm{size}(\mathcal{M})\to 0} \left|\sum_{\sigma\in\mathcal{E}} v_{K,\sigma} u_{\sigma,+}(\varphi(x_K) - \varphi(x_L)) + \int_\Omega u\mathbf{v} \cdot \nabla\varphi \, d\lambda\right| \leq C_7\varepsilon,$$

where C_7 does not depend on ε. This being true for any $\varepsilon > 0$, we deduce that

$$\sum_{\sigma\in\mathcal{E}} v_{K,\sigma} u_{\sigma,+}(\varphi(x_K) - \varphi(x_L)) \to -\int_\Omega u\mathbf{v} \cdot \nabla\varphi \, d\lambda \tag{58}$$

as size($\mathcal{M}$) → 0.

Step 5: Passage to the limit in the scheme. Using (49), (50), (51), and (58), we may pass to the limit in (48) to obtain

$$\int_\Omega \nabla u \cdot \nabla\varphi \, d\lambda - \int_\Omega u\mathbf{v} \cdot \nabla\varphi \, d\lambda + \int_\Omega bu\varphi \, d\lambda = \int_\Omega \varphi \, d\mu,$$

which proves that u is a solution to (14).

Step 6: Proof of the convergence of $(u_n)_{n\in\mathbb{N}}$. Thanks to Theorem 2.2 and to Propositions 2.3 and 2.4, we see that $(u_n)_{n\geq 1}$ is relatively compact in $L^q(\Omega)$ for all $q \in [1, \frac{d}{d-1})$ and that the adherence values (in $L^q(\Omega)$) of this sequence are in $W_0^{1,q}(\Omega)$ (for $q \in (1, \frac{d}{d-1})$). Up to a subsequence, we can thus suppose that $u_n \to u$ in $L^q(\Omega)$ for all $q \in [1, \frac{d}{d-1})$, with $u \in \cap_{q<\frac{d}{d-1}} W_0^{1,q}(\Omega)$; by Proposition 2.2 and Theorem 2.2, $(u_n)_{n\geq 1}$ is also bounded in $L^p(\Omega)$ for all $p < \frac{d}{d-2}$, so that, by an easy consequence of the Vitali convergence theorem, $u_n \to u$ in $L^p(\Omega)$ for all $p < \frac{d}{d-2}$.

By what we have just proved, we see that u is then a solution to (14); since this solution is unique, this proves that the whole sequence $(u_n)_{n\geq 1}$ converges to u.

As a by-product, this convergence entails the existence of a solution to (14) (which can be deduced from previous works, for instance, [2] and [9]). □

5. A scheme with jump of the fluxes.

Until now, we have considered, in the definition of "admissible mesh," a partition of Ω into convex polygonal (or polyhedral) sets. We then defined a finite volume scheme where the conservativity of the numerical fluxes can be written as $F_{K,\sigma} = -F_{L,\sigma}$ for all $\sigma = K|L \in \mathcal{E}_{\rm int}$.

There is, however, another manner of dealing with the discretization of a right-hand-side measure, which was implemented, for instance, in [17] for the numerical simulation of fuel cells. In this formulation, we write that if the support of the measure intersects a given edge, then there is a jump of the flux on this edge. This leads to the following scheme.

The mesh $\mathcal{M}$ that we consider now is defined by a finite family $\mathcal{T}$ of polygonal (or polyhedral) open disjoint subsets of Ω, by a finite family $\mathcal{E}$ of subsets of $\overline{\Omega}$ contained in affine hyperplanes, and by a finite family $\mathcal{P} = (x_K)_{K\in\mathcal{T}}$ of points of Ω such that

(a) $\mathcal{T} \cup \mathcal{E}$ is a partition of $\overline{\Omega}$;

(b) for each $\sigma \in \mathcal{E}$, there exists $K \in \mathcal{T}$ and a nonempty open subset O of ∂K such that $O \subset \sigma \subset \overline{O}$;

(c) items (iii)–(vi) of Definition 2.1 hold.

The notation concerning the mesh is the same as before, and the reader can easily verify that Propositions 2.1–2.4 are still true for such meshes.

Still defining $(b_K)_{K\in\mathcal{T}}$ and $(v_{K,\sigma})_{K\in\mathcal{T}\,,\,\sigma\in\mathcal{E}_K}$ by (7), the new scheme is

$$\text{for all } K \in \mathcal{T},\ \sum_{\sigma\in\mathcal{E}_K} F_{K,\sigma} + \sum_{\sigma\in\mathcal{E}_K} v_{K,\sigma} u_{\sigma,+} + \text{meas}(K) b_K u_K = \mu(K), \tag{59}$$

$$\begin{aligned} &\text{for all } \sigma = K|L \in \mathcal{E}_{\rm int}\,, \quad F_{K,\sigma} = -\tfrac{\text{meas}(\sigma)}{d_{K,\sigma}}(u_\sigma - u_K),\\ &\text{for all } \sigma \in \mathcal{E}_{\rm ext} \cap \mathcal{E}_K, \quad\ \ F_{K,\sigma} = \tau_\sigma u_K, \end{aligned} \tag{60}$$

$$\text{for all } \sigma = K|L \in \mathcal{E}_{\rm int}, \quad F_{K,\sigma} + F_{L,\sigma} = -\mu(\sigma), \tag{61}$$

$$\begin{aligned} &\text{for all } \sigma = K|L \in \mathcal{E}_{\rm int}, \quad u_{\sigma,+} = u_K \text{ if } v_{K,\sigma} \geq 0\,, \quad u_{\sigma,+} = u_L \text{ otherwise},\\ &\text{for all } \sigma \in \mathcal{E}_{\rm ext} \cap \mathcal{E}_K\,, \quad u_{\sigma,+} = u_K \text{ if } v_{K,\sigma} \geq 0\,, \quad u_{\sigma,+} = 0 \text{ otherwise}. \end{aligned} \tag{62}$$

Notice that the unknowns of this scheme are $(u_K)_{K\in\mathcal{T}}$ and $(u_\sigma)_{\sigma\in\mathcal{E}}$ (which represent approximate values on the edges), but that relation (61) allows us to eliminate the $(u_\sigma)_{\sigma\in\mathcal{E}}$; this scheme can thus be considered as a linear system on $(u_K)_{K\in\mathcal{T}}$.

In fact, the elimination of u_σ thanks to (61) gives, for $\sigma = K|L \in \mathcal{E}_{\rm int}$,

$$F_{K,\sigma} = \frac{\text{meas}(\sigma)}{d_\sigma}(u_K - u_L) - \frac{d_{L,\sigma}}{d_\sigma}\mu(\sigma).$$

Thus, this new scheme is in fact the scheme (8)–(10), where we have changed, for all $K \in \mathcal{T}$, $\mu(K)$ by $\widetilde{\mu}_K = \mu(K) + \sum_{\sigma\in\mathcal{E}_K} \frac{d_{L,\sigma}}{d_\sigma}\mu(\sigma)$ (with $\sigma = K|L$ if $\sigma \in \mathcal{E}_{\text{int}}$, and $d_{L,\sigma} = 0$ if $\sigma \in \mathcal{E}_{\text{ext}} \cap \mathcal{E}_K$), which is just another way to discretize the measure μ (forgetting the values of μ on the boundary of the domain, which does not modify the problem since we consider Dirichlet boundary conditions).

The matrix of (59)–(62) is thus the same as the matrix of (8)–(10), and, since $(\widetilde{\mu}_K)_{K\in\mathcal{T}}$ satisfies

$$\begin{aligned}\sum_{K\in\mathcal{T}} |\widetilde{\mu}_K| &\leq \sum_{K\in\mathcal{T}} |\mu(K)| + \sum_{\sigma\in\mathcal{E}} \left(\frac{d_{K,\sigma}}{d_\sigma} + \frac{d_{L,\sigma}}{d_\sigma}\right) |\mu(\sigma)| \\ &= \sum_{K\in\mathcal{T}} |\mu(K)| + \sum_{\sigma\in\mathcal{E}} |\mu(\sigma)| \leq ||\mu||_{M(\overline{\Omega})}\end{aligned}$$

(because $\mathcal{T} \cup \mathcal{E}$ is a partition of $\overline{\Omega}$), the a priori estimates on the solutions to (59)–(62) are obtained in exactly the same way as the estimates of the solutions to (8)–(10).

We also have, for $\varphi \in C_c(\Omega)$, for $\sigma = K|L \in \mathcal{E}_{\text{int}}$,

$$\left|\frac{d_{L,\sigma}}{d_\sigma}\varphi(x_K)\mu(\sigma) + \frac{d_{K,\sigma}}{d_\sigma}\varphi(x_L)\mu(\sigma) - \int_\sigma \varphi\, d\mu\right| \leq \omega(\varphi, \text{size}(\mathcal{M}))|\mu(\sigma)|,$$

where $\omega(\varphi, h)$ is the modulus of continuity of φ; thus,

$$\sum_{K\in\mathcal{T}} \varphi(x_K)\widetilde{\mu}_K \to \int_\Omega \varphi\, d\mu$$

as $\text{size}(\mathcal{M}) \to 0$, and the convergence of the solution of (59)–(62) as $\text{size}(\mathcal{M}) \to 0$ is obtained by the same technique as in the proof of Theorem 2.1.

6. Numerical results.

We performed a few simple numerical experiments on problems to which the exact solution is known, in order to try and obtain some rates of convergence of the finite volume scheme in the presence of a nonregular right-hand side. Numerical results were also shown in [11] in the noncoercive case with right-hand side in H^{-1}, so we shall concentrate here on tests in the irregular data case.

6.1. Comparison of the two finite volume schemes.

The first numerical experiment is concerned with the comparison of the treatment of the singularity in the one-dimensional case. In this case, the Dirac is not a very "mean" measure, in the sense that the solution of the problem is continuous; the jump is only on the derivative. In the first version of the finite volume scheme (scheme (8)–(10), which we shall call Scheme 1 in what follows), the Dirac measure is taken in its integral form in the right-hand side, while in the second version (scheme (59)–(62), which we shall call Scheme 2), the mesh is adapted so as to be able to write the numerical jump of the flux on a cell interface. We solve $-u'' = \delta_{1/2}$, $u(0) = 0$, $u(1) = 0$, on the interval $(0, 1)$; the exact solution is $u(x) = \frac{x}{2}$ for $x < .5$, $u(x) = \frac{(1-x)}{2}$ for $x \geq .5$. We use a uniform mesh and ensure that the number of cells is even, so that in the second scheme, the flux jump is located on a cell interface. The error function e is defined by $e(x) = u(x_K) - u_K$ for any $x \in K$, where $u(x_K)$ denotes the value of the exact solution of the continuous problem at point x_K, and $(u_K)_{K\in\mathcal{T}}$ denotes the solution to the finite volume scheme.

We analyze the rate of convergence by showing the L^1, L^2, and L^∞ norms of the error e versus the number of cells with a log-log scale in Figure 3.

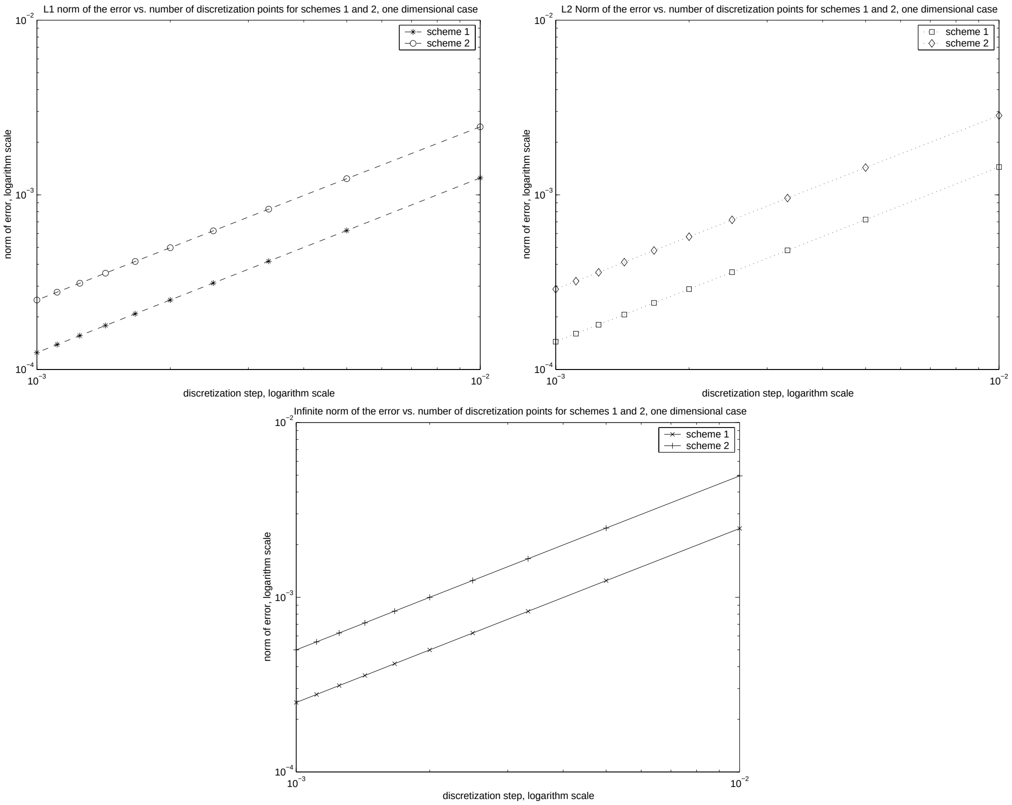


FIG. 3. *Convergence rates in the one-dimensional case.*

TABLE 1
Values of (C, α) *for Schemes* 1 *and* 2*, in the one-dimensional case.*

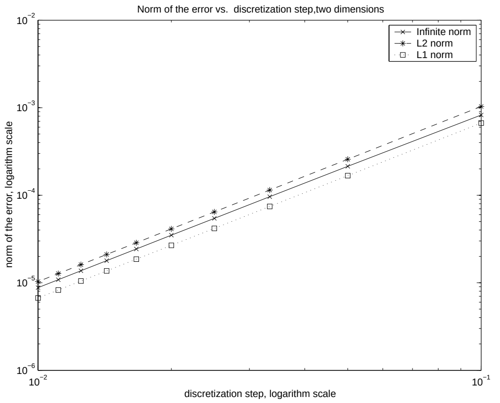
α	L^1 norm	L^2 norm	L^∞ norm
Scheme 1	1.0000	1.0000	0.9961
Scheme 2	0.9923	0.9941	0.9961

C	L^1 norm	L^2 norm	L^∞ norm
Scheme 1	0.1250	0.1443	0.2431
Scheme 2	0.2365	0.2768	0.4861

The results show straight lines for all three norms, so that it is natural to try and evaluate the norms of the error as $||e|| \equiv Ch^\alpha$. The computation of the coefficients C and α from the numerical results are given in Table 1. These coefficients are computed using the two finest meshes.

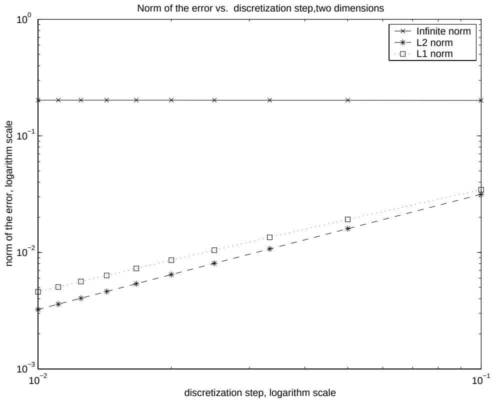
These results show that the two schemes have a rate of convergence which is roughly the same (close to one) and that the constant C is about twice as large for Scheme 2 (jump of flux) than for Scheme 1 (Dirac in one cell). This is quite in accordance with what can be seen from the implementation of the schemes, because Scheme 2 amounts to spreading the Dirac measure over two cells, instead of the one in Scheme 1.

6.2. Two- and three-dimensional tests on a Cartesian mesh. We also implemented the finite volume scheme on the square (resp., cubic) domain $\Omega = (-1, 1)^2$ (resp., $\Omega = (-1, 1)^3$). The domain is discretized with a uniform mesh, and the L^p norm of the error is computed for an increasing number of cells, so as to evaluate the



	α	C
L^1 norm	2.0000	.1031
L^2 norm	2.0000	.0428
L^∞ norm	1.7931	.0690

FIG. 4. *Convergence rate, two-dimensional case, regular right-hand side.*



	α	C
L^1 norm	.9047	.2421
L^2 norm	.9965	.3181

FIG. 5. *Convergence rate, two-dimensional case, right-hand side Dirac at zero, nonsymmetric discrete problem.*

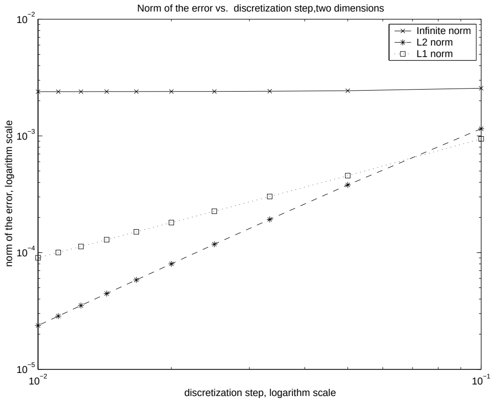
rate of convergence.

We first tested the two-dimensional code for regular data, obtaining the exact solution $u(x, y) = \sin x \sin y$; results are shown in Figure 4. In this case, since the mesh is rectangular and the exact solution regular, the consistency error on the flux is of order 2, and the rate of convergence in the L^2 norm can theoretically be shown to be of order 2 ([14], [20]; see also [5] for a related covolume scheme). The rate of convergence was computed for the piecewise constant error function defined by $e_K = u(x_K) - u_K$ for $K \in \mathcal{T}$, where u is the exact solution and $(u_K)_{K\in\mathcal{T}}$ is the solution to the finite volume scheme.

We then performed some tests with a right-hand side given by a Dirac measure at 0. The boundary conditions were taken such that the exact solution would be the restriction of the solution of $-\Delta u = \delta_0$ in the whole set $\mathbb{R}^2$ (resp., $\mathbb{R}^3$). It is well known that this function lies in $L^p(\mathbb{R}^2)$ for $p \in [1, +\infty)$ (resp., $L^p(\mathbb{R}^3)$ for $p \in [1, 3)$).

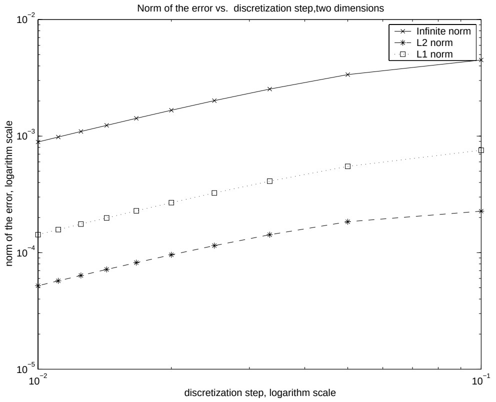
We obtain the results (in log-log scale) given in Figure 5. The coefficients C and α such that $||e|| = Ch^\alpha$ are again evaluated for the norms $L^1(\Omega)$ and $L^2(\Omega)$, and are also given in Figure 5.

In these tests, the mesh is such that the point $(0, 0)$ is located at the corner of the cell $[0, h] \times [0, h]$, where h is the discretization step of the mesh. Hence the radial



	α	C
L^1 norm	1.7740	.0073
L^2 norm	1.0010	.0837

FIG. 6. *Convergence rate, two-dimensional case, right-hand side Dirac at zero, symmetric discrete problem.*



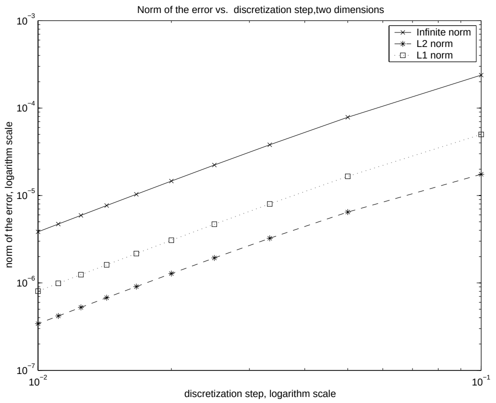
	α	C
L^1 norm	0.9131	.0035
L^2 norm	0.9350	.0086
L^∞ norm	0.9360	.0586

FIG. 7. *Convergence rate, two-dimensional case, right-hand side Dirac at zero, nonsymmetric discrete problem, norm computed on a "regular zone."*

symmetry of the solution is broken by the mesh. If we restore it by allocating one fourth of the Dirac measure to each of the four cells $[0,h]\times[0,h]$, $[0,h]\times[0,-h]$, $[-h,0]\times[0,h]$, and $[-h,0]\times[-h,0]$, we gain in the order of convergence, as can be seen in Figure 6. Hence the order of convergence depends on the singularity of the data, but also on the preservation of the symmetry of the solution.

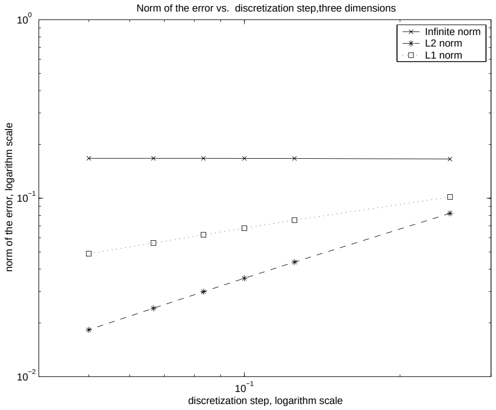
A question of interest is whether the singular data influences the rate of convergence outside of the region of singularity. In order to check this point, we compute the norm of the error between the exact and approximate solutions on the region $\{x \le -.5\} \times \{y \le -.5\}$. We find that, in this case, we recover an order of convergence close to 1 in all norms if the Dirac measure is located at the corner of a cell, in which case the symmetry of the solution is not preserved by the discretization (see Figure 7). In this case, the rate of convergence in the regular zone is perturbed by the singularity outside this zone. (Recall that the theoretical rate of convergence for regular solutions on rectangular meshes is 2; see [20], [14].) However, if we restore the symmetry of the problem as described above, then the rate of convergence is close to 2 (see Figure 8).

We then implemented a three-dimensional Cartesian mesh and found, for the nonsymmetric discrete problem (Dirac located at a corner of the cell $[0,h]^3$), a rate



	α	C
L^1 norm	1.9486	.0247
L^2 norm	1.9571	.0042
L^∞ norm	1.9305	.0025

FIG. 8. *Convergence rate, two-dimensional case, right-hand side Dirac at zero, located at the center of the center cell, norm computed on a "regular zone."*



	α	C
L^1 norm	0.9670	.3314
L^2 norm	0.4809	.2062

FIG. 9. *Convergence rate, three-dimensional case, right-hand side Dirac at zero, nonsymmetric discrete problem.*

of convergence close to 1 in norm L^1 and .5 in norm L^2, as shown in Figure 9. Recall that in this case the exact solution is in L^p for $1 \leq p < 3$.

If the Dirac measure is distributed on the eight cells neighboring the origin, in order to symmetrize the discrete problem, as was done in the two-dimensional case, then one obtains a rate of convergence of 1.631 in the L^1 norm and .504 in the L^2 norm. This seems to indicate a superconvergence in the L^1 norm, although not to the second order (see also Remark 6.1).

6.3. Two-dimensional tests on an unstructured mesh. We also tested our algorithm on an unstructured triangular mesh. Numerical experiments for the cell-centered scheme on triangular meshes were performed in [4] and in [7] in the case of coercive convection-diffusion equations and regular data. These experiments show a convergence rate of order 2, as in the finite element case, although this superconvergence is still, to our knowledge, an open problem in the finite volume case. We show in Figure 10 the rate of convergence which we obtain for the Poisson equation where the right-hand side is a Dirac measure at 0 and the boundary conditions are such that the exact solution is $u(x_1, x_2) = \ln(x_1^2 + x_2^2)$. The refined meshes are not imbedded,

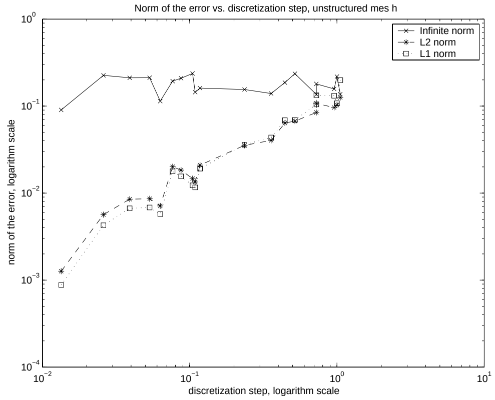
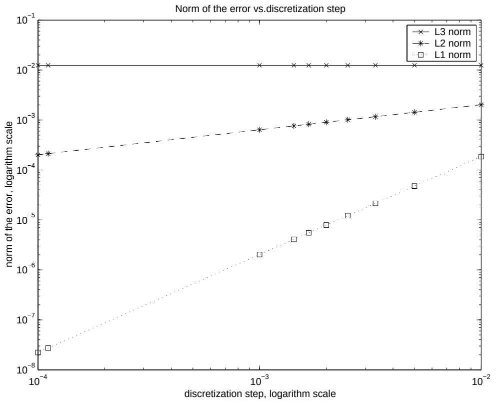


FIG. 10. *Convergence rate, two-dimensional case, right-hand side Dirac at zero, triangular mesh.*



	α	C
L^1 norm	1.9288	.4993
L^2 norm	0.5000	.1879

FIG. 11. *Convergence rate, three-dimensional case, right-hand side Dirac at zero, spherical case.*

so that the convergence lines are not straight, but one can figure out that the L^1 and L^2 norms of the error between the exact and approximate solutions are bounded by $0.1\text{size}(\mathcal{M})^{0.7}$.

6.4. Spherical domain and mesh. We also made some experiments for a three-dimensional spherical problem: We searched for the solution of $-\Delta u = \delta_0$ on the Euclidean unit ball $B(0,1)$ of $\mathbb{R}^3$, with boundary conditions such that the exact solution is the restriction of the solution of $-\Delta u = \delta_0$ in the whole set $\mathbb{R}^3$. The control volumes are defined by $K_i = \{x \in B(0,1);\, ih \le |x| \le (i{+}1)h\}$, for $i = 0, \dots, N$, where $h = \frac{1}{N+1/2}$. As we noted in Remark 2.1, such domains and meshes are not strictly contained in Definition 2.1 of an admissible mesh, since a sphere is hardly a polyhedral domain, but in fact, the discretization of the normal flux on the boundaries of such a spherical mesh is clearly consistent when looking at spherical solutions of (1). Indeed, the numerical flux at interface $i + 1/2$ is taken as $F_{i+1/2} = \frac{4\pi i^2 h^2}{h}(u_{i+1} - u_i)$, where the $(u_i)_{i=0,\dots,N}$ denote the discrete unknowns. In this case, the rate of convergence of the method was found to be 2 in norm L^1 and .5 in norm L^2: see Figure 11.

Hence the symmetry of the problem seems to improve the performance of the method, at least on the L^1 norm.

Remark 6.1. We recall that, in the three-dimensional case, the exact solution

Table 2

Convergence rate, three-dimensional case. Left: the right-hand side is a Dirac measure at zero, spherical case, norm computed on a "regular zone"; right: the right-hand side is a two-dimensional Lebesgue measure supported on the sphere of radius $1/2$*. The norm is computed on the whole set* Ω*.*

	α	C
L^1 norm	2.0506	.3411
L^2 norm	2.1164	.1720
L^∞ norm	2.1295	.2331

	α	C
L^1 norm	1.0506	.1874
L^2 norm	0.9993	.1787
L^∞ norm	0.9983	.2006

$-\Delta u = \delta_0$ is in $L^{3-\varepsilon}$ for any $\varepsilon > 0$; hence we can expect a convergence in L^p for $1 \le p < 3$. From a convergence in $L^{3-\varepsilon}$ for any $\varepsilon > 0$, and a convergence with a rate h^α in L^1, one may deduce (from Hölder's inequality) a convergence in the L^2 norm with a rate of at least $h^{\frac{\alpha}{4}-\varepsilon}$ for any $\varepsilon > 0$. The above numerical results are in accordance with this estimate, both in the spherical case and in the Cartesian case of section 6.2.

We also give in Table 2 the rate of convergence obtained when computing the norm of the error on a zone where the solution is regular, i.e., on the set $\{x \in \mathbb{R}^3, |x| > 1/2\}$. Again, we find in this case a rate of convergence of 2 (even a little more than 2) for all norms.

If we now search for the solution of $-\Delta u = \mu$ on the three-dimensional unit ball, with μ the two-dimensional Lebesgue measure supported on the sphere of radius .5, then the obtained convergence rate is again 1, even though the exact solution is more regular than the solution to the Dirac problem; see Figure 8. Note that, in this case, the exact solution is in L^∞ (and even in H^1).

7. Appendix.

Throughout this section, for any $q \in (1, +\infty)$ we denote by q' its conjugate exponent, that is, $q' \in (1, +\infty)$ such that $\frac{1}{q} + \frac{1}{q'} = 1$.

Proof of Proposition 2.1. The case $q = 2$ is addressed in [14]. We use the same method for $q \in [1, 2)$.

Define, for $\sigma \in \mathcal{E}$ and $(x, y) \in \mathbb{R}^d$, $\chi_\sigma(x, y) = 1$ if $\sigma \cap [x, y] \neq \emptyset$ and $\chi_\sigma(x, y) = 0$ otherwise. Let $\mathbf{d}$ be a unit vector and define, for $x \in \Omega$, $y(x)$ as the point on the semi-line, with origin x and direction $\mathbf{d}$, such that $y(x) \in \partial\Omega$ and $[x, y(x)] \subset \overline{\Omega}$. If $\sigma \in \mathcal{E}$, we let $c_\sigma = |\mathbf{n}_\sigma \cdot \mathbf{d}|$, where $\mathbf{n}_\sigma$ is a unit normal to σ.

For all $x \in \Omega$ such that x does not belong to an affine hyperplane generated by some $\sigma \in \mathcal{E}$, i.e., for almost every $x \in \Omega$, we have

$$|v_\mathcal{T}(x)| \le \sum_{\sigma \in \mathcal{E}} \chi_\sigma(x, y(x)) D_\sigma v_\mathcal{T}.$$

(Recall that $v_\mathcal{T}(x) = v_K$ for the $K \in \mathcal{T}$ such that $x \in K$.) Take such an x and suppose that, for some $\sigma \in \mathcal{E}$, $c_\sigma = 0$; we then have $\chi_\sigma(x, y(x)) = 0$ (indeed, otherwise x would belong to the affine hyperplane generated by σ). Thus, the preceding sum can be reduced to the $\sigma \in \mathcal{E}$ such that $c_\sigma \neq 0$, and we can write, thanks to Hölder's inequality, for almost every $x \in \Omega$,

(63)

$$|v_\mathcal{T}(x)|^q \le \left(\sum_{\sigma \in \mathcal{E} \,|\, c_\sigma \neq 0} \chi_\sigma(x, y(x)) d_\sigma c_\sigma^{-\frac{q}{q'}} \left(\frac{D_\sigma v_\mathcal{T}}{d_\sigma} \right)^q \right) \left(\sum_{\sigma \in \mathcal{E} \,|\, c_\sigma \neq 0} \chi_\sigma(x, y(x)) d_\sigma c_\sigma \right)^{\frac{q}{q'}}.$$

Since we have $\sum_{\sigma\in\mathcal{E}} \chi_\sigma(x,y(x))d_\sigma c_\sigma \le \operatorname{diam}(\Omega)$ for all $x\in\Omega$ (see [14]) and $\int_\Omega \chi_\sigma(x,y(x))\,d\lambda(x) \le \operatorname{diam}(\Omega)\operatorname{meas}(\sigma)c_\sigma$, we obtain, integrating (63) on Ω,

$$\int_\Omega |v_\mathcal{T}|^q\,d\lambda \le \operatorname{diam}(\Omega)^{\frac{q}{q'}} \sum_{\sigma\in\mathcal{E}\,|\,c_\sigma\neq 0} \operatorname{diam}(\Omega)\operatorname{meas}(\sigma)d_\sigma {c_\sigma}^{1-\frac{q}{q'}} \left(\frac{D_\sigma v_\mathcal{T}}{d_\sigma}\right)^q.$$

However, $q\le 2$, so that $1-\frac{q}{q'} = 2-q \ge 0$ and $c_\sigma^{2-q}\le 1$, which concludes this proof. □

Proof of Proposition 2.2. The case $d=2$ has already been covered in the course of the proof of the discrete Sobolev inequalities in [14, inequality (9.73), p. 791].

For $d=3$, the case $q=2$ may be found in [8]. The case of a general q is similar; we use the following inequality [14, inequality (9.75), p. 793]: for any $w_\mathcal{T}\in X(\mathcal{T})$,

$$\int_\Omega |w_\mathcal{T}|^{\frac{3}{2}}\,d\lambda \le \left(\sum_{\sigma\in\mathcal{E}} \operatorname{meas}(\sigma)D_\sigma w_\mathcal{T}\right)^{\frac{3}{2}}.$$

Applying this to $w_K = |v_K|^{\frac{2q}{3-q}}\operatorname{sgn}(v_K)$, and since $D_\sigma w_\mathcal{T} \le \frac{2q}{3-q}(|v_K|^{\frac{3(q-1)}{3-q}} + |v_L|^{\frac{3(q-1)}{3-q}})$ $\cdot\, D_\sigma v_\mathcal{T}$ (with $\sigma = K|L\in\mathcal{E}_{\rm int}$ or $v_L=0$ if $\sigma\in\mathcal{E}_{\rm ext}\cap\mathcal{E}_K$), we deduce, by the Hölder inequality,

$$\begin{aligned}\left(\int_\Omega |v_\mathcal{T}|^{\frac{3q}{3-q}}\,d\lambda\right)^{\frac{2}{3}} &\le \frac{2q}{3-q}\sum_{\sigma\in\mathcal{E}} \operatorname{meas}(\sigma)d_\sigma \left(|v_K|^{\frac{3(q-1)}{3-q}} + |v_L|^{\frac{3(q-1)}{3-q}}\right)\frac{D_\sigma v_\mathcal{T}}{d_\sigma}\\ &\le \frac{2q}{3-q}\left(\sum_{\sigma\in\mathcal{E}} \operatorname{meas}(\sigma)d_\sigma\left(\frac{D_\sigma v_\mathcal{T}}{d_\sigma}\right)^q\right)^{\frac{1}{q}}\\ &\quad\times\left(\sum_{\sigma\in\mathcal{E}} \operatorname{meas}(\sigma)d_\sigma\left(2^{q'-1}|v_K|^{\frac{3q}{3-q}} + 2^{q'-1}|v_L|^{\frac{3q}{3-q}}\right)\right)^{\frac{1}{q'}}.\end{aligned}$$

However, by the hypothesis on ζ,

$$\begin{aligned}\sum_{\sigma\in\mathcal{E}} \operatorname{meas}(\sigma)d_\sigma|v_K|^{\frac{3q}{3-q}} &= \sum_{K\in\mathcal{T}} |v_K|^{\frac{3q}{3-q}} \sum_{\sigma\in\mathcal{E}_K} \operatorname{meas}(\sigma)d_\sigma\\ &\le \frac{1}{\zeta}\sum_{K\in\mathcal{T}} |v_K|^{\frac{3q}{3-q}} \sum_{\sigma\in\mathcal{E}_K} \operatorname{meas}(\sigma)d_{K,\sigma}\\ &= \frac{3}{\zeta}\sum_{K\in\mathcal{T}} \operatorname{meas}(K)|v_K|^{\frac{3q}{3-q}}\\ &= \frac{3}{\zeta}||v_\mathcal{T}||^{\frac{3q}{3-q}}_{L^{\frac{3q}{3-q}}(\Omega)}.\end{aligned}$$

Thus, we finally have

$$\left(\int_\Omega |v_\mathcal{T}|^{\frac{3q}{3-q}}\,d\lambda\right)^{\frac{2}{3}} \le C||v_\mathcal{T}||_{1,q,\mathcal{M}}||v_\mathcal{T}||^{\frac{3(q-1)}{3-q}}_{L^{\frac{3q}{3-q}}(\Omega)},$$

where C depends only on (q,ζ), and this gives the desired estimate. □

Proof of Proposition 2.3. Define $\chi_\sigma(x,y)$ as at the beginning of the proof of Proposition 2.1.

Suppose first that $q > 1$, and take $h \in \mathbb{R}^d \backslash \{0\}$. Define, for $\sigma \in \mathcal{E}$, $c_\sigma = |\mathbf{n}_\sigma \cdot \frac{h}{|h|}|$ (where $\mathbf{n}_\sigma$ is a unit normal to σ).

We have, for almost every $x \in \Omega$ (in fact, for all x which do not belong to an affine hyperplane generated by some $\sigma \in \mathcal{E}$),

$$|w_\mathcal{T}(x+h) - w_\mathcal{T}(x)| \le \sum_{\sigma \in \mathcal{E}} \chi_\sigma(x+h,x) D_\sigma v_\mathcal{T}. \tag{64}$$

As in the proof of Proposition 2.1, this sum can be limited to those $\sigma \in \mathcal{E}$ such that $c_\sigma \neq 0$, and we have then, by Hölder, for almost every $x \in \Omega$,

$$|w_\mathcal{T}(x+h) - w_\mathcal{T}(x)| \le \left(\sum_{\sigma \in \mathcal{E} \,|\, c_\sigma \neq 0} \frac{\chi_\sigma(x+h,x) d_\sigma}{c_\sigma} \left(\frac{D_\sigma v_\mathcal{T}}{d_\sigma} \right)^q \right)^{\frac{1}{q}} \times \left(\sum_{\sigma \in \mathcal{E}} \chi_\sigma(x+h,x) d_\sigma c_\sigma^{\frac{q'}{q}} \right)^{\frac{1}{q'}}.$$

Since $q \le 2$ (and hence $q'/q \ge 1$) and $c_\sigma \in [0,1]$, we have $c_\sigma^{q'/q} \le c_\sigma$; however (see [14]), $\sum_{\sigma \in \mathcal{E}} \chi_\sigma(x+h,x) d_\sigma c_\sigma \le |h| + C\text{size}(\mathcal{M})$, where C depends only on Ω. Thus,

$$|w_\mathcal{T}(x+h) - w_\mathcal{T}(x)|^q \le (|h| + C\text{size}(\mathcal{M}))^{q-1} \sum_{\sigma \in \mathcal{E} \,|\, c_\sigma \neq 0} \frac{\chi_\sigma(x+h,x) d_\sigma}{c_\sigma} \left(\frac{D_\sigma v_\mathcal{T}}{d_\sigma} \right)^q.$$

Since $\int_{\mathbb{R}^d} \chi_\sigma(x+h,x)\, d\lambda(x) \le \text{meas}(\sigma) c_\sigma |h|$, we deduce, after integrating, the desired estimate (17).

If $q = 1$, we simply integrate (64), and this directly gives (bounding $\int_{\mathbb{R}^d} \chi_\sigma(x+h, x)\, d\lambda(x)$ by $\text{meas}(\sigma)|h|$) the estimate.

The compactness result is then an immediate application of Kolmogorov's theorem, with the use of Proposition 2.1 to obtain a bound in $L^q(\Omega)$. □

Proof of Proposition 2.4. Applying (17) to v_n and passing to the limit $n \to \infty$, we get, for $h \in \mathbb{R}^d \backslash \{0\}$,

$$\int_{\mathbb{R}^d} \frac{|w(x+h) - w(x)|^q}{h^q}\, d\lambda(x) \le C,$$

where C does not depend on h and w is the extension of v to $\mathbb{R}^d$ by 0 outside Ω.

Since $q > 1$, this estimate classically gives $w \in W^{1,q}(\mathbb{R}^d)$, and, by the regularity of Ω, since w is the extension of v by 0 outside Ω, $v \in W_0^{1,q}(\Omega)$. □

REFERENCES

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

[2] L. Boccardo and T. Gallouët, *Nonlinear elliptic and parabolic equations involving measure data*, J. Funct. Anal., 87 (1989), pp. 241–273.

[3] L. Boccardo, T. Gallouët, and J.-L. Vazquez, *Nonlinear elliptic equations in $\mathbb{R}^N$ without growth restrictions on the data*, J. Differential Equations, 105 (1993), pp. 334–363.

[4] S. Boivin, F. Cayré, and J.M. Hérard, *A finite volume method to solve the Navier–Stokes equations for incompressible flows on unstructured meshes*, Int. J. Therm. Sci., 39 (2000), pp. 806–825.

[5] S.H. Chou and P.S. Vassilevski, *A general mixed covolume framework for constructing conservative schemes for elliptic problems*, Math. Comp., 68 (1999), pp. 991–1011.

[6] Y. Coudière, T. Gallouët, and R. Herbin, *Discrete Sobolev inequalities and L^p error estimates for approximate finite volume solutions of convection diffusion equations*, M2AN Math. Model. Numer. Anal., 35 (2001), pp. 767–778.

[7] R. Eymard, T. Gallouët, and R. Herbin, *Finite volume approximation of elliptic problems and convergence of an approximate gradient*, Appl. Numer. Math., 37 (2001), pp. 31–53.

[8] Y. Coudière, J.P. Vila, and P. Villedieu, *Convergence rate of a finite volume scheme for a two dimensional convection diffusion problem*, M2AN Math. Model. Numer. Anal., 33 (1999), pp. 493–516.

[9] J. Droniou, *Noncoercive linear elliptic problems*, Potential Anal., to appear.

[10] J. Droniou, *Etude de Certaines Equations aux Dérivées Partielles*, Ph.D. thesis, Centre de Mathématiques et Informatique, Université de Provence, 2001.

[11] J. Droniou and T. Gallouët, *A finite volume scheme for noncoercive Dirichlet problems with right-hand side in H^{-1}*, in Finite Volume for Complex Applications III, R. Herbin and D. Kröner, eds., Hermes Penton Science, London, 2002, pp. 195–202.

[12] J. Droniou and T. Gallouët, *Finite volume methods for convection-diffusion equations with right-hand side in H^{-1}*, M2AN Math. Model. Numer. Anal., 36 (2002), pp. 705–724.

[13] J. Droniou and T. Gallouët, *A uniqueness result for quasilinear elliptic equations with measures as data*, Rend. Mat. Appl. (7), 21 (2001), pp. 57–86.

[14] R. Eymard, T. Gallouët, and R. Herbin, *Finite volume methods*, in Solutions of Equations in R_n. Part 3, Techniques of Scientific Computing, Handb. Numer. Anal. 7, P.G. Ciarlet and J.L. Lions, eds., North–Holland, Amsterdam, pp. 713–1020.

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

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

[17] J.M. Fiard and R. Herbin, *Comparison between finite volume finite element methods for the numerical simulation of an elliptic problem arising in electrochemical engineering*, Comput. Methods Appl. Mech. Engrg., 115 (1994), pp. 315–338.

[18] P.A. Forsyth and P.H. Sammon, *Quadratic convergence for cell-centered grids*, Appl. Numer. Math., 4 (1988), pp. 377–394.

[19] T. Gallouët and R. Herbin, *Finite volume methods for diffusion problems and irregular data*, in Finite Volumes for Complex Applications, Problems and Perspectives, II, F. Benkhaldoun, M. Hänel, and R. Vilsmeier, eds., Hermes, Paris, 1999, pp. 155–162.

[20] T. Gallouët, R. Herbin, and M.H. Vignal, *Error estimates for the approximate finite volume solution of convection diffusion equations with general boundary conditions*, SIAM J. Numer. Anal., 37 (2000), pp. 1935–1972.

[21] T. Gallouët and A. Monier, *On the regularity of solutions to elliptic equations*, Rend. Mat. Appl. (7), 19 (1999), pp. 471–488.

[22] P. Grisvard, *Elliptic Problems in Nonsmooth Domains*, Pitman, Boston, 1985.

[23] R. Herbin, *An error estimate for a finite volume scheme for a diffusion-convection problem on a triangular mesh*, Numer. Methods Partial Differential Equations, 11 (1995), pp. 165–173.

[24] R. Herbin, *Finite volume approximation of elliptic problems with irregular data*, in Finite Volumes for Complex Applications, Problems and Perspectives, F. Benkhaldoun, D. Hanel, and R. Vilsmeier, eds., Hermes, Paris, 1999, pp. 153–160.

[25] T. Gallouët and R. Herbin, *Finite volume methods for diffusion convection equations on general meshes*, in Finite Volumes for Complex Applications, Problems and Perspectives, F. Benkhaldoun and R. Vilsmeier, eds., Hermes, Paris, 1996, pp. 153–160.

[26] R.D. Lazarov, I.D. Mishev, and P.S. Vassilevski, *Finite volume methods for convection-diffusion problems*, SIAM J. Numer. Anal., 33 (1996), pp. 31–55.

[27] T.A. Manteuffel and A.B. White, *The numerical solution of second-order boundary value problems on nonuniform meshes*, Math. Comp., 47 (1986), pp. 511–535.

[28] N.G. Meyers, *An L^p estimate for the gradient of solutions of second order divergence equations*, Ann. Scuola Norm. Sup. Pisa, 17 (1963), pp. 189–206.

[29] I.D. Mishev, *Finite volume methods on Voronoï meshes*, Numer. Methods Partial Differential Equations, 14 (1998), pp. 193–212.

SIAM J. Numer. Anal.
Vol. 41, No. 6, pp. 2032–2055

A POSTERIORI ERROR ESTIMATORS FOR REGULARIZED TOTAL VARIATION OF CHARACTERISTIC FUNCTIONS[*]

FRANCESCA FIERRO[†] and ANDREAS VEESER[†]

Abstract. We consider a nonuniformly elliptic double obstacle problem arising from "convexifying" and regularizing a minimization of a functional with total variation in the set of characteristic functions. We derive a posteriori estimators for the discretization error with linear finite elements, which are uniform in the regularization, incorporate computable and local information on the conditioning, vanish in the intersection of discrete and exact contact sets, and are not affected by possible nonuniqueness. Moreover, we integrate these estimators in an adaptive algorithm and illustrate their properties by various numerical experiments.

Key words. a posteriori error estimates, adaptive finite element methods, total variation, mean curvature, obstacle problems

AMS subject classifications. Primary, 65N30, 65N15; Secondary, 35J85, 65K10

DOI. 10.1137/S0036142902408283

1. Motivation and introduction.

The minimization of the functional

$$\mathcal{I}(v) := \int_\Omega |Dv| - \int_\Omega \kappa v \quad \text{in} \quad \mathrm{BV}\big(\Omega; \{\pm 1\}\big) \tag{1.1}$$

constitutes a weak formulation of the prescribed mean curvature problem: the hyper surface $\partial\{\chi = 1\}$ of any minimum point χ has mean curvature $\kappa/(2d)$ and normal contact with $\partial\Omega$ in a weak sense; cf. Finn [12], and see section 2 for the notation. Such minimization has applications in capillary surfaces, time discretizations of mean curvature flow, and phase transition models; see, e.g., [2, 12, 13, 20].

To approximate minimum points of (1.1) numerically, Bellettini, Paolini, and Verdi [5, 6] observed that $\mathcal{I}$ can be "equivalently" minimized in the convex set $\mathrm{BV}\big(\Omega, [-1, 1]\big)$ and considered the additionally regularized minimization of

$$\mathcal{I}(v; \epsilon) := \int_\Omega \sqrt{\epsilon^2 + |Dv|^2} - \int_\Omega \kappa v \quad \text{in} \quad \mathrm{BV}\big(\Omega; [-1, 1]\big) \tag{1.2}$$

with parameter $\epsilon > 0$. Discretizing (1.2) with continuous linear finite elements offers the minimization of "semistrictly" convex functionals which Γ-converge in $L_1(\Omega)$ to $\mathcal{I}$ as ϵ and the mesh size decrease to 0 independently. Notice that the case $\epsilon = 1$ of (1.2) is also interesting in itself.

The *nonuniform* convexity of $\mathcal{I}(\cdot, \epsilon)$ and the presence of obstacles in (1.2) strongly suggest the use of *adaptive* methods; see also [10]. In this connection, it also seems worthwhile to consider a space-dependent regularization parameter ϵ. This paper concerns the adaptive computation of approximations to regular minimum points of (1.2), where ϵ is a strictly positive and piecewise constant function. In its theoretical

[*] Received by the editors May 24, 2002; accepted for publication February 5, 2003; published electronically November 25, 2003. This research was partially supported by the TMR network "Viscosity solutions and their applications," the CNR contract CU99.01713.CT01, Italian M.I.U.R. Cofin2000 "Calcolo Scientifico: Modelli e Metodi Numerici Innovativi," and Cofin2001 "Metodi numerici avanzati per equazioni alle derivate parziali di interesse applicativo."

http://www.siam.org/journals/sinum/41-6/40828.html

[†] Dipartimento di Matematica, Università degli Studi di Milano, Via C. Saldini 50, 20133 Milano, Italy (fierro@mat.unimi.it, veeser@mat.unimi.it).

part we derive a posteriori error estimators for the discretization of (1.2), which are used to guide adaptive algorithms in its computational part.

We outline the results and organization of this paper in more detail. In section 2 we introduce the notation used as well as regular and discrete minimum points of (1.2) together with their variational inequalities. We observe that both types of minimum points are not unique in general and recall a characterization of their uniqueness. Furthermore, similar to [15, 17, 18], we associate with the minimum points auxiliary functionals, which recover information lost by the variational inequalities. The "discrete auxiliary functional," i.e., the one associated with discrete minimum points, may however differ from those in [15, 17, 18]. This difference allows for a complete and thus better localization; see Remark 4.4.

In section 3 we introduce an error notion. It measures a "distance" between the classes of discrete and regular minimum points including an error in the auxiliary functional. In addition, it bounds the error in the approximation of the minimum value in (1.2) from above; see (3.10). The latter means that the error notion is useful when (1.2) is interpreted as an approximation of (1.1). Moreover, we relate the introduced error to an adaptation of the Galerkin functional in [17]; this relationship will guide us in deriving upper and lower a posteriori bounds.

In section 4 we derive a conditional a posteriori upper bound for the aforementioned error. Here "conditional" means that the upper bound holds under the assumption of an a posteriori condition, i.e., a condition involving only the computed minimum point and (discretization) data. The upper bound enjoys the following properties:

- Its indicators do not depend on the particular choice of the discrete minimum point if the latter is not unique.
- It is uniform in the regularization "parameter" ϵ.
- Its local indicators inside the discrete contact set are 0 whenever the discrete minimum point is unique and κ satisfies a local sign condition; the latter holds particularly in the exact contact set.
- Only weighted jump residuals, weighted data oscillation (which is of higher order), and indicators controlling a consistency error of the discrete auxiliary functional are involved; interior residuals do not appear.

Apart from the aforementioned ingredients, the upper bound's proof utilizes (and generalizes) ideas of [11, section 5] in regard to the handling of the underlying operator.

In section 5 we derive a posteriori local lower bounds complementing the upper bound. The lower bound involving the jump residual exhibits a gap with respect to the upper bound; see Remark 5.1. This gap is related to the quotient of the extreme eigenvalues of the underlying operator. Its avoidance probably requires information on the direction of the error. Important ingredients for the proofs of the lower bounds are section 3, Verfürth's constructive argument, and, for the lower bound related to the consistency error of the discrete auxiliary functional, an adaptation of [15, Lemma 6.4] whose proof is inspired by Lemma 3.3 of Chen and Nochetto [8].

In section 6 we formulate an adaptive algorithm guided by the derived estimators and present several numerical experiments. Studying situations with singular or nonunique solutions as well as "dynamic" regularization, we illustrate the properties of the derived estimators.

2. Regular and discrete minimum points.

After fixing some notation and the setting, we introduce regular and discrete minimum points of (1.2) as well as associated functionals. Furthermore, we discuss some of their properties.

The following notation is used throughout this article (and, partially, was already used in section 1). For $q \in [1, \infty]$ and an open set $U \subset \mathbb{R}^d$, the space of Lebesgue-measurable and q-integrable functions is denoted by $L_q(U)$. We shall write $\|\cdot\|_{0,q;U}$ for its norm. For any set $A \subset \mathbb{R}$, we define $L_q(U; A) := \{v \in L_q(U) \mid v(x) \in A$ for a.e. $x \in U\}$; this notation is used also for analogous subsets of other Banach spaces. $W^1_q(U)$ is the Sobolev space of $L_q(U)$-functions that have first weak derivatives in $L_q(U)$. For $\epsilon \in L_1(U; \mathbb{R}^+)$ and $v \in L_1(U)$, we set

$$\int_U \sqrt{\epsilon^2 + |Dv|^2} := \sup \left\{ \int_U \left(\epsilon g_{d+1} + v \sum_{i=1}^{d} \partial_i g_i \right) \;\middle|\; \begin{array}{l} g = (g_1, \dots, g_{d+1}) \text{ in} \\ C^1_0(U; \mathbb{R}^{d+1}) \text{ and } |g| \le 1 \end{array} \right\}. \tag{2.1}$$

Note that (2.1) is the total variation $\int_\Omega |Dv|$ of v if $\epsilon = 0$ in U. The subset of $L_1(U)$ of the functions with bounded total variation is denoted by $\mathrm{BV}(U)$.

Let $\mathcal{T}_h$ be a (conforming) triangulation of a bounded, connected, open set Ω in $\mathbb{R}^d$, $d \ge 2$, with Lipschitz boundary $\partial\Omega$. Moreover, let $\kappa \in L_\infty(\Omega)$ and $\epsilon \in L_\infty(\Omega; \mathbb{R}^+)$ be constant on each $T \in \mathcal{T}_h$. Finally, let u be a *regular minimum point*; i.e., u is a minimum point of (1.2) and $u \in W^1_1(\Omega)$. Denoting the gradient of $v \in W^1_1(\Omega)$ by ∇v, we have $\int_\Omega \sqrt{\epsilon^2 + |Dv|^2} = \int_\Omega \sqrt{\epsilon^2 + |\nabla v|^2}$ and u satisfies the variational inequality

$$\forall\, v \in W^1_1\big(\Omega; [-1, 1]\big), \quad \int_\Omega a(\nabla u; \epsilon) \cdot \nabla(u - v) \le \int_\Omega \kappa(u - v), \tag{2.2}$$

where $a(p; r) := p/\sqrt{r^2 + |p|^2}$ for all $p \in \mathbb{R}^d$ and $r > 0$. Note that u is not necessarily the only regular minimum point. However, if $\tilde{u}$ is also a regular minimum point, then

$$\exists\, c \in \mathbb{R}, \quad \tilde{u} = u + c \text{ in } \Omega, \tag{2.3a}$$

$$\left[\int_\Omega \kappa \ne 0 \text{ or } \left(\sup_\Omega u = 1 \text{ and } \inf_\Omega u = -1 \right) \right] \implies c = 0. \tag{2.3b}$$

If u is the only regular minimum point, then the condition in (2.3b) is implied.

We associate the functional $\sigma_\epsilon \in W^1_1(\Omega)^*$ defined by

$$\forall\, \varphi \in W^1_1(\Omega), \quad \langle \sigma_\epsilon, \varphi \rangle := \int_\Omega \kappa\varphi - \int_\Omega a(\nabla u; \epsilon) \cdot \nabla\varphi \tag{2.4}$$

with the class of regular minimum points. In fact, σ_ϵ does not depend on the particular choice of u thanks to (2.3a). Important properties of σ_ϵ are (they express that σ_ϵ is a subgradient in u of the convex potential associated with the constraint $|u| \le 1$)

$$\begin{gathered} \sigma_\epsilon = \kappa \le 0 \text{ in } \operatorname{int}\{u = -1\}, \qquad \sigma_\epsilon = \kappa \ge 0 \text{ in } \operatorname{int}\{u = 1\}, \\ \sigma_\epsilon = 0 \text{ in } \{-1 < u < 1\}. \end{gathered} \tag{2.5}$$

Here $\operatorname{int}\{u = -1\}$ $\big($or $\operatorname{int}\{u = 1\}\big)$ stands for the biggest open subset of Ω on which u is equal to -1 (or 1) in a distributional sense, while $\{-1 < u < 1\} := \bigcup_{\delta > 0} \operatorname{int}\{-1 + \delta \le u \le 1 - \delta\}$, where $\operatorname{int}\{-1 + \delta \le u \le 1 - \delta\}$ is the biggest open subset of Ω on which the given inequalities hold in a distributional sense.

In what follows, we shall use the letter h as lower index to indicate a "discrete object" or, often together with ϵ, to indicate the dependence on the triangulation $\mathcal{T}_h$. Let W_h be the space of continuous affine finite elements over $\mathcal{T}_h$, i.e.,

$$W_h := \big\{ w_h \in C^0(\overline{\Omega}) \mid \forall\, T \in \mathcal{T}_h \; w_h|_T \in P_1(T) \big\}.$$

Hereafter, $P_m(T)$ denotes the space of polynomials on T with degree smaller than or equal to $m \in \mathbb{N}_0$. Moreover, we set $\mathcal{K}_h := \{w_h \in W_h \mid |w_h| \le 1\}$, which is a subset of $W^1_1\big(\Omega, [-1,1]\big)$. An algorithm for the minimization of the functional $\mathcal{I}(\cdot,\epsilon)$ in $\mathcal{K}_h$ is described in [6]. Let u_h be a minimum point of $\mathcal{I}(\cdot,\epsilon)$ in $\mathcal{K}_h$, i.e., a *discrete minimum point.* Similarly to a regular minimum point, the discrete minimum point u_h satisfies the discrete variational inequality

$$\forall\, v_h \in \mathcal{K}_h, \quad \int_\Omega a(\nabla u_h;\epsilon)\cdot\nabla(u_h - v_h) \le \int_\Omega \kappa(u_h - v_h) \tag{2.6}$$

and may not be the only discrete minimum point; there holds

$$\exists\, c \in \mathbb{R}, \quad \tilde{u}_h = u_h + c \text{ in } \Omega, \tag{2.7a}$$

$$\left[\int_\Omega \kappa \ne 0 \text{ or } \left(\sup_\Omega u_h = 1 \text{ and } \inf_\Omega u_h = -1\right)\right] \implies c = 0 \tag{2.7b}$$

if $\tilde{u}_h$ is also a discrete minimum point.

Next, we define a counterpart $\sigma_{\epsilon h}$ of the functional σ_ϵ in (2.4). To this end, we denote the set of nodes of W_h (or vertices of simplices in $\mathcal{T}_h$) by $\mathcal{N}_h$ and recall that the hat functions $(\phi_z)_{z\in\mathcal{N}_h}$ defined by $\phi_z \in W_h$, $\phi_z(y) = 0$ for $y \in \mathcal{N}_h \setminus \{z\}$, and $\phi_z(z) = 1$ constitute a basis of W_h. The properties of σ_ϵ in (2.5) are imitated on the discrete level by

$$\begin{gathered} \Sigma_z \le 0 \text{ if } u_h(z) = -1, \quad \Sigma_z \ge 0 \text{ if } u_h(z) = 1, \\ \Sigma_z = 0 \text{ if } -1 < u_h(z) < 1, \end{gathered} \tag{2.8}$$

where

$$\Sigma_z := \left[\int_\Omega \phi_z\right]^{-1}\left[\int_\Omega \kappa\phi_z - \int_\Omega a(\nabla u_h;\epsilon)\cdot\nabla\phi_z\right]. \tag{2.9}$$

Note that we suppress the dependence of Σ_z on $\mathcal{T}_h$ and ϵ in the notation; this convention will be used also for other computable quantities in the following.

In [15, 17, 18] mass lumping was used to extend (2.8) in the spirit of (2.5). Here, we shall use another approach which exploits that the hat functions $(\phi_z)_{z\in\mathcal{N}_h}$ constitute a partition of unity on Ω. More precisely, let $\omega_z := \operatorname{supp}\phi_z$ be the support of ϕ_z for any $z \in \mathcal{N}_h$ and define a function $\sigma_{\epsilon h} \in L_\infty(\Omega)$ by

$$\sigma_{\epsilon h} = \sum_{z\in\mathcal{N}_h} \sigma_{\epsilon h,z}\phi_z \quad \text{with} \quad \sigma_{\epsilon h,z}(x) := \begin{cases} \kappa(x) & \text{if } z \in \mathcal{C}_{\epsilon h}, \\ \Sigma_z & \text{if } z \in \mathcal{N}_h \setminus \mathcal{C}_{\epsilon h}, \end{cases} \quad x \in \omega_z, \tag{2.10a}$$

where the set $\mathcal{C}_{\epsilon h}$ of "full-contact nodes" is defined as follows:

$$\begin{aligned} &\text{if the condition in (2.7b) holds, then} \\ &\quad \mathcal{C}_{\epsilon h} := \big\{z \in \mathcal{N}_h \mid (u_h = -1,\ \kappa \le 0 \text{ on } \omega_z) \text{ or } (u_h = 1,\ \kappa \ge 0 \text{ on } \omega_z)\big\}; \\ &\text{else } \mathcal{C}_{\epsilon h} := \emptyset. \end{aligned} \tag{2.10b}$$

Note that $\sigma_{\epsilon h}$ depends only on the class of discrete minimum points and not on the particular choice u_h. In general, $\sigma_{\epsilon h}$ is not a subgradient in u_h of the convex potential associated with the constraint $|u_h| \le 1$. However (we use the convention that simplices are closed),

$$\sigma_{\epsilon h} = 0 \text{ in } \bigcup\big\{T \in \mathcal{T}_h \,:\, u_h(T) \subset\,]{-1}, 1[\big\} \tag{2.11a}$$

and, provided that the condition in (2.7b) holds,

$$\sigma_{\epsilon h} = \kappa \leq 0 \text{ in } \bigcup \big\{ T \in \mathcal{T}_h \,:\, u_h = -1,\, \kappa \leq 0 \text{ in } \omega_T \big\}, \tag{2.11b}$$

$$\sigma_{\epsilon h} = \kappa \geq 0 \text{ in } \bigcup \big\{ T \in \mathcal{T}_h \,:\, u_h = 1,\, \kappa \geq 0 \text{ in } \omega_T \big\}, \tag{2.11c}$$

where ω_T denotes the union of all simplices touching $T \in \mathcal{T}_h$.

The sign conditions on κ in (2.10b) might appear artificial at first glance. They are crucial, however, because (2.8) provides less information than (2.5). In fact, $u_h = -1$ (or $u_h = 1$) in ω_z determines only the sign of $\int_\Omega \kappa \phi_z$ but not the sign of κ in ω_z. See also Remark 4.5 below.

3. Error and the Galerkin functional.

We introduce a functional $\mathcal{G}_{\epsilon h}$ that, for our purposes, plays the same role in the context of the variational inequality (2.2) as the residual in the context of unconstrained problems. Motivated by properties of $\mathcal{G}_{\epsilon h}$, we introduce an error notion for couples $(u_h, \sigma_{\epsilon h})$ of discrete minimum points u_h and associated functionals $\sigma_{\epsilon h}$. This error notion does not depend on the particular choice of the discrete minimum point u_h and bounds the error in the approximate minimum value $\mathcal{I}(u_h)$.

Let the functional $\mathcal{G}_{\epsilon h} \in W^1_1(\Omega)^*$ be defined by

$$\langle \mathcal{G}_{\epsilon h}, \varphi \rangle := \int_\Omega a(\nabla u_h; \epsilon) \cdot \nabla \varphi + \int_\Omega \sigma_{\epsilon h} \varphi - \int_\Omega \kappa \varphi, \tag{3.1}$$

where u_h is a discrete minimum point and $\sigma_{\epsilon h}$ is defined as in (2.10). Note that $\mathcal{G}_{\epsilon h}$ does not depend on the particular choice of u_h. Definition (3.1) adapts the Galerkin functional of [17] to the variational inequality (2.2) and the definition (2.10). In particular, it generalizes the residual $\mathcal{R}_{\epsilon h} := -\operatorname{div} a(\nabla u_h; \epsilon) - \kappa \in W^1_1(\Omega)^*$: if u_h touches neither the lower nor the upper obstacle, then $\mathcal{G}_{\epsilon h} = \mathcal{R}_{\epsilon h}$. The residual $\mathcal{R}_{\epsilon h}$ is a key quantity for the derivation of a posteriori estimators; see the books [1, 19] and, in particular, [11], where the unconstrained case with $\epsilon \equiv 1$ and Dirichlet boundary values is analyzed.

We observe that, for any regular minimum point u, definition (2.4) yields

$$\langle \mathcal{G}_{\epsilon h}, \varphi \rangle = \int_\Omega \big[a(\nabla u_h; \epsilon) - a(\nabla u; \epsilon) \big] \cdot \nabla \varphi + \langle \sigma_{\epsilon h} - \sigma_\epsilon, \varphi \rangle \tag{3.2}$$

for all $\varphi \in W^1_1(\Omega)$. The discretization error introduced below will consist of two parts. The first part is associated with the first term in the right-hand side of (3.2) and can be considered also in the unconstrained case, while the second part is associated with the second term and is only relevant if u or u_h touches at least one obstacle.

The following immediate generalization of Lemma 3.1 in [11] concerns the first term in the right-hand side of (3.2).

LEMMA 3.1 (monotonicity of a). *Let $p_1, p_2 \in \mathbb{R}^d$ and $r > 0$. We have*

$$\big[a(p_1; r) - a(p_2; r) \big] \cdot (p_1 - p_2) \;=\; \left| \frac{P_1}{|P_1|} - \frac{P_2}{|P_2|} \right|^2 \frac{|P_1| + |P_2|}{2} \tag{3.3}$$

with $P_i := (p_i, -r) \in \mathbb{R}^{d+1}$ for $i = 1, 2$.

Proof. We calculate (note that all dots after the first equal sign denote the scalar product in $\mathbb{R}^{d+1}$)

$$\begin{aligned}\big[a(p_1;r)-a(p_2;r)\big]\cdot(p_1-p_2) &= \left(\frac{P_1}{|P_1|}-\frac{P_2}{|P_2|}\right)\cdot(P_1-P_2)\\ &= \left(\frac{P_1}{|P_1|}-\frac{P_2}{|P_2|}\right)\cdot\frac{P_1}{|P_1|}\,|P_1| + \left(\frac{P_2}{|P_2|}-\frac{P_1}{|P_1|}\right)\cdot\frac{P_2}{|P_2|}\,|P_2|.\end{aligned}$$

Combining this with the identity $(N_1-N_2)\cdot N_1 = \frac{1}{2}|N_1-N_2|^2$ for $|N_1|=|N_2|=1$, we arrive at (3.3). □

Lemma 3.1 and the geometrical interpretation for $r=1$ of the right-hand side in (3.3) motivate the following definitions (see also [11, section 3]):

$$A(p;r) := \sqrt{|p|^2+|r|^2}, \quad N(p;r) := \frac{(p,-r)}{A(p;r)}$$

for all $p\in\mathbb{R}^d$ and $r>0$ and

$$e_{\epsilon h} := \left(\int_\Omega |N(\nabla u_h;\epsilon)-N(\nabla u;\epsilon)|^2\,\frac{A(\nabla u_h;\epsilon)+A(\nabla u;\epsilon)}{2}\right)^{1/2}. \tag{3.4}$$

Note that $e_{\epsilon h}$ does not depend on the particular choices of u_h and u and so measures an error between the classes of discrete and regular minimum points.

Testing (3.2) with $\varphi = u_h - u$ yields

$$e_{\epsilon h}^2 = \langle \mathcal{G}_{\epsilon h}, u_h-u\rangle - \langle \sigma_{\epsilon h}-\sigma_\epsilon, u_h-u\rangle. \tag{3.5}$$

This identity suggests estimating $e_{\epsilon h}$ by bounding appropriately $\langle \mathcal{G}_{\epsilon h}, u_h-u\rangle$ from above and $\langle \sigma_{\epsilon h}-\sigma_\epsilon, u_h-u\rangle$ from below. Both bounds are established in section 4 with the help of local computable quantities. The upper bound for $\langle \mathcal{G}_{\epsilon h}, u_h-u\rangle$ is established similarly to the one for $\langle \mathcal{R}_{\epsilon h}, u_h-u\rangle$ in the unconstrained case. The lower bound for $\langle \sigma_{\epsilon h}-\sigma_\epsilon, u_h-u\rangle = \langle \sigma_{\epsilon h}, u_h-u\rangle + \langle \sigma_\epsilon, u-u_h\rangle$ exploits that $\sigma_{\epsilon h}$ is an "approximate subgradient" in u_h; cf. (2.11).

We next discuss the error part associated with the second term in the right-hand side of (3.2). To this end, we observe $|a(\nabla u_h;\epsilon)-a(\nabla u;\epsilon)| \le |N(\nabla u_h;\epsilon)-N(\nabla u;\epsilon)|$ and introduce the seminorm

$$|\varphi|_{\epsilon h} := \left(\int_\Omega |\nabla\varphi|^2 A(\nabla u_h;\epsilon)^{-1}\right)^{1/2}$$

with the computable weight $A(\nabla u_h;\epsilon)^{-1}$. We thus obtain

$$\left|\int_\Omega \big[a(\nabla u_h;\epsilon)-a(\nabla u;\epsilon)\big]\cdot\nabla\varphi\right| \le \sqrt{2}e_{\epsilon h}|\varphi|_{\epsilon h}. \tag{3.6}$$

This suggests measuring the error of $\sigma_{\epsilon h}$ in the dual seminorm of $|\cdot|_{\epsilon h}$, that is,

$$|\Psi|_{\epsilon h;*} := \sup\big\{\langle\Psi,\varphi\rangle \mid \varphi\in W_2^1(\Omega),\ |\varphi|_{\epsilon h}\le 1\big\}. \tag{3.7}$$

In fact, (3.2) then implies

$$|\sigma_{\epsilon h}-\sigma_\epsilon|_{\epsilon h;*} \le |\mathcal{G}_{\epsilon h}|_{\epsilon h;*} + \sqrt{2}e_{\epsilon h}. \tag{3.8}$$

As $e_{\epsilon h}$ before, $|\sigma_{\epsilon h}-\sigma_\epsilon|_{\epsilon h;*}$ does not depend on the particular choices of u_h and u. The upper bound for $|\mathcal{G}_{\epsilon h}|_{\epsilon h;*}$ is established in section 4. It does not lead to additional computable quantities; see (4.19) in the proof of Theorem 4.6. Thus the upper bound of $e_{\epsilon h}$ essentially implies the one of $|\sigma_{\epsilon h}-\sigma_\epsilon|_{\epsilon h;*}$ and the computable quantities bounding $e_{\epsilon h}$ even estimate the combined error $e_{\epsilon h}+|\sigma_{\epsilon h}-\sigma_\epsilon|_{\epsilon h;*}$ from above.

Local estimates in the inverse direction will rely on

$$|\mathcal{G}_{\epsilon h}|_{\epsilon h,\omega;*}\le\sqrt{2}e_{\epsilon h}(\omega)+|\sigma_{\epsilon h}-\sigma_\epsilon|_{\epsilon h,\omega;*} \tag{3.9}$$

for any open set ω in Ω, which is again a consequence of (3.2) and (3.6). Here, $|\cdot|_{\epsilon h,\omega;*}$ and $e_{\epsilon h}(\omega)$ are the local counterparts of $|\cdot|_{\epsilon h;*}$ and $e_{\epsilon h}$; that is,

$$\begin{aligned}|\Psi|_{\epsilon h,\omega;*}&:=\sup\bigl\{\langle\Psi,\varphi\rangle\mid\varphi\in W_2^1(\Omega),\ \operatorname{supp}\varphi\subset\overline{\omega},\ |\varphi|_{\epsilon h}\le 1\bigr\},\\ e_{\epsilon h}(\omega)&:=\left(\int_\omega|N(\nabla u_h;\epsilon)-N(\nabla u;\epsilon)|^2\,\frac{A(\nabla u_h;\epsilon)+A(\nabla u;\epsilon)}{2}\right)^{1/2}.\end{aligned}$$

We conclude this section by showing that $e_{\epsilon h}$ is closely related to the error in the minimum value, i.e., to the difference $\mathcal{I}(u_h;\epsilon)-\mathcal{I}(u;\epsilon)$. On the one hand, we have $\mathcal{I}(u_h;\epsilon)-\mathcal{I}(u;\epsilon)\ge 0$, and on the other hand,

$$\begin{aligned}\mathcal{I}(u;\epsilon)-\mathcal{I}(u_h;\epsilon)&\ge\int_\Omega a(\nabla u_h;\epsilon)\cdot\nabla(u-u_h)-\kappa(u-u_h)\\&\ge\int_\Omega\bigl[a(\nabla u_h;\epsilon)-a(\nabla u;\epsilon)\bigr]\cdot\nabla(u-u_h)=-e_{\epsilon h}^2\end{aligned}$$

by the convexity of $A(\cdot;r)$, $\nabla A(\cdot;r)=a(\cdot;r)$, the variational inequality (2.2), and Lemma 3.1. Combining the two inequalities yields

$$0\le\mathcal{I}(u_h;\epsilon)-\mathcal{I}(u;\epsilon)\le e_{\epsilon h}^2. \tag{3.10}$$

4. Upper bound. We derive an a posteriori upper bound for the error $e_{\epsilon h}+|\sigma_{\epsilon h}-\sigma_\epsilon|_{\epsilon h;*}$ introduced in section 3. To this end, we follow the discussion therein: we first establish appropriate bounds involving computable quantities for

$$|\mathcal{G}_{\epsilon h}|_{\epsilon h;*},\ \langle\mathcal{G}_{\epsilon h},u_h-u\rangle,\quad\text{and }\langle\sigma_{\epsilon h}-\sigma_\epsilon,u_h-u\rangle \tag{4.1}$$

and then combine these bounds with (3.5) and (3.8).

The first two terms in (4.1) are estimated with the help of the following lemma, which involves the computable quantities $(\eta_z)_{z\in\mathcal{N}_h\setminus\mathcal{C}_{\epsilon h}}$ defined by

$$\eta_z^2=h_z\|J_{\epsilon h}\|_{0,2;\gamma_z}^2+h_z^d\|\kappa-\kappa_z\|_{0,d;\omega_z}^2. \tag{4.2}$$

Here, u_h satisfies (2.6), the index set $\mathcal{C}_{\epsilon h}$ is defined as in (2.10b), ω_z denotes the support of the hat function ϕ_z, h_z is the diameter of ω_z, $\kappa_z:=\int_\Omega\kappa\phi_z/\int_\Omega\phi_z$ is the mean value of κ with respect to the weight ϕ_z, γ_z is the union of all sides in ω_z, $\|\cdot\|_{0,2;\gamma_z}$ denotes the L_2-norm with respect to the $(d-1)$-dimensional Hausdorff measure restricted to γ_z, and $J_{\epsilon h}:=J(u_h;\epsilon)$ is defined as follows: for any side S between simplices T_1 and T_2,

$$J(u_h;\epsilon)|_S:=\bigl[a(\nabla u_h|_{T_1};\epsilon)-a(\nabla u_h|_{T_2};\epsilon)\bigr]\cdot n, \tag{4.3a}$$

where n is the normal of S pointing from T_2 to T_1 (note that this definition does not depend on the choice of T_1 and T_2); for any boundary side $S \subset \partial\Omega$,

$$J(u_h;\epsilon)|_S := -a(\nabla u_h|_T;\epsilon)\cdot n, \tag{4.3b}$$

where T is the simplex containing S, and n is the normal of S pointing outward of T. Note that $h_z^d\|\kappa-\kappa_z\|^2_{0,d;\omega_z}$ "scales" in h_z like $h_z^2\|\kappa-\kappa_z\|^2_{0,2;\omega_z}$; cf. [11, Remark 5.1].

In what follows, we shall use "$\preccurlyeq$" instead of "$\leq C$," where C may depend on d and the *shape-regularity* γ_h of $\mathcal{T}_h$ defined by

$$\gamma_h := \max_{T\in\mathcal{T}_h} h_T/\rho_T \in \,]1,\infty[,$$

where h_T denotes the diameter of the smallest ball containing T and ρ_T the diameter of the biggest ball contained in T.

LEMMA 4.1. *Let $\mathcal{G}_{\epsilon h}$ be defined as in* (3.1) *and $\varphi \in W_1^1(\Omega)$. Then*

$$\langle \mathcal{G}_{\epsilon h},\varphi\rangle \preccurlyeq \sum_{z\in\mathcal{N}_h\setminus\mathcal{C}_{\epsilon h}} h_z^{-d/2}\,\eta_z\,\|\nabla\varphi\|_{0,1;\omega_z}.$$

Proof. Simplexwise integration by parts, $\sum_{z\in\mathcal{N}_h}\phi_z = 1$ in Ω, and (2.10a) yield

$$\langle \mathcal{G}_{\epsilon h},\varphi\rangle = -\sum_{z\in\mathcal{N}_h}\left[\int_{\gamma_z} J_{\epsilon h}\varphi\phi_z + \int_{\omega_z}\kappa\varphi\phi_z - \int_{\omega_z}\sigma_{\epsilon h,z}\varphi\phi_z\right]. \tag{4.4}$$

Suppose that $z\in\mathcal{C}_{\epsilon h}$. Then $J_{\epsilon h}=0$ on γ_z and $\sigma_{\epsilon h,z}=\kappa$, and we obtain

$$\int_{\gamma_z} J_{\epsilon h}\varphi\phi_z + \int_{\omega_z}\kappa\varphi\phi_z - \int_{\omega_z}\sigma_{\epsilon h,z}\varphi\phi_z = 0.$$

Consequently, only nodes in $\mathcal{N}_h\setminus\mathcal{C}_{\epsilon h}$ contribute to the sum on the right-hand side in (4.4). Let $z\in\mathcal{N}_h\setminus\mathcal{C}_{\epsilon h}$ be such a node. Then $\sigma_{\epsilon h,z}=\Sigma_z$ is a constant, and the definition (2.9) of Σ_z implies $\int_{\omega_z}\sigma_{\epsilon h,z}\varphi\phi_z = \left(\int_{\gamma_z}J_{\epsilon h}\phi_z + \int_{\omega_z}\kappa\phi_z\right)\varphi_z$, with the mean value $\varphi_z := \int_\Omega\varphi\phi_z/\int_\Omega\phi_z$. Using in addition $\int_{\omega_z}(\varphi-\varphi_z)\phi_z = 0$, we arrive at

$$\int_{\gamma_z} J_{\epsilon h}\varphi\phi_z + \int_{\omega_z}\kappa\varphi\phi_z - \int_{\omega_z}\sigma_{\epsilon h,z}\varphi\phi_z = \int_{\gamma_z}J_{\epsilon h}(\varphi-\varphi_z)\phi_z + \int_{\omega_z}(\kappa-\kappa_z)(\varphi-\varphi_z)\phi_z.$$

The right-hand side can be estimated along standard lines. We apply the "scaled" trace theorem $\|\psi\|_{0,1;\gamma_z} \preccurlyeq h_z^{-1}\|\psi\|_{0,1;\omega_z} + \|\nabla\psi\|_{0,1;\omega_z}$ and the scaled Sobolev inequality $\|\psi\|_{0,d;\omega_z} \preccurlyeq h_z^{-1}\|\psi\|_{0,1;\omega_z} + \|\nabla\psi\|_{0,1;\omega_z}$ with $\psi=\varphi-\varphi_z$. We then exploit the invariance $\varphi-\varphi_z = (\varphi-c)-(\varphi-c)_z$ and the stability property $\|\psi_z\|_{0,1;\omega_z}\preccurlyeq\|\psi\|_{0,1;\omega_z}$, where $\psi=\varphi-c$ with some $c\in\mathbb{R}$. The variant $\inf_{c\in\mathbb{R}}\|\varphi-c\|_{0,1;\omega_z}\preccurlyeq h_z\|\nabla\varphi\|_{0,1;\omega_z}$ (cf. [16, (4.2)]) of the Bramble–Hilbert lemma finally implies

$$\left|\int_{\gamma_z}J_{\epsilon h}(\varphi-\varphi_z)\phi_z + \int_{\omega_z}(\kappa-\kappa_z)(\varphi-\varphi_z)\phi_z\right| \preccurlyeq h_z^{-d/2}\eta_z\|\nabla\varphi\|_{0,1;\omega_z}. \qquad \square$$

Remark 4.1 (no interior residuals). If $\sigma_{\epsilon h}=0$ on Ω, then $\mathcal{G}_{\epsilon h}=\mathcal{R}_{\epsilon h}$ and Lemma 4.1 improves upon inequality (5.7) in [11], since, as h_z decreases, the data oscillation indicator $h_z^{d/2}\,\|\kappa-\kappa_z\|_{0,d,\omega_z}$ vanishes asymptotically faster than the local interior residual $h_z^{d/2}\,\|\kappa\|_{0,d,\omega_z}$. Estimates, where the interior residual is replaced by data oscillation, seem to appear first in section 2 of Babuška and Miller [3]; if $\sigma_{\epsilon h} = 0$ in Ω, the proof of Lemma 4.1 exploits $\sum_{z\in\mathcal{N}_h} \phi_z = 1$ and the definition of the discrete solution in a similar way as Carstensen and Verfürth [7, Theorem 7.1] and Morin, Nochetto, and Siebert [14].

A straightforward consequence of Lemma 4.1 is the following upper bound for $|\mathcal{G}_{\epsilon h}|_{\epsilon h;*}$, which additionally involves the computable quantities

$$\underline{\Lambda}_z := \inf_{\omega_z}\big[A(\nabla u_h;\epsilon)^{-1}\big], \quad z \in \mathcal{N}_h \setminus \mathcal{C}_{\epsilon h}. \tag{4.5}$$

These quantities are related to the maximum eigenvalue

$$\Lambda(p;r) = \big(r^2 + |p|^2\big)^{-1/2} \tag{4.6}$$

of the matrix $Da(p;r)$, $p \in \mathbb{R}^d$, $r > 0$: if $\nabla u_h = p$ and $\epsilon = r$ on ω_z, then $\underline{\Lambda}_z = \Lambda(p;r)$.

Corollary 4.2. *The Galerkin functional $\mathcal{G}_{\epsilon h}$ defined in* (3.1) *satisfies*

$$|\mathcal{G}_{\epsilon h}|_{\epsilon h;*} \preccurlyeq \left(\sum_{z\in\mathcal{N}_h\setminus\mathcal{C}_{\epsilon h}} \underline{\Lambda}_z^{-1}\eta_z^2\right)^{1/2}.$$

Proof. We use $\|\nabla\varphi\|_{0,1;\omega_z} \preccurlyeq h_z^{d/2}\underline{\Lambda}_z^{-1/2}\big(\int_{\omega_z}|\nabla\varphi|^2 A(\nabla u_h;\epsilon)^{-1}\big)^{1/2}$ in Lemma 4.1 and then Cauchy–Schwarz as well as

$$\#\{z \in \mathcal{N}_h \mid \omega_z \supset T\} = d+1 \tag{4.7}$$

for any $T \in \mathcal{T}_h$. □

Next, we derive an upper bound of $\langle \mathcal{G}_{\epsilon h}, u_h - u\rangle$. To this end, the following inequality, which generalizes [11, Lemma 5.1], will be useful.

Lemma 4.3. *Let $p_1, p_2 \in \mathbb{R}^d$, $r > 0$, and set $P_i := (p_i, -r)$ for $i = 1, 2$. We have*

$$|p_1 - p_2|\,\frac{r^2}{|P_1|^2} \;\le\; 2\left|\frac{P_1}{|P_1|} - \frac{P_2}{|P_2|}\right|\sqrt{|P_1|}\,\frac{r}{\sqrt{|P_1|}} + \left|\frac{P_1}{|P_1|} - \frac{P_2}{|P_2|}\right|^2 |P_2|.$$

Proof. We first observe

$$|p_1 - p_2| \;=\; |P_1 - P_2| \;\le\; \left|\frac{P_1}{|P_1|} - \frac{P_2}{|P_2|}\right| |P_1| + \big||P_1| - |P_2|\big|.$$

Moreover, we estimate

$$\begin{aligned}\big||P_1| - |P_2|\big|\,\frac{r^2}{|P_1|^2} \;&\le\; r^2\left|\frac{|P_1| - |P_2|}{|P_1|\,|P_2|}\right| + r\left|\frac{|P_1| - |P_2|}{|P_1|}\left(\frac{r}{|P_1|} - \frac{r}{|P_2|}\right)\right| \\ &\le\; r\left|\frac{r}{|P_1|} - \frac{r}{|P_2|}\right| + \left|\frac{r}{|P_1|} - \frac{r}{|P_2|}\right|^2 |P_2|.\end{aligned}$$

We insert the last inequality in the first one multiplied with $r^2/|P_1|^2$ and establish the claim by observing $|P_1| \ge r$ and $\big|r/|P_1| - r/|P_2|\big| \le \big|P_1/|P_1| - P_2/|P_2|\big|$. □

To give an interpretation of the new quantities in the estimate of Lemma 4.3, we recall (4.6) and observe that, for $p \in \mathbb{R}^d$ and $r > 0$,

$$\lambda(p;r) = r^2\big(r^2 + |p|^2\big)^{-3/2} \quad\text{and}\quad Q(p;r) = \big(r^2 + |p|^2\big)r^{-2},$$

where $\lambda(p;r)$ is the minimum eigenvalue and $Q(p;r) = \Lambda(p;r)/\lambda(p;r)$ is the quotient of the extreme eigenvalues of the matrix $Da(p;r)$. The weight $r^2/|P_1|^2$ of the estimated term $|p_1 - p_2|$ is the inverse of the quotient $Q(p_1;r) = \Lambda(p_1;r)/\lambda(p_1;r)$. Moreover, the weight of the leading order term $\big|P_1/|P_1| - P_2/|P_2|\big|\sqrt{|P_1|}$ on the right-hand side is $r/|P_1|^{1/2}$. Thus the squared quotient of the two weights satisfies

$$\left(\frac{r^2/|P_1|^2}{r/|P_1|^{1/2}}\right)^2 = \lambda(p_1;r).$$

The following upper bound for $\langle \mathcal{G}_{\epsilon h}, u_h - u\rangle$, derived with the help of Lemmas 4.1 and 4.3, involves the computable quantities

$$Q_z := \sup_{\omega_z} \frac{\epsilon^2 + |\nabla u_h|^2}{\epsilon^2} \quad \text{and} \quad \lambda_z := Q_z^{-2} \inf_{\omega_z} \frac{\sqrt{\epsilon^2 + |\nabla u_h|^2}}{\epsilon^2}. \tag{4.8}$$

PROPOSITION 4.4. *If u_h is a discrete minimum point, u a regular one, $\mathcal{G}_{\epsilon h}$ the functional defined in* (3.1), *and $M_h := \max_{z\in\mathcal{N}_h\setminus\mathcal{C}_{\epsilon h}} Q_z h_z^{-d/2}\eta_z$, then*

$$\begin{aligned}\langle \mathcal{G}_{\epsilon h}, u_h - u\rangle \;-\; &\frac{1}{2}\int_\Omega |N(\nabla u_h;\epsilon) - N(\nabla u;\epsilon)|^2 \,\frac{A(\nabla u_h;\epsilon)}{2}\\ &\preccurlyeq \sum_{z\in\mathcal{N}_h\setminus\mathcal{C}_{\epsilon h}} \lambda_z^{-1}\eta_z^2 \;+\; M_h \int_\Omega |N(\nabla u_h;\epsilon) - N(\nabla u;\epsilon)|^2\,\frac{A(\nabla u;\epsilon)}{2}.\end{aligned}$$

Proof. Given Lemmas 4.1 and 4.3, the proof essentially follows from arguments presented in step 4 of [11]. For the convenience of the reader, we adapt those arguments. Thanks to Lemma 4.3, we have

$$\frac{|\nabla(u_h - u)|}{Q(\nabla u_h;\epsilon)} \le 2\epsilon|N(\nabla u_h;\epsilon) - N(\nabla u;\epsilon)| + |N(\nabla u_h;\epsilon) - N(\nabla u;\epsilon)|^2 A(\nabla u;\epsilon)$$

on ω_z for any $z \in \mathcal{N}_h \setminus \mathcal{C}_{\epsilon h}$, whence

$$\begin{aligned}\langle \mathcal{G}_{\epsilon h}, u_h - u\rangle \;\le\; &C \sum_{z\in\mathcal{N}_h\setminus\mathcal{C}_{\epsilon h}} Q_z\, h_z^{-d/2}\,\eta_z \int_{\omega_z} |N(\nabla u_h;\epsilon) - N(\nabla u;\epsilon)|\,\epsilon\\ &+ C \sum_{z\in\mathcal{N}_h\setminus\mathcal{C}_{\epsilon h}} Q_z\, h_z^{-d/2}\,\eta_z \int_{\omega_z} |N(\nabla u_h;\epsilon) - N(\nabla u;\epsilon)|^2\, A(\nabla u;\epsilon)\\ &=: \mathrm{I} + \mathrm{II}\end{aligned} \tag{4.9}$$

by means of Lemma 4.1. We first consider sum I. The inequality

$$\begin{aligned}&\int_{\omega_z} |N(\nabla u_h;\epsilon) - N(\nabla u;\epsilon)|\,\epsilon\\ &\preccurlyeq\; h_z^{d/2}\left[\sup_{\omega_z} \frac{\epsilon^2}{\sqrt{|\epsilon|^2 + |\nabla u_h|^2}}\right]^{1/2} \left[\int_{\omega_z} |N(\nabla u_h;\epsilon) - N(\nabla u;\epsilon)|^2\, A(\nabla u_h;\epsilon)\right]^{1/2},\end{aligned}$$

the identity $Q_z^2 \sup_{\omega_z} \epsilon^2/\sqrt{\epsilon^2 + |\nabla u_h|^2} \;=\; \lambda_z^{-1}$, and

$$\forall\, s,t \ge 0,\ \delta > 0, \qquad st \le \frac{\delta}{2}s^2 + \frac{1}{2\delta}t^2 \tag{4.10}$$

as well as (4.7) yield

$$\mathrm{I} - \frac{1}{2}\int_\Omega |N(\nabla u_h;\epsilon) - N(\nabla u;\epsilon)|^2\,\frac{A(\nabla u_h;\epsilon)}{2} \preccurlyeq \sum_{z\in\mathcal{N}_h\setminus\mathcal{C}_{\epsilon h}} \lambda_z^{-1}\,\eta_z^2. \tag{4.11}$$

It remains to consider sum II. Using again (4.7), we obtain

$$\mathrm{II} \preccurlyeq \left(\max_{z\in\mathcal{N}_h\setminus\mathcal{C}_{\epsilon h}} Q_z\, h_z^{-d/2}\,\eta_z\right)\int_\Omega |N(\nabla u_h;\epsilon) - N(\nabla u;\epsilon)|^2\,\frac{A(\nabla u;\epsilon)}{2} \tag{4.12}$$

and conclude by inserting (4.11) and (4.12) into (4.9). □

We still have to estimate the last term $\langle\sigma_{\epsilon h} - \sigma_\epsilon, u_h - u\rangle$ of (4.1). To this end, we have to encounter the fact that $\sigma_{\epsilon h}$ is not quite a subgradient in u_h of the convex potential associated with the constraint $|u_h| \le 1$. This consistency error will be controlled by means of the computable quantities $\Sigma_z d_z$, $z \in \mathcal{F}_{\epsilon h}$, where

$$\mathcal{F}_{\epsilon h} := \big\{z \in \mathcal{N}_h \mid z \notin \mathcal{C}_{\epsilon h} \text{ and } \Sigma_z \neq 0\big\} \tag{4.13}$$

are the "free boundary nodes," Σ_z is defined as in (2.9), and

$$d_z := \begin{cases} \int_{\omega_z}(1-u_h)\phi_z & \text{if } \Sigma_z > 0,\\ -\int_{\omega_z}(u_h+1)\phi_z & \text{if } \Sigma_z < 0. \end{cases} \tag{4.14}$$

Since Σ_z, $z \in \mathcal{N}_h$, and $\mathcal{C}_{\epsilon h}$ do not depend on the particular choice of u_h, the same holds for $\mathcal{F}_{\epsilon h}$ and d_z, $z \in \mathcal{F}_{\epsilon h}$. In view of (2.8), $\Sigma_z > 0$ implies $u_h(z) = 1$ and $\Sigma_z < 0$ implies $u_h(z) = -1$. The quantity $|d_z|$ therefore measures the detachment of u_h from a free boundary node $z \in \mathcal{F}_{\epsilon h}$. Furthermore, there holds $\Sigma_z d_z \ge 0$ for all $z \in \mathcal{F}_{\epsilon h}$.

Proposition 4.5 (quasi-monotonicity of auxiliary functionals). *If the pairs* $(u_h, \sigma_{\epsilon h})$ *and* (u, σ_ϵ) *satisfy* (2.6), (2.10), (2.2), *and* (2.4), *respectively, then*

$$\langle\sigma_{\epsilon h} - \sigma_\epsilon, u_h - u\rangle \ge -\sum_{z\in\mathcal{F}_{\epsilon h}} \Sigma_z d_z.$$

Proof. We may write $\langle\sigma_{\epsilon h} - \sigma_\epsilon, u_h - u\rangle = \langle\sigma_{\epsilon h}, u_h - u\rangle + \langle\sigma_\epsilon, u - u_h\rangle$. Since $u_h \in W_1^1\big(\Omega;[-1,1]\big)$, (2.4) and (2.2) imply $\langle\sigma_\epsilon, u - u_h\rangle \ge 0$, and it remains to consider

$$\int_\Omega \sigma_{\epsilon h}(u_h - u) = \sum_{z\in\mathcal{N}_h}\int_\Omega \sigma_{\epsilon h,z}(u_h - u)\phi_z.$$

Some terms of the right-hand side are nonnegative. In fact, if $z \in \mathcal{C}_{\epsilon h}$, then $|u| \le 1$ in Ω and (2.10a) yield $\sigma_{\epsilon h,z}(u_h - u) \ge 0$. Moreover, if $z \in \mathcal{N}_h \setminus \mathcal{C}_{\epsilon h}$ and $\Sigma_z = 0$, then (2.10a) implies $\sigma_{\epsilon h,z} = 0$. We can therefore estimate

$$\int_\Omega \sigma_{\epsilon h}(u_h - u) \ge \sum_{z\in\mathcal{F}_{\epsilon h}}\int_\Omega \sigma_{\epsilon h,z}(u_h - u)\phi_z. \tag{4.15}$$

The remaining terms are treated by considering two cases. Let $z \in \mathcal{F}_{\epsilon h}$ with $\Sigma_z < 0$. Then (2.10a) and $u \ge -1$ give

$$\int_\Omega \sigma_{\epsilon h,z}(u_h - u)\phi_z = \int_{\omega_z}\Sigma_z(u_h+1)\phi_z + \int_{\omega_z}\Sigma_z(-1-u)\phi_z \ge -\Sigma_z d_z.$$

Since also $u \leq 1$, the same inequality holds if we replace $\Sigma_z < 0$ by $\Sigma_z > 0$. Thus we finish the proof by inserting these inequalities into (4.15). $\square$

Finally, we combine the derived estimates to obtain an a posteriori upper bound for the combined error $e_{\epsilon h} + |\sigma_{\epsilon h} - \sigma_\epsilon|_{\epsilon h;*}$ under an a posteriori condition. The involved computable quantities are defined in (2.9), (2.10b), (4.2), (4.8), (4.14), and (4.13).

Theorem 4.6 (conditional upper bound). *Let u_h and u be a discrete and regular minimum point, and let $\sigma_{\epsilon h}$ and σ_ϵ be defined as in* (2.4) *and* (2.10). *There exists a constant C depending only on the shape-regularity γ_h of the triangulation $\mathcal{T}_h$ such that the following holds: if*

$$\max_{z \in \mathcal{N}_h \setminus \mathcal{C}_{\epsilon h}} Q_z h_z^{-d/2} \eta_z \leq C, \tag{4.16}$$

then the error $e_{\epsilon h} + |\sigma_{\epsilon h} - \sigma_\epsilon|_{\epsilon h;*}$ defined by (3.4) *and* (3.7) *is bounded from above in terms of computable quantities and a multiplicative constant depending on γ_h:*

$$e_{\epsilon h} + |\sigma_{\epsilon h} - \sigma_\epsilon|_{\epsilon h;*} \preccurlyeq \left(\sum_{z \in \mathcal{N}_h \setminus \mathcal{C}_{\epsilon h}} \lambda_z^{-1} \eta_z^2 + \sum_{z \in \mathcal{F}_{\epsilon h}} \Sigma_z d_z \right)^{1/2}. \tag{4.17}$$

Proof. Let $1/(2C)$ be the constant hidden in "$\preccurlyeq$" of (4.12), and suppose that (4.16) holds with this choice of C. We then obtain

$$e_{\epsilon h} \preccurlyeq \left(\sum_{z \in \mathcal{N}_h \setminus \mathcal{C}_{\epsilon h}} \lambda_z^{-1} \eta_z^2 + \sum_{z \in \mathcal{F}_{\epsilon h}} \Sigma_z d_z \right)^{1/2} \tag{4.18}$$

by using Propositions 4.4 and 4.5 in (3.5). In view of (3.8), Corollary 4.2, and

$$\underline{\Lambda}_z^{-1} \leq \lambda_z^{-1}, \tag{4.19}$$

adding $|\sigma_{\epsilon h} - \sigma_\epsilon|_{\epsilon h;*}$ to the left-hand side of (4.18) changes only the constant hidden in "$\preccurlyeq$." $\square$

Remark 4.2 (a posteriori condition). The upper bound in Theorem 4.6 cannot be expected to be valid in general, i.e., without any condition as, e.g., (4.16); see [11, (5.1) and Example 5.1], which shows that (the size of) the discrete gradient defining λ_z may be quite different from the exact one.

Moreover, we point out that (4.16) is of "a posteriori" nature: if we knew C explicitly, then (4.16) could be verified since its left-hand side is computable. In any case, the latter quantity can be monitored during computations as in [11, section 7.1], where numerical experiments show that it decreases for exact solutions with moderate gradients within a reasonable number of unknowns.

Remark 4.3 (uniform in regularization). We stress that the relationship between error and computable quantities in the upper bound of Theorem 4.6 is independent of the regularization "parameter" ϵ.

Remark 4.4 (complete localization). It is worthwhile to mention that the upper bound in Theorem 4.6 involves only indicators related to the typically proper subset $\bigcup\{\omega_z : z \in \mathcal{N}_h \setminus \mathcal{C}_{\epsilon h}\}$ of Ω. We refer to this property as "complete localization" to the discrete non-full-contact set. Remarkably, the complement of the latter subset contains at least $\bigcup\{T : u_h = u = 1 \text{ or } u_h = u = -1 \text{ in } \omega_T\}$, that is, approximately the intersection of the discrete and exact contact sets. In fact, for the involved nodes, the sign conditions on κ in (2.10b) are satisfied due to (2.5).

The upper bounds in [17, 18] exhibit a "partial localization" in that there are indicators in the discrete contact set with higher order than the ones in the discrete noncontact set. This improvement in the upper bound for the first error part $e_{\epsilon h}$ is due to the new definition (2.10) of the auxiliary functional.

Remark 4.5 (importance of sign conditions). Suppose that (4.16) and the condition in (2.7b) hold. Then the upper bound in Theorem 4.6 does not hold in general if one replaces $\mathcal{C}_{\epsilon h}$ by $\tilde{\mathcal{C}}_{\epsilon h} := \{z \in \mathcal{N}_h \mid u_h = -1 \text{ on } \omega_z \text{ or } u_h = 1 \text{ on } \omega_z\}$. This reveals the example of Chen and Nochetto [8, Remark 4.2], where, in the case of the obstacle problem for the Laplacian, the discrete solution is different from the exact one and the only nonzero indicator corresponds to a node in $\tilde{\mathcal{C}}_{\epsilon h} \setminus \mathcal{C}_{\epsilon h}$.

5. Lower bounds.

We derive a posteriori local lower bounds for the error introduced in section 3. More precisely, we estimate the sum of the error $e_{\epsilon h}(\omega_z) + |\sigma_{\epsilon h} - \sigma_\epsilon|_{\epsilon h,\omega_z;*}$ in a star ω_z and associated data oscillation $h_y^{d/2} \|\kappa - \kappa_y\|_{0,d;\omega_y}$ from below in terms of

$$h_z^{1/2} \|J_{\epsilon h}\|_{0,2;\gamma_z} \quad \text{and} \quad \sqrt{\Sigma_z d_z}. \tag{5.1}$$

To this end, we shall use (3.9). The computable quantities (5.1) appear in the upper bound of section 4. Consequently, the following bounds concern their efficiency up to data oscillation, which is formally of higher order.

We begin with a lower bound involving the jump residual, i.e., the first term in (5.1). Apart from (3.9), the main step is an adaptation of Verfürth's constructive argument (see, e.g., [19, section 1.2]) to the representation formula of $\mathcal{G}_{\epsilon h}$ in Lemma 4.1. The use of (3.9) entails the computable weight

$$\Lambda_z := \sup_{\omega_z} \big[A(\nabla u_h; \epsilon)^{-1}\big], \tag{5.2}$$

differing from $\underline{\Lambda}_z$ defined in (4.5). In addition, the following index set is useful:

$$\mathcal{N}_h(z) := \big(\mathcal{N}_h \setminus \mathcal{C}_{\epsilon h}\big) \cap \omega_z. \tag{5.3}$$

THEOREM 5.1 (lower bound I). *Let $z \in \mathcal{N}_h \setminus \mathcal{C}_{\epsilon h}$ be a node, and let $\|J_{\epsilon h}\|_{0,2;\gamma_z}$ be defined as in* (4.2). *There holds*

$$\frac{h_z^{1/2}}{\Lambda_z^{1/2}} \|J_{\epsilon h}\|_{0,2;\gamma_z} \preccurlyeq e_{\epsilon h}(\omega_z) + |\sigma_{\epsilon h} - \sigma_\epsilon|_{\epsilon h,\omega_z;*} + \sum_{y \in \mathcal{N}_h(z)} \frac{h_y^{d/2}}{\Lambda_z^{1/2}} \|\kappa - \kappa_y\|_{0,d;\omega_y}.$$

Proof. We first prove a more local estimate for the jump residual $\|J_{\epsilon h}\|_{0,2;S}$ of a side S with $S \subset \gamma_z$. To this end, we recall the representation formula

$$\langle \mathcal{G}_{\epsilon h}, \varphi \rangle = -\sum_{y \in \mathcal{N}_h \setminus \mathcal{C}_{\epsilon h}} \left[\int_{\gamma_y} J_{\epsilon h}(\varphi - \varphi_y)\phi_y + \int_{\omega_y} (\kappa - \kappa_y)(\varphi - \varphi_y)\phi_y \right] \tag{5.4}$$

established in the proof of Lemma 4.1 and construct an appropriate test function φ. Let $\psi_S := \prod_{x \in \mathcal{N}_h \cap S} \phi_x$ and $\omega_S := \operatorname{supp} \psi_S$. Moreover, for any $T \in \mathcal{T}_h$ with $T \subset \omega_S$, let $\psi_T := \prod_{x \in \mathcal{N}_h \cap T} \phi_x$, and choose $\alpha_{T,x} \in \mathbb{R}$, $x \in \mathcal{N}_h \cap T$, such that

$$\sum_{x \in \mathcal{N}_h \cap T} \alpha_{T,x} \int_T \psi_T \phi_x \phi_y = \int_T \psi_S \phi_y \tag{5.5}$$

for any $y \in \mathcal{N}_h \cap T$. Finally, note that $J_{\epsilon h}|_S$ is a constant, and set

$$\varphi := J_{\epsilon h}|_S \left(\psi_S - \sum_{T\subset\omega_S, x\in\mathcal{N}_h\cap T} \alpha_{T,x}\psi_T\phi_x \right).$$

The support of φ is ω_S. Let $\mathcal{N}_h(S) := (\mathcal{N}_h \setminus \mathcal{C}_{\epsilon h}) \cap \omega_S$ and $\Lambda_S := \sup_{\omega_S} A(\nabla u_h;\epsilon)^{-1}$. We obtain

$$\begin{aligned}\|J_{\epsilon h}\|^2_{0,2;S} &\preccurlyeq \sum_{y\in\mathcal{N}_h\setminus\mathcal{C}_{\epsilon h}} \int_S J_{\epsilon h}\varphi\phi_y = -\langle \mathcal{G}_{\epsilon h}, \varphi\rangle - \sum_{y\in\mathcal{N}_h\setminus\mathcal{C}_{\epsilon h}} \int_{\omega_y} (\kappa-\kappa_y)\varphi\phi_y \\ &\preccurlyeq \Lambda_S^{1/2} |\mathcal{G}_{\epsilon h}|_{\epsilon h,\omega_S;*} \|\nabla\varphi\|_{0,2;\omega_S} + \sum_{y\in\mathcal{N}_h(S)} \|\kappa-\kappa_y\|_{0,2;\omega_S} \|\varphi\|_{0,2;\omega_S}\end{aligned}$$

by $\varphi|_S = \big(J_{\epsilon h}\psi_S\big)|_S$ and $\int_S 1 \preccurlyeq \int_S \psi_S\phi_y$ for $y \in \mathcal{N}_h \cap S$, by $\int_{\omega_y} \varphi\phi_y = 0$ for any $y \in \mathcal{N}_h$ and (5.4), and by (3.7). Hence, if h_S is the diameter of the side S,

$$\|J_{\epsilon h}\|_{0,2;S} \preccurlyeq \Lambda_S^{1/2} h_S^{-1/2} |\mathcal{G}_{\epsilon h}|_{\epsilon h,\omega_S;*} + h_S^{1/2} \sum_{y\in\mathcal{N}_h(S)} \|\kappa-\kappa_y\|_{0,2;\omega_S}$$

with the help of $\|\nabla\varphi\|_{0,2;\omega_S} + h_S\|\varphi\|_{0,2;\omega_S} \preccurlyeq h_S^{-1/2}\|\varphi\|_{0,2;S} \le h_S^{-1/2}\|J_{\epsilon h}\|_{0,2;S}$, which in turn follows from the fact that $\alpha_{T,x}$ in (5.5) does not depend on h_S. Employing (3.9), we finally arrive at

$$h_S^{1/2}\|J_{\epsilon h}\|_{0,2;S} \;\preccurlyeq\; \Lambda_S^{1/2}\big[e_{\epsilon h}(\omega_s) + |\sigma_{\epsilon h}-\sigma_\epsilon|_{\epsilon h,\omega_S;*}\big] + h_S \sum_{y\in\mathcal{N}_h(S)} \|\kappa-\kappa_y\|_{0,2;\omega_S}.$$

In view of (4.7), $\Lambda_S \le \Lambda_z$ for any $z \in \mathcal{N}_h \cap S$, and $h_S \approx h_y$ for all $y \in \mathcal{N}_h \cap \omega_S$, and the claimed estimate follows. □

Remark 5.1 (gap for jump residual). The jump residual $h_z^{1/2}\|J_{\epsilon h}\|_{0,2;\gamma_z}$ and the data oscillation are accompanied by different weights in the bounds of Theorems 4.6 and 5.1—by $\lambda_z^{-1/2}$ in the upper bound and by $\Lambda_z^{-1/2}$ in the lower bound. This gap between the two bounds seems to be unavoidable as long as the a posteriori error analysis does not take the unknown direction of the error into account; see also section 5.2 of [11], where different numerical experiments indicate that both bounds constitute possible worst cases.

We next bound the computable quantities $\sqrt{\Sigma_z d_z}$, $z \in \mathcal{F}_{\epsilon h}$, controlling the consistency error of $\sigma_{\epsilon h}$. To this end, we derive separate estimates for the two factors in the square root. For the estimate of Σ_z, $z \in \mathcal{F}_{\epsilon h}$, Verfürth's constructive argument and (3.9) are again important ingredients.

LEMMA 5.2. *Let $z \in \mathcal{F}_{\epsilon h}$ and Σ_z be defined as in* (2.9). *If $d_z \neq 0$, then*

$$\Lambda_z^{-1/2} h_z^{(d+2)/2} |\Sigma_z| \;\preccurlyeq\; e_{\epsilon h}(\omega_z) + |\sigma_{\epsilon h}-\sigma_\epsilon|_{\epsilon h,\omega_z;*} + \Lambda_z^{-1/2} h_z^{d/2} \|\kappa-\kappa_z\|_{0,d;\omega_z}.$$

Proof. Suppose that $d_z < 0$ for $z \in \mathcal{F}_{\epsilon h}$. Then $\Sigma_z < 0$ thanks to (4.14), and there are a simplex $T \in \mathcal{T}_h$ containing z and another node $y \in \mathcal{N}_h \cap T$ such that $u_h(y) > -1$. The latter implies $\Sigma_y \ge 0$ thanks to (2.8). Moreover, since all nodes $x \in \mathcal{N}_h \cap T$ are in $\mathcal{N}_h \setminus \mathcal{C}_{\epsilon h}$, the definition (2.10) of $\sigma_{\epsilon h}$ yields $\sigma_{\epsilon h} \in P_1(T)$. Denoting the diameter of T by h_T and using an inverse estimate, we thus obtain

$$\begin{aligned}|\Sigma_z| &\le \sigma_{\epsilon h}(y) - \sigma_{\epsilon h}(z) = \nabla\sigma_{\epsilon h|T}\cdot(y-z) \le h_T\|\nabla\sigma_{\epsilon h}\|_{0,\infty;T} \\ &= h_T\|\nabla(\sigma_{\epsilon h}-\kappa_z)\|_{0,\infty;T} \preccurlyeq h_T^{-d/2}\|\sigma_{\epsilon h}-\kappa_z\|_{0,2;T}.\end{aligned} \tag{5.6}$$

The same inequality can be derived for $d_z > 0$ in a similar manner.

Consequently, it remains to estimate $\|\sigma_{\epsilon h} - \kappa_z\|_{0,2;T}$ for $T \in \mathcal{T}_h$ with $T \subset \omega_z$ appropriately. We proceed similarly to the estimation of the jump residual in the proof of Theorem 5.1. The function $\psi_T = \prod_{x \in \mathcal{N}_h \cap T} \phi_x$ satisfies

$$\int_T |w_h|^2 \preccurlyeq \int_T |w_h|^2 \psi_T \quad \text{and} \quad \|\nabla(w_h \psi_T)\|_{0,2;T} \preccurlyeq h_T^{-1} \|w_h \psi_T\|_{0,2;T} \tag{5.7}$$

for all $w_h \in P_1(T)$; see [19, Lemma 3.3]. Setting $\varphi := (\sigma_{\epsilon h} - \kappa_z)\psi_T$, we derive

$$\begin{aligned}\|\sigma_{\epsilon h} - \kappa_z\|_{0,2;T}^2 &\preccurlyeq \int_T (\sigma_{\epsilon h} - \kappa_z)\varphi = \langle \mathcal{G}_{\epsilon h}, \varphi\rangle + \int_T (\kappa - \kappa_z)\varphi \\ &\preccurlyeq \Lambda_z^{1/2} |\mathcal{G}_{\epsilon h}|_{\epsilon h, T; *} \|\nabla \varphi\|_{0,2;T} + \|\kappa - \kappa_z\|_{0,2;T} \|\varphi\|_{0,2;T}\end{aligned}$$

with the help of (5.7), (3.1), and $\int_T a(\nabla u_h) \cdot \nabla \varphi = 0$ (integrate by parts). Therefore,

$$\Lambda_z^{-1/2} h_T \|\sigma_{\epsilon h} - \kappa_z\|_{0,2;T} \preccurlyeq e_{\epsilon h}(T) + |\sigma_{\epsilon h} - \sigma_\epsilon|_{\epsilon h, T; *} + \Lambda_z^{-1/2} h_T \|\kappa - \kappa_z\|_{0,2;T} .$$

Using this inequality in (5.6), $h_T \approx h_z$, and a Hölder inequality finishes the proof. □

For the estimation of d_z, $z \in \mathcal{F}_{\epsilon h}$, we adapt [15, Lemma 6.4] to the operator and boundary conditions of (2.2). The adaptation's proof is different from [15], and its second part resembles the one of Lemma 3.3 in Chen and Nochetto [8], which measures the deviation of positivity preserving interpolation from a projection.

LEMMA 5.3 (discrete growth around critical point). *Let $w_h \in W_h$ and $z \in \mathcal{N}_h$. If $w_h \geq 0$ in the star ω_z and $w_h(z) = 0$, then*

$$\int_{\omega_z} w_h \ \preccurlyeq \ \underline{\Lambda}_z(\nabla w_h; \epsilon)^{-1}\, h_z^2\, \|J(w_h; \epsilon)\|_{0,1;\gamma_z} ,$$

where $\underline{\Lambda}_z(\nabla w_h; \epsilon) := \inf_{\omega_z} \left[A(\nabla w_h; \epsilon)^{-1}\right]$, $J(\cdot; \epsilon)$ is defined as in (4.3), and γ_z is the union of all sides in ω_z.

Proof. We first replace the domain ω_z of integration by a part of a ball; this will be useful in what follows. Let B be the biggest ball with center z and $B \cap \Omega \subset \omega_z$. The equivalence of norms in $P_1(\hat{T})$, where $\hat{T}$ is the reference simplex, implies $\|w_h\|_{0,1;T} \preccurlyeq \|w_h\|_{0,1;T \cap B}$ for any $T \in \mathcal{T}_h$ with $T \subset \omega_z$. Combining this with $w_h \geq 0$ in ω_z yields

$$\int_{\omega_z} w_h \preccurlyeq \int_{B \cap \Omega} w_h \leq \underline{\Lambda}_z(\nabla w_h; \epsilon)^{-1} \int_{B \cap \Omega} \frac{w_h}{A(\nabla w_h; \epsilon)}. \tag{5.8}$$

The function w_h can be represented with the help of its gradient ∇w_h. In fact, since $w_h(z) = 0$ and ∇w_h is piecewise constant, we have $w_h(x) = \nabla w_h|_T \cdot (x - z)$ for all $x \in T$ and any $T \in \mathcal{T}_h$ with $T \subset \omega_z$. Therefore,

$$\int_{B \cap \Omega} \frac{w_h}{A(\nabla w_h; \epsilon)} = \int_{B \cap \Omega} a(\nabla w_h; \epsilon) \cdot (\mathrm{id} - z), \tag{5.9}$$

where id stands for the identity of Ω. To relate the integral on the right-hand side to the jumps $J(w_h; \epsilon)$, we observe that $\mathrm{id} - z = \frac{1}{2} \nabla |\mathrm{id} - z|^2$ and integrate by parts:

$$2 \int_{B \cap \Omega} a(\nabla w_h; \epsilon) \cdot (\mathrm{id} - z) = - \int_{\gamma_z \cap B} J(w_h; \epsilon) |\mathrm{id} - z|^2 + \int_{\partial B \cap \Omega} a(\nabla w_h; \epsilon) \cdot n |\mathrm{id} - z|^2,$$

where n denotes the outer normal of $B \cap \Omega$. Since B is a ball, the function $|\operatorname{id} - z|^2$ is constant on $\partial B \cap \Omega$, and it equals the radius r of B. The second integral of the right-hand side can thus be rewritten by means of the divergence theorem as

$$\int_{\partial B \cap \Omega} a(\nabla w_h; \epsilon) \cdot n |\operatorname{id} - z|^2 = r^2 \int_{\partial B \cap \Omega} a(\nabla w_h; \epsilon) \cdot n = r^2 \int_{\gamma_z \cap B} J(w_h; \epsilon).$$

Consequently, we obtain

$$2 \int_{B \cap \Omega} a(\nabla w_h; \epsilon) \cdot (\operatorname{id} - z) = \int_{\gamma_z \cap B} J(w_h; \epsilon) \big(r^2 - |\operatorname{id} - z|^2 \big) \leq h_z^2 \left\| J(w_h; \epsilon) \right\|_{0,1;\gamma_z}$$

by using $r \leq h_z$ and conclude by recalling (5.8) and (5.9). □

An immediate consequence of Lemma 5.3 is an estimate for d_z, $z \in \mathcal{F}_{\epsilon h}$.

COROLLARY 5.4. *Let $z \in \mathcal{F}_{\epsilon h}$. For d_z as in* (4.14)*, there holds*

$$\underline{\Lambda}_z h_z^{-(d+2)/2} |d_z| \preccurlyeq h_z^{1/2} \left\| J_{\epsilon h} \right\|_{0,2;\gamma_z}$$

with $\underline{\Lambda}_z$ and $J_{\epsilon h}$ as in (4.5) *and* (4.2)*, respectively.*

Proof. Let $z \in \mathcal{F}_{\epsilon h}$. It holds that $\Sigma_z \neq 0$, which, due to (2.8), implies $u_h(z) \in \{\pm 1\}$. Since $-1 \leq u_h \leq 1$ in Ω, we can apply Lemma 5.3 with $w_h = 1 - u_h$ or $w_h = u_h + 1$. The estimate $\left\| J_{\epsilon h} \right\|_{0,1;\gamma_z} \preccurlyeq h_z^{(d-1)/2} \left\| J_{\epsilon h} \right\|_{0,2;\gamma_z}$ completes the proof. □

The combination of Theorem 5.1, Lemma 5.2, and Corollary 5.4 finally yields the desired estimate for $\sqrt{\Sigma_z d_z}$, $z \in \mathcal{F}_{\epsilon h}$.

THEOREM 5.5 (lower bound II). *Let $z \in \mathcal{F}_{\epsilon h}$. For $\sqrt{\Sigma_z d_z}$ with Σ_z as in* (2.9) *and d_z as in* (4.14)*, there holds*

$$\frac{\underline{\Lambda}_z^{1/2}}{\Lambda_z^{1/2}} \sqrt{\Sigma_z d_z} \preccurlyeq e_{\epsilon h}(\omega_z) + |\sigma_{\epsilon h} - \sigma_\epsilon|_{\epsilon h, \omega_z; *} + \Lambda_z^{-1/2} \sum_{y \in \mathcal{N}_h(z)} h_y^{d/2} \left\| \kappa - \kappa_y \right\|_{0,d;\omega_y},$$

where $\underline{\Lambda}_z$, Λ_z, and $\mathcal{N}_h(z)$ are defined as in (4.5), (5.2), *and* (5.3).

Proof. We can assume $d_z \neq 0$ without loss of generality. Then,

$$\begin{aligned} 2 \frac{\underline{\Lambda}_z^{1/2}}{\Lambda_z^{1/2}} \sqrt{\Sigma_z d_z} &\leq \Lambda_z^{-1/2} h_z^{(d+2)/2} |\Sigma_z| + \underline{\Lambda}_z \Lambda_z^{-1/2} h_z^{-(d+2)/2} |d_z| \\ &\preccurlyeq e_{\epsilon h}(\omega_z) + |\sigma_{\epsilon h} - \sigma_\epsilon|_{\epsilon h, \omega_z; *} + \Lambda_z^{-1/2} h_z^{d/2} \left\| \kappa - \kappa_z \right\|_{0,d;\omega_z} \\ &\quad + \Lambda_z^{-1/2} h_z^{1/2} \left\| J_{\epsilon h} \right\|_{0,2;\gamma_z} \end{aligned} \tag{5.10}$$

with the help of (4.10) with $\delta = 1$, Lemma 5.2, and Corollary 5.4. Applying Theorem 5.1 to the last term on the right-hand side concludes the proof. □

Remark 5.2 (moderate gap for $\Sigma_z d_z$). The gap concerning $\sqrt{\Sigma_z d_z}$ between the upper and lower bounds is moderate with respect to the one for the jump residual $h_z^{1/2} \left\| J_{\epsilon h} \right\|_{0,2;\omega_z}$. In particular, if $\nabla u_h \approx p$ and $\epsilon \approx r$, then $\underline{\Lambda}_z^{1/2} / \Lambda_z^{1/2} \approx 1$.

6. Adaptive algorithm and numerical experiments. We derive a simplified upper bound, and, relying on this simplification, we formulate an adaptive algorithm for fixed regularization parameter ϵ. After some remarks on our implementation, we present our numerical results illustrating the various properties of the indicators derived in sections 4 and 5.

6.1. Simplified upper bound. A refined version of (5.10) allows us to eliminate the indicators $\Sigma_z d_z$ from the upper bound in Theorem 4.6 at the expense of some sharpness and a slightly modified weight λ_z^{-1}. The new version of λ_z, defined in (4.8), is given by

$$\tilde{\lambda}_z = \begin{cases} \min\{\lambda_z, \underline{\Lambda}_z^2/\Lambda_z\} & \text{if } z \in \mathcal{F}_{\epsilon h}, \\ \lambda_z & \text{otherwise}, \end{cases} \qquad z \in \mathcal{N}_h \setminus \mathcal{C}_{\epsilon h}. \tag{6.1}$$

Note that if the regularization ϵ is constant in Ω, then $\tilde{\lambda}_z = \lambda_z$ for all $z \in \mathcal{N}_h \setminus \mathcal{C}_{\epsilon h}$.

THEOREM 6.1 (simple upper bound). *Suppose that the assumptions including* (4.16) *of Theorem* 4.6 *are fulfilled. Then there holds*

$$e_{\epsilon h} + |\sigma_{\epsilon h} - \sigma_\epsilon|_{\epsilon h;*} \preccurlyeq \left(\sum_{z \in \mathcal{N}_h \setminus \mathcal{C}_{\epsilon h}} \tilde{\lambda}_z^{-1} \eta_z^2 \right)^{1/2}.$$

Proof. Proceeding similarly as in the proof of (5.10) but also using the full flexibility of (4.10), we get

$$\begin{aligned} \Sigma_z d_z &\preccurlyeq \delta \Big[e_{\epsilon h}(\omega_z)^2 + |\sigma_{\epsilon h} - \sigma_\epsilon|^2_{\epsilon h, \omega_z;*} + \Lambda_z^{-1} h_z^d \, \|\kappa - \kappa_z\|^2_{0,d;\omega_z} \Big] \\ &\quad + \delta^{-1} \Lambda_z \underline{\Lambda}_z^{-2} h_z \, \|J_{\epsilon h}\|^2_{0,2;\gamma_z} \end{aligned} \tag{6.2}$$

for any $\delta > 0$ and all $z \in \mathcal{F}_{\epsilon h}$. In order to sum over $z \in \mathcal{F}_{\epsilon h}$, we first show that

$$\sum_{z \in \mathcal{F}_{\epsilon h}} |\sigma_{\epsilon h} - \sigma_\epsilon|^2_{\epsilon h, \omega_z;*} \preccurlyeq |\sigma_{\epsilon h} - \sigma_\epsilon|^2_{\epsilon h;*}. \tag{6.3}$$

To establish this, let $w \in \mathring{H}^1(\Omega) := \{v \in W_2^1(\Omega) \mid v = 0 \text{ on } \partial\Omega\}$ be defined by the Riesz representation theorem through

$$\forall\, \varphi \in \mathring{H}^1(\Omega), \qquad \int_\Omega \frac{\nabla w \cdot \nabla \varphi}{A(\nabla u_h; \epsilon)} = \langle \sigma_{\epsilon h} - \sigma_\epsilon, \varphi \rangle.$$

We then have

$$|\sigma_{\epsilon h} - \sigma_\epsilon|_{\epsilon h, \omega_z;*} \leq \left(\int_{\omega_z} \frac{|\nabla w|^2}{A(\nabla u_h; \epsilon)} \right)^{1/2}, \qquad |\sigma_{\epsilon h} - \sigma_\epsilon|_{\epsilon h;*} = |w|_{\epsilon h},$$

and therefore (6.3) follows from the following consequence of (4.7):

$$\sum_{z \in \mathcal{F}_{\epsilon h}} \int_{\omega_z} \frac{|\nabla w|^2}{A(\nabla u_h; \epsilon)} \preccurlyeq \int_\Omega \frac{|\nabla w|^2}{A(\nabla u_h; \epsilon)}.$$

Summing (6.2) over $z \in \mathcal{F}_{\epsilon h}$ and using (6.3), (4.7), and $\Lambda_z \underline{\Lambda}_z^{-2} \leq \tilde{\lambda}_z^{-1}$ yield

$$\begin{aligned} \sum_{z \in \mathcal{F}_{\epsilon h}} \Sigma_z d_z &\preccurlyeq \delta \Bigg(e_{\epsilon h}^2 + |\sigma_{\epsilon h} - \sigma_\epsilon|^2_{\epsilon h;*} + \sum_{z \in \mathcal{F}_{\epsilon h}} \lambda_z^{-1} h_z^d \, \|\kappa - \kappa_z\|^2_{0,d;\omega_z} \Bigg) \\ &\quad + \delta^{-1} \sum_{z \in \mathcal{F}_{\epsilon h}} \tilde{\lambda}_z^{-1} h_z \, \|J_{\epsilon h}\|^2_{0,2;\gamma_z} \end{aligned} \tag{6.4}$$

for the second sum in the right-hand side of the upper bound (4.17). If (4.16) holds, the claimed upper bound follows from (4.17), $\sqrt{s^2 + t^2} \leq s + t$ for $s, t \geq 0$, (6.4), $\lambda_z^{-1} \leq \tilde{\lambda}_z^{-1}$, and choosing $\delta = \delta(\gamma_h) > 0$ sufficiently small. □

The local lower bound that complements the simplified upper bound in Theorem 6.1 is already given in Theorem 5.1.

6.2. Adaptive algorithm and implementation. We describe the main steps of an adaptive algorithm for the convexified and regularized minimization (1.2). To this end, we rely on the a posteriori bounds in Theorems 5.1 and 6.1 by using the computable quantities η_z, Λ_z, and $\tilde{\lambda}_z$ defined in (4.2), (5.2), and (6.1), respectively. Moreover, we replace the subscript "h" by an iteration counter "l."

ALGORITHM 6.1. *Let a tolerance* tol > 0*, a marking parameter* $\theta \in \,]0, 1]$*, and an initial triangulation* $\mathcal{T}_0$ *be given. Set* $l := 0$*, and iterate the following steps:*

1. *Compute a minimum point of* $\mathcal{I}(\cdot; \epsilon)$ *in* $\mathcal{K}_l$ *over* $\mathcal{T}_l$*.*
2. *Compute* η_z*,* $\tilde{\lambda}_z$*, and* Λ_z *for* $z \in \mathcal{N}_l \setminus \mathcal{C}_{\epsilon l}$*.*
3. *If* $\sum_{z\in\mathcal{N}_l\setminus\mathcal{C}_{\epsilon l}} \tilde{\lambda}_z^{-1}\eta_z^2 \leq \mathrm{tol}^2$*, then STOP.*
4. *Choose the smallest* $\hat{\mathcal{N}}_l \subset \mathcal{N}_l \setminus \mathcal{C}_{\epsilon l}$ *such that*

$$\sum_{z\in\hat{\mathcal{N}}_l} \Lambda_z^{-1}\eta_z^2 \geq \theta^2 \sum_{z\in\mathcal{N}_l\setminus\mathcal{C}_{\epsilon l}} \Lambda_z^{-1}\eta_z^2.$$

5. *Refine all triangles in* $\mathcal{T}_l$ *with a node in* $\hat{\mathcal{N}}_l$ *to obtain a new triangulation* $\mathcal{T}_{l+1}$ *in such a way that the shape-regularities are bounded uniformly in* l*.*
6. *Increment* l *and go to step* 1*.*

The marking parameter $\theta \in \,]0, 1[$ leads to adaptive refinement, while $\theta = 1$ entails nonadaptive uniform refinement. The stopping test in step 3 is motivated by Theorem 6.1 and may be extended as indicated in Remark 4.2. The aim of the iteration in Algorithm 6.1 is to satisfy the stopping test "as soon as possible," that is, to reduce $\sum_{z\in\mathcal{N}_l\setminus\mathcal{C}_{\epsilon l}} \lambda_z^{-1}\eta_z^2$ efficiently. To this end, one might use $\tilde{\lambda}_z^{-1}\eta_z^2$ also as marking indicators in Dörfler's fixed fraction strategy [9] in step 4. However, this can produce overrefinement in regions where the gradient of the discrete solution is big and, consequently, a slowdown of the convergence speed; see [11, section 7.2]. We therefore replace $\tilde{\lambda}_z^{-2}\eta_z^2$ by $\Lambda_z^{-1}\eta_z^2 = \Lambda_z^{-1}h_z \,\|J_{\epsilon h}\|_{0,2;\gamma_z}^2 + \Lambda_z^{-1}h_z^d \,\|\kappa - \kappa_z\|_{0,d;\omega_z}^2$. The higher order of the second part and Theorem 5.1 guarantee that this indicator does not lead to the aforementioned overrefinement. For the convergence properties of this strategy or a very similar one, we refer to section 6.3 and [11, sections 7.2–7.3].

We conclude this section with some comments on our two-dimensional implementation of Algorithm 6.1. Step 1 is performed with the help of the constrained quasi-Newton method introduced in [6]. In all experiments the tolerance tol > 0 was so small that the stopping test in step 3 was not satisfied if the number of unknowns $N_l = \#\mathcal{N}_l$ was below $50,000$. Following [9] and thus slightly reducing complexity, we construct an approximation of $\hat{\mathcal{N}}_l$ in step 4. In all adaptive experiments we used the marking parameter $\theta = 0.5$. The triangles to be refined in step 5 are bisected by means of the algorithm of Bänsch [4].

6.3. Uniformity in regularization and singular solutions. We present numerical experiments corroborating that the hidden constants in Theorems 4.6, 5.1, and 6.1 do not depend on the regularization ϵ. Moreover, we illustrate that the performance of adaptive refinement is superior to the one of nonadaptive uniform refinement in the presence of singularities. We start by introducing the following example.

Example 6.1 (transition layer). Given $\epsilon_0 > 0$, we consider minimization (1.2) with constant regularization $\epsilon \equiv \epsilon_0$ in $\Omega := \,]-2, 2[\, \times \,]-1, 1[$ and $\kappa(x_1, x_2) = \operatorname{sgn} x_1$ for $(x_1, x_2) \in \Omega$. The unique exact solution does not depend on x_2 and is odd in x_1;

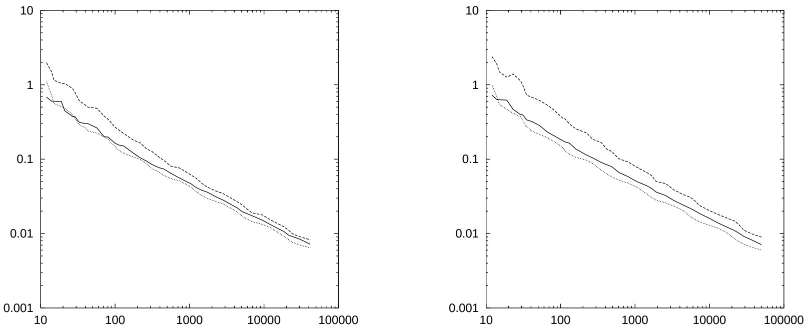


FIG. 6.1. *Example* 6.1 *with* $\epsilon_0 = 2$ *(left) and* $\epsilon_0 = 1.5$ *(right):* $e_{\epsilon l}$ *(solid),* $0.175\,\overline{\xi}_l$ *(dotted), and* $0.175\,\underline{\xi}_l$ *(dashed) as functions of* N_l *in* log-log *scale.*

setting

$$B(\epsilon_0) := \begin{cases} \sqrt{2\epsilon_0 - 1}/\epsilon_0 & \text{if } \epsilon_0 \geq 1, \\ 1 & \text{if } 0 < \epsilon_0 \leq 1, \end{cases}$$

it is given on $]0, 2[\, \times \,]{-1}, 1[$ by

$$u(x_1, x_2; \epsilon) = \begin{cases} 1 & \text{if } x_1 \geq B(\epsilon_0), \\ 1 + \epsilon_0 \Big[\sqrt{1 - \big(B(\epsilon_0) - x\big)^2} - 1 \Big] & \text{if } 0 < x_1 \leq B(\epsilon_0). \end{cases}$$

All experiments concerning Example 6.1 start from the same triangulation $\mathcal{T}_0$. The triangles of $\mathcal{T}_0$ and all its refinements do not cross the line $\{0\} \times \,]{-1}, 1[$, where κ and possibly $u(\cdot;\epsilon)$ have a jump; this facilitates the precise computation of indicators and error. Moreover, $\mathcal{T}_0$ and all its refinements do not exhibit symmetries in x_1 and x_2 corresponding to those of the exact solution $u(\cdot;\epsilon)$.

Providing numerical support that the hidden constants in Theorems 4.6, 5.1, and 6.1 do not depend on the regularization ϵ is complicated by the fact that the error part $|\sigma_{\epsilon h} - \sigma_\epsilon|_{\epsilon h;*}$ is not computable. We propose proceeding as follows. We apply Algorithm 6.1 to Example 6.1 with given ϵ_0, determine *one* constant $C_0 > 0$ such that the computable part $e_{\epsilon l}$ of the error satisfies

$$C_0 \underline{\xi}_l \leq e_{\epsilon l} \ \leq \ C_0 \overline{\xi}_l \tag{6.5}$$

with $\underline{\xi}_l^2 := \sum_{z \in \mathcal{N}_l \setminus \mathcal{C}_{\epsilon l}} \Lambda_z^{-1} \eta_z^2$ and $\overline{\xi}_l^2 := \sum_{z \in \mathcal{N}_l \setminus \mathcal{C}_{\epsilon l}} \tilde{\lambda}_z^{-1} \eta_z^2$ for all l with $N_l \geq 5\,000$, and then investigate whether (6.5) with the determined value of C_0 still holds for other choices of ϵ_0. The requirement of one common constant for both inequalities in (6.5) constrains the possible choices significantly and should enforce a sharp choice in each single inequality. Figure 6.1 (left) shows that $C_0 = 0.175$ is a possible choice for $\epsilon_0 = 2$. Figure 6.1 (right) depicts the case $\epsilon_0 = 1.5$ and reveals the validity of (6.5) with the chosen C_0. We affirmatively tested also other choices of ϵ_0. Since the oscillation of κ is of higher order (in particular, it is only nonzero for nodes in the line $\{0\} \times \,]{-1}, 1[$), the observed validity of the first inequality in (6.5) corroborates that the hidden constant in Theorem 5.1 does not depend on ϵ. Moreover, the observed validity of the second inequality corroborates that also the hidden constant in (4.18) does not depend on ϵ. This in turn supports the same statement for the hidden constants in Theorems 4.6 and 6.1.

Next, we illustrate the superior performance of adaptive refinement for singular solutions. If $\epsilon_0 = 1$ in Example 6.1, then $u(\cdot;\epsilon)$ is singular in the sense that $|\nabla u(x_1, x_2; \epsilon)| \to \infty$ as $x_1 \to 0$ in such a way that $u(\cdot;\epsilon) \notin W_2^1(\Omega)$. If $0 < \epsilon_0 < 1$, then $u(\cdot;\epsilon)$ has even a jump of size $2(1-\epsilon_0)$ across the line $\{0\} \times]-1,1[$, and thus it barely holds that $u(\cdot;\epsilon) \in \mathrm{BV}(\Omega)$. We measure the performance by investigating the relationship of the computable part $e_{\epsilon l}$ and the number of unknowns N_l as l increases. More precisely, we suppose the relationship $e_{\epsilon l} = DN_l^{-p}$ or, equivalently, $\log e_{\epsilon l} = \log D - p \log N_l$, with unknown constants $D, p > 0$, and determine p as slope of the corresponding regression line. The following table displays the (rounded) values of the experimental convergence order p obtained for adaptive ($\theta = 0.5$) and uniform ($\theta = 1$) refinement and regularization parameter $\epsilon_0 \in \{0.5, 1, 1.5\}$; the regression line was determined by using the data corresponding to l with $5000 \leq N_l \lesssim 50,000$ if $\theta = 0.5$ or $4000 \leq N_l \lesssim 50,000$ if $\theta = 1$.

	$\epsilon_0 = 0.5$	$\epsilon_0 = 1$	$\epsilon_0 = 1.5$
$\theta = 0.5$	0.43	0.45	0.51
$\theta = 1$	0.26	0.36	0.50

These values suggest the following. Adaptive refinement offers a significantly higher asymptotic convergence speed in the presence of singularities. In the case of regular solutions, the asymptotic convergence speed is not improved by adaptivity. However, it typically leads to a smaller constant D; see also [11, section 7.3].

6.4. Nonuniqueness.

We study effects of nonuniqueness in the original minimization (1.1) on the convexified and regularized minimization (1.2). In the course of the discussion, we also observe robustness of Algorithm 6.1 with respect to instabilities in the computational minimization of step 1.

The following example was proposed to us by Stefan Luckhaus.

Example 6.2 (a time step of mean curvature flow). Let $\epsilon_0 \geq 0$ and $\tau > 0$, and consider minimization (1.2) with constant regularization $\epsilon \equiv \epsilon_0$ in $\Omega :=]-1,1[^2$ and $\kappa(x_1, x_2) := \frac{1}{\tau}\,\mathrm{sgn}(x_1 x_2) \min\{|x_1|, |x_2|\}$ for $(x_1, x_2) \in \Omega$.

Example 6.2 corresponds to a convexified (and regularized if $\epsilon_0 > 0$) time step with length τ of the mean curvature flow evolving the region $]-1,0[^2 \cup]0,1[^2$; see [2, 10, 13]. The undiscretized flow, which shortens the boundary $\{(x_1,x_2) \in \Omega \mid x_1 x_2 = 0\}$, is not unique; in fact, the evolved region may become connected or not connected. The same holds for the first time step using (1.1), and thus Example 6.2 with $\epsilon_0 = 0$ has several solutions.

We apply Algorithm 6.1 to Example 6.2 with $\epsilon_0 = 0.175$ and $\tau = 0.2$ starting from a symmetric or an “asymmetric” initial triangulation. We shall use the letters “S” and “A” to distinguish the two initial triangulations and their corresponding runs. Triangulations S and A are depicted in the first row of Figure 6.2 on the left- and right-hand sides, respectively.

Note that $\int_\Omega \kappa = 0$ and, consequently, the exact and discrete solutions are unique iff they touch both obstacles. However, numerical integration of κ may not yield 0, and the computed solution is therefore pushed toward one of the obstacles. This is illustrated in Figure 6.3 for run A: on the left-hand side the computed solution touches the upper obstacle, while on the right-hand side it touches the opposite one. In spite of this unstable behavior of the computed solution, the indicators of Algorithm 6.1 yield

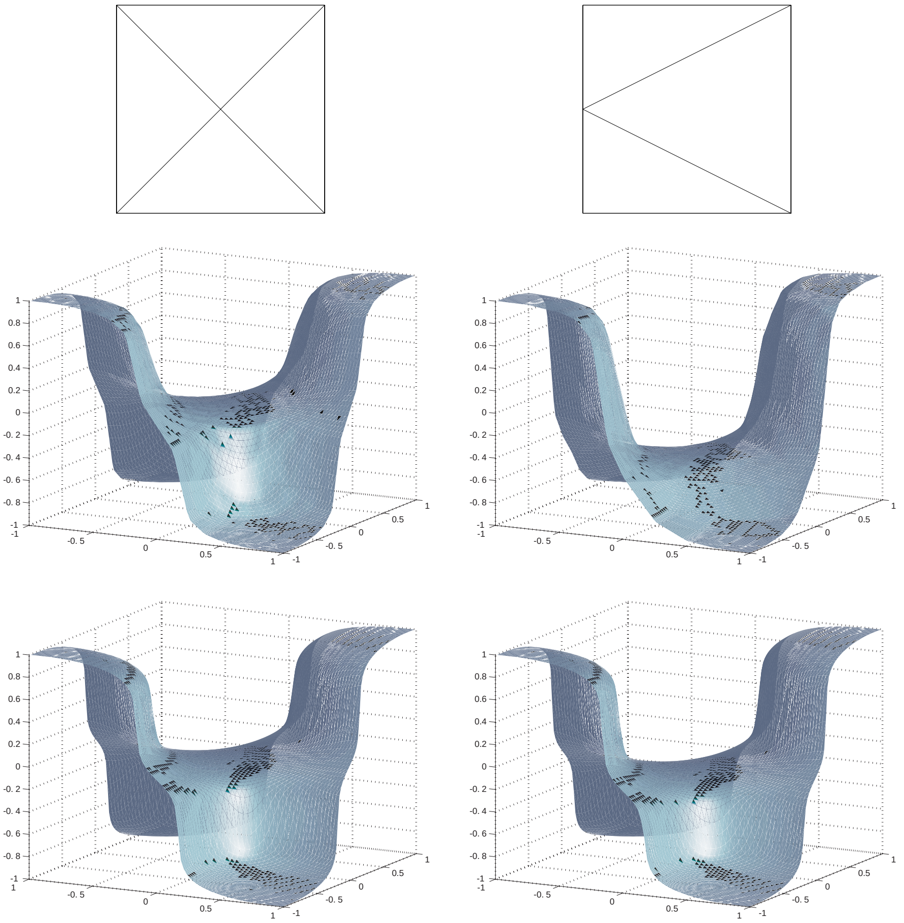


FIG. 6.2. *Example* 6.2 *with* $\epsilon_0 = 0.175$ *and* $\tau = 0.2$*: Initial triangulations (first row), solutions corresponding to* $N_l \approx 560$ *(second row), and* $N_l \approx 50\,000$ *(third row) of run S (left column) and run A (right column).*

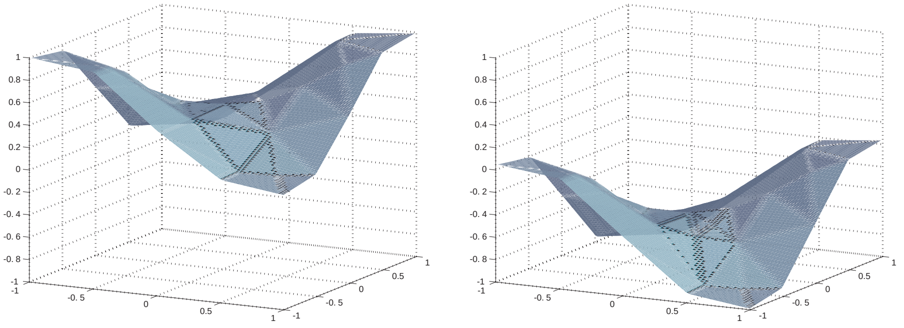


FIG. 6.3. *Example* 6.2 *with* $\epsilon_0 = 0.175$ *and* $\tau = 0.2$*: Unstable computational minimization illustrated by two successive (*$N_l = 33$ *(left) and* $N_l = 37$ *(right)) solutions of run A.*

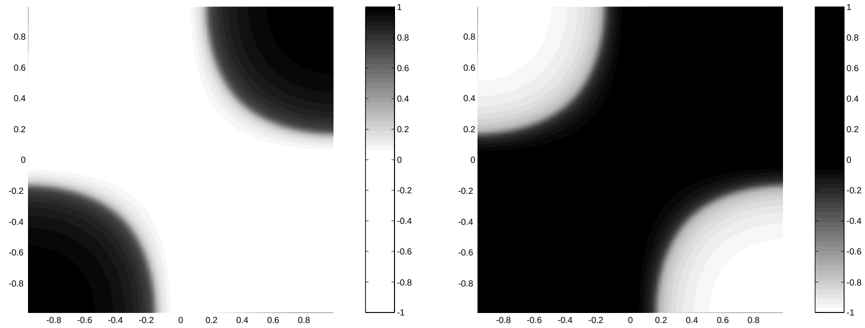


FIG. 6.4. *Example* 6.2 *with* $\epsilon_0 = 0.175$ *and* $\tau = 0.2$*: Contour plot of positive (left) and negative (right) levels of the last solution in run S.*

a reasonable refinement, and, for about 100 nodes, the computed solution touches both obstacles, thus being unique and stable.

For about 560 nodes, the computed solution of run S has a saddle point close to the origin, while the saddle point of the corresponding solution of run A is clearly below the origin; see the second row of Figure 6.2. (The opposite, a saddle point above the origin, can be produced by starting from another "asymmetric" initial triangulation.) However, the last row of Figure 6.2 suggests that this sensibility with respect to the initial triangulation disappears with further adaptive refinement according to Algorithm 6.1.

Let us conclude this subsection with an interpretation of the obtained approximate solutions. To this end, we recall the following statement from [5, section 1.1]: *if* u *is a minimum point of the convexified minimization* (1.2) *with* $\epsilon \equiv 0$*, then, for almost all* $t \in [0, 1]$*, the characteristic function of* $\{x \in \Omega \mid u(x) \geq t\}$ *is a minimum point of the non-convex minimization* (1.1)*.* This fact suggests the following interpretations: the approximate solutions in the last row of Figure 6.2 indicate the two minimum points of the limiting nonconvex minimization (1.1); see Figure 6.4, where positive and negative level lines of the last solution of run S are depicted. The temporary "asymmetry" in run A is a dominance of one minimum point depending on the underlying triangulation.

6.5. Nonconstant regularization. We study an example with nonconstant regularization. Here the meaning of "nonconstant" is twofold: the regularization will depend on space and on the iteration counter "l" in Algorithm 6.1. We choose the following example.

Example 6.3 (quarter of the 2-circle). Consider minimization (1.2) with $\epsilon \equiv 0$ and $\kappa(x) := -|x| + 5/2$ for $x \in \Omega :=]0, 4[^2$. The unique solution is the characteristic function of the sector $\{x \in \Omega \mid |x| \leq 2\}$. In order to approximate $\epsilon \equiv 0$ "dynamically," we use the piecewise constant regularization ϵ_l defined by

$$\forall T \in \mathcal{T}_l, \quad \epsilon_l|_T = \frac{1 + \big||x_T| - 2\big|}{2^{2+l/20}}, \quad \text{where } x_T \text{ is the barycenter of } T,$$

in iteration l of Algorithm 6.1.

We apply Algorithm 6.1 to Example 6.3 starting from an initial triangulation with 41 nodes. In order to measure the quality of the solution u_l computed in iteration l,

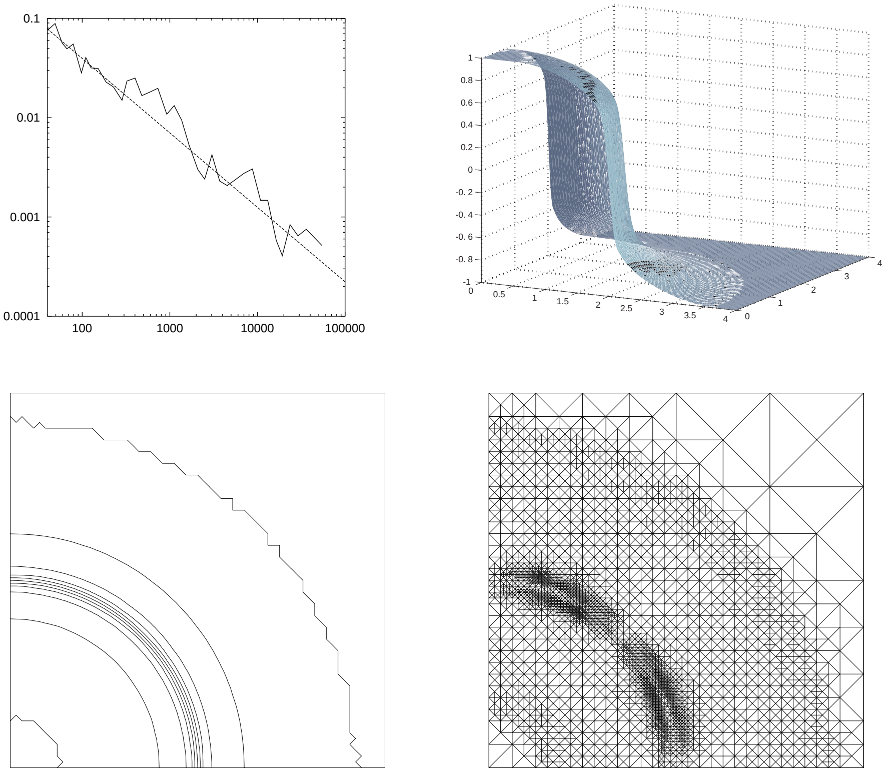


FIG. 6.5. *Example* 6.3: *δ_l as a function of N_l in* log-log *scale accompanied by a line with slope* 0.75 *(top left), solution (top right), its level lines $-1 + 0.2k$, $k = 0, \dots, 10$ (lower left), and the triangulation (lower right) for $N_l = 3042$.*

we introduce

$$\delta_l := \max\Big\{ \big| |x| - 2 \big| \mid x \in \Omega \text{ and } u_l(x) = 0 \Big\},$$

which measures the distance between the "interface" $\Gamma := \{x \in \Omega \mid |x| = 2\}$ of the sector $\{x \in \Omega \mid |x| \leq 2\}$ and its approximation $\Gamma_l := \{x \in \Omega \mid u_l(x) = 0\}$. Figure 6.5 (top left) shows δ_l as a function of $N_l = \#\mathcal{N}_l$ as l increases; the slope of the dashed line in the same picture is 0.75 and represents an approximation of the experimental convergence order. For $N_l = 3024$, the graph and level lines of u_l as well as the triangulation $\mathcal{T}_l$ are also depicted in Figure 6.5. In accordance with Remark 4.4, we observe a very coarse triangulation close to the corner points $(0, 0)$ and $(4, 4)$ due to the complete localization to the discrete non-full-contact set. Inside the discrete noncontact set, the triangulation is still graded. It is relatively coarse off the interface Γ and, with an exceptional thin stripe, relatively fine close to the interface Γ. In light of the rigidity of the used linear finite elements, this may be interpreted as follows: relatively big second derivatives of the exact solution, which are indicated by big jump residuals, require a relatively high density of the degrees of freedom. However, this relationship is affected by the size of the gradient of the exact solution through the weight Λ_z^{-1}, which provides local information on the conditioning. Note that, in the considered situation, incorporating the weight Λ_z^{-1} into the marking indicator intensifies the grading within the discrete noncontact set. This is correct because, as discussed in section 6.2, the marking indicator $\Lambda_z^{-1}\eta_z^2$ does not lead to an overrefinement where the gradient of the solution is big. Finally, note that, in the discrete noncontact set, the mesh is slightly finer close to the free boundaries than away from the interface Γ. This is a consequence of the spatial dependence of the regularization ϵ_l; with constant regularization $\tilde{\epsilon}_l \equiv 2^{-(2+l/20)}$, the triangulation is not finer there.

REFERENCES

[1] M. Ainsworth and J. T. Oden, *A Posteriori Error Estimation in Finite Element Analysis*, John Wiley, New York, 2000.

[2] F. Almgren, J. E. Taylor, and L. Wang, *Curvature-driven flows: A variational approach*, SIAM J. Control Optim., 31 (1993), pp. 387–438.

[3] I. Babuška and A. Miller, *A feedback finite element method with a posteriori error estimation* I. *The finite element method and some basic properties of the a posteriori error estimator*, Comput. Methods Appl. Mech. Engrg., 61 (1987), pp. 1–40.

[4] E. Bänsch, *Local mesh refinement in* 2 *and* 3 *dimensions*, Impact Comput. Sci. Engrg., 3 (1991), pp. 181–191.

[5] G. Bellettini, M. Paolini, and C. Verdi, *Convex approximations of functionals with curvature*, Atti Accad. Naz. Lincei Cl. Sci. Fis. Mat. Natur. Rend. Lincei (9) Mat. Appl., 2 (1991), pp. 297–306.

[6] G. Bellettini, M. Paolini, and C. Verdi, *Numerical minimization of functionals with curvature by convex approximations*, in Progress in Partial Differential Equations: Calculus of Variations, Applications. 1st European Conference on Elliptic and Parabolic Problems, Pitman Res. Notes Math. Ser. 267, C. Bandle et al., eds., Longman Scientific & Technical, Harlow, UK, 1993, pp. 124–138.

[7] C. Carstensen and R. Verfürth, *Edge residuals dominate a posteriori error estimates for low order finite element methods*, SIAM J. Numer. Anal., 36 (1999), pp. 1571–1587.

[8] Z. Chen and R. H. Nochetto, *Residual type a posteriori error estimates for elliptic obstacle problems*, Numer. Math., 84 (2000), pp. 527–548.

[9] W. Dörfler, *A convergent adaptive algorithm for Poisson's equation*, SIAM J. Numer. Anal., 33 (1996), pp. 1106–1124.

[10] F. Fierro, *Numerical approximation of the mean curvature flow with nucleation using implicit time-stepping: An adaptive algorithm*, Calcolo, 35 (1998), pp. 205–224.

[11] F. Fierro and A. Veeser, *On the a posteriori error analysis for equations of prescribed mean curvature*, Math. Comp., 72 (2003), pp. 1611–1635.

[12] R. Finn, *Equilibrium Capillary Surfaces*, Grundlehren Math. Wiss. 284, Springer-Verlag, New York, 1986.

[13] S. Luckhaus and T. Sturzenhecker, *Implicit time discretization for the mean curvature flow equation*, Calc. Var. Partial Differential Equations, 3 (1995), pp. 253–271.

[14] P. Morin, R. H. Nochetto, and K. G. Siebert, *Local problems on stars: A posteriori error estimators, convergence, and performance*, Math. Comp., 72 (2003), pp. 1067–1097.

[15] R. H. Nochetto, K. G. Siebert, and A. Veeser, *Pointwise a posteriori error control for elliptic obstacle problems*, Numer. Math., 95 (2003), pp. 163–195.

[16] L. R. Scott and S. Zhang, *Finite element interpolation of nonsmooth functions satisfying boundary conditions*, Math. Comp., 54 (1990), pp. 483–493.

[17] A. Veeser, *Efficient and reliable a posteriori error estimators for elliptic obstacle problems*, SIAM J. Numer. Anal., 39 (2001), pp. 146–167.

[18] A. Veeser, *On a posteriori error estimation for constant obstacle problems*, in Numerical Methods for Viscosity Solutions and Applications, Ser. Adv. Math. Appl. Sci. 59, M. Falcone and C. Makridakis, eds., World Scientific Publishing, River Edge, NJ, 2001, pp. 221–234.

[19] R. Verfürth, *A Review of A Posteriori Error Estimation and Adaptive Mesh-Refinement Techniques*, Adv. Numer. Math., John Wiley, Chichester, UK, 1996.

[20] A. Visintin, *Models of Phase Transitions*, Progress in Nonlinear Differential Equations and Their Applications 28, Birkhäuser Boston, Boston, 1996.

SIAM J. Numer. Anal.
Vol. 41, No. 6, pp. 2056–2080

GENERALIZATIONS AND ACCELERATIONS OF LIONS' NONOVERLAPPING DOMAIN DECOMPOSITION METHOD FOR LINEAR ELLIPTIC PDE[*]

W. GUO[†] AND L. S. HOU[‡]

Abstract. In this paper, a one-parameter generalization of Lions' nonoverlapping domain decomposition method for linear elliptic PDEs is proposed and studied. The generalized methods are shown to be descent-direction methods for minimizing an interface bias functional. Iteration convergence of both the continuous and finite element versions of the proposed methods is established. It is theoretically and numerically demonstrated that for generic choices of the parameter the generalized methods converge faster than Lions' original method. Algorithms are given and numerical results are presented.

Key words. domain decomposition, Lions' nonoverlapping method, optimization-based domain decomposition, convergence acceleration, descent-direction method, finite element, parallel computation

AMS subject classifications. 65N55, 65N30, 65Y10, 35J20, 65K10

DOI. 10.1137/S0036142902407150

1. Introduction.

Domain decompositon methods (DDMs) have been a flourishing area of research in scientific computing in the last two decades; see, e.g., the proceedings or monographs [11, 12, 13, 14, 24, 26, 27, 33, 50] and the web site http://www.ddm.org. The interests and research efforts in this subject have continued to expand in recent years; see, e.g., [8, 9, 15, 17, 20, 21, 28, 30, 32, 34, 42, 44, 47, 48, 49]. The various DDMs can be loosely classified into two categories based on the decomposition of subdomains: overlapping and nonoverlapping. The classic Schwarz alternating method [43, 25, 35], which is based on successive exchanges of Dirichlet data, is a century-old example of overlapping DDMs and is still an active topic of research [37, 38, 39]. The history of nonoverlapping DDMs is much shorter than that of overlapping ones. Early works on nonoverlapping DDMs include [2, 3, 4, 5, 6, 7, 10, 18]. One of the most popular nonoverlapping DDMs is the well-known Lions' method [36], which is based on successive exchanges of interface Robin data. The objective of this paper is to design and analyze a one-parameter generalization of Lions' nonoverlapping method for solutions of linear elliptic partial differential equations (PDEs) and establish the acceleration properties of the generalized methods.

Though our generalized method at each iteration step appears to simply involve a weighted average of Lions' interface updates and the Robin data of the previous iteration step, the analysis of those methods will be based on viewing Lions' method as a descent-direction method with a fixed step length for minimizing an interface bias functional. The accelerations are achieved by choosing suitable variable step lengths in the descent-direction method. In the sense of minimizing an interface bias functional our methods are akin to the optimization-based DDMs studied in [20, 21, 28, 30] (see also the earlier works [18, 25] and a relevant reference [45]). However, our methods

[*] Received by the editors May 4, 2002; accepted for publication (in revised form) February 27, 2003; published electronically November 25, 2003.
http://www.siam.org/journals/sinum/41-6/40715.html

[†] Eban Commerce Inc., 200A-219 Dufferin St., Toronto, Ontario, M6K 3J1 Canada (wayne@centraldeposit.com).

[‡] Department of Mathematics, Iowa State University, Ames, IA 50011 (hou@math.iastate.edu).

make use of a descent direction which in general is not the gradient direction; this is different from the work of [20, 21, 28, 30], which used gradient methods or optimality systems to find the minima of interface bias functionals.

Noteworthy features of the generalized methods include (i) true parallelism in the sense that each new iterate makes use of all subdomain solutions obtained in the previous step and thus calls for multiple processors (whereas Lions' method is an alternating domain method); (ii) $L^2(\Gamma)$ norm convergence for the interface bias of the Robin data (this is an improvement over the well-known $H^{-1/2}(\Gamma)$ norm convergence result for Lions' method—see [17, 20]); (iii) an explicit range of acceptable step lengths that is independent of the underlying elliptic operator (in contrast, the range of step lengths for the gradient methods typically depends on the spectrum of the elliptic operator—see [21]).

This paper is organized as follows. In section 2 we describe Lions' nonoverlapping method and define its generalizations for a linear elliptic problem. In section 3 we study the properties of certain operators associated with Robin boundary value problems, prove that Lions' method and the generalized methods are descent-direction algorithms for minimizing an interface bias functional, and demontstrate the convergence of the generalized methods. In section 4 we establish the convergence and acceleration properties of the finite element versions of the generalized methods. Finally, in section 5 we propose two algorithms (one with locally optimal step lengths and the other with a fixed step length) and present the results of numerical experiments.

2. Lions' DDM and its generalizations. In this section we review Lions' nonoverlapping DDM and define our generalizations of that method.

We consider the following linear elliptic PDE with a homogeneous boundary condition:

$$-\operatorname{div}[A(\mathbf{x})\nabla u] = f \quad \text{in } \Omega, \qquad u = 0 \quad \text{on } \partial\Omega, \tag{2.1}$$

where Ω is a two- or three-dimensional Lipschitz domain, f is a given function in $L^2(\Omega)$, and A is a symmetric-matrix-valued $C^1(\overline{\Omega})$ function that is uniformly positive definite. For simplicity and clarity of exposition we will describe Lions' method and our generalized methods on a partition of Ω into two disjoint subdomains Ω_1 and Ω_2. The extensions of these methods to partitions of Ω into multiple subdomains are straightforward [17].

We denote $\Gamma = \overline{\Omega_1} \cap \overline{\Omega_2}$ and $\Gamma_i = \partial\Omega_i \setminus \Gamma$, $i = 1, 2$. Let $\mathbf{n}_i$ be the unit outward normal to $\partial\Omega_i$ along the interface Γ (see Figure 2.1).

In a nonoverlapping DDM for (2.1) one solves

$$-\operatorname{div}[A(\mathbf{x})\nabla u] = f \quad \text{in } \Omega_i, \qquad u|_{\Gamma_i} = 0 \tag{2.2}$$

for $i = 1, 2$ separately with certain boundary conditions on the interface Γ. The boundary conditions on Γ are updated iteratively and (2.2) is solved repeatedly until convergence. Iterative boundary conditions on Γ are often designed based on an equivalent form of the transmission conditions:

$$u_1 - u_2 = 0 \quad \text{on } \Gamma \qquad \text{and} \qquad [A(\mathbf{x})\nabla u_1]\cdot\mathbf{n}_1 + [A(\mathbf{x})\nabla u_2]\cdot\mathbf{n}_2 = 0 \quad \text{on } \Gamma. \tag{2.3}$$

The well-known alternating domain methods [14, 19, 22, 40, 41] can be viewed as examples of iterations that arise from (2.3).

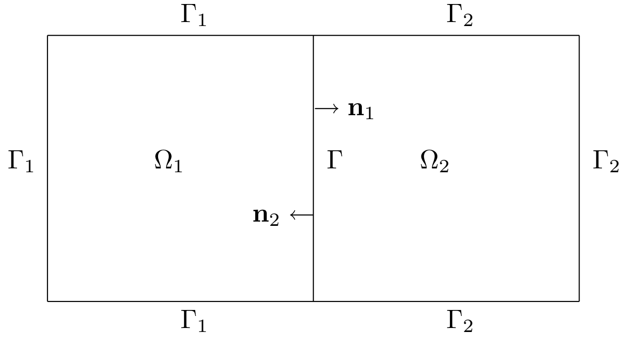


FIG. 2.1.

It can be easily checked that (2.3) is equivalent to the following Robin-type transmission conditions:

$$\begin{aligned} &u_1 + \lambda[A(\mathbf{x})\nabla u_1]\cdot\mathbf{n}_1 = u_2 - \lambda[A(\mathbf{x})\nabla u_2]\cdot\mathbf{n}_2 \quad \text{on } \Gamma,\\ &u_2 + \lambda[A(\mathbf{x})\nabla u_2]\cdot\mathbf{n}_2 = u_1 - \lambda[A(\mathbf{x})\nabla u_1]\cdot\mathbf{n}_1 \quad \text{on } \Gamma, \end{aligned} \tag{2.4}$$

where $\lambda > 0$ is a constant. An iterative implementation of (2.4) yields the following Lions' method: choose initial guesses $u_1^{(1)}$ in Ω_1 and $u_2^{(1)}$ in Ω_2; for $k = 1, 2, 3, \dots,$ solve

$$\left\{ \begin{aligned} &-\operatorname{div}[A(\mathbf{x})\nabla u_1^{(k+1)}] = f \quad \text{in } \Omega_1\,, \qquad u_1^{(k+1)} = 0 \quad \text{on } \Gamma_1,\\ &u_1^{(k+1)} + \lambda[A(\mathbf{x})\nabla u_1^{(k+1)}]\cdot\mathbf{n}_1 = u_2^{(k)} - \lambda[A(\mathbf{x})\nabla u_2^{(k)}]\cdot\mathbf{n}_2 \quad \text{on } \Gamma \end{aligned} \right. \tag{2.5}$$

and

$$\left\{ \begin{aligned} &-\operatorname{div}[A(\mathbf{x})\nabla u_2^{(k+1)}] = f \quad \text{in } \Omega_2\,, \qquad u_2^{(k+1)} = 0 \quad \text{on } \Gamma_2\,,\\ &u_2^{(k+1)} + \lambda[A(\mathbf{x})\nabla u_2^{(k+1)}]\cdot\mathbf{n}_2 = u_1^{(k)} - \lambda[A(\mathbf{x})\nabla u_1^{(k)}]\cdot\mathbf{n}_1 \quad \text{on } \Gamma\,. \end{aligned} \right. \tag{2.6}$$

By setting

$$g_i^{(k)} = u_i^{(k)} + \lambda[A(\mathbf{x})\nabla u_i^{(k)}]\cdot\mathbf{n}_i\,, \qquad i = 1, 2\,, \tag{2.7}$$

Lions' iterations (2.5)–(2.6) were recast into the following form that is more amenable to implementations due to the avoidance of normal derivative calculations (see [17]): choose initial guesses $g_1^{(1)}$ on Γ and $g_2^{(1)}$ on Γ; for $k = 1, 2, 3, \dots,$ solve

$$\left\{ \begin{aligned} &-\operatorname{div}[A(\mathbf{x})\nabla u_i^{(k)}] = f \quad \text{in } \Omega_i\,,\\ &u_i^{(k)} = 0 \quad \text{on } \Gamma_i\,, \qquad u_i^{(k)} + \lambda[A(\mathbf{x})\nabla u_i^{(k)}]\cdot\mathbf{n}_i = g_i^{(k)} \quad \text{on } \Gamma, \quad i = 1, 2, \end{aligned} \right. \tag{2.8}$$

and update

$$g_1^{(k+1)} = 2u_2^{(k)} - g_2^{(k)}\,, \qquad g_2^{(k+1)} = 2u_1^{(k)} - g_1^{(k)}\,. \tag{2.9}$$

Lions' method (2.5)–(2.6) was proposed in [36] with a proof for its convergence. The equivalent variant (2.8)–(2.9) along with finite element approximations was studied in [17]. Lions' method (2.5)–(2.6) and its variant (2.8)–(2.9) were also derived in

[20] through an optimization approach. An overlapping version of Lions' method was studied in [46].

We define a one-parameter generalization of Lions' method as follows: choose initial guesses $u_1^{(1)}$ in Ω_1 and $u_2^{(1)}$ in Ω_2; for $k = 1, 2, 3, \dots$, solve

$$\left\{\begin{aligned} &-\mathrm{div}\,[A(\mathbf{x})\nabla u_1^{(k+1)}] = f \quad \text{in } \Omega_1\,, \qquad u_1^{(k+1)} = 0 \quad \text{on } \Gamma_1\,,\\ &u_1^{(k+1)} + \lambda[A(\mathbf{x})\nabla u_1^{(k+1)}]\cdot \mathbf{n}_1\\ &= (1-\delta_k)\big(u_1^{(k)} + \lambda[A(\mathbf{x})\nabla u_1^{(k)}]\cdot \mathbf{n}_1\big) + \delta_k\big(u_2^{(k)} - \lambda[A(\mathbf{x})\nabla u_2^{(k)}]\cdot \mathbf{n}_2\big) \text{ on } \Gamma \end{aligned}\right. \tag{2.10}$$

and

$$\left\{\begin{aligned} &-\mathrm{div}\,[A(\mathbf{x})\nabla u_2^{(k+1)}] = f \quad \text{in } \Omega_2\,, \qquad u_2^{(k+1)} = 0 \quad \text{on } \Gamma_2\,,\\ &u_2^{(k+1)} + \lambda[A(\mathbf{x})\nabla u_2^{(k+1)}]\cdot \mathbf{n}_2\\ &= (1-\delta_k)\big(u_2^{(k)} + \lambda[A(\mathbf{x})\nabla u_2^{(k)}]\cdot \mathbf{n}_2\big) + \delta_k\big(u_1^{(k)} - \lambda[A(\mathbf{x})\nabla u_1^{(k)}]\cdot \mathbf{n}_1\big) \text{ on } \Gamma, \end{aligned}\right. \tag{2.11}$$

where $\{\delta_k\} \subset [\delta_{\min}, \delta_{\max}] \subset (0,1]$. If $\delta_k = 1$ for all k, then (2.10)–(2.11) reduce to Lions' method (2.5)–(2.6). Also, by introducing $g_i^{(k)}$ as defined by (2.7), we may recast (2.10)–(2.11) into the following form that avoids the calculations of normal derivatives: choose initial guesses $g_1^{(1)}$ on Γ and $g_2^{(1)}$ on Γ; for $k = 1, 2, 3, \dots$, solve

$$\left\{\begin{aligned} &-\mathrm{div}\,[A(\mathbf{x})\nabla u_i^{(k)}] = f \quad \text{in } \Omega_i\,,\\ &u_i^{(k)} = 0 \quad \text{on } \Gamma_i\,, \quad u_i^{(k)} + \lambda[A(\mathbf{x})\nabla u_i^{(k)}]\cdot \mathbf{n}_i = g_i^{(k)} \quad \text{on } \Gamma, \quad i = 1, 2, \end{aligned}\right. \tag{2.12}$$

and update

$$\left\{\begin{aligned} g_1^{(k+1)} &= (1-\delta_k) g_1^{(k)} + \delta_k (2u_2^{(k)} - g_2^{(k)})\,,\\ g_2^{(k+1)} &= (1-\delta_k) g_2^{(k)} + \delta_k (2u_1^{(k)} - g_1^{(k)})\,. \end{aligned}\right. \tag{2.13}$$

Remark 2.1. An observation of Lions' method is that it is not a truly parallel method. In fact, the sequence of iterative solutions

$$\{(u_1^{(1)}, u_2^{(1)}), (u_1^{(2)}, u_2^{(2)}), (u_1^{(3)}, u_2^{(3)}), \dots\}$$

on the two subdomains can be separated into two independent sequences:

$$\{u_1^{(1)}, u_2^{(2)}, u_1^{(3)}, u_2^{(4)}, \dots\}\,, \qquad \{u_2^{(1)}, u_1^{(2)}, u_2^{(3)}, u_1^{(4)}, \dots\},$$

each consisting of iterates of solutions alternating on the two subdomains. One needs only to compute one of the two separate sequences of solutions; thus at each iteration, only one processor is needed. In contrast, the generalized method when $\delta_k \in (0,1)$ naturally calls for two processors at each iteration with the single iteration sequence $\{(u_1^{(1)}, u_2^{(1)}), (u_1^{(2)}, u_2^{(2)}), (u_1^{(3)}, u_2^{(3)}), (u_1^{(4)}, u_2^{(4)}), \dots\}$.

To analyze the generalized method (2.12)–(2.13) in a proper mathematical framework, we introduce some notation and recast Lions' method and the generalized methods into weak formulations.

Let $H^s(\mathcal{D})$ denote the standard Sobolev space of order s on a set $\mathcal{D}$ with the norm $\|\cdot\|_{s,\mathcal{D}}$. Vector-valued Sobolev spaces are denoted by $\mathbf{H}^s(\mathcal{D})$ with the norm still denoted by $\|\cdot\|_{s,\mathcal{D}}$. Of course, $H^0(\mathcal{D}) = L^2(\mathcal{D})$ and $\mathbf{H}^0(\mathcal{D}) = \mathbf{L}^2(\mathcal{D})$. We denote the $L^2(\mathcal{D})$- and $\mathbf{L}^2(\mathcal{D})$-inner products by $[\cdot,\cdot]_{\mathcal{D}}$, i.e.,

$$[u, v]_{\mathcal{D}} = \int_{\mathcal{D}} u\, v \, d\mathcal{D} \quad \forall\, u, v \in L^2(\mathcal{D}) \quad \text{and} \quad [\mathbf{u}, \mathbf{v}]_{\mathcal{D}} = \int_{\mathcal{D}} \mathbf{u}\cdot\mathbf{v}\, d\mathcal{D} \quad \forall\, \mathbf{u}, \mathbf{v} \in \mathbf{L}^2(\mathcal{D}).$$

Also, we use the standard notation for the space $H^1_0(\Omega) = \{v \in H^1(\Omega) \mid v = 0 \text{ on } \partial\Omega\}$.

The weak formulation for the elliptic boundary value problem (2.1) is given by

$$a[u, v] = [f, v]_\Omega \qquad \forall\, v \in H^1_0(\Omega), \tag{2.14}$$

where the bilinear form $a[\cdot,\cdot]$ is defined by

$$a[u, v] = \int_\Omega A(\mathbf{x})\nabla u \cdot \nabla v \, d\mathbf{x} \qquad \forall\, u, v \in H^1(\Omega).$$

For $i = 1, 2$ we define the space $X_i = \{v \in H^1(\Omega_i) \mid v = 0 \text{ on } \Gamma_i\}$ equipped with the norm $\|\cdot\|_{X_i} = \|\cdot\|_{1,\Omega_i}$. We also introduce the subdomain bilinear forms

$$a_i[u, v] = \int_{\Omega_i} A(\mathbf{x})\nabla u \cdot \nabla v \, d\mathbf{x} \qquad \forall\, u, v \in X_i,\ i = 1, 2.$$

In terms of weak formulations Lions' method (2.8)–(2.9) can be stated as follows: choose initial guesses $g_1^{(k)} \in L^2(\Gamma)$ and $g_2^{(k)} \in L^2(\Gamma)$; for $k = 1, 2, 3, \dots$, solve for $u_i^{(k)} \in X_i$ $(i = 1, 2)$ from

$$a_i[u_i^{(k)}, v_i] + \lambda^{-1}[u_i^{(k)}, v_i]_\Gamma = [f, v_i]_{\Omega_i} + \lambda^{-1}[g_i^{(k)}, v_i]_\Gamma \qquad \forall\, v_i \in X_i\,,\ i = 1, 2, \tag{2.15}$$

and update

$$g_1^{(k+1)} = 2u_2^{(k)} - g_2^{(k)} \qquad \text{and} \qquad g_2^{(k+1)} = 2u_1^{(k)} - g_1^{(k)}\,. \tag{2.16}$$

The weak formulation of the generalized method (2.12)–(2.13) is as follows: choose initial guesses $g_1^{(1)} \in L^2(\Gamma)$ and $g_2^{(1)} \in L^2(\Gamma)$; for $k = 1, 2, 3, \dots$, solve for $u_i^{(k)} \in X_i$ $(i = 1, 2)$ from

$$a_i[u_i^{(k)}, v_i] + \lambda^{-1}[u_i^{(k)}, v_i]_\Gamma = [f, v_i]_{\Omega_i} + \lambda^{-1}[g_i^{(k)}, v_i]_\Gamma \qquad \forall\, v_i \in X_i\,,\ i = 1, 2, \tag{2.17}$$

and update

$$\begin{cases} g_1^{(k+1)} = (1-\delta_k)g_1^{(k)} + \delta_k(2u_2^{(k)} - g_2^{(k)})\,, \\ g_2^{(k+1)} = (1-\delta_k)g_2^{(k)} + \delta_k(2u_1^{(k)} - g_1^{(k)})\,. \end{cases} \tag{2.18}$$

We will reveal that Lions' method and the generalized methods are descent-direction algorithms for minimizing the interface bias functional

$$\frac{1}{2}\int_\Gamma \Big(|u_1 - u_2|^2 + \lambda^2 \Big| [A(\mathbf{x})\nabla u_1]\cdot\mathbf{n}_1 + [A(\mathbf{x})\nabla u_2]\cdot\mathbf{n}_2 \Big|^2 \Big), \tag{2.19}$$

where $u_i \in X_i$ $(i = 1, 2)$ are solutions of the subdomain Robin boundary value problem

$$a_i[u_i, v_i] + \lambda^{-1}[u_i, v_i]_\Gamma = [f, v_i]_{\Omega_i} + \lambda^{-1}[g_i, v_i]_\Gamma \qquad \forall\, v_i \in X_i\,,\ i = 1, 2. \tag{2.20}$$

Obviously, (2.19) attains the minimum value 0 when $u_1 = \widehat{u}|_{\Omega_1}$ and $u_2 = \widehat{u}|_{\Omega_2}$, where $\widehat{u} \in H_0^1(\Omega)$ is the solution of the problem (2.1) or (2.14) on the entire domain Ω. We will show that the generalized method (2.17)–(2.18) converges to the minimum of the functional (2.19); i.e., the subdomain solutions defined by (2.17)–(2.18) converges to the global solution $\widehat{u}$. In the course of the convergence proofs, the acceleration properties of the generalized methods will become clear.

3. Convergence of the generalized method. In this section we will first study some properties of a Robin–Robin map (which maps a Robin-type boundary value on Γ into another Robin-type boundary value) and then prove the convergence of iterations (2.17)–(2.18).

3.1. Solution operators for the Robin boundary value problems. It is well known that Robin-type boundary value problems on subdomains (for a fixed $\lambda > 0$)

$$\begin{cases} -\mathrm{div}\,[A(\mathbf{x})\nabla u_i] = f \quad \text{in } \Omega_i, \\ u_i = 0 \quad \text{on } \Gamma_i\,, \qquad u_i + \lambda[A(\mathbf{x})\nabla u_i]\cdot \mathbf{n}_i = g_i \quad \text{on } \Gamma, \end{cases} \tag{3.1}$$

$i = 1, 2$, admit a unique solution in the sense of the following weak formulation: seek a $u_i \in X_i$, $i = 1, 2$, such that

$$a_i[u_i, v_i] + \lambda^{-1}[u_i, v_i]_\Gamma = [f, v_i]_{\Omega_i} + \lambda^{-1}[g_i, v_i]_\Gamma \qquad \forall\, v_i \in X_i\,. \tag{3.2}$$

For $i = 1, 2$ we denote by $S_i^f : L^2(\Gamma) \to X_i$ the solution operator for the Robin boundary value problem (3.2); i.e., $u_i = S_i^f g_i$ for $g_i \in L^2(\Gamma)$ if and only if u_i and g_i satisfy (3.2). We define the operator $S^f : \mathbf{L}^2(\Gamma) \to X_1 \times X_2$ by $S^f\mathbf{g} = (S_1^f g_1, S_2^f g_2)$ for all $\mathbf{g} = (g_1, g_2) \in \mathbf{L}^2(\Gamma)$. If $f = 0$, we write S_i^0 and S^0 in place of S_i^f and S^f, respectively.

We also denote by $T : L^2(\Omega) \to X_1 \times X_2$ the solution operator for (3.2) with homogeneous Robin boundary condition; i.e., for every $f \in L^2(\Omega)$, $(u_1, u_2) = Tf = (T_1 f, T_2 f)$ if and only if $(u_1, u_2) \in X_1 \times X_2$ is the solution of

$$a_i[u_i, v_i] + \lambda^{-1}[u_i, v_i]_\Gamma = [f, v_i]_{\Omega_i} \qquad \forall\, v_i \in X_i\,,\ i = 1, 2. \tag{3.3}$$

The operator T is obviously linear.

LEMMA 3.1. *If* $\mathbf{g}, \widetilde{\mathbf{g}} \in \mathbf{L}^2(\Gamma)$ *and* c_1, c_2 *are constants, then*

(a) $S^0(c_1\mathbf{g} + c_2\widetilde{\mathbf{g}}) = c_1 S^0\mathbf{g} + c_2 S^0\widetilde{\mathbf{g}}$;

(b) $S^f = S^0 + Tf$;

(c) $S^0(\mathbf{g} - \widetilde{\mathbf{g}}) = S^f\mathbf{g} - S^f\widetilde{\mathbf{g}}$.

Proof. By definition, the operator S^0 is defined as follows: for every $\mathbf{g} = (g_1, g_2) \in \mathbf{L}^2(\Gamma)$, $(u_1, u_2) = (S_1^0 g_1, S_2^0 g_2)$ if and only if

$$a_i[u_i, v_i] + \lambda^{-1}[u_i, v_i]_\Gamma = \lambda^{-1}[g_i, v_i]_\Gamma \qquad \forall\, v_i \in X_i. \tag{3.4}$$

Thus it is obvious that S_i^0 and S^0 are linear operators and (a) holds.

The solution of (3.2) evidently can be split into the sum of the solution for (3.3) and the solution for (3.4). In other words, $S^f\mathbf{g} = Tf + S^0\mathbf{g}$ for every $\mathbf{g} \in \mathbf{L}^2(\Gamma)$. This proves (b).

As an easy consequence of (b) and (a), we obtain (c):

$$S^f\mathbf{g} - S^f\widetilde{\mathbf{g}} = S^0\mathbf{g} + Tf - S^0\widetilde{\mathbf{g}} - Tf = S^0(\mathbf{g} - \widetilde{\mathbf{g}}). \qquad \square$$

3.2. The Robin–Robin map and its basic properties. We define the Robin–Robin map $R^f : \mathbf{L}^2(\Gamma) \to \mathbf{L}^2(\Gamma)$ as follows. For any $\mathbf{g} = (g_1, g_2) \in \mathbf{L}^2(\Gamma)$,

$$R^f \mathbf{g} = (R_1^f \mathbf{g}, R_2^f \mathbf{g}) \equiv \Big(2(S_2^f g_2)|_\Gamma - g_2 \, , 2(S_1^f g_1)|_\Gamma - g_1 \Big). \tag{3.5}$$

If $f = 0$, we write R^0 in place of R^f.

We denote by $\widehat{u}$ the unique exact solution of (2.14) in $H_0^1(\Omega)$. Following the proofs of [23, Theorem I.2.5 and Corollary I.2.6] we may justify that $[A(\mathbf{x})\nabla \widehat{u}] \cdot \mathbf{n}_i|_\Gamma \in H^{-1/2}(\Gamma)$. We make the regularity assumption

$$[A(\mathbf{x})\nabla \widehat{u}_i] \cdot \mathbf{n}_i|_\Gamma \in L^2(\Gamma), \qquad i = 1, 2, \tag{3.6}$$

so that

$$a_i[\widehat{u}_i, v_i] = [f, v_i]_{\Omega_i} + [(A(\mathbf{x})\nabla \widehat{u}_i) \cdot \mathbf{n}_i, v_i]_\Gamma \qquad \forall\, v_i \in X_i\,. \tag{3.7}$$

We define $\widehat{u}_i$ and $\widehat{g}_i$ by

$$\widehat{u}_i = \widehat{u}|_{\Omega_i}, \qquad \widehat{g}_i = \widehat{u}_i + \lambda [A(\mathbf{x})\nabla \widehat{u}_i] \cdot \mathbf{n}_i\,, \qquad i = 1, 2. \tag{3.8}$$

Of course, $\widehat{u}_1$ and $\widehat{u}_2$ satisfy the transmission conditions

$$\widehat{u}_1|_\Gamma = \widehat{u}_2|_\Gamma \qquad \text{and} \qquad [A(\mathbf{x})\nabla \widehat{u}_1] \cdot \mathbf{n}_1 = -[A(\mathbf{x})\nabla \widehat{u}_2] \cdot \mathbf{n}_2 \quad \text{on } \Gamma. \tag{3.9}$$

Also, (3.7) implies

$$a_i[\widehat{u}_i, v_i] + \lambda^{-1}[\widehat{u}_i, v_i]_\Gamma = [f, v_i]_{\Omega_i} + \lambda^{-1}[\widehat{g}_i, v_i]_\Gamma \qquad \forall\, v_i \in X_i\,. \tag{3.10}$$

It is easily verified that Lions' iterations (2.15)–(2.16) can be simply written as fixed point iterations

$$\begin{cases} \mathbf{g}^{(1)} \text{ given,} \\ \mathbf{g}^{(k+1)} = R^f \mathbf{g}^{(k)}, \qquad k = 1, 2, 3, \dots. \end{cases}$$

It is well known that Lions' iterations converge to the exact solution $\widehat{u}$ (see [17, 36]). Thus we expect $\widehat{\mathbf{g}} = (\widehat{g}_1, \widehat{g}_2)$ defined by (3.8) to be a fixed point of R^f, as will be proved in the following lemma.

LEMMA 3.2. $\mathbf{g} \in \mathbf{L}^2(\Gamma)$ *satisfies* $R^f \mathbf{g} = \mathbf{g}$ *if and only if* $\mathbf{g} = \widehat{\mathbf{g}}$.

Proof. If $\mathbf{g} = \widehat{\mathbf{g}}$, using definition (3.8) and the transmission conditions (3.9), we have

$$\begin{aligned} R^f \widehat{\mathbf{g}} &= (2\widehat{u}_2 - \widehat{g}_2 \, , 2\widehat{u}_1 - \widehat{g}_1) = \Big(\widehat{u}_2 - \lambda[A(\mathbf{x})\nabla \widehat{u}_2] \cdot \mathbf{n}_2 \, , \widehat{u}_1 - \lambda[A(\mathbf{x})\nabla \widehat{u}_1] \cdot \mathbf{n}_1 \Big) \\ &= \Big(\widehat{u}_1 + \lambda[A(\mathbf{x})\nabla \widehat{u}_1] \cdot \mathbf{n}_1 \, , \widehat{u}_2 + \lambda[A(\mathbf{x})\nabla \widehat{u}_2] \cdot \mathbf{n}_2 \Big) = \widehat{\mathbf{g}}\,. \end{aligned}$$

To prove the converse, we assume that $\mathbf{g} \in \mathbf{L}^2(\Gamma)$ satisfies $R^f \mathbf{g} = \mathbf{g}$, i.e.,

$$2u_2 - g_2 = g_1 \quad \text{and} \quad 2u_1 - g_1 = g_2 \quad \text{on } \Gamma, \tag{3.11}$$

where $u_i = S_i^f g_i$, $i = 1, 2$. Subtracting the two equations of (3.11), we immediately obtain

$$u_1 = u_2 \quad \text{on } \Gamma. \tag{3.12}$$

Thus the function $u \in L^2(\Omega)$ defined by

$$u = \begin{cases} u_1 & \text{in } \Omega_1, \\ \\ u_2 & \text{in } \Omega_2 \end{cases}$$

satisfies $u \in H_0^1(\Omega)$. By definition $u_i = S_i^f g_i$ is determined by (3.2). Let an arbitrary $v \in H_0^1(\Omega)$ be given. Setting $v_i = v|_{\Omega_i}$ in (3.2) for $i = 1, 2$ and adding the two equations, we obtain

$$a[u, v] = [f, v]_\Omega + \lambda^{-1}[g_1 + g_2 - u_1 - u_2, v]_\Gamma \,.$$

Substituting (3.11) into the last equation, we are led to

$$a[u, v] = [f, v]_\Omega \qquad \forall\, v \in H_0^1(\Omega)\,.$$

Thus $u = \widehat{u}$. Then a comparison of (3.2) (now $u = \widehat{u}$) and (3.10) yields $(g_1, g_2) = (\widehat{g}_1, \widehat{g}_2)$. □

We next derive some basic properties of R^f. We easily see that R^f and R^0 satisfy the following relations in terms of the operators S^f and T:

$$R^0\mathbf{g} = \Big(2(S^0\mathbf{g})|_\Gamma - \mathbf{g}\Big)\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \tag{3.13}$$

and

$$\begin{aligned} R^f\mathbf{g} &= \Big(2(S^f\mathbf{g})|_\Gamma - \mathbf{g}\Big)\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = \Big(2(S^0\mathbf{g})|_\Gamma - \mathbf{g} + 2(Tf)|_\Gamma\Big)\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \\ &= R^0\mathbf{g} + 2(Tf)|_\Gamma \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}. \end{aligned} \tag{3.14}$$

Relations (3.13)–(3.14) and Lemma 3.1 trivially yield the following lemma.

LEMMA 3.3. *If* $\mathbf{g}, \widetilde{\mathbf{g}} \in \mathbf{L}^2(\Gamma)$ *and* c_1, c_2 *are constants, then*

(a) $R^0(\mathbf{g} - \widetilde{\mathbf{g}}) = R^f\mathbf{g} - R^f\widetilde{\mathbf{g}}$*;*

(b) $R^0(\mathbf{g} - \widehat{\mathbf{g}}) = R^f\mathbf{g} - \widehat{\mathbf{g}}$*, where* $\widehat{g}$ *is defined in* (3.8)*;*

(c) $R^0(c_1\mathbf{g} + c_2\widetilde{\mathbf{g}}) = c_1R^0\mathbf{g} + c_2R^0\widetilde{\mathbf{g}}$*.*

PROPOSITION 3.4. *Let* $\mathbf{g} = (g_1, g_2) \in \mathbf{L}^2(\Gamma)$ *and* $u_i = S_i^f g_i$*. Then*

$$\|R^f\mathbf{g}\|_{0,\Gamma}^2 = \|\mathbf{g}\|_{0,\Gamma}^2 - 4\lambda \sum_{i=1}^{2} \Big(a_i[u_i, u_i] - [f, u_i]_{\Omega_i}\Big).$$

In particular, if $f = 0$*, then*

$$\|R^0\mathbf{g}\|_{0,\Gamma}^2 = \|\mathbf{g}\|_{0,\Gamma}^2 - 4\lambda \sum_{i=1}^{2} a_i[S_i^0 g_i, S_i^0 g_i]\,.$$

Proof. The proof is adapted from that of [17]. We recall that $u_i = S_i^f\mathbf{g}$ is defined by (3.2). Setting $v_i = u_i$ in (3.2), we obtain

$$\int_\Gamma (u_iu_i - g_iu_i) = -\lambda\Big(a_i[u_i, u_i] - [f, u_i]_{\Omega_i}\Big), \quad i = 1, 2.$$

Thus

$$\begin{aligned}\|R^f\mathbf{g}\|_{0,\Gamma}^2 &= \int_\Gamma |2u_1 - g_1|^2 + \int_\Gamma |2u_2 - g_2|^2 \\ &= \int_\Gamma (g_1^2 + g_2^2) + 4\int_\Gamma (u_1^2 - u_1 g_1) + 4\int_\Gamma (u_2^2 - u_2 g_2) \\ &= \|\mathbf{g}\|_{0,\Gamma}^2 - 4\lambda \sum_{i=1}^2 \Big(a_i[u_i, u_i] - [f, u_i]_{\Omega_i} \Big). \qquad \square\end{aligned}$$

3.3. An interface bias functional and a descent direction.

We define the interface bias functional $E^f(\mathbf{g})$ associated with subdomain Robin-type boundary value problem (3.2) by

$$E^f(\mathbf{g}) = \frac{1}{2}\int_\Gamma \Big(|u_1 - u_2|^2 + |u_1 + u_2 - g_1 - g_2|^2 \Big), \tag{3.15}$$

where $(u_1, u_2) = S^f\mathbf{g}$ is defined through (3.2). If $[A(\mathbf{x})\nabla u_i] \cdot \mathbf{n}_i|_\Gamma \in L^2(\Gamma)$, $i = 1, 2$, then E^f can be rewritten as

$$E^f(\mathbf{g}) = \frac{1}{2}\int_\Gamma \Big(|u_1 - u_2|^2 + \lambda^2 \Big| [A(\mathbf{x})\nabla u_1] \cdot \mathbf{n}_1 + [A(\mathbf{x})\nabla u_2] \cdot \mathbf{n}_2 \Big|^2 \Big);$$

i.e., E^f indeed measures the interface bias.

The minimum value 0 of E^f is uniquely attained at $\widehat{\mathbf{g}}$ which is defined in (3.8). It is easily verified that the generalized method (2.17)–(2.18) can be simply expressed by

$$\begin{cases} \mathbf{g}^{(1)} \text{ given}, \\ \mathbf{g}^{(k+1)} = \mathbf{g}^{(k)} + \delta_k (R^f \mathbf{g}^{(k)} - \mathbf{g}^{(k)}), \qquad k = 1, 2, \ldots. \end{cases} \tag{3.16}$$

We will show that $R^f\mathbf{g} - \mathbf{g}$ provides a descent direction for E^f at any $\mathbf{g}$ so that the generalized method (3.16) is a descent-direction method for solving the following minimization problem:

$$\min_{\mathbf{g} \in \mathbf{L}^2(\Gamma)} E^f(\mathbf{g}) \text{ subject to (3.2)}, \ i = 1, 2. \tag{3.17}$$

We first establish the following identities.

PROPOSITION 3.5. *Let* $\mathbf{g} = (g_1, g_2) \in \mathbf{L}^2(\Gamma)$. *Then*

$$E^f(\mathbf{g}) = \frac{1}{4}\int_\Gamma |\mathbf{g} - R^f\mathbf{g}|^2 = \frac{1}{4}\int_\Gamma \Big(|g_1 - R_1^f\mathbf{g}|^2 + |g_2 - R_2^f\mathbf{g}|^2 \Big) \tag{a}$$

and

$$E^f(R^f\mathbf{g}) = E^f(\mathbf{g}) - 4\lambda \sum_{i=1}^2 a_i[S_i^f(g_i) - S_i^f(R_i^f\mathbf{g}), S_i^f(g_i) - S_i^f(R_i^f\mathbf{g})]\,. \tag{b}$$

Proof. We set $u_i = S_i^f g_i$ and $\widetilde{u}_i = S_i^f(R_i^f\mathbf{g})$, $i = 1, 2$. From the definition of R^f, i.e., (3.5), we have

$$(R_1^f\mathbf{g} - g_1) - (R_2^f\mathbf{g} - g_2) = 2(u_2 - u_1) \quad \text{on } \Gamma$$

and

$$(R_1^f\mathbf{g} - g_1) + (R_2^f\mathbf{g} - g_2) = 2(u_1 + u_2) - 2(g_1 + g_2) \quad \text{on } \Gamma.$$

Squaring both sides of the last two equations and adding them together, we deduce

$$(R_1^f\mathbf{g} - g_1)^2 + (R_2^f\mathbf{g} - g_2)^2 = 2(u_1 - u_2)^2 + 2|u_1 + u_2 - g_1 - g_2|^2 \quad \text{on } \Gamma,$$

which readily yields (a).

To prove (b) we note that identity (a), Lemma 3.3 (a), Proposition 3.4, and Lemma 3.1 (c) lead us to

$$\begin{aligned} E^f(R^f\mathbf{g}) &= \|R^f\mathbf{g} - R^f(R^f\mathbf{g})\|_{0,\Gamma}^2 = \|R^0(\mathbf{g} - R^f\mathbf{g})\|_{0,\Gamma}^2 \\ &= \|\mathbf{g} - R^f\mathbf{g}\|_{0,\Gamma}^2 - 4\lambda \sum_{i=1}^2 \int_{\Omega_i} \left|\nabla\Big(S_i^0(g_i - R_i^f\mathbf{g})\Big)\right|^2 \\ &= \|\mathbf{g} - R^f\mathbf{g}\|_{0,\Gamma}^2 - 4\lambda \sum_{i=1}^2 a_i[S_i^f(g_i) - S_i^f(R_i^f\mathbf{g}), S_i^f(g_i) - S_i^f(R_i^f\mathbf{g})]\,. \qquad \square \end{aligned}$$

PROPOSITION 3.6. *Let* $\mathbf{g} \in \mathbf{L}^2(\Gamma)$. *Then* $\Lambda(\mathbf{g}) \equiv \|\mathbf{g} - 2R^f\mathbf{g} + R^fR^f\mathbf{g})\|_{0,\Gamma}^2 = 0$ *if and only if* $\mathbf{g} = \widehat{\mathbf{g}}$.

Proof. If $\mathbf{g} = \widehat{\mathbf{g}}$, then Lemma 3.2 implies $R^f\widehat{\mathbf{g}} = \widehat{\mathbf{g}}$ so that

$$\Lambda(\widehat{\mathbf{g}}) = \|\widehat{\mathbf{g}} - 2R^f\widehat{\mathbf{g}} + R^fR^f\widehat{\mathbf{g}}\|_{0,\Gamma}^2 = 0\,.$$

Conversely, we assume that $\Lambda(\mathbf{g}) = 0$ and proceed to show that $R^f\mathbf{g} = \mathbf{g}$. Obviously $\Lambda(\mathbf{g}) = 0$ implies $\mathbf{g} - R^f\mathbf{g} = R^f\mathbf{g} - R^f(R^f\mathbf{g})$. Using this relation and employing Proposition 3.5 (a) repeatedly, we have

$$E^f(\mathbf{g}) = \frac{1}{4}\|\mathbf{g} - R^f\mathbf{g}\|_{0,\Gamma}^2 = \frac{1}{4}\|R^f\mathbf{g} - R^fR^f\mathbf{g})\|_{0,\Gamma}^2 = E^f(R^f\mathbf{g})$$

so that by Proposition 3.5 (b) we deduce

$$a_1[u_1 - \widetilde{u}_1, u_1 - \widetilde{u}_1] + a_2[u_2 - \widetilde{u}_2, u_2 - \widetilde{u}_2] = 0,$$

where $u_i = S_i^f g_i$ and $\widetilde{u}_i = S_i^f(R_i^f\mathbf{g})$, $i = 1, 2$. Thus, by virtue of the Poincaré inequality, we obtain

$$u_i - \widetilde{u}_i = 0 \quad \text{in } \Omega_i \qquad \text{and} \qquad u_i - \widetilde{u}_i = 0 \quad \text{on } \Gamma \tag{3.18}$$

for $i = 1, 2$. By the definition of S_i^f we have, for $i = 1, 2$,

$$a_i[u_i, v_i] + \lambda^{-1}[u_i, v_i]_\Gamma = [f, v_i]_{\Omega_i} + \lambda^{-1}[g_i, v_i]_\Gamma \qquad \forall\, v_i \in X_i$$

and

$$a_i[\widetilde{u}_i, v_i] + \lambda^{-1}[\widetilde{u}_i, v_i]_\Gamma = [f, v_i]_{\Omega_i} + \lambda^{-1}[R_i^f\mathbf{g}, v_i]_\Gamma \qquad \forall\, v_i \in X_i\,.$$

Subtracting the last two equations and applying (3.18), we obtain $g_i = R_i^f\mathbf{g}$, $i = 1, 2$, i.e., $\mathbf{g} = R^f\mathbf{g}$. Thus by Lemma 3.2 we conclude $\mathbf{g} = \widehat{\mathbf{g}}$. $\square$

A review of the proof of Proposition 3.6 reveals the following.

COROLLARY 3.7. *Let* $\mathbf{g} \in \mathbf{L}^2(\Gamma)$. *The following statements are equivalent:*

(i) $\Lambda(\mathbf{g}) = 0$*;*

(ii) $\mathbf{g} = \widehat{\mathbf{g}}$*;*

(iii) $\mathbf{g} = R^f\mathbf{g}$*;*

(iv) $E^f(\mathbf{g}) = 0$*;*

(v) $E^f(\mathbf{g}) = E^f(R^f\mathbf{g})$.

As a consequence of Proposition 3.5 and Corollary 3.7, we have the following.

COROLLARY 3.8. $E^f(R^f\mathbf{g}) < E^f(\mathbf{g})$ *whenever* $\mathbf{g} \in \mathbf{L}^2(\Gamma)$ *and* $\mathbf{g} \neq \widehat{\mathbf{g}}$.

The main result of this section is the following.

THEOREM 3.9. *Assume that* $\mathbf{g} \in \mathbf{L}^2(\Gamma)$ *and* $\Lambda(\mathbf{g}) \equiv \|\mathbf{g}-2R^f\mathbf{g}+R^fR^f\mathbf{g})\|^2_{0,\Gamma} \neq 0$. *Then there is a* $\delta_0 = \delta_0(\mathbf{g}) \geq 1/2$ *such that*

$$E^f(\mathbf{g}) > E^f(\mathbf{g} + \delta(R^f\mathbf{g} - \mathbf{g})) \qquad \forall\, \delta \in (0, 2\delta_0)\,.$$

Moreover, $E^f(\mathbf{g} + \delta(R^f\mathbf{g} - \mathbf{g}))$ *as a function of* δ *is strictly decreasing on* $[0, \delta_0]$ *and strictly increasing on* $[\delta_0, 2\delta_0]$.

Proof. Assume that $\mathbf{g} \in \mathbf{L}^2(\Gamma)$ and $\Lambda(\mathbf{g}) \equiv \|\mathbf{g} - 2R^f\mathbf{g} + R^fR^f\mathbf{g})\|^2_{0,\Gamma} \neq 0$. Using Lemma 3.3 (a) repeatedly, we have

$$\begin{aligned} R^f(\mathbf{g} + \delta(R^f\mathbf{g} - \mathbf{g})) &= R^f(\mathbf{g} + \delta(R^f\mathbf{g} - \mathbf{g})) - R^f\mathbf{g} + R^f\mathbf{g} \\ &= R^0(\delta(R^f\mathbf{g} - \mathbf{g})) + R^f\mathbf{g} = \delta(R^f(R^f\mathbf{g}) - R^f\mathbf{g}) + R^f\mathbf{g}. \end{aligned}$$

The last equation and Proposition 3.5 imply

$$\begin{aligned} \phi(\delta) \equiv 4E^f(\mathbf{g} + \delta(R^f\mathbf{g} - \mathbf{g})) &= \|\mathbf{g} + \delta(R^f\mathbf{g} - \mathbf{g}) - (\delta(R^f(R^f\mathbf{g}) - R^f\mathbf{g}) + R^f\mathbf{g})\|^2_{0,\Gamma} \\ &= \|\mathbf{g} - R^f\mathbf{g}\|^2_{0,\Gamma} + \delta^2\|\mathbf{g} - 2R^f\mathbf{g} + R^f(R^f\mathbf{g})\|^2_{0,\Gamma} \\ &\qquad - 2\delta[\mathbf{g} - R^f\mathbf{g}, \mathbf{g} - 2R^f\mathbf{g} + R^f(R^f\mathbf{g})]_\Gamma\,. \end{aligned}$$

Thus $\phi(\delta)$ attains its minimum at

$$\delta_0 = \delta_0(\mathbf{g}) = \frac{[\mathbf{g} - R^f\mathbf{g}, \mathbf{g} - 2R^f\mathbf{g} + R^f(R^f\mathbf{g})]_\Gamma}{\|\mathbf{g} - 2R^f\mathbf{g} + R^f(R^f\mathbf{g})\|^2_{0,\Gamma}} \tag{3.19}$$

with the minimum value

$$\phi(\delta_0) = 4E^f(\mathbf{g} + \delta_0(R^f\mathbf{g} - \mathbf{g})) = \|\mathbf{g} - R^f\mathbf{g}\|^2_{0,\Gamma} - \frac{[\mathbf{g} - R^f\mathbf{g}, \mathbf{g} - 2R^f\mathbf{g} + R^f(R^f\mathbf{g})]^2_\Gamma}{\|\mathbf{g} - 2R^f\mathbf{g} + R^f(R^f\mathbf{g})\|^2_{0,\Gamma}}.$$

We verify $\delta_0 > 0$ as follows:

$$\begin{aligned} &[\mathbf{g} - R^f\mathbf{g}, \mathbf{g} - 2R^f\mathbf{g} + R^f(R^f\mathbf{g})]_\Gamma \\ &\qquad = \|\mathbf{g} - R^f\mathbf{g}\|^2_{0,\Gamma} - [\mathbf{g} - R^f\mathbf{g}, R^f\mathbf{g} - R^f(R^f\mathbf{g})]_\Gamma \\ &\qquad \geq \|\mathbf{g} - R^f\mathbf{g}\|^2_{0,\Gamma} - \frac{1}{2}\|\mathbf{g} - R^f\mathbf{g}\|^2_{0,\Gamma} - \frac{1}{2}\|R^f\mathbf{g} - R^f(R^f\mathbf{g})\|^2_{0,\Gamma} \\ &\qquad = \frac{1}{2}\|\mathbf{g} - R^f\mathbf{g}\|^2_{0,\Gamma} - \frac{1}{2}\|R^f\mathbf{g} - R^f(R^f\mathbf{g})\|^2_{0,\Gamma} = 2[E^f(\mathbf{g}) - E^f(R^f\mathbf{g})] > 0, \end{aligned}$$

where in the last step we used Proposition 3.5 (a) and Corollary 3.8.

The function $\phi(\delta) \equiv E^f(\mathbf{g}+\delta(R^f\mathbf{g}-\mathbf{g}))$, a quadratic function of δ, is strictly decreasing on $(-\infty,\delta_0]$ and strictly increasing on $\in [\delta_0,\infty)$. Straightforward calculations reveal

$$\phi(2\delta_0) = \|\mathbf{g} - R^f\mathbf{g}\|^2_{0,\Gamma} = 4E^f(\mathbf{g}) \qquad \text{and} \qquad \phi(1) = 4E^f(R^f\mathbf{g}).$$

Thus an application of Corollary 3.8 yields $\phi(2\delta_0) > \phi(1)$ so that $2\delta_0 \geq 1$. □

3.4. Convergence of the generalized method. By virtue of Theorem 3.9 the generalized method (3.16), which is equivalent to iterations (2.17)–(2.18), is a descent-direction method for solving the minimization problem (3.17). We have the following convergence result for iterations (3.16).

THEOREM 3.10. *Assume that* $\mathbf{g}^{(1)} \in \mathbf{L}^2(\Gamma)$ *and* $\{\delta_k\}_{k=1}^\infty \subset [\delta_{\min},1] \subset (0,1]$. *Let* $\{\mathbf{g}^{(k)}\}_{k=2}^\infty$ *be defined by* (3.16) *and* $(u_1^{(k)},u_2^{(k)}) = S^f\mathbf{g}^{(k)}$, $k=1,2,3,\ldots.$ *Then*

$$\|u_i^{(k)} - \widehat{u}_i\|_{1,\Omega_i} \to 0 \quad \text{as } k\to\infty,\ i=1,2.$$

Proof. Recall that $\widehat{\mathbf{g}} = R^f\widehat{\mathbf{g}}$ and $R^f\mathbf{g}^{(k)} - R^f\widehat{\mathbf{g}} = R^0(\mathbf{g}^{(k)}-\widehat{\mathbf{g}})$. Thus

$$\mathbf{g}^{(k+1)} - \widehat{\mathbf{g}} = (1-\delta_k)(\mathbf{g}^{(k)}-\widehat{\mathbf{g}}) + \delta_k R^0(\mathbf{g}^{(k)}-\widehat{\mathbf{g}})\,.$$

It follows that

$$\begin{aligned}\|\mathbf{g}^{(k+1)} - \widehat{\mathbf{g}}\|^2_{0,\Gamma} &= (1-\delta_k)^2\|\mathbf{g}^{(k)}-\widehat{\mathbf{g}}\|^2_{0,\Gamma} + \delta_k^2\|R^0(\mathbf{g}^{(k)}-\widehat{\mathbf{g}})\|^2_{0,\Gamma}\\ &\qquad + 2\delta_k(1-\delta_k)[\mathbf{g}^{(k)}-\widehat{\mathbf{g}}, R^0(\mathbf{g}^{(k)}-\widehat{\mathbf{g}})]_{0,\Gamma}\\ &\leq (1-\delta_k)^2\|\mathbf{g}^{(k)}-\widehat{\mathbf{g}}\|^2_{0,\Gamma} + \delta_k^2\|R^0(\mathbf{g}^{(k)}-\widehat{\mathbf{g}})\|^2_{0,\Gamma}\\ &\qquad + \delta_k(1-\delta_k)\{\|\mathbf{g}^{(k)}-\widehat{\mathbf{g}}\|^2_{0,\Gamma} + \|R^0(\mathbf{g}^{(k)}-\widehat{\mathbf{g}})\|^2_{0,\Gamma}\}\\ &= (1-\delta_k)\|\mathbf{g}^{(k)}-\widehat{\mathbf{g}}\|^2_{0,\Gamma} + \delta_k\|R^0(\mathbf{g}^{(k)}-\widehat{\mathbf{g}})\|^2_{0,\Gamma}\,.\end{aligned}$$

On the other hand, using Proposition 3.4 and Lemma 3.1 (c) and noting that $(\widehat{u}_1,\widehat{u}_2) = S^f\widehat{\mathbf{g}}$, we have

$$\|R^0(\mathbf{g}^{(k)}-\widehat{\mathbf{g}})\|^2_{0,\Gamma} = \|\mathbf{g}^{(k)}-\widehat{\mathbf{g}}\|^2_{0,\Gamma} - 4\lambda\sum_{i=1}^2 a_i[u_i^{(k)}-\widehat{u}_i, u_i^{(k)}-\widehat{u}_i],$$

where $(u_1^{(k)},u_2^{(k)}) = S^f\mathbf{g}^{(k)}$. Combining the last two relations, we obtain

$$\begin{aligned}\|\mathbf{g}^{(k+1)} - \widehat{\mathbf{g}}\|^2_{0,\Gamma} &\leq \|\mathbf{g}^{(k)}-\widehat{\mathbf{g}}\|^2_{0,\Gamma} - 4\lambda\delta_k\sum_{i=1}^2\int_{\Omega_i}|\nabla(u_i^{(k)}-\widehat{u}_i)|^2\\ &\leq \|\mathbf{g}^{(k)}-\widehat{\mathbf{g}}\|^2_{0,\Gamma} - 4\lambda\delta_{\min}\sum_{i=1}^2\int_{\Omega_i}|\nabla(u_i^{(k)}-\widehat{u}_i)|^2\end{aligned}$$

or, equivalently,

$$4\lambda\delta_{\min}\sum_{i=1}^2 a_i[u_i^{(k)}-\widehat{u}_i, u_i^{(k)}-\widehat{u}_i] \leq \|\mathbf{g}^{(k)}-\widehat{\mathbf{g}}\|^2_{0,\Gamma} - \|\mathbf{g}^{(k+1)}-\widehat{\mathbf{g}}\|^2_{0,\Gamma}.$$

Summing in k from 1 to an arbitrary $N \geq 1$, we have

$$4\lambda\delta_{\min}\sum_{k=1}^N\sum_{i=1}^2 a_i[u_i^{(k)}-\widehat{u}_i, u_i^{(k)}-\widehat{u}_i] \leq \|\mathbf{g}^{(1)}-\widehat{\mathbf{g}}\|^2_{0,\Gamma} - \|\mathbf{g}^{(N+1)}-\widehat{\mathbf{g}}\|^2_{0,\Gamma} \leq \|\mathbf{g}^{(1)}-\widehat{\mathbf{g}}\|^2_{0,\Gamma}\,.$$

This implies that

$$\sum_{i=1}^{2} a_i[u_i^{(k)} - \widehat{u}_i, u_i^{(k)} - \widehat{u}_i] \to 0 \quad \text{as } k \to \infty.$$

From the Poincaré inequality, we conclude that

$$\|u_i^{(k)} - \widehat{u}_i\|_{1,\Omega_i} \to 0 \quad \text{as } k \to \infty. \qquad \square$$

Next, we will demonstrate that $E^f(\mathbf{g}^{(k)}) \to 0$ as $k \to \infty$. We will need the following lemma.

LEMMA 3.11. *Assume that* $\{r_k\}_{k=1}^\infty$, $\{b_k\}_{k=1}^\infty$, *and* $\{a_k\}_{k=1}^\infty$ *satisfy* $\lim_{k\to\infty} b_k = 0$,

$$0 \le r_k \le r_{\max} < 1\,, \quad |a_{k+1}| \le r_k |a_k| + |b_k|\,, \quad |a_k| \le M \qquad \forall\, k,$$

where $r_{\max}$ *and* M *are constants independent of* k. *Then* $\lim_{k\to\infty} a_k = 0$.

Proof. Let an $\epsilon > 0$ be given. As $\lim_{k\to\infty} b_k = 0$, we may choose an integer $K_1 > 0$ such that $|b_k| < (1 - r_{\max})\epsilon/2$ for all $k \ge K_1$. Since $r_{\max} \in [0, 1)$, we may choose an integer $K_2 > 0$ such that $|r_{\max}|^k < \epsilon/(2M)$ for all $k \ge K_2$. Hence, for $k > K_1 + K_2$, using the relation

$$|a_{m+1}| \le r_{\max}\, |a_m| + |b_m|$$

recursively, we obtain

$$\begin{aligned} |a_k| &\le |r_{\max}|^{K_2} |a_{k-K_2}| + |r_{\max}|^{K_2-1} |b_{k-K_2}| + \cdots + |r_{\max}|\, |b_{k-2}| + |b_{k-1}| \\ &\le |r_{\max}|^{K_2} M + (|r_{\max}|^{K_2-1} + \cdots + |r_{\max}| + 1) \frac{1 - r_{\max}}{2}\, \epsilon \\ &< \frac{\epsilon}{2M} \cdot M + \frac{1}{1 - r_{\max}} \frac{1 - r_{\max}}{2}\, \epsilon = \epsilon. \qquad \square \end{aligned}$$

THEOREM 3.12. *Assume that* $\mathbf{g}^{(1)} \in \mathbf{L}^2(\Gamma)$ *and* $\{\delta_k\}_{k=1}^\infty \subset [\delta_{\min}, \delta_{\max}] \subset (0, 1)$. *Let* $\{\mathbf{g}^{(k)}\}_{k=2}^\infty$ *be defined by* (3.16) *and* $(u_1^{(k)}, u_2^{(k)}) = S^f \mathbf{g}^{(k)}$, $k = 1, 2, 3, \ldots$. *Then* $E^f(\mathbf{g}^{(k)}) \to 0$ *as* $k \to \infty$.

Proof. Theorem 3.10 and the trace theorem imply

$$\|u_i^{(k)} - \widehat{u}_i\|_{0,\Gamma}^2 \to 0 \quad \text{as } k \to \infty$$

so that using a triangle inequality and the fact that $\widehat{u}_1|_\Gamma = \widehat{u}_2|_\Gamma$, we have

$$\|u_1^{(k)} - u_2^{(k)}\|_{0,\Gamma}^2 \to 0 \quad \text{as } k \to \infty. \tag{3.20}$$

We also have, for every $k \ge 1$,

$$\begin{aligned} &u_1^{(k+1)} + u_2^{(k+1)} - g_1^{(k+1)} - g_2^{(k+1)} \\ &\quad = u_1^{(k+1)} + u_2^{(k+1)} - (1 - \delta_k)(g_1^{(k)} + g_2^{(k)}) - \delta_k (2u_1^{(k)} + 2u_2^{(k)} - g_1^{(k)} - g_2^{(k)}) \\ &\quad = u_1^{(k+1)} - u_1^{(k)} + u_2^{(k+1)} - u_2^{(k)} + (1 - 2\delta_k)(u_1^{(k)} + u_2^{(k)} - g_1^{(k)} - g_2^{(k)}) \end{aligned}$$

so that

$$\begin{aligned}&\|u_1^{(k+1)} + u_2^{(k+1)} - g_1^{(k+1)} - g_2^{(k+1)}\|_{0,\Gamma}^2\\ &\le \|u_1^{(k+1)} - u_1^{(k)}\|_{0,\Gamma}^2 + \|u_2^{(k+1)} - u_2^{(k)}\|_{0,\Gamma}^2 + |1-2\delta_k|\,\|u_1^{(k)} + u_2^{(k)} - g_1^{(k)} - g_2^{(k)}\|_{0,\Gamma}^2 .\end{aligned}$$

Thus, by setting

$$\begin{aligned}&r_k = |1-2\delta_k|, \quad a_k = \|u_1^{(k)} + u_2^{(k)} - g_1^{(k)} - g_2^{(k)}\|_{0,\Gamma}^2,\\ &\text{and} \quad b_k = \|u_1^{(k+1)} - u_1^{(k)}\|_{0,\Gamma}^2 + \|u_2^{(k+1)} - u_2^{(k)}\|_{0,\Gamma}^2 ,\end{aligned}$$

we have $\lim_{k\to\infty} b_k = 0$,

$$a_{k+1} \le |1-2\delta_k|\, a_k + b_k\,, \quad 0 \le r_k \le r_{\max} \qquad \forall\, k,$$

where $r_{\max} \equiv \max\{|1-2\delta_{\min}|, |1-2\delta_{\max}|\}$. The fact that $(1-2\delta)^2 < 1$ for all $\delta \in (0,1)$ implies $r_{\max} < 1$. Also, by virtue of Theorem 3.9, $E^f(\mathbf{g}^{(k)})$ is a nonincreasing nonnegative sequence and is thus bounded. This in turn yields the boundedness of $\{a_k\}$. Hence an application of Lemma 3.11 yields

$$a_k = \|u_1^{(k)} + u_2^{(k)} - g_1^{(k)} - g_2^{(k)}\|_{0,\Gamma}^2 \to 0 \quad \text{as } k \to \infty. \tag{3.21}$$

Combining (3.20) and (3.21), we conclude $E^f(\mathbf{g}^{(k)}) \to 0$. □

Remark 3.13. If $\delta_k = 1$ for all k, it is well known [17] that $\|g_i^{(k)} - \widehat{g}_i\|_{H^{-\frac{1}{2}}(\Gamma)} \to 0$. Triangle inequalities and the fact that $\|\widehat{u}_1 + \widehat{u}_2 - \widehat{g}_1 - \widehat{g}_2\|_{-1/2,\Gamma}^2 = 0$ yield

$$\begin{aligned}&\|u_1^{(k)} + u_2^{(k)} - g_1^{(k)} - g_2^{(k)}\|_{-1/2,\Gamma}^2\\ &\quad\le \|u_1^{(k)} - \widehat{u}_1\|_{-1/2,\Gamma}^2 + \|u_2^{(k)} - \widehat{u}_2\|_{-1/2,\Gamma}^2 + \|\widehat{g}_1 - g_1^{(k)}\|_{-1/2,\Gamma}^2 + \|\widehat{g}_1 - g_1^{(k)}\|_{-1/2,\Gamma}^2\end{aligned}$$

so that $\|u_1^{(k)} + u_2^{(k)} - g_1^{(k)} - g_2^{(k)}\|_{-1/2,\Gamma}^2 \to 0$ as $k \to \infty$. However, this does not imply $E^f(\mathbf{g}^{(k)}) \to 0$ (the assumptions of Theorem 3.12 exclude the case $\delta_k = 1$).

Remark 3.14. The acceleration properties of the generalized methods can now be explained. First, if δ_k is chosen between 1 and the optimal $\delta_0(\mathbf{g}^{(k)})$, then Theorem 3.9 implies that at this step, the interface bias functional E^f descends faster with the generalized methods than with Lions' method. Second, the interface bias in Theorem 3.12 is measured by $\|u_1 - u_2\|_{0,\Gamma}^2 + \|u_1 + u_2 - g_1 - g_2\|_{0,\Gamma}^2$, whereas the interface bias for Lions' method can only be measured by $\|u_1 - u_2\|_{0,\Gamma}^2 + \|u_1 + u_2 - g_1 - g_2\|_{-1/2,\Gamma}^2$ (see Remark 3.13); hence when the interface bias is measured in the same norms (i.e., measured in terms of $\|u_1 - u_2\|_{0,\Gamma}^2 + \|u_1 + u_2 - g_1 - g_2\|_{-1/2,\Gamma}^2$), we expect the convergence for the generalized methods to be faster than that for Lions' method.

4. The finite element version of the generalized methods. The finite element version of the generalized methods can be analyzed in essentially the same way as the continuous version. We will state without proofs all results parallel to those of the continuous case. We will also establish mesh-dependent geometric convergence for the finite element version of the generalized methods.

We assume that Ω is a two-dimensional polygon or a three-dimensional polyhedron partitioned into two subdomains Ω_1 and Ω_2 as shown in Figure 2.1. Let $\mathcal{T}^h(\Omega)$ be a family of regular triangulations of Ω such that no element of the triangulations crosses the interface Γ. Let $X^h \subset H_0^1(\Omega)$ be a family of finite element spaces, and we set $X_i^h \subset X^h|_{\Omega_i}$ $(i = 1, 2)$ and

$$G^h = G_1^h \times G_2^h \equiv X_1^h|_\Gamma \times X_2^h|_\Gamma \,.$$

We assume X^h and X_i^h satisfy standard approximation properties [16].

The finite element version of the generalized method (2.17)–(2.18) is described as follows: choose initial guesses $g_1^{h,1} \in G_1^h$ and $g_2^{h,1} \in G_2^h$; for $k = 1, 2, 3, \ldots$, solve for $u_i^{h,k} \in X_i^h$ $(i = 1, 2)$ from

$$(4.1)\quad a_i[u_i^{h,k}, v_i^h] + \lambda^{-1}[u_i^{h,k}, v_i^h]_\Gamma = [f, v_i^h]_{\Omega_i} + \lambda^{-1}[g_i^{h,k}, v_i^h]_\Gamma \quad \forall\, v_i^h \in X_i^h,\ i = 1, 2,$$

and update

$$(4.2)\qquad \begin{aligned} g_1^{h,k+1} &= (1 - \delta_k) g_1^{h,k} + \delta_k (2u_2^{h,k} - g_2^{h,k}) \,, \\ g_2^{h,k+1} &= (1 - \delta_k) g_2^{h,k} + \delta_k (2u_1^{h,k} - g_1^{h,k}) \,. \end{aligned}$$

4.1. Solution operators for the discrete Robin boundary value problems. For given $f \in L^2(\Omega)$ and $\mathbf{g} = (g_1, g_2) \in \mathbf{L}^2(\Gamma)$, the discrete Robin-type boundary value problems on subdomains are defined as follows: seek a $u_i^h \in X_i^h$, $i = 1, 2$, such that

$$(4.3)\qquad a_i[u_i^h, v_i^h] + \lambda^{-1}[u_i^h, v_i^h]_\Gamma = [f, v_i^h]_{\Omega_i} + \lambda^{-1}[g_i, v_i^h]_\Gamma \qquad \forall\, v_i^h \in X_i^h.$$

For $i = 1, 2$ we denote by $S_{i,h}^f : L^2(\Gamma) \to X_i^h$ the solution operator for the discrete Robin boundary value problem (4.3); i.e., $u_i^h = S_{i,h}^f\, g_i$ for all $g_i \in L^2(\Gamma)$ if and only if u_i^h and g_i satisfy (4.3). We define the operator $S_h^f : \mathbf{L}^2(\Gamma) \to X_1^h \times X_2^h$ by $S_h^f \mathbf{g} = (S_{1,h}^f\, g_1, S_{2,h}^f\, g_2)$ for all $\mathbf{g} \in \mathbf{L}^2(\Gamma)$. If $f = 0$, we write S_h^0 in place of S_h^f.

We also denote by $T_h : L^2(\Omega) \to X_1^h \times X_2^h$ the solution operator for (4.3) with homogeneous Robin boundary condition; i.e., for every $f \in L^2(\Omega)$, $(u_1^h, u_2^h) = T_h f = (T_{1,h}\, f, T_{2,h}\, f)$ if and only if $(u_1^h, u_2^h) \in X_1^h \times X_2^h$ is the solution of

$$(4.4)\qquad a_i[u_i^h, v_i^h] + \lambda^{-1}[u_i^h, v_i^h]_\Gamma = [f, v_i^h]_{\Omega_i} \qquad \forall\, v_i^h \in X_i^h,\ i = 1, 2.$$

The operator T_h is obviously linear.

Lemma 4.1. *If* $\mathbf{g}, \widetilde{\mathbf{g}} \in \mathbf{L}^2(\Gamma)$ *and* c_1, c_2 *are constants, then*

(a) $S_h^0(c_1\mathbf{g} + c_2\widetilde{\mathbf{g}}) = c_1 S_h^0\mathbf{g} + c_2 S_h^0\widetilde{\mathbf{g}}$*;*

(b) $S_h^f = S_h^0 + T_h f$*;*

(c) $S_h^0(\mathbf{g} - \widetilde{\mathbf{g}}) = S_h^f\mathbf{g} - S_h^f\widetilde{\mathbf{g}}$.

4.2. The discrete Robin–Robin map and its basic properties. We define the finite element Robin–Robin map $R_h^f : G^h \to G^h$ as follows. For any $\mathbf{g}^h = (g_1^h, g_2^h) \in G^h$,

$$(4.5)\qquad R_h^f\mathbf{g}^h = (R_{1,h}^f\mathbf{g}^h, R_{2,h}^f\mathbf{g}^h) \equiv (2u_2^h|_\Gamma - g_2^h, 2u_1^h|_\Gamma - g_1^h),$$

where $u_i^h = S_{i,h}^f g_i^h$, $i = 1, 2$. If $f = 0$, we write R_h^0 in place of R_h^f.

We denote by $\widehat{u}^h \in X^h$ the unique finite element solution of the discrete elliptic problem, i.e.,

$$(4.6)\qquad a[\widehat{u}^h, v^h] = [f, v^h]_\Omega \qquad \forall\, v^h \in X^h\,.$$

Then we have

$$a_i[\widehat{u}^h, v_i^h] = [f, v_i^h]_{\Omega_i} \qquad \forall\, v_i^h \in X_i^h \cap H_0^1(\Omega_i)\,.$$

The results of [29] allow us to find a $\widehat{t}_i^h \in G_i^h = X_i^h|_\Gamma$ such that

$$a_i[\widehat{u}^h, v_i^h] = [f, v_i^h]_\Omega + [\widehat{t}_i^h, v_i^h]_\Gamma \qquad \forall\, v_i^h \in X_i^h\,. \tag{4.7}$$

We define

$$\widehat{u}_i^h = \widehat{u}^h|_{\Omega_i}, \qquad \widehat{g}_i^h = \widehat{u}_i^h + \lambda \widehat{t}_i^h, \qquad i = 1, 2. \tag{4.8}$$

For an arbitrary $v^h \in X^h$, setting $v_i^h = v^h|_{\Omega_i}$ $(i = 1, 2)$ in (4.7) and adding the two equations, we obtain

$$a[\widehat{u}^h, v^h] = [f, v^h]_\Omega + [\widehat{t}_1^h + \widehat{t}_2^h, v^h]_\Gamma \qquad \forall\, v^h \in X^h\,.$$

A comparison of the last equation with (4.6) yields

$$[\widehat{t}_1^h + \widehat{t}_2^h, v^h]_\Gamma = 0 \qquad \forall\, v^h \in X^h$$

so that

$$\widehat{t}_1^h + \widehat{t}_2^h = 0\,. \tag{4.9}$$

Of course, we have

$$\widehat{u}_1^h|_\Gamma = \widehat{u}_2^h|_\Gamma\,. \tag{4.10}$$

Also, (4.7) and (4.8) imply

$$a_i[\widehat{u}_i^h, v_i^h] + \lambda^{-1}[\widehat{u}_i^h, v_i^h]_\Gamma = [f, v_i^h]_{\Omega_i} + \lambda^{-1}[\widehat{g}_i^h, v_i^h]_\Gamma \qquad \forall\, v_i^h \in X_i^h\,,\ i = 1, 2\,. \tag{4.11}$$

Lemma 4.2. *$\mathbf{g}^h \in G^h$ satisfies $R_h^f \mathbf{g}^h = \mathbf{g}^h$ if and only if $\mathbf{g}^h = \widehat{\mathbf{g}}^h$.*

We easily see that R_h^f and R_h^0 satisfy the following relations:

$$R_h^0 \mathbf{g}^h = \Big(2(S_h^0 \mathbf{g}^h)|_\Gamma - \mathbf{g}^h\Big) \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \tag{4.12}$$

and

$$\begin{aligned} R_h^f \mathbf{g}^h &= \Big(2(S_h^f \mathbf{g}^h)|_\Gamma - \mathbf{g}^h\Big) \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \\ &= \Big(2(S_h^0 \mathbf{g}^h)|_\Gamma - \mathbf{g}^h + 2(T_h f)|_\Gamma\Big) \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = R_h^0 \mathbf{g}^h + 2(T_h f)|_\Gamma \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}. \end{aligned} \tag{4.13}$$

Relations (4.12)–(4.13) and Lemma 4.1 trivially yield the following.

Lemma 4.3. *If $\mathbf{g}^h, \widetilde{\mathbf{g}}^h \in G^h$ and c_1, c_2 are constants, then*

(a) $R_h^0(\mathbf{g}^h - \widetilde{\mathbf{g}}^h) = R_h^f \mathbf{g}^h - R_h^f \widetilde{\mathbf{g}}^h$;

(b) $R_h^0(\mathbf{g}^h - \widehat{\mathbf{g}}^h) = R_h^f \mathbf{g}^h - \widehat{\mathbf{g}}^h$;

(c) $R_h^0(c_1 \mathbf{g}^h + c_2 \widetilde{\mathbf{g}}^h) = c_1 R_h^0 \mathbf{g}^h + c_2 R_h^0 \widetilde{\mathbf{g}}^h$.

Proposition 4.4. *Let $\mathbf{g}^h = (g_1^h, g_2^h) \in G^h$ and $u_i^h = S_{i,h}^f\, g_i^h$. Then*

$$\|R_h^f \mathbf{g}^h\|_{0,\Gamma}^2 = \|\mathbf{g}^h\|_{0,\Gamma}^2 - 4\lambda \sum_{i=1}^{2} \Big(a_i[u_i^h, u_i^h] - [f, u_i^h]_{\Omega_i}\Big).$$

In particular, if $f = 0$, then

$$\|R^0 \mathbf{g}^h\|_{0,\Gamma}^2 = \|\mathbf{g}^h\|_{0,\Gamma}^2 - 4\lambda \sum_{i=1}^{2} a_i[S_{i,h}^0\, g_i^h, S_{i,h}^0\, g_i^h]\,.$$

4.3. A discrete interface bias functional and a descent direction. We define the interface bias functional $E_h^f(\mathbf{g}^h)$ associated with discrete subdomain Robin-type boundary value problem (4.3) by

$$E_h^f(\mathbf{g}^h) = \frac{1}{2}\int_\Gamma \Big(|u_1^h - u_2^h|^2 + |u_1^h + u_2^h - g_1^h - g_2^h|^2\Big), \tag{4.14}$$

where $(u_1^h, u_2^h) = S_h^f \mathbf{g}^h$.

The minimum value 0 of E_h^f is uniquely attained at $\widehat{\mathbf{g}}^h$ which is defined in (4.8). It is easily verified that the generalized method (4.1)–(4.2) can be simply expressed by

$$\begin{cases} \mathbf{g}^{h,1} \text{ given,} \\ \mathbf{g}^{h,k+1} = \mathbf{g}^{h,k} + \delta_k(R_h^f \mathbf{g}^{h,k} - \mathbf{g}^{h,k}), \quad k = 1,2,3,\dots. \end{cases} \tag{4.15}$$

We may show that $R_h^f \mathbf{g}^h - \mathbf{g}^h$ provides a descent direction for E_h^f at any $\mathbf{g}^h$ so that the generalized method (4.15) is a descent-direction method for solving the following minimization problem:

$$\min_{\mathbf{g}^h \in G^h} E_h^f(\mathbf{g}^h) \text{ subject to (4.3)}, \ i = 1, 2. \tag{4.16}$$

The following identities hold.

Proposition 4.5. *Let* $\mathbf{g}^h = (g_1^h, g_2^h) \in G^h$. *Then*

$$\text{(a)} \quad E_h^f(\mathbf{g}^h) = \frac{1}{4}\|\mathbf{g}^h - R_h^f \mathbf{g}^h\|_{0,\Gamma}^2 \equiv \frac{1}{4}\int_\Gamma \Big(|g_1^h - R_{1,h}^f\, \mathbf{g}^h|^2 + |g_2^h - R_{2,h}^f\, \mathbf{g}^h|^2\Big)$$

and

$$\text{(b)} \ E_h^f(R_h^f \mathbf{g}^h) = E_h^f(\mathbf{g}^h) - 4\lambda \sum_{i=1}^2 a_i[S_{i,h}^f(g_i^h) - S_{i,h}^f(R_{i,h}^f\, \mathbf{g}^h), S_{i,h}^f(g_i^h) - S_{i,h}^f(R_{i,h}^f\, \mathbf{g}^h)]\,.$$

Proposition 4.6. *Let* $\mathbf{g}^h \in G^h$. *Then* $\Lambda_h(\mathbf{g}^h) \equiv \|\mathbf{g}^h - 2R_h^f \mathbf{g}^h + R_h^f R_h^f \mathbf{g}^h)\|_{0,\Gamma}^2 = 0$ *if and only if* $\mathbf{g}^h = \widehat{\mathbf{g}}^h$.

Corollary 4.7. *Let* $\mathbf{g}^h \in G^h$. *The following statements are equivalent:*

(i) $\Lambda_h(\mathbf{g})^h = 0$;

(ii) $\mathbf{g}^h = \widehat{\mathbf{g}}$;

(iii) $\mathbf{g}^h = R_h^f \mathbf{g}$;

(iv) $E_h^f \mathbf{g}^h = 0$;

(v) $E_h^f(\mathbf{g}^h) = E_h^f(R_h^f \mathbf{g}^h)$.

As a consequence of Proposition 4.5 and Corollary 4.7, we have the following.

Corollary 4.8. $E_h^f(R_h^f \mathbf{g}^h) < E_h^f(\mathbf{g}^h)$ *whenever* $\mathbf{g}^h \in G^h$ *and* $\mathbf{g}^h \neq \widehat{\mathbf{g}}^h$.

The main result of this subsection is the following theorem concerning the optimal step length in the descent direction.

Theorem 4.9. *Assume* $\mathbf{g}^h \in G^h$ *and* $\Lambda_h(\mathbf{g}^h) \equiv \|\mathbf{g}^h - 2R_h^f \mathbf{g}^h + R_h^f R_h^f \mathbf{g}^h)\|_{0,\Gamma}^2 \neq 0$. *Then there is a* $\delta_0^h = \delta_0^h(\mathbf{g}^h) \geq 1/2$ *such that*

$$E_h^f(\mathbf{g}^h) > E_h^f(\mathbf{g}^h + \delta(R_h^f \mathbf{g}^h - \mathbf{g}^h)) \qquad \forall\, \delta \in (0, 2\delta_0^h).$$

Moreover, $E_h^f(\mathbf{g}^h + \delta(R_h^f \mathbf{g}^h - \mathbf{g}^h))$ *as a function of* δ *is strictly decreasing on* $[0, \delta_0^h]$ *and strictly increasing on* $[\delta_0^h, 2\delta_0^h]$.

4.4. Convergence of the discrete version of the generalized method. Recall that the discrete version of the generalized method (4.1)–(4.2) can be expressed by (4.15). By virtue of Theorem 4.9 the generalized method (4.15) is a descent-direction method for solving the discrete minimization problem (4.16).

We have the following convergence result for iterations (4.15).

THEOREM 4.10. *Assume that* $\mathbf{g}^{h,1} \in G^h$ *and* $\{\delta_k\} \subset [\delta_{\min}, 1] \subset (0,1]$. *Let* $\{\mathbf{g}^{h,k}\}_{k=2}^{\infty}$ *be defined by* (4.15) *and* $(u_1^{h,k}, u_2^{h,k}) = S_h^f \mathbf{g}^{h,k}$, $k = 1,2,3,\ldots$. *Then*

$$\|u_i^{h,k} - \widehat{u}_i^h\|_{1,\Omega_i} \to 0 \quad \textit{as } k \to \infty,\ i = 1,2.$$

THEOREM 4.11. *Assume that* $\{\delta_k\} \subset [\delta_{\min}, \delta_{\max}] \subset (0,1)$ *and* $\mathbf{g}^{h,1} \in G^h$. *Let* $\{\mathbf{g}^{h,k}\}_{k=2}^{\infty}$ *be defined by* (4.15) *and* $(u_1^{h,k}, u_2^{h,k}) = S_h^f \mathbf{g}^{h,k}$, $k = 1,2,3,\ldots$. *Then* $E_h^f(\mathbf{g}^{h,k}) \to 0$ *as* $k \to \infty$.

To conclude this finite element section, we prove the geometric convergence for algorithm (4.15).

THEOREM 4.12. *Assume that* $\{\delta_k\} \subset [\delta_{\min}, \delta_{\max}] \subset (0,1]$ *and* $\mathbf{g}^{h,1} \in G^h$. *Let* $\{\mathbf{g}^{h,k}\}_{k=2}^{\infty}$ *be defined by* (4.15), $k = 1,2,3,\ldots$. *Then* $\|\mathbf{g}^{h,k} - \widehat{\mathbf{g}}^h\|_{0,\Gamma} \to 0$ *geometrically as* $k \to \infty$.

Proof. Mimicking the proof of Theorem 3.10 in the finite element context, we have

$$(4.17)\qquad \|\mathbf{g}^{h,k+1} - \widehat{\mathbf{g}}^h\|_{0,\Gamma}^2 = \|\mathbf{g}^{h,k} - \widehat{\mathbf{g}}^h\|_{0,\Gamma}^2 - 4\lambda\delta_{\min} \sum_{i=1}^{2} a_i[u_i^{h,k} - \widehat{u}_i^h, u_i^{h,k} - \widehat{u}_i^h]\,.$$

Invoking [29, Lemma 11] (see also [1]), we may find a $\widetilde{v}_i^h \in X_i^h$ such that

$$\widetilde{v}_i^h|_\Gamma = g_i^{h,k} - \widehat{g}_i^h \quad \text{and} \quad \|\widetilde{v}_i^h\|_{1,\Omega_i} \le C\|g_i^{h,k} - \widehat{g}_i^h\|_{1/2,\Gamma}\,.$$

Subtracting (4.3) from (4.11), setting $v_i^h = \widetilde{v}_i^h$, and then using the Poincaré inequality, trace theorems, and inverse inequalities, we deduce that

$$\begin{aligned}\|g_i^{h,k} - \widehat{g}_i^h\|_{0,\Gamma}^2 &= a_i[u_i^h - \widehat{u}_i^h, \widetilde{v}_i^h] + \lambda^{-1}[u_i^h - \widehat{u}_i^h, \widetilde{v}_i^h]_\Gamma \\ &\le C\|u_i^h - \widehat{u}_i^h\|_{1,\Omega_i}\|\widetilde{v}_i^h\|_{1,\Omega_i} \le C\Big(a_i[u_i^h - \widehat{u}_i^h, u_i^h - \widehat{u}_i^h]\Big)^{1/2}\|g_i^{h,k} - \widehat{g}_i^h\|_{1/2,\Gamma} \\ &\le Ch^{-1/2}\Big(a_i[u_i^h - \widehat{u}_i^h, u_i^h - \widehat{u}_i^h]\Big)^{1/2}\|g_i^{h,k} - \widehat{g}_i^h\|_{0,\Gamma}, \qquad i = 1,2\,.\end{aligned}$$

This leads to

$$(4.18)\qquad a_i[u_i^h - \widehat{u}_i^h, u_i^h - \widehat{u}_i^h] \ge Ch\|g_i^{h,k} - \widehat{g}_i^h\|_{0,\Gamma}^2, \qquad i = 1,2\,.$$

We choose a $\widetilde{\delta} \in (0, \delta_{\min}]$ such that $0 < 1 - 8\lambda C\widetilde{\delta}h < 1$, where C is the constant in (4.18). Combining (4.18) with (4.17), we obtain

$$\|\mathbf{g}^{h,k+1} - \widehat{\mathbf{g}}^h\|_{0,\Gamma}^2 \le (1 - 8\lambda C\delta_{\min}h)\|\mathbf{g}^{h,k} - \widehat{\mathbf{g}}^h\|_{0,\Gamma}^2 \le (1 - 8\lambda C\widetilde{\delta}h)\|\mathbf{g}^{h,k} - \widehat{\mathbf{g}}^h\|_{0,\Gamma}^2\,.$$

Hence $\|\mathbf{g}^{h,k} - \widehat{\mathbf{g}}^h\|_{0,\Gamma} \to 0$ geometrically as $k \to \infty$. □

5. Algorithms and numerical experiments. The generalized method was put into the concise form (3.16), which can be rewritten as

$$(5.1)\qquad \begin{cases} \mathbf{g}^{(1)} \text{ given,} \\ \mathbf{g}^{(k+1)} = (1 - \delta_k)\mathbf{g}^{(k)} + \delta_k R^f \mathbf{g}^{(k)}, \qquad k = 1,2,3,\ldots. \end{cases}$$

According to Theorem 3.9, at the kth iteration step with a known $\mathbf{g}^{(k)}$, the choice of step length $\delta_k = \delta_0(\mathbf{g}^{(k)})$ defined by (3.19) will lead to maximum descent in the descent direction $(R^f\mathbf{g}^{(k)} - \mathbf{g}^{(k)})$. Algorithm I given below will be based on such choices of locally optimal step lengths. Here we give only the continuous version of the algorithm. The corresponding discrete version is obvious.

ALGORITHM I (locally optimal step lengths).

```
specify $(g_1^{(1)}, g_2^{(1)}) \in \mathbf{L}^2(\Gamma)$;
for $k = 1, 2, \dots,$
        – solve for $(u_1^{(k)}, u_2^{(k)}) \in X_1 \times X_2$ from
            $a_i[u_i^{(k)}, v_i] + \lambda^{-1}[u_i^{(k)}, v_i]_\Gamma = [f, v_i]_{\Omega_i} + [g_i^{(k)}, v_i]_\Gamma \quad \forall\, v_i \in X_i\,,\ i = 1, 2;$
        – set $(\widetilde{g}_1^{(k)}, \widetilde{g}_2^{(k)}) = (2u_2^{(k)} - g_2^{(k)}, 2u_1^{(k)} - g_1^{(k)})$ on $\Gamma$;
        – solve for $(\widetilde{u}_1^{(k)}, \widetilde{u}_2^{(k)}) \in X_1 \times X_2$ from
            $a_i[\widetilde{u}_i^{(k)}, v_i] + \lambda^{-1}[\widetilde{u}_i^{(k)}, v_i]_\Gamma = [f, v_i]_{\Omega_i} + [\widetilde{g}_i^{(k)}, v_i]_\Gamma \quad \forall\, v_i \in X_i\,,\ i = 1, 2;$
        – set $(\widetilde{\widetilde{g}}_1^{(k)}, \widetilde{\widetilde{g}}_2^{(k)}) = (2\widetilde{u}_2^{(k)} - \widetilde{g}_2^{(k)}, 2\widetilde{u}_1^{(k)} - \widetilde{g}_1^{(k)})$ on $\Gamma$;
        – calculate
            $\delta_k = \dfrac{[\mathbf{g}^{(k)} - \widetilde{\mathbf{g}}^{(k)}, \mathbf{g}^{(k)} - 2\widetilde{\mathbf{g}}^{(k)} + \widetilde{\widetilde{\mathbf{g}}}^{(k)}]_\Gamma}{\|\mathbf{g}^{(k)} - 2\widetilde{\mathbf{g}}^{(k)} + \widetilde{\widetilde{\mathbf{g}}}^{(k)}\|_{0,\Gamma}^2}\,;$
        – set $(g_1^{(k+1)}, g_2^{(k+1)}) = \delta_k(\widetilde{g}_1^{(k)}, \widetilde{g}_2^{(k)}) + (1 - \delta_k)(g_1^{(k)}, g_2^{(k)})$ on $\Gamma$;
end for
```

Remark 5.1. In Algorithm I the $\widetilde{\widetilde{\mathbf{g}}}^{(k)}$ step is an additional Lions' step, and it is performed purely for the purpose of calculating the optimal step length δ_k.

In Algorithm II below a fixed step length δ is used. In general, one should choose a $\delta \in [1/2, 1)$ since the optimal step lengths are always greater than or equal to 1/2. In the absence of any further guidance as to a good choice of a constant δ, we suggest using the golden ratio constant $(\sqrt{5} - 1)/2 \approx 0.618$. Again, we describe only the continuous version of the algorithm.

ALGORITHM II (fixed step length δ).

```
specify $(g_1^{(1)}, g_2^{(1)}) \in \mathbf{L}^2(\Gamma)$;
for $k = 1, 2, \dots,$
        – solve for $(u_1^{(k)}, u_2^{(k)}) \in X_1 \times X_2$ from
            $a_i[u_i^{(k)}, v_i] + \lambda^{-1}[u_i^{(k)}, v_i]_\Gamma = [f, v_i]_{\Omega_i} + [g_i^{(k)}, v_i]_\Gamma \quad \forall\, v_i \in X_i\,,\ i = 1, 2;$
        – set $(g_1^{(k+1)}, g_2^{(k+1)}) = \delta(2u_2^{(k)} - g_2^{(k)}, 2u_1^{(k)} - g_1^{(k)}) + (1 - \delta)(g_1^{(k)}, g_2^{(k)})$ on
          $\Gamma$;
end for
```

Stopping criteria. Based on Corollary 3.7, we may use one or more of the following as stopping criteria for Algorithms I and II:

$$\|\mathbf{g}^{(k+1)} - \mathbf{g}^{(k)}\|_{0,\Gamma}^2 < \text{tol}; \qquad E^f(\mathbf{g}^{(k)}) - E^f(R^f\mathbf{g}^{(k)}) < \text{tol};$$

$$E^f(\mathbf{g}^{(k)}) < \text{tol}; \qquad \|R^f\mathbf{g}^{(k)} - \mathbf{g}^{(k)}\|_{0,\Gamma}^2 < \text{tol}.$$

Computational results. In Remark 3.14 we explained that we expect the generalized methods with suitable choices of step lengths $\{\delta_k\}$ to converge faster than

Lions' method. We will demonstrate numerically that this is indeed the case. We implemented both Algorithms I and II for the boundary value problem

$$-\Delta u = f \quad \text{in } \Omega\,, \qquad u|_{\partial\Omega} = 0\,.$$

The following three examples were used in our numerical testing.

Example 1. An example with a known exact solution (from [30]):

$$\begin{aligned}&\Omega = (0,2)\times(0,1), \quad \Omega_1 = (0,1)\times(0,1), \quad \Omega_2 = (1,2)\times(0,1),\\ &f = \Delta[(x-2)y(\sin x)\cos(\pi y/2)], \quad \mathbf{g}^{(1)} = (1,1),\\ &16\times 16 \text{ grids for each } \Omega_i,\ i=1,2.\end{aligned}$$

Example 2. An example with a known exact solution having multiple humps:

$$\begin{aligned}&\Omega = (0,2)\times(0,1), \quad \Omega_1 = (0,1)\times(0,1), \quad \Omega_2 = (1,2)\times(0,1),\\ &f = -\Delta[(x-2)y(\sin 2\pi x)\cos(3\pi y/2)], \quad \mathbf{g}^{(1)} = (1,1),\\ &32\times 32 \text{ grids for each } \Omega_i,\ i=1,2.\end{aligned}$$

Example 3. An example with generic choices of f and initial guesses:

$$\begin{aligned}&\Omega = (0,2)\times(0,1), \quad \Omega_1 = (0,1)\times(0,1), \quad \Omega_2 = (1,2)\times(0,1),\\ &f = x^2+y^2+\sin(\pi xy), \quad \mathbf{g}^{(1)} = (2,3),\\ &16\times 16 \text{ grids for each } \Omega_i,\ i=1,2.\end{aligned}$$

For each example we implemented Algorithms I and II with a number of choices of fixed step lengths δ. We used $E_h^f(\mathbf{g}^{h,k}) <$ tol as the stopping criterion. The calculation of $E_h^f(\mathbf{g}^{(k)})$ involved integrals on Γ which were approximated by composite Simpson rules. Also, we choose $\lambda = 1$ in all examples. For each example and for each algorithm (Algorithms I and II with different choices of δ), the computed values of the interface bias $E_h(k) \equiv E_h^f(\mathbf{g}^{h,k})$ for iterations 1 through 5 are given in Tables 5.1–5.3. The optimal step lengths for Examples 1–3, which were calculated in the implementations of Algorithm I, are given in Table 5.4. To illustrate graphically the convergence acceleration of the generalized methods compared to Lions' method, we plot in Figure 5.1 the curves $E_h(k)$ for Algorithm II with $\delta = 1$ (the Lions' method) and Algorithm I (optimal δ_k).

Dependence on mesh sizes and initial guesses. Our computation results suggest that the iterative reduction in $E^f(\mathbf{g}^{(k)})$ for Algorithm I (optimal step lengths) is essentially unaffected by mesh refinements, whereas the reduction in $E^f(\mathbf{g}^{(k)})$ for Lions' original method (i.e., $\delta_k = 1$) markedly slowed down as the mesh was refined; computational results with different mesh sizes for Examples 1 and 2 are plotted in Figure 5.2. See [31] for more computational examples. Also, the performance of Algorithm I is less sensitive to choices of initial guesses than that of Lions' original algorithm. See [31] for further illustrations.

Comparison with gradient methods. For comparison we derive below the gradient method for minimizing the particular interface functional (3.15) subject to the subdomain Robin boundary value problems (3.2). We introduce Lagrange multipliers ξ_i $(i = 1,2)$ and form the Lagrangian functional

$$\begin{aligned}\mathcal{L}(g_1,g_2,u_1,u_2,\xi_1,\xi_2) &= \frac{1}{2}\int_\Gamma |u_1-u_2|^2 + \frac{1}{2}\int_\Gamma |u_1+u_2-g_1-g_2|^2\\ &\quad - \sum_{i=1}^{2}\Big(a_i[u_i,\xi_i] + \lambda^{-1}[u_i,\xi_i]_\Gamma - [f,\xi_i]_{\Omega_i} - \lambda^{-1}[g_i,\xi_i]_\Gamma\Big)\,.\end{aligned}$$

Table 5.1

Implementation results for Algorithms I *and* II *with different* δ *(Example* 1*).*

	$E_h(1)$	$E_h(2)$	$E_h(3)$	$E_h(4)$	$E_h(5)$
Alg. I (opt. δ)	6.669087	0.047283	0.001074	0.000274	0.000085
Alg. II ($\delta = 1$)	6.669087	2.079631	0.838677	0.437171	0.269485
Alg. II ($\delta = 0.8$)	6.669087	0.438941	0.058890	0.011521	0.002526
Alg. II ($\delta = 0.7$)	6.669087	0.098330	0.007416	0.000780	0.000136
Alg. II ($\delta = 0.618$)	6.669087	0.057681	0.001302	0.000321	0.000145
Alg. II ($\delta = 0.5$)	6.669087	0.376576	0.025433	0.002160	0.000397
Alg. II ($\delta = 0.3$)	6.669087	1.934113	0.571690	0.171613	0.052275

Table 5.2

Implementation results for Algorithms I *and* II *with different* δ *(Example* 2*).*

	$E_h(1)$	$E_h(2)$	$E_h(3)$	$E_h(4)$	$E_h(5)$
Alg. I (opt. δ)	6.825659	1.083740	0.366894	0.139247	0.057358
Alg. II ($\delta = 1$)	6.825088	2.438442	1.115827	0.642416	0.430271
Alg. II ($\delta = 0.8$)	6.825088	1.287785	0.453566	0.194676	0.092025
Alg. II ($\delta = 0.7$)	6.825088	1.092706	0.492522	0.249100	0.132673
Alg. II ($\delta = 0.618$)	6.825088	1.121899	0.584488	0.324944	0.185895
Alg. II ($\delta = 0.5$)	6.825088	1.463038	0.773638	0.475356	0.301023
Alg. II ($\delta = 0.3$)	6.825088	2.847364	1.527228	0.982614	0.697551

Table 5.3

Implementation results for Algorithms I *and* II *with different* δ *(Example* 3*).*

	$E_h(1)$	$E_h(2)$	$E_h(3)$	$E_h(4)$	$E_h(5)$
Alg. I (opt. δ)	26.827161	0.317631	0.041461	0.011125	0.003360
Alg. II ($\delta = 1$)	26.827161	9.480301	4.384460	2.509528	1.618675
Alg. II ($\delta = 0.8$)	26.827161	2.254784	0.364381	0.077773	0.018560
Alg. II ($\delta = 0.7$)	26.827161	0.647391	0.072631	0.016409	0.006017
Alg. II ($\delta = 0.618$)	26.827161	0.326878	0.041768	0.018543	0.009033
Alg. II ($\delta = 0.5$)	26.827161	1.443123	0.140078	0.035929	0.017449
Alg. II ($\delta = 0.3$)	26.827161	7.586218	2.233397	0.692824	0.233357

Table 5.4

Optimal step lengths in Algorithm I *for Examples* 1–3.

Iteration k	1	2	3	4	5
δ_k (Example 1)	–	0.643500	0.573260	0.827441	0.961937
δ_k (Example 2)	–	0.629759	0.618903	1.197064	0.667865
δ_k (Example 3)	–	0.673044	1.247752	0.670146	1.282857

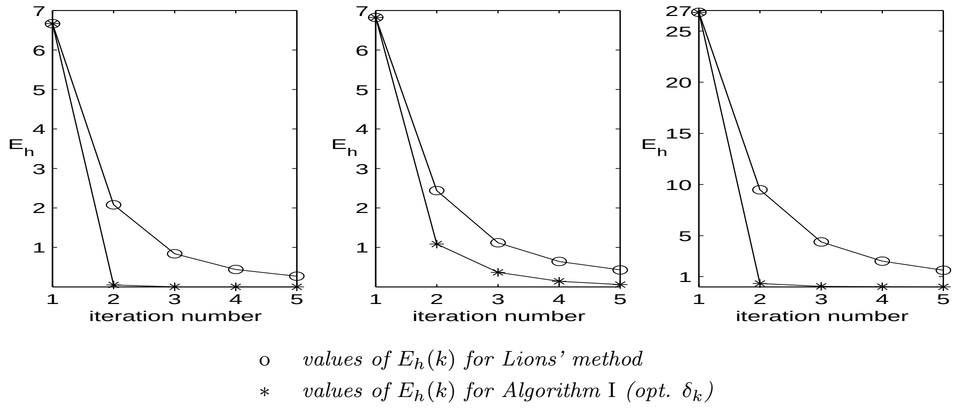


FIG. 5.1. *$E_h(k)$ curve for Example* 1 *(left), Example* 2 *(center), and Example* 3 *(right).*

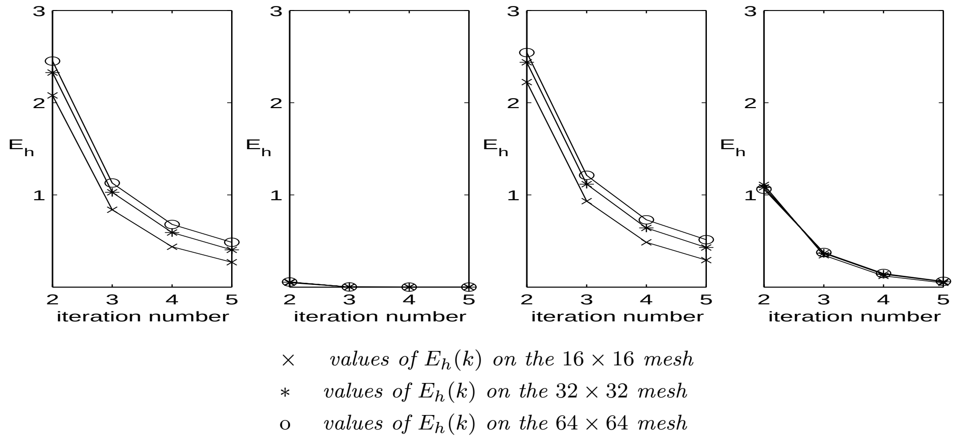


FIG. 5.2. *$E_h(k)$ curves for Examples* 1–2 *by Lions' method and Algorithm* I *on three different meshes. First plot from left: Example* 1*, Lions' method; second from left: Example* 1*, Algorithm* I; *third plot from left: Example* 2*, Lions' method; fourth from left: Example* 2*, Algorithm* I.

Taking variations of the Lagrangian $\mathcal{L}$ with respect to u_1 and u_2, respectively, we obtain

$$a_1[\xi_1, w_1] + \lambda^{-1}[\xi_1, w_1]_\Gamma = [2u_1 - g_1 - g_2, w_1]_\Gamma \qquad \forall\, w_1 \in X_1 \tag{5.2}$$

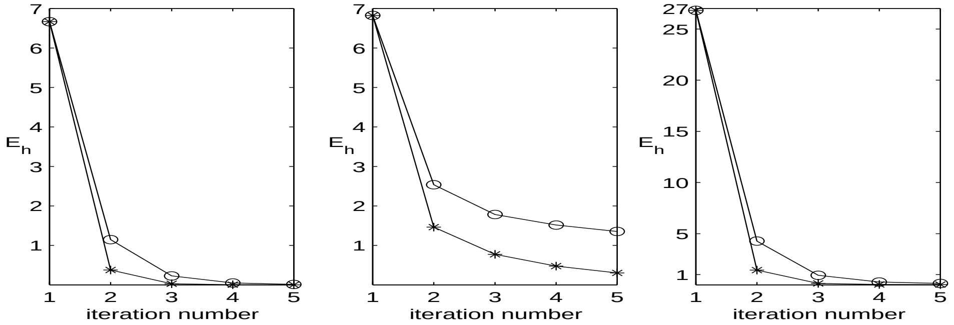
and

$$a_2[\xi_2, w_2] + \lambda^{-1}[\xi_2, w_2]_\Gamma = [2u_2 - g_1 - g_2, w_2]_\Gamma \qquad \forall\, w_2 \in X_2\,. \tag{5.3}$$

By taking variations of the Lagrangian $\mathcal{L}$ with respect to g_1 and g_2, respectively, we obtain the Frechét derivative of $E^f(g_1, g_2)$:

$$\begin{aligned}\langle (E^f)'(g_1, g_2), (z_1, z_2)\rangle = &- [u_1 + u_2 - g_1 - g_2, z_1]_\Gamma - [u_1 + u_2 - g_1 - g_2, z_2]_\Gamma \\ &+ \lambda^{-1}[z_1, \xi_1]_\Gamma + \lambda^{-1}[z_2, \xi_2]_\Gamma \qquad \forall\,(z_1, z_2) \in \mathbf{L}^2(\Gamma)\,,\end{aligned}$$

i.e., $(E^f)'(g_1, g_2) = (\lambda^{-1}\xi_1 - u_1 - u_2 + g_1 + g_2, \lambda^{-1}\xi_2 - u_1 - u_2 + g_1 + g_2)$, where ξ_1 and ξ_2 are defined by (5.2) and (5.3), respectively. Hence the kth step iteration formulae



o *values of* $E_h(k)$ *obtained by gradient method* ($\delta_k \equiv \delta = 0.5$)

$*$ *values of* $E_h(k)$ *obtained by Algorithm* II ($\delta = 0.5$)

FIG. 5.3. $E_h(k)$ *curves for Example* 1 *(left), Example* 2 *(center), and Example* 3 *(right).*

for the variable step length gradient method is given by

$$\begin{aligned}(g_1^{(k+1)}, g_2^{(k+1)}) &= (g_1^{(k)}, g_2^{(k)})\\ &\quad - \delta_k(\lambda^{-1}\xi_1^{(k)} - u_1^{(k)} - u_2^{(k)} + g_1^{(k)} + g_2^{(k)}, \lambda^{-1}\xi_2^{(k)} - u_1^{(k)} - u_2^{(k)} + g_1^{(k)} + g_2^{(k)}),\end{aligned}$$

where $u_i^{(k)}$, $i = 1, 2$, are defined by

$$a_i[u_i^{(k)}, v_i] + \lambda^{-1}[u_i^{(k)}, v_i]_\Gamma = [f, v_i]_{\Omega_i} + \lambda^{-1}[g_i^{(k)}, v_i]_\Gamma \qquad \forall\, v_i \in X_i\,,\ i = 1, 2\,,$$

and $\xi_i^{(k)}$, $i = 1, 2$, are defined by

$$a_i[\xi_i^{(k)}, w_i] + \lambda^{-1}[\xi_i^{(k)}, w_i]_\Gamma = [2u_i^{(k)} - g_1^{(k)} - g_2^{(k)}, w_i]_\Gamma \qquad \forall\, w_i \in X_i\,.$$

We implemented the gradient method with a fixed step length $\delta = 0.5$ for Examples 1–3. The results are plotted in Figure 5.3. For comparison, computational results for Algorithm II with a fixed step length $\delta = 0.5$ are also plotted in Figure 5.3.

Neumann boundary value problems. Algorithms I and II can be easily adapted to treat the case of a Neumann boundary value problem

$$-\text{div}\,[A(\mathbf{x})\nabla u] = f \quad \text{in } \Omega, \qquad [A(\mathbf{x})\nabla u]\cdot\mathbf{n} = d \quad \text{on } \partial\Omega\,,$$

where $f \in L^2(\Omega)$ and $d \in L^2(\partial\Omega)$. The results of sections 3 and 4 also can be carried over to this case with straightforward modifications, e.g., replacing $[f, v]$ by $[f, v] + [d, v]_{\partial\Omega}$ and $[f, v_i]_{\Omega_i}$ by $[f, v_i]_{\Omega_i} + [d, v_i]_{\Gamma_i}$.

Acknowledgment. The authors thank anonymous referees for various useful suggestions.

REFERENCES

[1] M. BERGGREN, *Approximations of very weak solutions to boundary-value problems*, SIAM J. Numer. Anal., to appear.

[2] P. BJØRSTAD AND O. B. WIDLUND, *Solving elliptic problems on regions partitioned into substructures*, in Elliptic Problem Solvers II, G. Birkhoff and A. Schoenstadt, eds., Academic Press, New York, 1984, pp. 245–255.

[3] P. BJØRSTAD AND O. B. WIDLUND, *Iterative methods for the solution of elliptic problems on regions partitioned into substructures*, SIAM J. Numer. Anal., 23 (1986), pp. 1097–1120.

[4] J. Bramble, J. Pasciak, and A. Schatz, *The construction of preconditioners for elliptic problems by substructuring* I, Math. Comp., 47 (1986), pp. 103–134.

[5] J. Bramble, J. Pasciak, and A. Schatz, *The construction of preconditioners for elliptic problems by substructuring* II, Math. Comp., 49 (1987), pp. 1–16.

[6] J. Bramble, J. Pasciak, and A. Schatz, *The construction of preconditioners for elliptic problems by substructuring* III, Math. Comp., 51 (1988), pp. 415–430.

[7] J. Bramble, J. Pasciak, and A. Schatz, *The construction of preconditioners for elliptic problems by substructuring* IV, Math. Comp., 53 (1989), pp. 1–24.

[8] J. Bramble, J. Pasciak, and A. Vassilev, *Analysis of non-overlapping domain decomposition algorithms with inexact solves*, Math. Comp., 67 (1998), pp. 1–19.

[9] S. C. Brenner, *The condition number of the Schur complement in domain decomposition*, Numer. Math., 83 (1999), pp. 187–203.

[10] T. Chan and D. C. Resasco, *A Survey of Preconditioners for Domain Decomposition*, Technical Report, /DCS/RR-414, Yale University, New Haven, CT, 1985.

[11] T. Chan, R. Glowinski, J. Périaux, and O. B. Widlund, *Domain Decomposition Methods*, SIAM, Philadelphia, 1989.

[12] T. Chan, R. Glowinski, J. Périaux, and O. B. Widlund, *Proceedings of the Third International Symposium on Domain Decomposition Methods for Partial Differential Equations*, SIAM, Philadelphia, PA, 1990.

[13] T. Chan, D. Keyes, G. Meurant, J. Scroggs, and R. Voigt, *Proceedings of the Fifth International Symposium on Domain Decomposition Methods for Partial Differential Equations*, SIAM, Philadelphia, PA, 1992.

[14] T. Chan and T. Mathew, *Domain decomposition algorithms*, in Acta Numerica 1994, Acta Numer., Cambridge University Press, Cambridge, UK, 1994, pp. 61–143.

[15] T. Chan and J. Zou, *A convergence theory of multilevel additive Schwarz methods on unstructured meshes*, Numer. Algorithms, 13 (1996), pp. 365–398.

[16] P. Ciarlet, *Finite Element Methods for Elliptic Problems*, North–Holland, Amsterdam, 1978.

[17] Q.-P. Deng, *An analysis for a nonoverlapping domain decomposition iterative procedure*, SIAM J. Sci. Comput., 18 (1997), pp. 1517–1525.

[18] Q. Dinh, R. Glowinski, and J. Périaux, *Solving elliptic problems by domain decomposition methods with application*, in Elliptic Problem Solvers II, G. Birkoff and A. Schoenstadt, eds., Academic Press, New York, 1984, pp. 395–426.

[19] M. Dryja, B. Smith, and O. B. Widlund, *Schwarz analysis of iterative substructuring algorithms for elliptic problems in three dimensions*, SIAM J. Numer. Anal., 31 (1994), pp. 1662–1694.

[20] Q. Du, *Optimization based nonoverlapping domain decomposition algorithms and their convergence*, SIAM J. Numer. Anal., 39 (2001), pp. 1056–1077.

[21] Q. Du and M. D. Gunzburger, *A gradient method approach to optimization-based multidisciplinary simulations and nonoverlapping domain decomposition algorithms*, SIAM J. Numer. Anal., 37 (2000), pp. 1513–1541.

[22] D. Funaro, A. Quarteroni, and P. Zanolli, *An iterative procedure with interface relaxation for domain decomposition methods*, SIAM J. Numer. Anal., 25 (1988), pp. 1213–1236.

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

[24] R. Glowinski, G. Golub, G. Meurant, and J. Périaux, *Proceedings of the First International Symposium on Domain Decomposition Methods for Partial Differential Equations*, SIAM, Philadelphia, 1988.

[25] R. Glowinski and P. Le Tallec, *Augmented Lagrangian interpretation of the nonoverlapping Schwarz alternating method*, in Third International Symposium on Domain Decomposition Methods for Partial Differential Equations, T. Chan, R. Glowinski, J. Périaux, and O. B. Wildlund, eds., SIAM, Philadelphia, PA, 1990, pp. 224–231.

[26] R. Glowinski, Y. Kuznetsov, G. Meurant, J. Périaux, and O. B. Widlund, *Proceedings of the Fourth International Symposium on Domain Decomposition Methods for Partial Differential Equations*, SIAM, Philadelphia, PA, 1991.

[27] R. Glowinski, J. Périaux, Z. Shi, and O. B. Widlund, *Domain Decomposition Methods in Scientific and Engineering Computing*, John Wiley & Sons, New York, 1996.

[28] M. D. Gunzburger, M. Heinkenschloss, and H. K. Lee, *Solution of elliptic partial differential equations by an optimization-based domain decomposition method*, Appl. Math. Comput., 113 (2000), pp. 111–139.

[29] M. D. Gunzburger and S. Hou, *Treating inhomogeneous essential boundary conditions in finite element methods and the calculation of boundary stresses*, SIAM J. Numer. Anal., 29 (1992), pp. 390–424.

[30] M. D. Gunzburger, H. Lee, and J. Peterson, *An optimization based domain decomposition method for partial differential equations*, Comput. Math. Appl., 37 (1999), pp. 77–93.

[31] W. Guo, *Domain Decomposition Methods for the Solutions of Partial Differential Equations*, Ph.D. thesis, York University, Toronto, Canada, 2001.

[32] A. Hadjidimos, D. Noutsos, and M. Tzoumas, *Nonoverlapping domain decomposition: A linear algebra viewpoint*, Math. Comput. Simulation, 51 (2000), pp. 597–625.

[33] D. Keyes and J. Xu, *Domain Decomposition Methods in Scientific and Engineering Computing*, AMS, Providence, RI, 1994.

[34] E. Larsson, *A domain decomposition method for the Helmholtz equation in a multilayer domain*, SIAM J. Sci. Comput., 20 (1999), pp. 1713–1731.

[35] P.-L. Lions, *On the Schwarz alternating method* I, in First International Symposium on Domain Decomposition Methods for Partial Differential Equations, R. Glowinski, G. H. Golub, G. A. Meurant, and J. Périaux, eds., SIAM, Philadelphia, PA, 1988, pp. 1–42.

[36] P.-L. Lions, *On the Schwarz alternating methods* III: *A variant for nonoverlapping subdomains*, in Third International Symposium on Domain Decomposition Methods for Partial Differential Equations, T. F. Chan, R. Glowinski, J. Périaux, and O. B. Wildlund, eds., SIAM, Philadelphia, 1990, pp. 202–223.

[37] S. H. Lui, *On Schwarz alternating methods for nonlinear elliptic PDEs*, SIAM J. Sci. Comput., 21 (2000), pp. 1506–1523.

[38] S. H. Lui, *On Schwarz alternating methods for the incompressible Navier–Stokes equations*, SIAM J. Sci. Comput., 22 (2001), pp. 1974–1986.

[39] S. H. Lui, *On linear monotone and Schwarz alternating methods for nonlinear elliptic PDEs*, Numer. Math., 93 (2002), pp. 109–129.

[40] L. Marini and A. Quarteroni, *A relaxation procedure for domain decomposition methods using finite elements*, Numer. Math., 55 (1989), pp. 575–598.

[41] A. Matsokin and S. Nepomnyashchikh, *On the convergence of the alternating subdomain Schwartz method without intersections*, Sov. J. Numer. Anal. Math. Modelling, 4 (1989), pp. 479–485.

[42] J. R. Rice, E. A. Vavalis, and D. Y. Yang, *Convergence analysis of a nonoverlapping domain decomposition method for elliptic PDEs*, J. Comput. Appl. Math., 87 (1997), pp. 11–19.

[43] H. A. Schwarz, *Über einige Abbildungsdufgaben*, J. Reine Angew. Math., 70 (1869), pp. 105–120.

[44] H. Sun and W.-P. Tang, *An overdetermined Schwarz alternating method*, SIAM J. Sci. Comput., 17 (1996), pp. 884–905.

[45] X.-C. Tai and J. Xu, *Global convergence of subspace correction methods for convex optimization problems*, Math. Comp., 71 (2002), pp. 105–124.

[46] W.-P. Tang, *Generalized Schwarz splittings*, SIAM J. Sci. Statist. Comput., 13 (1992), pp. 573–595.

[47] J. Xu and J. Zou, *Some nonoverlapping domain decomposition methods*, SIAM Rev., 40 (1998), pp. 857–914.

[48] D. Q. Yang, *A parallel iterative nonoverlapping domain decomposition algorithm for elliptic problems*, IMA J. Numer. Anal., 16 (1996), pp. 75–91.

[49] D. Q. Yang, *A nonoverlapping subdomain algorithm with Lagrange multipliers and its object oriented implementation for interface problems*, in Domain Decomposition Methods (Boulder, CO, 1997), Contemp. Math. 218, AMS, Providence, RI, 1998, pp. 389–397.

[50] J. Xu, *Iterative methods by space decomposition and subspace correction*, SIAM Rev., 34 (1992), pp. 581–613.

SIAM J. NUMER. ANAL.
Vol. 41, No. 6, pp. 2081–2095

A FAST NUMERICAL METHOD FOR THE BLACK–SCHOLES EQUATION OF AMERICAN OPTIONS*

HOUDE HAN† AND XIAONAN WU‡

Abstract. This paper introduces a fast numerical method for computing American option pricing problems governed by the Black–Scholes equation. The treatment of the free boundary is based on some properties of the solution of the Black–Scholes equation. An artificial boundary condition is also used at the other end of the domain. The finite difference method is used to solve the resulting problem. Computational results are given for some American call option problems. The results show that the new treatment is very efficient and gives better accuracy than the normal finite difference method.

Key words. American option, free boundary, artificial boundary condition, finite difference method

AMS subject classifications. 35A35, 35A40, 65N99

DOI. 10.1137/S0036142901390238

1. Introduction.

In option pricing theory, the Black–Scholes equation is one of the effective models for option pricing [2]. For European options, the Black–Scholes equation results in a boundary value problem of a diffusion equation. For American options, the Black–Scholes equation results in a free boundary value problem. There are usually two ways to solve the option pricing problem—the analytic and numerical approaches. For European options, the analytic solution is relatively easier to obtain. For the analytic approach, efforts have been mainly on the American options. Johnson [16] and MacMillan [18] use analytical approximation for American puts on a nondividend paying stock. For American options on dividend paying stock, Geske and Johnson [10] give an analytic solution in a series form. When closed form solutions cannot be obtained, or when the formulas for the exact solutions are too difficult to be practically usable, numerical solution is a natural way to solve the problem. The binomial method is a simple and very effective method for solving American options; this is introduced by Cox, Ross, and Rubinstein [7], and the convergence of the binomial method for American options is proved by Amin and Khanna [1]. Brennan and Schwartz [3], [4] and Schwartz [19] introduced finite difference methods for solving American options. Jaillet, Lamberton, and Lapeyre [15] show the convergence of the finite difference method. A comparison of different numerical methods for option pricing can be found in [5], [11].

In solving the Black–Scholes equation for American options, a natural approach is to transform the original equation to a standard forward diffusion equation over an infinite domain. The finite difference method is applied to the equation over a truncated finite domain, and the original asymptotic infinite boundary conditions are shifted to the ends of the truncated finite domain. To avoid generating large errors

*Received by the editors May 31, 2001; accepted for publication (in revised form) April 1, 2003; published electronically November 25, 2003.
http://www.siam.org/journals/sinum/41-6/39023.html

†Department of Mathematical Sciences, Tsinghua University, Beijing, 100084, China (hhan@math.tsinghua.edu.cn). This research was supported in part by the Special Funds for Major State Basic Research Projects of China.

‡Department of Mathematics, Hong Kong Baptist University, Kowloon Tong, Hong Kong (xwu@hkbu.edu.hk). This research was supported in part by RGC of Hong Kong and FRG of Hong Kong Baptist University.

in the solution due to this approximation of the boundary conditions, the truncated domain must be large enough, which results in a large cost. Obviously, a large part of the finite difference solution is actually useless, and the reason to compute these is only to guarantee the accuracy of the rest of the solutions. Kangro and Nicolaides [17] give error estimates of the numerical solutions with far field boundary conditions. Artificial boundary conditions have been applied to different problems on infinite domains; see, for examples, [8], [12], [13], [14]. In this paper, we find the accurate boundary conditions on the far boundary for the American option problem, which is actually a relation between the function and its partial derivatives. Then this boundary condition is discretized and combined with the finite difference discretization for the partial differential equation. With these boundary conditions, we can make the computational domain small and obtain accurate solutions. For the free boundary, we give some properties of the solution of the Black–Scholes equations. Using these properties, we design a simple numerical method to determine the location of the free boundary. Some computational results are given for the American options with dividend paying, and the results are compared with approximations using standard finite difference methods.

The computational results show that these algorithms give more accurate numerical results than the standard finite difference approximation. With a relatively small truncated domain, the standard finite difference method usually cannot give satisfactory results. Our algorithms give more accurate numerical results, and the option price can be obtained for all the asset prices.

2. Some properties of the solution of the Black–Scholes equation. Assume that S is the asset price, t is the time, and C is the call option value. Let r denote the risk-free interest rate, let σ denote the volatility of the asset price, and let D_0 denote the continuous dividend yield. Then the call value of the American option is given by the free boundary value problem of the Black–Scholes equation [20],

$$\begin{aligned} &\frac{\partial C}{\partial t} + \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 C}{\partial S^2} + (r - D_0)S\frac{\partial C}{\partial S} - rC = 0, \\ &0 < S < S_f(t), \qquad 0 \le t < T, \end{aligned} \tag{1}$$

where $S_f(t)$ is the free boundary of early exercise. The final and boundary conditions are given by

$$C(S,T) = h(S), \qquad 0 \le S \le S_f(T) = S_0, \tag{2}$$

$$C(S_f(t),t) = h(S_f(t)), \quad \frac{\partial C}{\partial S}(S_f(t),t) = 1, \qquad 0 \le t \le T, \tag{3}$$

$$C(S,t) \to 0 \quad \text{as} \quad S \to 0, \qquad 0 \le t \le T, \tag{4}$$

where $h(S) = \max(S - E, 0)$, with $E > 0$ and $S_0 = \max(E, rE/D_0)$.

We introduce the change of variable for t:

$$t = T - \frac{2\tau}{\sigma^2}.$$

Denote

$$\begin{aligned}&C^*(S,\tau)=C(S,t)=C(S,T-2\tau/\sigma^2),\\&S_f^*(\tau)=S_f(T-2\tau/\sigma^2),\\&r^*=\frac{2r}{\sigma^2},\\&D^*=\frac{2D_0}{\sigma^2},\\&\tau^*=\frac{\sigma^2T}{2}.\end{aligned}$$

Then the free boundary value problem (1)–(4) is equivalent to the following problem:

$$LC^*\equiv-\frac{\partial C^*}{\partial\tau}+S^2\frac{\partial^2C^*}{\partial S^2}+(r^*-D^*)S\frac{\partial C^*}{\partial S}-r^*C^*=0,\tag{5}$$
$$0<S<S_f^*(\tau),\qquad 0<\tau<\tau^*,$$
$$C^*(S,0)=h(S),\qquad 0\le S\le S_f^*(0),\tag{6}$$
$$C^*(S_f^*(\tau),\tau)=h(S_f^*(\tau)),\qquad\frac{\partial C^*}{\partial S}(S_f^*(\tau),\tau)=1,\qquad 0\le\tau\le\tau^*,\tag{7}$$
$$C^*(S,\tau)\to0\quad\text{as}\quad S\to0.\tag{8}$$

Let

$$\begin{aligned}&k'=r^*-D^*,\\&\alpha=\frac{-(k'-1)}{2},\qquad\beta=\frac{-(k'-1)^2}{4}-r^*.\end{aligned}$$

Furthermore, we introduce the change of variables

$$\begin{aligned}&S=Ee^x,\\&C^*(S,\tau)=Ee^{\alpha x+\beta\tau}u(x,\tau).\end{aligned}$$

Then the free boundary value problem (5)–(8) is equivalent to the following problem:

$$\frac{\partial u}{\partial\tau}=\frac{\partial^2u}{\partial x^2},\qquad-\infty<x<x_f(\tau),\tag{9}$$
$$u(x,0)=g(x,0),\qquad-\infty<x\le x_f(0),\tag{10}$$
$$u(x_f(\tau),\tau)=g(x_f(\tau),\tau),\ \alpha u(x_f(\tau),\tau)+\frac{\partial u(x_f(\tau),\tau)}{\partial x}\tag{11}$$
$$=e^{(1-\alpha)x_f(\tau)-\beta\tau},\ 0\le\tau\le\tau^*,$$
$$u(x,\tau)\to0\text{ as }x\to-\infty,\tag{12}$$

where

$$g(x,\tau)=e^{-\alpha x-\beta\tau}\max(e^x-1,0).$$

The free boundary $x_f(\tau)=\ln(S_f^*(\tau)/E)$. It is known that $x_f(\tau)>0$ for $\tau>0$.

We now consider the problem (5)–(8). Let

$$W=\frac{\partial C^*(S,\tau)}{\partial S};$$

then W satisfies

$$\begin{gathered}-\frac{\partial W}{\partial \tau}+S^2\frac{\partial^2 W}{\partial S^2}+(2+r^*-D^*)S\frac{\partial W}{\partial S}-D^*W=0,\\ 0<S<S_f^*(\tau),\qquad 0<\tau\le\tau^*,\end{gathered}\tag{13}$$

$$W(S,0)=0,\qquad 0<S<E,\qquad \text{and } W(S,0)=1,\qquad E<S<S_f^*(0),\tag{14}$$

$$W(S_f^*(\tau),\tau)=1,\qquad 0<\tau\le\tau^*.\tag{15}$$

For $W(S,\tau)$, we have the following lemma.

Lemma 1.

$$W(S,\tau)=\frac{\partial C^*(S,\tau)}{\partial S}\to 0 \textit{ when } S\to 0^+.$$

Proof. Since

$$\begin{aligned}W(S,\tau)&=\frac{\partial C^*(S,\tau)}{\partial S}\\ &=\frac{\partial}{\partial S}\left(Ee^{\alpha x+\beta\tau}u(x,\tau)\right)\\ &=\frac{\partial}{\partial x}\left(Ee^{\alpha x+\beta\tau}u(x,\tau)\right)\frac{dx}{dS}\\ &=e^{(\alpha-1)x+\beta\tau}\left(\frac{\partial u(x,\tau)}{\partial x}+\alpha u(x,\tau)\right),\end{aligned}\tag{16}$$

where $u(x,\tau)$ satisfies (9)–(12), let $\phi(\tau)=u(0,\tau)$, $\phi(0)=0$, and then [6]

$$\begin{aligned}u(x,\tau)&=\frac{-x}{2\sqrt{\pi}}\int_0^\tau e^{-\frac{x^2}{4(\tau-\lambda)}}\frac{\phi(\lambda)d\lambda}{(\tau-\lambda)^{3/2}},\qquad x<0,\\ |u(x,\tau)|&\le\frac{|x|}{2\sqrt{\pi}}\Phi\int_0^\tau e^{-\frac{x^2}{4(\tau-\lambda)}}\frac{d\lambda}{(\tau-\lambda)^{3/2}}\\ &\le\frac{4\Phi}{\sqrt{\pi}}\int_0^\tau e^{-\frac{x^2}{8(\tau-\lambda)}}\frac{d\lambda}{\sqrt{\tau-\lambda}}\\ &=\frac{4\Phi}{\sqrt{\pi}}e^{-\frac{x^2}{8\tau}}\int_0^\tau e^{-\frac{x^2}{8}\left(\frac{1}{\tau-\lambda}-\frac{1}{\tau}\right)}\frac{d\lambda}{\sqrt{\tau-\lambda}}\\ &\le\frac{8\Phi}{\sqrt{\pi}}\sqrt{\tau}e^{-\frac{x^2}{8\tau}},\qquad x\le-1,\quad 0\le\tau\le\tau^*,\end{aligned}\tag{17}$$

where

$$\Phi=\max_{0\le\lambda\le\tau^*}|\phi(\lambda)|.$$

Similarly,

$$\left|\frac{\partial u(x,\tau)}{\partial x}\right|\le C_0\Phi\sqrt{\tau}e^{-\frac{x^2}{8\tau}},\qquad x\le-1,\quad 0\le\tau\le\tau^*,\tag{18}$$

where C_0 is a constant. Combining (16)–(18), we obtain

$$\lim_{S\to 0^+}W(S,\tau)=\lim_{S\to 0^+}\frac{\partial C^*(S,\tau)}{\partial S}=0.\qquad\square\tag{19}$$

Finally, we know that $W(S,\tau)$ is the solution of problem (13)–(15) and (19). By the strong maximum principle of the parabolic equation [9] we have the following theorem.

THEOREM 1. *$W(S,\tau)$ satisfies the following inequality:*

$$0 < W(S,\tau) < 1, \quad 0 < S < S_f^*(\tau), \quad 0 < \tau \le \tau^*.$$

Namely,

$$0 < \frac{\partial C^*(S,\tau)}{\partial S} < 1, \quad 0 < S < S_f^*(\tau), \quad 0 < \tau \le \tau^*,$$

and

$$0 < e^{(\alpha-1)x+\beta\tau}\left(\frac{\partial u(x,\tau)}{\partial x} + \alpha u(x,\tau)\right) < 1, \quad 0 < x < x_f(\tau), \quad 0 < \tau \le \tau^*. \tag{20}$$

For the solution $\{C^*(S,\tau), S_f^*(\tau)\}$ of the problem (5)–(8), we extend $C^*(S,\tau)$ to the domain

$$S_f^*(\tau) < S < +\infty, \quad 0 \le \tau \le \tau^*,$$

by

$$C^*(S,\tau) = h(S), \quad S_f^*(\tau) < S < +\infty, \quad 0 \le \tau \le \tau^*.$$

For a given smooth boundary $S = \hat{S}(\tau)$ and a given τ_j, $0 < \tau_j < \tau^*$, satisfying

$$S_f^*(\tau) < \hat{S}(\tau), \quad \tau_j < \tau \le \tau^*,$$

consider the following auxiliary problem:

$$L\hat{C}(S,\tau) = 0, \quad 0 < S < \hat{S}(\tau), \quad \tau_j < \tau \le \tau^*, \tag{21}$$

$$\hat{C}(S,\tau) \to 0, \quad S \to 0, \tag{22}$$

$$\hat{C}(\hat{S}(\tau),\tau) = h(\hat{S}(\tau)), \quad \tau_j < \tau \le \tau^*, \tag{23}$$

$$\hat{C}(S,\tau_j) = C^*(S,\tau_j), \quad 0 \le S \le \hat{S}(\tau_j). \tag{24}$$

Problem (21)–(24) has a solution $\hat{C}(S,\tau)$ on

$$\Omega = \{(S,\tau) \mid 0 < S < \hat{S}(\tau),\ \tau_j \le \tau \le \tau^*\}.$$

Let

$$\varepsilon(S,\tau) = C^*(S,\tau) - \hat{C}(S,\tau).$$

A computation shows that

$$L\varepsilon(S,\tau) = \begin{cases} 0, & 0 < S < S_f^*(\tau),\ \tau_j < \tau \le \tau^*, \\ -D^*S + r^*E \le 0, & S_f^*(\tau) \le S \le \hat{S}(\tau),\ \tau_j < \tau \le \tau^*, \end{cases}$$

and

$$\frac{\partial \varepsilon(S,\tau)}{\partial S} \text{ is continuous on } S = S_f^*(\tau), \quad \tau_j < \tau \le \tau^*.$$

From the strong maximum principle we get

$$\varepsilon(S,\tau) > 0, \qquad 0 < S < \hat{S}(\tau), \quad \tau_j < \tau \le \tau^*.$$

When $S_f^*(\tau) < S < \hat{S}(\tau)$, we have

$$\varepsilon(S,\tau) = C^*(S,\tau) - \hat{C}(S,\tau) = h(S) - \hat{C}(S,\tau) > 0,$$

namely,

$$\hat{C}(S,\tau) < h(S), \qquad S_f^*(\tau) < S < \hat{S}(\tau), \quad \tau_j < \tau \le \tau^*. \tag{25}$$

Let

$$\hat{C}(S,\tau) = Ee^{\alpha x+\beta\tau}\hat{u}(x,\tau)$$

with $S = Ee^x$. Then the auxiliary problem (21)–(24) is equivalent to the problem

$$\frac{\partial \hat{u}}{\partial \tau} = \frac{\partial^2 \hat{u}}{\partial x^2}, \qquad -\infty < x < \hat{x}_f(\tau), \tag{26}$$

$$\hat{u}(x,0) = u(x,0), \qquad -\infty < x \le \hat{x}_f(\tau), \tag{27}$$

$$\hat{u}(\hat{x}_f(\tau),\tau) = g(\hat{x}_f(\tau),\tau), \tag{28}$$

$$\hat{u}(x,\tau) \to 0 \text{ as } x \to -\infty, \tag{29}$$

where $\hat{x}_f(\tau) = \ln(\hat{S}(\tau)/E) \ge x_f(\tau)$, $\tau_j < \tau \le \tau^*$.

From inequality (25) we obtain the following theorem.

THEOREM 2. *For the solution $\hat{u}(x,\tau)$ of the auxiliary problem* (26)–(29) *the following inequality holds:*

$$\hat{u}(x,\tau) < g(x,\tau), \qquad x_f(\tau) < x < \hat{x}_f(\tau), \quad \tau_j < \tau \le \tau^*. \tag{30}$$

The inequality (30) is very useful for determining the location of the free boundary in the numerical schemes.

3. An exact boundary condition on the artificial boundary. We now return to the problem (9)–(12), which is defined on an unknown unbounded domain $\bar{\Omega}$:

$$\bar{\Omega} = \{(x,\tau) \mid -\infty < x < x_f(\tau),\ 0 < \tau \le \tau^*\}.$$

It is known that the free boundary $x_f(\tau) > 0$, $0 < \tau \le \tau^*$. Let $a < 0$ be a real number. We introduce an artificial boundary Γ_a:

$$\Gamma_a = \{(x,\tau) \mid x = a,\ 0 < \tau \le \tau^*\}.$$

The artificial boundary Γ_a divides the domain Ω into two parts, the bounded part Ω_i and the unbounded part Ω_e:

$$\Omega_i = \{(x,\tau) \mid a < x < x_f(\tau),\ 0 < \tau \le \tau^*\},$$
$$\Omega_e = \{(x,\tau) \mid -\infty < x < a,\ 0 < \tau \le \tau^*\}.$$

If we can find a suitable boundary condition on the artificial boundary Γ_a, then the problem (9)–(12) can be reduced on the bounded domain Ω_i. On Ω_e, the solution of (9)–(12), $u(x,\tau)$, satisfies

$$\frac{\partial u}{\partial \tau} = \frac{\partial^2 u}{\partial x^2}, \qquad -\infty < x < a, \quad 0 < \tau \le \tau^*, \tag{31}$$

$$u(x,0) = 0, \qquad -\infty < x < a. \tag{32}$$

The problem (31)–(32) is an incompletely posed problem. If we know the value of $u(x,\tau)$ on the boundary Γ_a,

$$u(a,\tau) = \phi(\tau) \tag{33}$$

with $\phi(0)=0$, then the solution of (31)–(33) is given by [6]

$$u(x,\tau) = \frac{-(x-a)}{2\sqrt{\pi}} \int_0^\tau e^{\frac{-(x-a)^2}{4(\tau-\lambda)}} \frac{\phi(\lambda)d\lambda}{(\tau-\lambda)^{3/2}}. \tag{34}$$

Introducing the new variable

$$\mu = \frac{x-a}{2\sqrt{\tau-\lambda}},$$

we get

$$u(x,\tau) = \frac{2}{\sqrt{\pi}} \int_{-\infty}^{\frac{x-a}{2\sqrt{\tau}}} \phi\left(\tau - \frac{(x-a)^2}{4\mu^2}\right) e^{-\mu^2} d\mu. \tag{35}$$

Then we have

$$\begin{aligned}
\frac{\partial u(x,\tau)}{\partial x} &= \frac{2}{\sqrt{\pi}}\phi(0)e^{-\frac{(x-a)^2}{4\tau}} \cdot \frac{1}{2\sqrt{\tau}} \\
&\quad + \frac{2}{\sqrt{\pi}} \int_{-\infty}^{\frac{x-a}{2\sqrt{\tau}}} \phi'\left(\tau - \frac{(x-a)^2}{4\mu^2}\right)\left(-\frac{x-a}{2\mu^2}\right) e^{-\mu^2} d\mu \\
&= -\frac{2}{\sqrt{\pi}} \int_{-\infty}^{\frac{x-a}{2\sqrt{\tau}}} \phi'\left(\tau - \frac{(x-a)^2}{4\mu^2}\right)\left(\frac{x-a}{2\mu^2}\right) e^{-\mu^2} d\mu \\
&= \frac{1}{\sqrt{\pi}} \int_0^\tau e^{-\frac{(x-a)^2}{4(\tau-\lambda)}} \frac{\phi'(\lambda)d\lambda}{\sqrt{\tau-\lambda}}, \quad x < a.
\end{aligned}$$

Thus we have

$$\left.\frac{\partial u}{\partial x}\right|_{x=a} = \frac{1}{\sqrt{\pi}} \int_0^\tau \frac{\partial u(a,\lambda)}{\partial \lambda} \frac{d\lambda}{\sqrt{t-\lambda}}. \tag{36}$$

The relationship in (36) is an exact boundary condition satisfied by $u(x,\tau)$, the solution of problem (9)–(12), on the artificial boundary Γ_a. By the exact boundary condition (36), the free boundary value problem for an American call option with dividend paying in an unbounded domain Ω is reduced to a problem in a bounded domain Ω_i:

$$\frac{\partial u}{\partial \tau} = \frac{\partial^2 u}{\partial x^2}, \qquad a < x < x_f(\tau), \quad 0 < \tau \le \tau^*, \tag{37}$$

$$u(x,0) = g(x,0), \qquad a < x < x_f(0), \tag{38}$$

$$u(x_f(\tau),\tau) = g(x_f(\tau),\tau), \qquad 0 < \tau \le \tau^*, \tag{39}$$

$$e^{(\alpha-1)x_f(\tau)+\beta\tau}\left[\frac{\partial u(x_f(\tau),\tau)}{\partial x} + \alpha u(x_f(\tau),\tau)\right] = 1, \qquad 0 \le \tau \le \tau^*, \tag{40}$$

$$\left.\frac{\partial u}{\partial x}\right|_{x=a} = \frac{1}{\sqrt{\pi}} \int_0^\tau \frac{\partial u(a,\lambda)}{\partial \lambda} \frac{d\lambda}{\sqrt{t-\lambda}}. \tag{41}$$

It is straightforward to check that the problem (9)–(12) is equivalent to the problem (37)–(41) in the following sense: If $\{u(x,\tau), x_f(\tau)\}$ is the solution of problem (9)–(12), then $\{u(x,\tau), x_f(\tau)\}$ is the solution of problem (37)–(41). If $\{u^*(x,\tau), x_f^*(\tau)\}$ is the solution of problem (37)–(41), let

$$\begin{aligned}
&x_f(\tau) = x_f^*(\tau),\\
&u(x,\tau) = \left\{ \begin{array}{cc} u^*(x,\tau), & a \le x \le x_f(\tau), \quad 0 \le \tau \le \tau^*,\\ -\frac{(x-a)}{2\sqrt{\pi}} \int_0^\tau \frac{u^*(a,\lambda)e^{-(x-a)^2/4(\tau-\lambda)}}{(\tau-\lambda)^{3/2}} d\lambda, & x < a, \quad 0 \le \tau \le \tau^*, \end{array} \right.
\end{aligned}$$

and then $\{u(x,\tau), x_f(\tau)\}$ is the solution of problem (9)–(12).

4. Finite difference approximation. In this section, we consider the numerical approximation of the problem (37)–(41). Let $\delta\tau$ and δx denote the step sizes of the finite difference approximation. Let $x_n = a + n\delta x$ and $\tau_m = m\delta\tau$, and denote the approximate solution of $u(x_n, \tau_m)$ by u_n^m. Using the Crank–Nicolson finite difference for (37), we get

$$\begin{aligned}
&\frac{u_n^m - u_n^{m-1}}{\delta\tau} = \frac{1}{2}\left(\frac{u_{n+1}^{m-1} - 2u_n^{m-1} + u_{n-1}^{m-1}}{(\delta x)^2} + \frac{u_{n+1}^m - 2u_n^m + u_{n-1}^m}{(\delta x)^2} \right),\\
&n = 1, 2, \ldots, \qquad m = 0, 1, \ldots.
\end{aligned}$$

Letting $\rho = \delta t/(\delta x)^2$, we have

$$(1+\rho)u_1^m - \frac{\rho}{2}u_2^m = b_1, \tag{42}$$

$$-\frac{\rho}{2}u_{n-1}^m + (1+\rho)u_n^m - \frac{\rho}{2}u_{n+1}^m = b_n, \qquad n = 2, 3, \ldots, \tag{43}$$

where

$$\begin{aligned}
b_1 &= \frac{\rho}{2}(u_0^{m-1} + u_2^{m-1}) + (1-\rho)u_1^{m-1} + \frac{\rho}{2}u_0^m,\\
b_n &= \frac{\rho}{2}(u_{n-1}^{m-1} + u_{n+1}^{m-1}) + (1-\rho)u_n^{m-1}, \qquad n = 2, 3, \ldots.
\end{aligned}$$

The solution u_n^m can be obtained as follows. Let $s_1 = 1 + \rho$ and $y_1 = b_1$; then we have

$$s_1 u_1^m - \frac{\rho}{2}u_2^m = y_1.$$

Solving for u_1^m and substituting into (43), we get

$$s_2 u_2^m - \frac{\rho}{2}u_2^m = y_2,$$

where

$$s_2 = 1 + \rho - \frac{\rho^2}{4s_1}, \qquad y_2 = b_2 + \frac{\rho y_1}{2s_1}.$$

In general, we have

$$s_n u_n^m - \frac{\rho}{3}u_{n+1}^m = y_n,$$

where

$$s_n = 1 + \rho - \frac{\rho^2}{4s_{n-1}}, \qquad y_n = b_n + \frac{\rho y_{n-1}}{2s_{n-1}}.$$

If the boundary condition

$$u^m_{N_e+1} = g_{N_e+1}$$

is given at certain point, then u^m_n, $n \leq N_e$, can be obtained by back substitution. From Theorems 1 and 2 we know that for a given τ the free boundary is the only point satisfying the partial differential equation and the condition

$$e^{(\alpha-1)x+\beta\tau}\left(\frac{\partial u(x,\tau)}{\partial x} + \alpha u(x,\tau)\right) = 1$$

or, equivalently,

$$\frac{\partial C(S,t)}{\partial S} = 1,$$

and if the boundary condition $u(x,\tau) = g(x,\tau)$ is given at $x > x_f(\tau)$, then $u(x,\tau) < g(x,\tau)$ will occur on the left of the boundary. Let N_e be the largest number such that

$$u^m_{N_e} \geq g_{N_e};$$

then we have

$$\begin{aligned} u^m_{N_e} &= \frac{1}{s_{N_e}}\left(b_{N_e} + \frac{\rho}{2}g_{N_e+1} + \frac{\rho y_{N_e-1}}{2s_{N_e-1}}\right), \\ u^m_n &= \frac{1}{s_n}\left(y_n + \frac{\rho}{2}u^m_{n+1}\right) \qquad \text{for } n = N_e-1, N_e-2, \ldots, 1. \end{aligned}$$

For the artificial boundary condition, since

$$\begin{aligned} \int_0^{\tau_m} \frac{\partial u(a,\lambda)}{\partial \lambda}\frac{d\lambda}{\sqrt{\tau_m-\lambda}} &= \sum_{j=1}^m \int_{\tau_{j-1}}^{\tau_j} \frac{\partial u(a,\lambda)}{\partial \lambda}\frac{d\lambda}{\sqrt{\tau_m-\lambda}} \\ &= \sum_{j=1}^m \frac{\partial u(a,\xi_j)}{\partial \tau}\int_{\tau_{j-1}}^{\tau_j} \frac{d\lambda}{\sqrt{\tau_m-\lambda}} \\ &= \sum_{j=1}^m \frac{2(\tau_j-\tau_{j-1})u_\tau(a,\xi_j)}{\sqrt{\tau_m-\tau_j}+\sqrt{\tau_m-\tau_{j-1}}}, \end{aligned}$$

equation (41) is approximated by

$$\frac{u^m_1 - u^m_{-1}}{2\delta x} = \frac{1}{\sqrt{\pi}}\sum_{j=1}^m \frac{2(u^j_0 - u^{j-1}_0)}{\sqrt{\tau_m-\tau_j}+\sqrt{\tau_m-\tau_{j-1}}}. \tag{44}$$

Approximating (37) at $x = a$, we get

$$\frac{u^m_0 - u^{m-1}_0}{\delta\tau} = \frac{u^m_1 - 2u^m_0 + u^m_{-1}}{(\delta x)^2}. \tag{45}$$

From (44) and (45) we obtain the boundary condition

$$u_0^m = \frac{\theta H_1 + \sqrt{\pi} H_2/4}{\theta + \sqrt{\pi}(1+2\theta^2)/4}, \tag{46}$$

where $\theta = \sqrt{\delta\tau}/\delta x$, $H_2 = u_0^{m-1} + \theta^2 u_1^m$, and

$$H_1 = u_0^{m-1} + \sqrt{\pi}\theta u_1^m/4 - \sum_{j=1}^{m-1} \frac{u_0^j - u_0^{j-1}}{\sqrt{m-j} + \sqrt{m-j+1}}.$$

Thus we have the following algorithm.

Algorithm.

At each time step, do the following:

1. Set up the linear system using the Crank–Nicolson finite difference method for $x \geq a$.
2. Combine the artificial boundary condition (46) and (42) to eliminate u_0^m.
3. Use the elimination for the linear system in the interval $[a, b]$. Move the right boundary b until the free boundary is found.
4. Use back substitution to find all solutions in $[a, b]$.

At the end τ^*, use (34) to find all solutions to the left of a.

5. Computational results.

To compare the above algorithm with the standard finite difference approximations, we computed two examples of call options. The second example was also computed by Broadie and Detemple [5]. The comparisons are based on the accuracy of the approximate option values and the total computation cost, i.e., the CPU time. Since the exact option values are unknown, we use the binomial method with large steps (15000) to find the option values. The results of the binomial method with large steps are considered very accurate. Thus we take these values as the exact option values for the purpose of comparison. All the algorithms are implemented using MATLAB for testing purposes, and the computations are carried out on an IBM RS/6000 43P Model 260 workstation.

In both examples, ABF stands for the numerical method given in the previous section, artificial boundary condition with free boundary treatment. FDP stands for the Crank–Nicolson finite difference approximation with projected SOR iteration to impose the free boundary condition. FDE stands for the Crank–Nicolson finite difference approximation with elimination-backsubstitution. In both FDP and FDE methods, the systems are set up in the interval $[x_m, x_p]$, where $x_m = a < 0$ and $x_p > x_f(\tau) > 0$ for all $\tau > 0$. The asymptotic boundary conditions are applied at both ends $x = x_m$ and $x = x_p$.

Example 1. Consider a six-month American call option with a dividend rate $D_0 = 0.03$. The exercise price is \$100, the risk-free interest rate is $r = 0.03$, and the volatility is 40% per annum. The right boundary is set to $x_p = 0.8$. (The largest value of $x_f(\tau)$ is about 0.62.) A step size $m = 400$ with ratio $\rho = 1$ is taken for all methods. Table 1 shows the results. When $a = -0.2$, the corresponding asset price is about 81.87. Thus the option values corresponding to $S \leq 80$ are not shown for FDP and FDE methods. Similarly, when $a = -0.6$, the corresponding asset price is about 54.88, and the option values corresponding to $S \leq 50$ are not shown. However, the ABF method can give the option values corresponding to any asset prices.

From the computational results shown in Table 1, it is clear that the accuracy of the option values are largely affected by the choice of the left boundary $x = a$. To obtain an accurate option value for the asset price $S = 80$, a must be smaller than

Table 1
American call option value (maturity T= 0.5, M= 400).

a	Asset price	FDP	FDE	ABF	True value
-0.2	40			0.0028	0.002792
	50			0.0456	0.045594
	60			0.3013	0.301387
	70			1.1459	1.145799
	80			3.0435	3.041536
	90	4.3058	4.3058	6.3643	6.328677
	100	10.1228	10.1228	11.1267	11.108407
	110	16.7980	16.7980	17.2772	17.266726
	120	24.3457	24.3458	24.5710	24.565972
CPU		21.2000	8.8300	6.2900	
-0.6	40			0.0028	0.002792
	50			0.0455	0.045594
	60	0.2493	0.2492	0.3011	0.301387
	70	1.1365	1.1366	1.1464	1.145799
	80	3.0396	3.0398	3.0416	3.041536
	90	6.3282	6.3283	6.3287	6.328677
	100	11.1066	11.1067	11.1068	11.108407
	110	17.2664	17.2665	17.2665	17.266726
	120	24.5654	24.5655	24.5655	24.565972
CPU		29.4000	12.2600	8.9100	
-1.0	40	0.0025	0.0025	0.0028	0.002792
	50	0.0457	0.0457	0.0457	0.045594
	60	0.3014	0.3015	0.3015	0.301387
	70	1.1459	1.1461	1.1461	1.145799
	80	3.0414	3.0415	3.0415	3.041536
	90	6.3285	6.3287	6.3287	6.328677
	100	11.1066	11.1068	11.1068	11.108407
	110	17.2664	17.2665	17.2665	17.266726
	120	24.5654	24.5655	24.5655	24.565972
CPU		37.1300	15.7600	11.5000	

-0.4, and for $S = 70$, a must be smaller than -0.5. The ABF method gives much more accurate solutions. Compared with the FDP and FDE methods, to obtain an accurate option value for $S = 80$, $a = -0.2$ is enough, and for $S = 70$, $a = -0.4$ is enough.

Figure 1 shows the comparison of error and CPU for the FDE and ABF methods. The error e is measured by

$$e = \left[\frac{1}{K} \sum_{i=1}^{K} (C_i - \bar{C}_i)^2 \right]^{1/2},$$

where C_i is the binomial value, $\bar{C}_i$ is the value of the FDE method, or of the value of the ABF method, and K is the total number of option values taken. The figure shows clearly that the ABF method is much more efficient than the FDE method.

When the maturity time is longer, the error generated due to the rough approximation at the left boundary can be more serious. In example 2, we compare the different algorithms for option value with longer maturity time.

Example 2. Consider a three-year American call option with a dividend rate $D_0 = 0.07$. The exercise price is \$100, the risk-free interest rate is $r = 0.03$, and the volatility is 40% per annum. The right boundary is set to $x_p = 0.8$. (The largest value of $x_f(\tau)$ is about 0.7.) A step size $m = 400$ with ratio $\rho = 1$ is taken for all methods. Table 2 shows the results. The corresponding asset price for $a = -0.4$ is

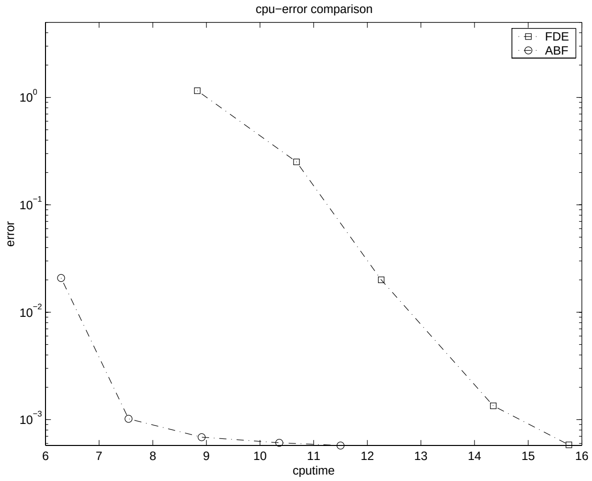


Fig. 1.

Table 2
American call option value (maturity T= 3.0, M= 400).

	Asset				True
a	price	FDP	FDE	ABF	value
-0.4	20			0.053	0.053
	40			1.129	1.127
	60			4.729	4.719
	80	6.378	6.378	11.354	11.326
	100	17.639	17.639	20.853	20.793
	120	30.744	30.744	32.824	32.781
CPU		13.32	4.34	5.22	
-0.8	20			0.053	0.053
	40			1.127	1.127
	60	3.845	3.845	4.720	4.719
	80	10.942	10.942	11.329	11.326
	100	20.610	20.610	20.801	20.793
	120	32.686	32.686	32.784	32.781
CPU		18.64	6.19	6.32	
-1.2	20			0.053	0.053
	40	0.977	0.977	1.126	1.127
	60	4.684	4.684	4.717	4.719
	80	11.314	11.314	11.323	11.326
	100	20.786	20.786	20.790	20.793
	120	32.781	32.781	32.783	32.781
CPU		22.53	7.72	7.36	
-1.6	20			0.053	0.053
	40	1.124	1.124	1.127	1.127
	60	4.720	4.720	4.720	4.719
	80	11.327	11.327	11.327	11.326
	100	20.796	20.795	20.796	20.793
	120	32.783	32.783	32.783	32.781
CPU		27.53	9.29	8.44	
-2.0	20	0.052	0.052	0.053	0.053
	40	1.128	1.128	1.127	1.127
	60	4.720	4.720	4.720	4.719
	80	11.328	11.328	11.327	11.326
	100	20.796	20.796	20.796	20.793
	120	32.781	32.781	32.781	32.781
CPU		31.21	10.70	9.64	

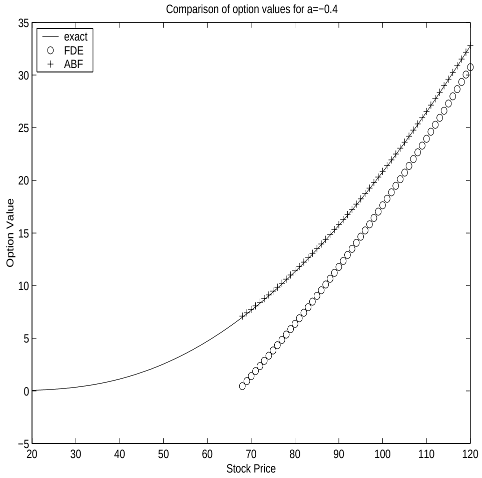


Fig. 2.

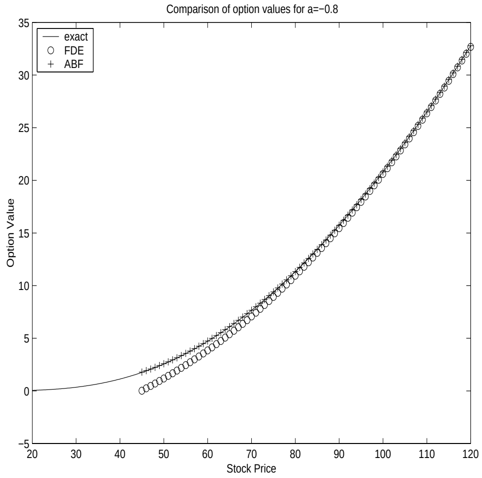


Fig. 3.

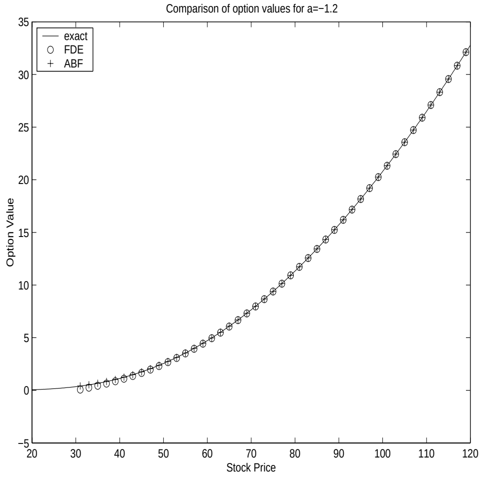


Fig. 4.

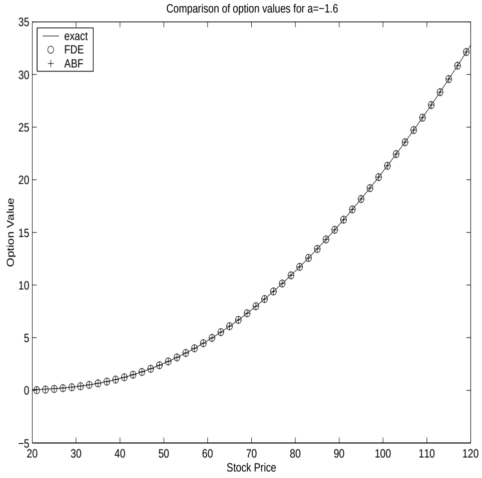


Fig. 5.

about 67.03, that for $a = -0.8$ is about 44.93, that for $a = -1.2$ is about 30.12, and that for $a = -1.6$ is about 20.19. Option values corresponding to asset prices smaller than these values for the FDP and FDE methods are not shown.

For the FDP and FDE methods, the option values are totally wrong for $a = -0.4$. When $a = -0.8$, the option values are still not accurate for asset prices up to $S = 90$, although the left boundary is about $S = 44.93$. To obtain accurate option values, the left boundary needs to be $a = -2.0$. The ABF method improved the results greatly. Even for $a = -0.4$, the option values are close to the true values.

In comparison of the efficiency of all algorithms, we can see from the table that if more option values are needed, then a large saving can be resulted by using artificial boundary conditions. For example, if the option values for $S = 20$ to $S = 120$ are needed, then for the FDP and FDE methods, a must be at least -2.0, and the corresponding CPU time is about 31.21 seconds for the FDP method and 10.7

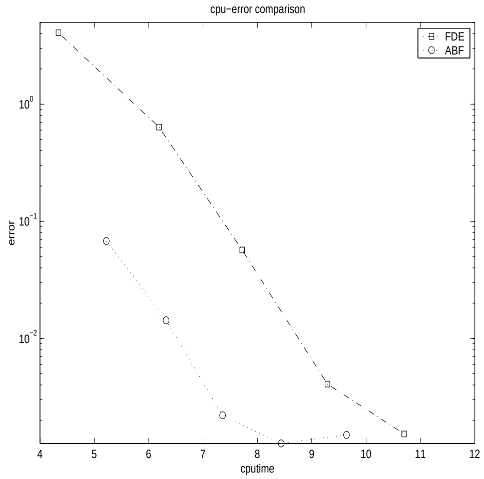


Fig. 6.

seconds for the FDE method. However, the ABF method with $a = -0.4$ can give more accurate option values, while the CPU time is only 5.22 seconds. For this example, the savings in CPU time is nearly 50%. The savings in CPU time will be different for different cases, but it is clear that the savings are significant.

Figures 2–5 show close comparisons of the option values for the FDE method and the ABF method, where the exact option values are obtained by the ABF method with $m = 2000$ on a large interval, $x \geq -2$; the results are very accurate. The four figures show the comparison results for $a = -0.4$, $a = -0.8$, $a = -1.2$, and $a = -1.6$. It is clear from the figures that the results of the FDE method are not acceptable for $a = -0.4$ and $a = -0.8$. When $a = -1.2$, the error can still be seen for the FDE method, while the ABF method gives accurate values for all the cases.

Figure 6 shows the comparison of error and CPU for the FDE and ABF methods. Again, we can see that the ABF method is much more efficient than the FDE method.

6. Conclusion. The artificial boundary conditions give accurate relations of the solutions at the boundary. By using the artificial boundary conditions in the finite difference approximation, we obtained the solution in a small truncated domain. Numerical results show that the solution is very accurate. The treatment for the free boundary is also very efficient. The computational cost is greatly reduced.

Acknowledgment. The authors wish to thank the referees for many valuable suggestions.

REFERENCES

[1] K. Amin and A. Khanna, *Convergence of American option values from discrete- to continuous-time financial models*, Math. Finance, 4 (1994), pp. 289–304.

[2] F. Black and M. Scholes, *The pricing of options and corporate liabilities*, J. Pol. Econ., 81 (1973), pp. 637–659.

[3] M. Brennan and E. Schwartz, *The valuation of American put options*, J. Fin., 32 (1977), pp. 449–462.

[4] M. Brennan and E. Schwartz, *Finite difference methods and jump processes arising in the pricing of contingent claims: A synthesis*, J. Fin. Quan. Anal., 13 (1978), pp. 461–474.

[5] M. Broadie and J. Detemple, *American option valuation: New bounds, approximations, and a comparison of existing methods*, Rev. Fin. Stud., 9 (1996), pp. 1211–1250.

[6] H. S. Carslaw and J. C. Jaeqer, *Conduction of Heat in Solids*, Clarendon Press, Oxford, UK, 1959.

[7] J. C. Cox, S. A. Ross, and M. Rubinstein, *Option pricing: A simplified approach*, J. Fin. Econ., 7 (1979), pp. 229–263.

[8] B. Enguist and A. Majda, *Absorbing boundary conditions for the numerical simulation of waves*, Math. Comp., 31 (1977), pp. 629–651.

[9] A. Friedman, *Partial Differential Equations of Parabolic Type*, Robert E. Krieger Publishing, Huntington, NY, 1983.

[10] R. Geske and H. E. Johnson, *The American put options valued analytically*, J. Fin., 39 (1984), pp. 1511–1524.

[11] R. Geske and K. Shastri, *Valuation by approximation: A comparison of alternative option valuation techniques*, J. Fin. Quan. Anal., 20 (1985), pp. 45–71.

[12] H. Han and Z. Huang, *A class of artificial boundary conditions for heat equation in unbounded domains*, Comput. Math. Appl., 43 (2002), pp. 889–900.

[13] H. Han and X. Wu, *Approximation of infinite boundary conditions and its application to finite element method*, J. Comput. Math., 3 (1985), pp. 179–192.

[14] T. M. Hagstrom and H. B. Keller, *Exact boundary conditions at an artificial boundary for partial differential equations in cylinders*, SIAM J. Math. Anal., 17 (1986), pp. 322–341.

[15] P. Jaillet, D. Lamberton, and B. Lapeyre, *Variational inequalities and the pricing of American options*, Acta Appl. Math., 21 (1990), pp. 263–289.

[16] H. Johnson, *An analytic approximation for the American put price*, J. Fin. Quan. Anal., 18 (1983), pp. 141–148.

[17] R. Kangro and R. Nicolaides, *Far field boundary conditions for Black–Scholes equations*, SIAM J. Numer. Anal., 38 (2000), pp. 1357–1368.

[18] L. W. MacMillan, *An analytic approximation for the American put price*, Advances in Futures and Options Research, 1 (1986), pp. 119–139.

[19] E. S. Schwartz, *The valuation of warrants: Implementing a new approach*, J. Fin. Econ., 4 (1977), pp. 79–93.

[20] P. Wilmott, J. Dewynne, and S. Howison, *Option Pricing: Mathematical Models and Computation*, Oxford Financial Press, Oxford, UK, 1993.

SIAM J. Numer. Anal.
Vol. 41, No. 6, pp. 2096–2113

ATTRACTION RATES, ROBUSTNESS, AND DISCRETIZATION OF ATTRACTORS*

LARS GRÜNE†

Abstract. We investigate necessary and sufficient conditions for the convergence of attractors of discrete time dynamical systems induced by numerical one-step approximations of ODEs to an attractor for the approximated ODE. We show that both the existence of uniform attraction rates (i.e., uniform speed of convergence toward the attractors) and uniform robustness with respect to perturbations of the numerical attractors are necessary and sufficient for this convergence property. In addition, we can conclude estimates for the rate of convergence in the Hausdorff metric.

Key words. ODE, numerical one-step approximation, attractor, attraction rate, robustness

AMS subject classifications. 65L20, 65L06, 34D45, 34E10

DOI. 10.1137/S003614290139411X

1. Introduction.

The long time behavior of dissipative dynamical systems is essentially determined by the attractors of these systems, since for large times its trajectories will typically stay on or near an attractor. Even for moderately complex finite dimensional systems, however, it is rarely possible to determine attractors by analytic methods. Hence numerical approximations form a natural part of a systematic analysis. It is therefore important to know about the effects of discretization errors on attractors in order to give a reasonable interpretation to numerical experiments and to justify numerical findings.

For dynamical systems induced by ODEs, numerical one-step approximations like Runge–Kutta or Taylor schemes form an important class of schemes. It follows from a result of Kloeden and Lorenz in 1986 [18] for attracting sets that if the ODE possesses an attractor A, then the discrete time dynamical system induced by such a numerical scheme also has an attractor contained in a neighborhood N of A, where the size of N shrinks down to A as the time-step of the discretization approaches 0.

A number of examples (see, e.g., [7, Example (2.12)], [9, Example 1.1.1], or [11] for the case of finite dimensional approximations of infinite dimensional systems) shows that the limit set for a convergent sequence of numerical attractors for vanishing time-steps may be strictly smaller than A. This fact, however, imposes a major problem for the interpretation of numerical results, since it implies that in general one cannot conclude the existence of a real attractor close to a numerical attractor. Hence it is important to derive techniques or conditions which allow us to conclude convergence of numerical attractors to a real attractor.

There are three main approaches for tackling this problem: The first is to impose suitable conditions on the approximated system which ensure a faithful numerical approximation and exclude the appearance of numerical artifacts. Typical examples of this approach are, for instance, results on the numerical approximation of Morse–Smale systems by Garay [5, 6], on the discretization of gradient systems by Hale and Raugel [12] or Stuart and Humphries [22, section 7.7], and a result on hyperbolic attractors by the author [8, Remark 2.10(ii)].

[*] Received by the editors August 23, 2001; accepted for publication (in revised form) May 1, 2003; published electronically November 25, 2003.
http://www.siam.org/journals/sinum/41-6/39411.html

[†] Mathematisches Institut, Fakultät für Mathematik und Physik, Universität Bayreuth, 95440 Bayreuth, Germany (lars.gruene@uni-bayreuth.de).

The second approach is to design algorithms which can be shown to converge to the right objects under no or very mild conditions on the approximated system. An example for this approach is the subdivision algorithm for the computation of attractors originally developed by Dellnitz and Hohmann [3] using a rigorous discretization as proposed by Junge [14, 15]; see also [9, section 6.3] for a description and a quantitative convergence analysis of this method based on robust Lyapunov functions.

The idea of the third approach is the formulation of conditions on the behavior of the numerical systems under which we can ensure convergence of the respective sets or the existence of respective nearby sets for the approximated system. A typical example are the sufficient conditions for the convergence of numerical attracting sets in the Galerkin approximation to Navier–Stokes equations by Kloeden [16]. For finite dimensional systems, in [8] a necessary and sufficient condition of this type was developed based on uniform attraction properties of the numerical attractors, which were characterized using (uniformly) shrinking families of neighborhoods which are mapped onto each other by suitable perturbations of the numerical flows.

The present paper follows this third approach. As in [8] we are going to formulate necessary and sufficient conditions for the convergence of numerical attractors to a real attractor in terms of uniform attraction properties. The difference from [8] lies in the type of uniformity properties used for this purpose, because (i) here we formulate the properties directly in terms of the numerical attractors instead of using auxiliary attracting sets, (ii) we use comparison functions for characterizing attraction properties instead of using geometrical characterizations by means of shrinking neighborhoods which are difficult to identify in a numerical simulation, and (iii) most importantly, instead of using a condition on the rate of attraction (i.e., the speed of convergence toward the attractor) for perturbed numerical systems, here we "decouple" the rate and the perturbation and give two different conditions—one based on the attraction rate for the unperturbed numerical systems and the other based on the robustness against perturbations.

More precisely, we prove that a sequence of numerical attractors converges to a real attractor

- if and only if the numerical attractors are attracting with uniformly bounded attraction rates (cf. Theorem 6.2(iii));
- if and only if the numerical attractors are robust against perturbation with uniformly bounded robustness gains (cf. Theorem 6.2(ii) and Theorem 6.4).

In addition, in Theorem 6.5 we give estimates for the discretization error based on the local error of the numerical scheme and the robustness gains of the respective attractors.

The tools we need in order to obtain these results are developed step by step in this paper, which is organized as follows. After defining the setup and stating some preliminary results in section 2, in section 3 we define a suitable robustness concept for attracting sets with respect to perturbations, describe the concept of embedding systems into each other, and show how this applies to numerical one-step approximations. In section 4 we study the relation between the robustness of attracting sets and their rate of attraction. In section 5 we prove some useful results on the relation between a continuous time system and its time-h map, and finally, in section 6 we state the main results on attractors under one-step discretization.

2. Setup and preliminaries.

We consider ODEs given by

$$\dot{x} = f(x) \tag{2.1}$$

and, for some time-step $h > 0$, discrete time systems of the form

$$x(t+h) = \Phi_h(x(t)), \tag{2.2}$$

where f and Φ_h are maps from $\mathbb{R}^d$ to $\mathbb{R}^d$, $d \in \mathbb{N}$.

For simplicity of exposition (cf. Remark 2.1, below) we assume global Lipschitz properties of the respective systems; i.e., we assume that there exists a constant $L > 0$ such that the inequalities

$$\|f(x) - f(y)\| \leq L\|x - y\| \text{ for all } x,\, y \in \mathbb{R}^d \tag{2.3}$$

and

$$\|F_h(x) - F_h(y)\| \leq L\|x - y\| \text{ for all } x,\, y \in \mathbb{R}^d \tag{2.4}$$

hold, where $F_h(x) = (\Phi_h(x) - x)/h$.

The trajectories of system (2.1) with initial value $x_0 \in \mathbb{R}^d$ for initial time $t = 0$ are denoted by $\varphi(t, x_0)$ for $t \in \mathbb{R}_0^+$, and the respective trajectories for (2.2) are denoted by $\Phi_h(t, x)$ for $t \in h\mathbb{N}_0 := \{hk \,|\, k \in \mathbb{N}_0\}$.

It will often be convenient to combine continuous and discrete time trajectories in one notation. For this we use the notation $\Phi(t, x)$, which either denotes $\varphi(t, x)$ or $\Phi_h(t, x)$, where the precise meaning will be clear from the context. Whenever we consider a discrete time system with time-step $h > 0$, the time t is implicitly assumed to be in the respective discrete time-scale $h\mathbb{N}_0$.

For sets $C \subset \mathbb{R}^d$ we use the notation

$$\Phi(t, C) = \bigcup_{x \in C} \{\Phi(t, x)\}.$$

A special type of a discrete time system is the time-h map of (2.1) which is defined by the discrete time system

$$x(t+h) = \varphi(h, x(t)). \tag{2.5}$$

The trajectories of (2.5) are denoted by $\varphi_h(t, x)$. Note that if (2.1) satisfies (2.3) for some $L > 0$, then Gronwall's lemma implies that the time-h map (2.5) satisfies (2.4) for the Lipschitz constant $\widetilde{L} = Le^{hL} > L$.

Another special type of a discrete time system (2.2) is a numerical one-step approximation $\widetilde{\Phi}_h$ of (2.1) which is supposed to satisfy (2.4) and is such that there exist constants c, $q > 0$ with

$$\|\widetilde{\Phi}_h(x) - \varphi(h, x)\| \leq ch^{q+1} \text{ for all } x \in \mathbb{R}^d. \tag{2.6}$$

Here q is called the order of the scheme. Examples for such approximations are Taylor and Runge–Kutta schemes; for details we refer, e.g., to the textbooks [4, 13, 21].

Remark 2.1. The global estimates in the inequalities (2.3), (2.4), and (2.6) are in general quite restrictive. However, since we are interested in the behavior on bounded subsets of the state space, one can always assume these properties without loss of generality by applying standard cutoff techniques.

Since we are going to measure distances between different sets, we need the following definitions.

DEFINITION 2.2. *Let C, $D \subset \mathbb{R}^d$ be nonempty compact sets, $x \in \mathbb{R}^d$, and let $\|\cdot\|$ be the Euclidean norm on $\mathbb{R}^d$. We define the distance from a point to a set by*

$$\|x\|_D := \min_{y\in D} \|x - y\|,$$

the nonsymmetric Hausdorff distance between two compact sets by

$$\operatorname{dist}(C, D) := \max_{x\in C} \min_{y\in D} \|x - y\|,$$

and the Hausdorff metric for compact sets by

$$d_H(C, D) := \max\{\operatorname{dist}(C, D), \operatorname{dist}(D, C)\}.$$

We use these distances for arbitrary bounded sets C, $D \subset \mathbb{R}^d$ by defining

$$\|x\|_D := \|x\|_{\operatorname{cl} D},\ \ \operatorname{dist}(C, D) := \operatorname{dist}(\operatorname{cl} C, \operatorname{cl} D) \text{ and } d_H(C, D) := d_H(\operatorname{cl} C, \operatorname{cl} D).$$

For $\varepsilon > 0$ we denote the (open) ε-ball around C by $\mathcal{B}(\varepsilon, C) := \{y \in \mathbb{R}^d \mid \|y\|_C < \varepsilon\}$. If $C = \{x\}$, we also write $\mathcal{B}(\varepsilon, x)$.

Note that for all bounded sets $C, D, E \subset \mathbb{R}^d$, the equivalences $\operatorname{dist}(C, D) = 0 \Leftrightarrow C \subseteq \operatorname{cl} D$ and $d_H(C, D) = 0 \Leftrightarrow \operatorname{cl} C = \operatorname{cl} D$ and the implication $C \subseteq E \Rightarrow \operatorname{dist}(C, D) \leq \operatorname{dist}(E, D)$ hold.

Our main objects of interest are attracting sets and attractors as given by the following definition.

DEFINITION 2.3. *Let Φ denote the trajectories of a system of type* (2.1) *or* (2.2). *Consider a compact set $A \subset \mathbb{R}^d$ and an open and bounded set $B \subset \mathbb{R}^d$ with $A \subset B$. Then A is called an* attracting set *with* attracted neighborhood B *if it is forward invariant, i.e.,*

$$\Phi(t, A) \subseteq A \text{ for all } t \geq 0,$$

and satisfies

$$\lim_{t\to\infty} \operatorname{dist}(\Phi(t, B), A) \to 0.$$

An attracting set is called an attractor *if it is invariant, i.e.,*

$$\Phi(t, A) = A \text{ for all } t \geq 0.$$

In order to characterize quantitative properties of attracting sets and attractors, we make use of comparison functions as introduced by Hahn [10].

DEFINITION 2.4. *We define the following classes of* comparison functions.

$$\begin{aligned}
\mathcal{K} \quad &:= \quad \{\gamma : \mathbb{R}_0^+ \to \mathbb{R}_0^+ \mid \gamma \text{ is continuous, strictly increasing and } \gamma(0) = 0\},\\
\mathcal{L} \quad &:= \quad \{\sigma : \mathbb{R}_0^+ \to \mathbb{R}_0^+ \mid \sigma \text{ is continuous, strictly decreasing and } \lim_{r\to\infty} \sigma(r) = 0\},\\
\mathcal{KL} \quad &:= \quad \{\beta : \mathbb{R}_0^+ \times \mathbb{R}_0^+ \to \mathbb{R}_0^+ \mid \beta(\cdot, r) \in \mathcal{K} \text{ and } \beta(r, \cdot) \in \mathcal{L} \text{ for each } r \geq 0\}.
\end{aligned}$$

Remark 2.5. The functions $\beta \in \mathcal{KL}$ are closely related to the usual ε-δ definition of asymptotic stability. More precisely, for any function $a : \mathbb{R}_0^+ \times \mathbb{R}_0^+ \to \mathbb{R}_0^+$ satisfying the two properties

(i) for all $\varepsilon > 0$ there exists $\delta > 0$ such that if $r \leq \delta$, then $a(r, t) < \varepsilon$ for all $t \geq 0$,

(ii) for all $\varepsilon > 0$ and for all $R > 0$ there exists $T > 0$ such that $a(r,t) < \varepsilon$ for all $0 \le r \le R$ and for all $t \ge T$,

there exists a function $\beta \in \mathcal{KL}$ with $a(r,t) \le \beta(r,t)$ for all r, $t \ge 0$.

This fact was already implicitly used in Hahn's book [10]; in this form it is stated (but not proved) in Albertini and Sontag [1, Lemma 4.1] and proved (but not explicitly stated) by Lin, Sontag, and Wang [19, section 3].

Using class $\mathcal{KL}$ functions we can define rates of attraction for attracting sets.

DEFINITION 2.6. *Let Φ denote the trajectories of a system of type* (2.1) *or* (2.2), *and let A be an attracting set with attracted neighborhood B. Then $\beta \in \mathcal{KL}$ is called* rate of attraction *of A if the inequality*

$$\|\Phi(t,x)\|_A \le \beta(\|x\|_A, t)$$

holds for each $x \in B$ and each $t \ge 0$.

The following lemma shows that each attracting set possesses a rate of attraction.

LEMMA 2.7. *Let Φ denote the trajectories of a system of type* (2.1) *or* (2.2), *and let A be an attracting set with attracted neighborhood B. Then there exists a rate of attraction $\beta \in \mathcal{KL}$ for A.*

Proof. Using the forward invariance and attractivity properties of A and the (uniform) continuous dependence of a trajectory on the initial value (as induced by Gronwall's lemma for (2.1) or by induction for (2.2)), one easily verifies that the function $a : \mathbb{R}_0^+ \times \mathbb{R}_0^+ \to \mathbb{R}_0^+$ defined by $a(0,t) = 0$ for $t \in \mathbb{R}_0^+$ and

$$a(r,t) := \sup_{\tau \ge t} \operatorname{dist}\left(\Phi\left(\tau, \mathcal{B}(r,A) \cap B\right),\, A\right)$$

satisfies the properties (i) and (ii) of Remark 2.5. Hence there exists $\beta \in \mathcal{KL}$ with $a \le \beta$ and consequently

$$\|\Phi(x,t)\|_A \le \sup_{\tau \ge t} \operatorname{dist}\left(\Phi\left(\tau, \mathcal{B}(\|x\|_A, A) \cap B\right),\, A\right) = a(\|x\|_A, t) \le \beta(\|x\|_A, t)$$

for all $x \in B$ and all $t \ge 0$. This shows the claim. □

We end this section by stating some useful properties of attractors which we will need in what follows.

LEMMA 2.8. *Let Φ denote the trajectories of a system of type* (2.1) *or* (2.2), *and let A be an attracting set with attracted neighborhood B. Then A contains an attractor with attracted neighborhood B.*

Proof. Verification of the desired properties shows that

$$\widetilde{A} := \bigcap_{T \ge 0} \operatorname{cl} \bigcup_{t \ge T} \Phi(t,B)$$

is the desired attractor; see [22, Theorem 2.7.4] for details. □

LEMMA 2.9. *Let Φ denote the trajectories of a system of type* (2.1) *or* (2.2). *Then a compact forward invariant attracting set A for Φ with attracted neighborhood B is an attractor with attracted neighborhood B if and only if it is the minimal compact forward invariant attracting set (with respect to set inclusion) with attracted neighborhood B. In particular, for each open and bounded set $B \subset \mathbb{R}^d$, there exists at most one attractor with attracted neighborhood B.*

Proof. Let A be an attractor with attracted neighborhood B. Then, in particular, A is invariant. Now assume that $\tilde{A} \subset A$, $\tilde{A} \ne A$ is a forward invariant attracting set.

Then there exists a neighborhood $\mathcal{N} \supset \tilde{A}$ with $A \not\subseteq \mathcal{N}$ such that $\Phi(t, B) \subset \mathcal{N}$ for some $t \geq 0$, i.e., in particular, $\Phi(t, A) \neq A$, which contradicts the invariance of A.

Let conversely A be a minimal forward invariant attracting set. Then, by Lemma 2.8, A contains an attractor which again is a forward invariant attracting set. Hence by minimality it coincides with A. □

The next lemma shows that the attractor is also the maximal compact invariant set contained in int$B$.

LEMMA 2.10. *Let Φ denote the trajectories of a system of type* (2.1) *or* (2.2), *and let A be an attractor with attracted neighborhood B for Φ. Then each compact invariant set $D \subset B$ for Φ is contained in A.*

Proof. Let $D \subset B$ be an invariant set for Φ. Then $D = \Phi(t, D) \subset \Phi(t, B)$ for all $t \geq 0$. On the other hand, for each neighborhood $\mathcal{N} \supset A$ we know that $\Phi(t, B) \subset \mathcal{N}$ for all $t \geq 0$ sufficiently large. Hence $D \subset \mathcal{N}$, which implies the assertion. □

3. Inflation, robustness, and embedding.

The main technique that we will use in this paper in order to obtain results on convergence of numerical attractors is the embedding of the numerical approximation into a perturbed continuous time system, and vice versa. In this section we define suitable perturbed systems, the corresponding attracting sets, and a useful robustness concept for attracting sets. In addition, we give a mathematically precise meaning for the embedding property.

The following definition gives the appropriate perturbed systems (see also [17] for an equivalent definition using differential inclusions).

DEFINITION 3.1. *For $\alpha \in \mathbb{R}$, $\alpha \geq 0$, we define the set of perturbation values*

$$W_\alpha := \{w \in \mathbb{R}^d \,|\, \|w\| \leq \alpha\}$$

and the space of measurable functions with values in W_α by

$$\mathcal{W}_\alpha := \{w : \mathbb{R} \to \mathbb{R}^d \,|\, w \textit{ measurable with } w(t) \in W_\alpha \textit{ for almost all } t \in \mathbb{R}\}.$$

For functions $w \in \mathcal{W}_\alpha$ and real values $a < b$, we define $\|w\|_{[a,b]} := \operatorname{ess\,sup}_{t\in[a,b]}\|w(t)\|$.

For a continuous time system (2.1) *we define the α-*inflated system *by*

$$\dot{x} = f(x) + w, \quad w \in W_\alpha, \tag{3.1}$$

and for a discrete time system we define it by

$$x(t+h) = \Phi_h(x(t)) + \int_t^{t+h} w(t)dt, \quad w \in \mathcal{W}_\alpha. \tag{3.2}$$

For each initial value $x \in \mathbb{R}^d$ and each $w \in \mathcal{W}_\alpha$ we denote the corresponding trajectory by $\varphi(t, x, w)$ or $\Phi_h(t, x, w)$, respectively. It should be noted that the discrete time inflation (3.2) *of the time-h map* (2.5) *of a continuous time system* (2.1) *differs from the time-h map of the continuous time inflation* (3.1) *of system* (2.1) *defined by*

$$x(t+h) = \varphi(h, x(t), w(t + \cdot)), \tag{3.3}$$

where $\varphi(h, x(t), w(t+\cdot))$ denotes the solution of $\dot{y}(s) = f(y(s)) + w(t+s)$ with $y(0) = x(t)$.

Throughout this paper, the term inflated time-h map *refers to system* (3.3), *whose trajectories will be denoted by $\varphi_h(t, x, w)$.*

As for the unperturbed systems, we use $\Phi(t, x, w)$ to denote both discrete and continuous time trajectories, depending on the context. Furthermore, for $\alpha > 0$, $x \in \mathbb{R}^d$, and a subset $C \subseteq \mathbb{R}^d$, we use the notation

$$\Phi^\alpha(t, x) := \bigcup_{w \in \mathcal{W}_\alpha} \{\Phi(t, x, w)\} \quad \textit{and} \quad \Phi^\alpha(t, C) := \bigcup_{x \in C} \Phi^\alpha(t, x).$$

Next we define suitable attracting sets for inflated systems and a robustness property of attracting sets.

Definition 3.2. *Consider an inflated continuous time system* (3.1)*, an inflated discrete time system* (3.2)*, or an inflated time-h map* (3.3) *with trajectories denoted by $\Phi(t, x, w)$. Then a compact set A_α with open neighborhood B is called an* α-attracting set *with* attracted neighborhood B *if it is α-forward invariant, i.e.,*

$$\Phi^\alpha(t, A) \subseteq A \textit{ for all } t \geq 0,$$

and satisfies

$$\lim_{t \to \infty} \operatorname{dist}(\Phi^\alpha(t, B), A) \to 0.$$

Let $\alpha_0 > 0$ and $\gamma \in \mathcal{K}$. Then an attracting set (or attractor) A with attracted neighborhood B for an unperturbed system (2.1) *or* (2.2) *is called* γ-robust *for γ and α_0 if for each $\alpha \in (0, \alpha_0]$ there exists an α-attracting set A_α with attracted neighborhood B for the corresponding inflated system with $A \subseteq A_\alpha$ and*

$$d_H(A, A_\alpha) \leq \gamma(\alpha).$$

Here $\gamma \in \mathcal{K}$ is called robustness gain.

Remark 3.3. Analogous to Lemma 2.7, one sees that for each α-attracting set A_α with attracted neighborhood B there exists $\beta \in \mathcal{KL}$ such that

$$\|\Phi(t, x, w)\|_{A_\alpha} \leq \beta(\|x\|_{A_\alpha}, t)$$

for all $x \in B$, $t \geq 0$, and $w \in \mathcal{W}_\alpha$.

We now define what we mean by an embedded system. For our purpose it is sufficient to define this concept for discrete time systems.

Definition 3.4. *Consider two inflated discrete time systems of type* (3.2) *with perturbations from $\mathcal{W}_{\tilde{\alpha}_0}$ and $\mathcal{W}_{\alpha_0}$, respectively. Denote the trajectories of the systems by Φ_h and Ψ_h, respectively, and let $\alpha \geq 0$ and $C \geq 1$. Then we say that the second system Ψ_h is* (α, C)-embedded *in the first Φ_h if for each $x \in \mathbb{R}^d$ and each $w \in \mathcal{W}_{\alpha_0}$ there exist $\tilde{w} \in \mathcal{W}_{\tilde{\alpha}_0}$ with $\|\tilde{w}\|_{[t,t+h]} \leq \alpha + C\|w\|_{[t,t+h]}$ and*

$$\Phi_h(t, x, \tilde{w}) = \Psi_h(t, x, w)$$

for all $t \in h\mathbb{N}_0$.

Here we call Φ_h the embedding system *and Ψ_h the* embedded system.

Lemma 3.5. *Consider three discrete time inflated systems Φ_h, Ψ_h, and Θ_h of type* (2.2)*, and assume that Ψ_h is $(\alpha_1,\ C_1)$-embedded in Θ_h and Θ_h is $(\alpha_2,\ C_2)$-embedded in Φ_h. Then Ψ_h is $(\alpha_1 + C_1\alpha_2,\ C_1C_2)$-embedded in Φ_h.*

Proof. The proof is straightforward using Definition 3.4. □

The following proposition shows how the inflated numerical system

$$x(t + h) = \widetilde{\Phi}_h(t, x(t)) + \int_t^{t+h} w(s)ds, \tag{3.4}$$

with trajectories denoted by $\widetilde{\Phi}_h(t,x,w)$, can be embedded into the inflated time-h map (3.3), and vice versa.

PROPOSITION 3.6. *Consider the numerical approximation $\widetilde{\Phi}_h$ of system* (2.1) *for some $h > 0$. Let $\alpha_0 > 0$, and consider the constants L and c from* (2.3) *and* (2.6).

Then the α_0-inflated numerical system $\widetilde{\Phi}_h(t,x,w)$ from (3.4) *is $(ch^q,\ 1+hL)$-embedded in the $ch^q + (1+hL)\alpha_0$-inflated time-h map φ_h from* (3.3).

Conversely, the α_0-inflated time-h map φ_h from (3.3) *is $(e^{hL}ch^q,\ e^{hL})$-embedded in the $e^{hL}ch^q + e^{hL}\alpha_0$-inflated numerical system $\widetilde{\Phi}_h(t,x,w)$ from* (3.4).

Proof. Consider the auxiliary system defined by

$$x(t+h) := \varphi(h,x(t)) + \int_t^{t+h} w(s)ds$$

for $t \in h\mathbb{N}_0$, and denote the trajectories with initial value $x \in \mathbb{R}^d$ at initial time $t = 0$ by $\tilde{\varphi}_h(t,x,w)$. It is immediate from Definition 3.4 and inequality (2.6) that $\tilde{\varphi}_h$ is $(ch^q, 1)$-embedded in $\widetilde{\Phi}_h$ and that $\widetilde{\Phi}_h$ is $(ch^q, 1)$-embedded in $\tilde{\varphi}_h$.

We claim that the system $\tilde{\varphi}_h$ is $(0,\ 1+Lh)$-embedded in φ_h and that φ_h is $(0,\ e^{Lh})$-embedded in $\tilde{\varphi}_h$. Then the assertion follows from Lemma 3.5.

In order to prove the embedding relation between φ_h and $\tilde{\varphi}_h$, fix some $w \in \mathcal{W}_{\alpha_0}$. It is sufficient to show the embedding for $t = h$ since then we can continue by induction. We first construct $\tilde{w}$ such that $\varphi_h(h,x,\tilde{w}) = \tilde{\varphi}_h(h,x,w)$.

Consider the perturbation

$$\tilde{w}(t) = f(\varphi(t,x)) + w(t) - f\left(\varphi(t,x) + \int_0^t w(\tau)d\tau\right)$$

for $t \in [0,h]$. Then we obtain

$$\begin{aligned}\frac{d}{dt}\left(\varphi(t,x) + \int_0^t w(\tau)d\tau\right) &= f(\varphi(t,x)) + w(t)\\ &= f\left(\varphi(t,x) + \int_0^t w(\tau)d\tau\right) + \tilde{w}(t)\end{aligned}$$

and

$$\frac{d}{dt}\varphi(t,x,\tilde{w}) = f(\varphi(t,x,\tilde{w})) + \tilde{w}(t),$$

which by the uniqueness of the solution to this differential equation implies

$$\varphi_h(h,x,\tilde{w}) = \varphi(h,x,\tilde{w}) = \varphi(h,x) + \int_0^h w(\tau)d\tau = \tilde{\varphi}_h(h,x,w).$$

From the Lipschitz estimate (2.3) we obtain for almost all $\tau \in [0,h]$ the inequality

$$\|\tilde{w}(\tau)\| \le \|w(\tau)\| + L\left\|\int_0^\tau w(s)ds\right\|,$$

which implies

$$\|\tilde{w}\|_{[0,h]} \le \|w\|_{[0,h]} + Lh\|w\|_{[0,h]}$$

and thus shows the claim.

Conversely, given again $w \in \mathcal{W}_{\alpha_0}$, we now construct $\tilde{w}$ such that $\tilde{\varphi}_h(h, x, \tilde{w}) = \varphi_h(h, x, w)$.

For this purpose, consider $\tilde{w}$ given by

$$\tilde{w}(t) = f(\varphi(t, x, w)) + w(t) - f(\varphi(t, x))$$

for $t \in [0, h]$. Then arguments similar to those above yield the equality $\tilde{\varphi}_h(h, x, \tilde{w}) = \varphi_h(h, x, w)$. By Gronwall's lemma one easily obtains $\|\varphi(\tau, x, w) - \varphi(\tau, x)\| \le \|w\|_{[0,\tau]}(e^{L\tau} - 1)/L$, which shows that $\|f(\varphi(\tau, x, w)) - f(\varphi(\tau, x))\| \le \|w\|_{[0,\tau]}(e^{L\tau} - 1)$ and thus

$$\|\tilde{w}(\tau)\| \le \|w(\tau)\| + \|w\|_{[0,\tau]}(e^{L\tau} - 1)$$

for almost all $\tau \in [0, h]$, implying

$$\|\tilde{w}\|_{[0,h]} \le \|w\|_{[0,h]} + \|w\|_{[0,h]}(e^{Lh} - 1) = e^{Lh}\|w\|_{[0,h]},$$

i.e., the desired estimate. □

In the following two propositions we show the relations between attracting sets of embedding and embedded systems.

PROPOSITION 3.7. *Consider a discrete time system with trajectories Ψ_h which is (α, C)-embedded in some other discrete time system with trajectories denoted by Φ_h for some $\alpha \ge 0$, $C \ge 1$. Assume that the embedding system Φ_h has an attracting set A which is γ-robust for some $\gamma \in \mathcal{K}$ and some $\alpha_0 \ge \alpha$. Then the embedded system Ψ_h has an attracting set $\widetilde{A}$ with attracted neighborhood B which satisfies*

$$d_H(\widetilde{A}, A) \le \gamma(\alpha).$$

Proof. By the embedding property we obtain $\Psi_h(t, B) \subseteq \Phi_h^\alpha(t, B)$. Hence the α-attracting set A_α for the inflated embedding system Φ_h^α is an attracting set for the embedded system Ψ_h. Hence $\widetilde{A} = A_\alpha$ is the desired set. □

The next proposition shows that we can even conclude the existence of robust attracting sets for the embedded system if we are willing to allow a larger distance between $\widetilde{A}$ and A.

PROPOSITION 3.8. *Consider a discrete time system with trajectories Ψ_h which is (α, C)-embedded in some other discrete time system with trajectories denoted by Φ_h for some $\alpha \ge 0$, $C \ge 1$. Assume that the embedding system Φ_h has an attracting set A which is γ-robust for some $\gamma \in \mathcal{K}$ and some $\alpha_0 > \alpha$. Then for each $D > 1$ with $D\alpha \le \alpha_0$ the embedded system Ψ_h has an attracting set $\widetilde{A}$, which is $\gamma(CD \cdot /(D-1))$-robust for $\alpha_1 = \alpha_0(D-1)/(CD)$ and satisfies $d_H(\widetilde{A}, A) \le \gamma(D\alpha)$.*

Proof. We set $\widetilde{A} = A_{D\alpha}$. The assumption on the (α, C)-embedding implies the inclusions

$$\Psi^{\alpha'}(t, x) \subseteq \Phi^{D\alpha}(t, x) \text{ for all } \alpha' \in [0, (D-1)\alpha/C]$$

and

$$\Psi^{\alpha'}(t, x) \subseteq \Phi^{CD\alpha'/(D-1)}(t, x) \text{ for all } \alpha' \ge (D-1)\alpha/C.$$

Hence setting $\widetilde{A}_{\alpha'} = A_{D\alpha}$ for $\alpha' \in [0, (D-1)\alpha/C]$ and $\widetilde{A}_{\alpha'} = A_{CD\alpha'/(D-1)}$ for $\alpha' \ge (D-1)\alpha/C$ gives attracting sets $\widetilde{A}_{\alpha'}$ for $\Psi^{\alpha'}$ satisfying

$$d_H(\widetilde{A}_{\alpha'}, \widetilde{A}) \le d_H(\widetilde{A}_{\alpha'}, A) \le CD\alpha'/(D-1) \text{ for all } \alpha' \ge 0.$$

This shows the claim. □

4. Robustness and attraction rates. In this section we investigate the relation between the robustness gain γ and the attraction rate β. We start by showing that we can find an upper bound for the robustness gain of an attracting set A which is essentially determined by its rate of attraction.

THEOREM 4.1. *There exist maps*

$$\mu : \mathcal{KL} \times \mathbb{R}^+ \times \mathbb{R}^+ \to \mathcal{K} \quad \textit{and} \quad \sigma : \mathcal{KL} \times \mathbb{R}^+ \times \mathbb{R}^+ \to \mathbb{R}^+$$

such that each compact attracting set $A \subset \mathbb{R}^d$ with attraction rate $\beta \in \mathcal{KL}$ and attracted neighborhood B for a system of type (2.1) *or* (2.2) *is γ-robust for $\gamma = \mu(\beta, d_H(B,A), L)$ and all $\alpha_0 > 0$ satisfying $\alpha_0 \le \sigma(\beta, d_H(B,A), L)$ and $\mathcal{B}(\gamma(\alpha_0), A) \subset B$, where L is the Lipschitz constant from* (2.3) *or* (2.4), *respectively.*

Proof. Set $r_0 = d_H(A,B)$. For all $r \in (0, r_0]$ we can define

$$T_\beta(r) = \min\left\{ t \ge 0 \;\middle|\; \beta(s,t) \le \frac{s}{4} \text{ for all } s \in [r, r_0] \right\}.$$

Note that T_β is finite for all $r > 0$ (because $\beta(s,t) \le \beta(r_0,t) \to 0$ as $t \to \infty$), monotone decreasing, and continuous from above; i.e., for $r_n \searrow r$ it follows that $T_\beta(r_n) \to T_\beta(r)$ as a consequence of the continuity of β. This definition implies $\beta(s, T_\beta(r) + t) \le r/4$ for all $t \ge 0$ and all $s \in [0, r]$. We set

$$\alpha_0 = \sigma(\beta, r_0, L) := e^{-LT_\beta(r_0)} \min\{r_0, \beta(r_0, 0)\}/4.$$

Now for all $\alpha \in (0, \alpha_0]$ consider the sets

$$D_\alpha := \operatorname{cl} \mathcal{B}(r(\alpha), A),$$

where $r(\alpha)$ is chosen minimal such that $e^{LT_\beta(r(\alpha))}\alpha \le r(\alpha)/4$. The function $r(\alpha)$ is well defined because of the continuity from above of T_β. Observe that r depends only on β, r_0, and L and that it is monotone increasing with $r(\alpha) \to 0$ as $\alpha \to 0$. By Gronwall's lemma for continuous time systems or by induction for discrete time systems we obtain for $t \le T_\beta(\|x\|_A)$

$$\|\Phi(t,x,w)\|_A \le \beta(\|x\|_A, t) + e^{Lt}\alpha \tag{4.1}$$

for all $w \in \mathcal{W}_\alpha$, which implies that for each point $x \in D_\alpha$ we obtain

$$\Phi(T_\beta(r(\alpha)), x, w) \in D_\alpha \tag{4.2}$$

and

$$\|\Phi(t,x,w)\|_A \le \beta(r(\alpha), 0) + r(\alpha)/4 \text{ for all } t \in [0, T_\beta(r(\alpha))]. \tag{4.3}$$

Furthermore, for any $w \in \mathcal{W}_\alpha$ and any $x \in B$ inequality (4.1) implies that the trajectory satisfies

$$\|\Phi(i\,T_\beta(r(\alpha)), x, w)\|_A \le \max\{r_0/2^i, r(\alpha)\} \text{ for all } i \in \mathbb{N} \tag{4.4}$$

and hence hits D_α in some uniformly bounded finite time. Now we set

$$A_\alpha := \bigcup_{t \in [0, T_\beta(r(\alpha))]} \operatorname{cl} \Phi^\alpha(t, D_\alpha).$$

These sets are α-forward invariant by construction and by (4.2) and α-attracting by (4.4). Furthermore, they satisfy $A \subseteq A_\alpha$ for $\alpha \in (0, \alpha_0]$, $\mathcal{B}(r(\alpha), A) \subseteq A_\alpha$, and because of (4.3) one obtains $d_H(A_\alpha, A) \leq \gamma(\alpha)$ with

$$\gamma(\alpha) = \mu(\beta, r_0, L)(\alpha) := \beta(r(\alpha), 0) + r(\alpha)/4.$$

This shows the desired robustness property. □

In general, this construction of γ might not yield the best possible robustness gain for a given system and attracting set. However, the importance of this theorem is the uniformity that can be deduced from it: knowing the attraction rate, the distance between B and A, and the Lipschitz constant of the system allows us to give an upper bound for the robustness gain, no matter what the geometric structure of A or the behavior of Φ around or on A look like. In particular, when uniform attraction holds for a family of systems with uniform Lipschitz properties and uniform distance between the attracting sets and their attracted neighborhoods, then uniform robustness can also be deduced.

Let us illustrate one special case in which the proof of Theorem 4.1 yields an explicit expression for μ.

Example 4.2. Assume that A attracts exponentially; i.e., there exist constants $\rho > 0$ and $\lambda > 0$ such that $\beta(r, t) = \rho e^{-\lambda t} r$. In this case we obtain $T_\beta(r) = \ln(4\rho)/\lambda$, and thus $r(\alpha) = c_1\alpha$ for $c_1 = 4(4\rho)^{L/\lambda}$, and consequently $\mu(\beta, d_H(B, A), L)(\alpha) = c_2\alpha$ for $c_2 = c_1(\rho + 1/4)$. Hence exponential attraction yields γ-robustness with linear robustness gain.

Another interesting consequence of Theorem 4.1 is the following corollary.

COROLLARY 4.3. *Consider an attracting set A for a system of type* (2.1) *or* (2.2)*. Then there exist $\alpha_0 > 0$ and $\gamma \in \mathcal{K}$ such that A is γ-robust for γ and α_0.*

Proof. By Lemma 2.7 there exists an attraction rate $\beta \in \mathcal{KL}$ for A. Hence Theorem 4.1 immediately gives the assertion. □

Knowing that any attracting set admits a robustness gain, we can easily find an upper bound for a robustness gain for nested attracting sets.

LEMMA 4.4. *Let A be an attracting set with attracted neighborhood B for a system of type* (2.1) *or* (2.2)*, and let $\widehat{A} \supset A$ be an attracting set which is contained in B and is $\hat{\gamma}$-robust for the α_0-inflated system for some $\hat{\gamma} \in \mathcal{K}$ and some $\alpha_0 > 0$. Let $\rho :=$ dist$(\widehat{A}, A)$. Then A is γ-robust for this α_0 and some $\gamma \in \mathcal{K}$ satisfying $\gamma(r) \leq \hat{\gamma}(r) + \rho$.*

Proof. By Corollary 4.3 there exist $\tilde{\alpha}_0 > 0$ and $\tilde{\gamma} \in \mathcal{K}$ such that A is $\tilde{\gamma}$-robust for the $\tilde{\alpha}_0$-inflated system. Without loss of generality, we may assume $\tilde{\alpha_0} \leq \alpha_0$ and $\tilde{\gamma}(\tilde{\alpha}_0) \geq \hat{\gamma}(\tilde{\alpha}_0) + \rho$. Now for each $\alpha \in (0, \alpha_0]$ there exists an α-attracting set $\widehat{A}_\alpha \supseteq \widehat{A}$ with dist$(\widehat{A}_\alpha, \widehat{A}) \leq \hat{\gamma}(\alpha)$. Since this implies

$$\operatorname{dist}(\widehat{A}_\alpha, A) \leq \operatorname{dist}(\widehat{A}_\alpha, \widehat{A}) + \operatorname{dist}(\widehat{A}, A) \leq \hat{\gamma}(\alpha) + \rho,$$

we can conclude that A is γ-robust with γ defined by

$$\gamma(\alpha) := \begin{cases} \min\{\tilde{\gamma}(\alpha), \hat{\gamma}(\alpha) + \rho\}, & \alpha \in [0, \tilde{\alpha}_0], \\ \hat{\gamma}(\alpha) + \rho, & \alpha \in [\tilde{\alpha}_0, \alpha_0]. \end{cases}$$

This γ is easily verified to be of class $\mathcal{K}$; thus the assertion follows. □

We end this section by proving a "uniform attraction" property of the α-attracting sets appearing in the definition of the γ-robustness property.

PROPOSITION 4.5. *Consider an attracting set A with attracted neighborhood B for a system of type* (2.1) *or* (2.2) *which is γ-robust for some $\gamma \in \mathcal{K}$ and some $\alpha_0 > 0$. Then for each $\varepsilon > 0$ there exists a function $\beta \in \mathcal{KL}$ such that the inequality*

$$\|\Phi(t,x,w)\|_A \le \beta(\|x\|_A, t) + (1+\varepsilon)\gamma(\alpha) \tag{4.5}$$

holds for all $t \ge 0$, all $x \in B$, all $\alpha \in [0,\alpha_0]$, all $w \in \mathcal{W}_\alpha$, and the trajectories of the corresponding inflated system.

Proof. It is easily seen that there exists a monotone decreasing sequence $\alpha_n \to 0$ such that α_0 is the inflation parameter from the assumption and $\gamma(\alpha) < (1+\varepsilon)\gamma(\alpha_{n+1})$ for all $\alpha \in [\alpha_{n+1}, \alpha_n]$. We set $d_n = (1+\varepsilon)\gamma(\alpha_{n+1})$ and $r_0 = d_H(B,A)$. Now for each $r \in (0, r_0]$ we define the functions

$$\sigma_n(r,t) := \sup_{\tau \ge t} \operatorname{dist}(\Phi^{\alpha_n}(\tau, \mathcal{B}(r,A) \cap B), A)$$

and

$$\mu_n(r,t) := \max\{\sigma_n(r,t) - d_n, 0\}.$$

It is immediate that for all $r,t > 0$ the sequences $\sigma_n(r,t)$ and $\mu_n(r,t) + d_n$ are monotone decreasing in n and monotone increasing in r. From Remark 3.3 we obtain the existence of functions $\beta_n \in \mathcal{KL}$ such that

$$\operatorname{dist}(\Phi^{\alpha_n}(t,x), A_{\alpha_n}) \le \beta_n(\|x\|_{A_{\alpha_n}}, t). \tag{4.6}$$

This implies

$$\limsup_{t\to\infty} \operatorname{dist}(\Phi^{\alpha_n}(\tau, \mathcal{B}(r_0,A) \cap B), A) \le \gamma(\alpha_n) < d_n,$$

and thus for each $n \in \mathbb{N}$ there exists $T > 0$ such that

$$\mu_k(r,t) = 0 \text{ for all } k = 1,\dots,n, \text{ all } r \in (0,r_0], \text{ and all } t \ge T. \tag{4.7}$$

Furthermore, since $A \subseteq A_\alpha$ for all $\alpha \in (0,\alpha_0]$, from (4.6) for each $n \in \mathbb{N}$ and all $r > 0$ sufficiently small (depending on n) we obtain

$$\sigma_n(r,0) \le \beta_n(r,0) + \gamma(\alpha_n) \le d_n.$$

Hence for each $n \in \mathbb{N}$ there exists $R > 0$ such that

$$\mu_k(r,t) = 0 \text{ for all } k = 1,\dots,n, \text{ all } r \in [0,R], \text{ and all } t \ge 0. \tag{4.8}$$

Now consider the function $a(r,t) := \sup_{n\in\mathbb{N}_0} \mu_n(r,t)$. From the definition of the μ_n we obtain

$$\|\Phi(t,x,w)\|_A \le \mu_n(\|x\|_A, t) + d_n \le a(\|x\|_A, t) + d_n \le a(\|x\|_A, t) + (1+\varepsilon)\gamma(\alpha)$$

for all $t \ge 0$, all $\alpha \in [\alpha_{n+1}, \alpha_n]$, and all $w \in \mathcal{W}_\alpha$. Furthermore, $a(r,t)$ is monotone increasing in r and monotone decreasing in t. If we fix $n \in \mathbb{N}$ and choose $T > 0$ such that (4.7) holds, then (4.7) and the monotonicity of $\mu_n(r,t) + d_n$ in n imply

$$a(r,t) \le \sup_{k\in\mathbb{N}_0} \mu_k(r,t) \le \sup_{k\ge n} \mu_k(r,t) \le \sup_{k\ge n} \mu_k(r,t) + d_k \le \mu_n(r,t) + d_n \le d_n$$

for all $t \geq T$. Similarly, from (4.8) one sees that for each $n \in \mathbb{N}$ and each $t \geq 0$ there exists $R \geq 0$ such that $a(r,t) \leq d_n$ for all $r \leq R$. Thus, since $d_n \to 0$ as $n \to \infty$, we obtain $a(r,t) \to 0$ if either $r \to 0$ or $t \to \infty$. Hence by Remark 2.5 a can be bounded from above by some function $\beta \in \mathcal{KL}$, which shows the claim. □

Remark 4.6. Inequality (4.5) describes a property which in nonlinear control theory is known as *input-to-state stability (ISS);* see, e.g., the survey [20]. For a detailed comparative study of various robustness concepts for attracting sets including their characterization via Lyapunov functions, we refer to [9].

5. Discrete and continuous time systems. By its very nature, a numerical one-step approximation with time-step $h > 0$ gives only an approximation to the time-h map φ_h (2.5) of the continuous time system φ (2.1). It is therefore necessary to obtain information on the dynamical behavior of φ from its time-h map φ_h. In this section we give two results for this purpose.

PROPOSITION 5.1. *Consider a sequence of time-steps $h_n \to 0$ and a sequence of γ_n-robust attracting sets A_n for $\gamma_n \in \mathcal{K}$ and $\alpha_n > 0$, each with the same attracted neighborhood B for the inflated time-h_n maps* (3.3)*, where $\alpha_n \to \alpha_0 > 0$ as $n \to \infty$. Assume there exist $\gamma \in \mathcal{K}$ such that* $\limsup_{n\to\infty} \gamma_n(\alpha) \leq \gamma(\alpha)$ *and a compact set $A \subset B$ such that* $\lim_{n\to\infty} d_H(A_n, A) = 0$.

Then A is a γ-robust attracting set for the continuous time system (2.1) *for γ and each $\tilde{\alpha}_0 \in (0, \alpha_0)$.*

Proof. We first show that A is an attracting set for φ. For this, fix $\varepsilon > 0$ and consider $n \in \mathbb{N}$ such that $d_H(A_n, A) < \varepsilon/3$ and $h_n M < \varepsilon/3$, where M is a bound on $\|f(x)\|$ for x in a sufficiently large neighborhood of B. Then it is easily seen that there exists $T > 0$ such that $\varphi(ih_n, B) \subset \mathcal{B}(\varepsilon/3, A_n)$ for all $i \in \mathbb{N}$ with $ih_n \geq T$. Consequently, we obtain $\varphi(t, B) \subset \mathcal{B}(\varepsilon, A)$ for all $t \geq T$, and since $\varepsilon > 0$ was arbitrary, this shows the desired convergence $\operatorname{dist}(\varphi(t,B), A) \to 0$ as $t \to \infty$.

It remains to show the γ-robustness. To this end, fix some $\alpha \in (0, \alpha_0)$ and consider the set

$$A^\alpha = \bigcap_{n\geq 0} \operatorname{cl} \bigcup_{k\geq n} A_n^\alpha,$$

where the A_n^α denote the α-attracting sets for the inflated time-h_n maps (3.3). Using the fact that

$$d_H\left(\operatorname{cl} \bigcup_{k\geq n} A_n^\alpha,\, A^\alpha\right) \to 0 \text{ as } k \to \infty$$

(cf. [2, Proposition 1.1.5]), with the same argument as above one sees that this set is α-attracting for the inflated system. Since for each $\varepsilon > 0$ we find $N \in \mathbb{N}$ such that for all $n \geq N$ the inequalities

$$d_H(A_n, A) < \varepsilon/2 \quad \text{and} \quad d_H(A_n^\alpha, A_n) \leq \gamma_n(\alpha) \leq \gamma(\alpha) + \varepsilon/2$$

hold, we can conclude that $d_H(A_n^\alpha, A) \leq \gamma(\alpha) + \varepsilon$ for all $n \geq N$ and thus

$$d_H(A_n^\alpha, A) \leq \gamma(\alpha) + \varepsilon,$$

which shows the desired distance since $\varepsilon > 0$ was arbitrary. □

While in general an attracting set for the time-h map is not an attracting set for the continuous time system, this property is always true for attractors, as the following lemma shows.

LEMMA 5.2. *Let $h > 0$ and A_h be an attractor with attracted neighborhood B for the time-h map φ_h* (2.5) *of the continuous time system* (2.1). *Then A_h is also an attractor with attracted neighborhood B for the continuous time system* (2.1).

Proof. We first show forward invariance of A_h for φ, i.e., $\varphi(t, A_h) \subseteq A_h$ for each $t \geq 0$. By invariance of A_h for φ_h, for each $t \geq 0$ we know that $\varphi_h(h, \varphi(t, A_h)) = \varphi(t, \varphi_h(h, A_h)) = \varphi(t, A_h)$; hence $\varphi(t, A_h)$ is a compact invariant set for φ_h, and by Lemma 2.10 it is contained in A_h.

Now we show that A_h is an attracting set for the continuous time system φ. Forward invariance of A_h and continuous dependence on the initial value imply that for each $\delta > 0$ there exists $\varepsilon > 0$ such that

$$d_H(D, A_h) < \varepsilon \quad \Rightarrow \quad d_H(\varphi(t, D), A_h) < \delta$$

for all $t \in [0, h]$ and arbitrary bounded sets $D \subset \mathbb{R}^d$. Since attractivity of A_h for φ_h implies $\lim_{i \to \infty, i \in \mathbb{N}} \operatorname{dist}(\varphi(ih, B), A) = 0$, we can conclude $\lim_{t \to \infty} \operatorname{dist}(\varphi(t, B), A_h) = 0$; i.e., A_h is also an attracting set for φ with attracted neighborhood B.

It remains to show that A_h is an attractor for φ. By Lemma 2.8 there exists an attractor $A \subseteq A_h$ for φ. This, in turn, is also an attractor set for φ_h; hence by Lemma 2.9 it must coincide with A_h. Thus A_h is an attractor for φ. □

6. Numerical discretization. In this section we combine the results from the previous sections in order to derive criteria under which one can conclude the existence of attracting sets and attractors from numerical approximations. We start with sufficient conditions for the existence of attracting sets.

PROPOSITION 6.1. *Consider the continuous time system* (2.1) *and a numerical one-step approximation $\widetilde{\Phi}_h$ for $h > 0$ satisfying* (2.6) *for $c\,q > 0$. Let L denote a Lipschitz constant for both systems from* (2.3) *and* (2.4), *respectively.*

(a) *Let A be a γ-robust attracting set for* (2.1) *for $\gamma \in \mathcal{K}$ and $\alpha_0 \geq e^{hL}ch^q$. Then there exists an attracting set A_h for the discrete time system induced by the numerical approximation $\widetilde{\Phi}_h$ satisfying*

$$d_H(A_h, A) \leq \gamma(e^{hL}ch^q).$$

(b) *Let A_h be a γ-robust attracting set for $\widetilde{\Phi}_h$, for $\gamma \in \mathcal{K}$, and $\alpha_0 \geq ch^q$. Then there exists an attracting set $\widetilde{A}_h$ for the time-h map* (2.5) *of the continuous time system satisfying*

$$d_H(\widetilde{A}_h, A_h) \leq \gamma(ch^q).$$

Proof. This follows directly from Propositions 3.6 and 3.7. □

THEOREM 6.2. *Consider the continuous time system* (2.1) *and a family of numerical one-step approximations $\widetilde{\Phi}_{h_n}$ satisfying* (2.6) *for a sequence of time-steps $h_n \to 0$ as $n \to \infty$. Let A_n be attractors for the discrete time systems induced by these numerical approximations, each with the same attracted neighborhood B, and assume that there exists a compact set $A \subset B$ with $d_H(A_n, A) \to$ as $n \to \infty$.*

Then the following properties are equivalent:

(i) *A is an attractor for* (2.1) *with attracted neighborhood B.*

(ii) *There exist $N \in \mathbb{N}$, $\gamma \in \mathcal{K}$, $\alpha_0 > 0$, and sequences $C_n \to 1$ and $\rho_n \to 0$ such that for each $n \geq N$ the attractor A_n is γ_n-robust for the numerical system $\widetilde{\Phi}_{h_n}$ for α_0 and $\gamma_n(r) \leq \gamma(C_n r) + \rho_n$.*

(iii) *There exist $N \in \mathbb{N}$, $\beta \in \mathcal{KL}$, and a sequence $\varepsilon_n \to 0$ such that for each $n \geq N$ the attractor A_n for the numerical system $\widetilde{\Phi}_{h_n}$ has attraction rate $\beta_n \in \mathcal{KL}$ satisfying $\beta_n(r,t) \leq \beta(r+\varepsilon_n, t) + \varepsilon_n$.*

In addition, if (ii) *holds for $\gamma \in \mathcal{K}$ and $\alpha_0 > 0$, then A is γ-robust for this γ and each $\tilde{\alpha}_0 \in (0, \alpha_0)$, and if* (iii) *holds for $\beta \in \mathcal{KL}$, then A is attracting with this rate β for this continuous time system* (2.1).

Proof. (i)⇒(ii): Since A is an attracting set by Corollary 4.3, it is also γ-robust for some suitable $\gamma \in \mathcal{K}$ and $\tilde{\alpha}_0 > 0$. Since by Proposition 3.6 the α_0-inflated numerical system (3.4) for $\alpha_0 = \tilde{\alpha}_0/2$ and n sufficiently large is embedded in the $\tilde{\alpha}_0$-inflated time-h_n map (3.3), Proposition 3.8 applied with $\alpha = ch_n^q$, $C = 1 + h_n L$, and $D = 1/\sqrt{ch_n^q}$ implies the existence of $\tilde{\gamma}_n$-robust attracting sets $\widetilde{A}_n$ with $d_H(A, \widetilde{A}_n) \leq \gamma(\sqrt{ch_n^q})$ and

$$\tilde{\gamma}_n(r) \leq \gamma\left(\frac{1 + h_n L}{1 - \sqrt{ch_n^q}} r\right).$$

Since the attractors A_n converge to A and—by minimality—are contained in the attracting sets $\widetilde{A}_n$, we can conclude that $\rho_n = \operatorname{dist}(A_n, \widetilde{A}_n) \to 0$ as $n \to \infty$; hence by Lemma 4.4 the A_n are γ_n-robust with $\gamma_n(r) \leq \tilde{\gamma}_n(r) + \rho_n$ and α_0, which shows the claim.

(i) ⇒ (iii): Since A is an attracting set, by Corollary 4.3 and Proposition 4.5, it satisfies inequality (4.5) for suitable β, γ, and ε. Thus by Proposition 3.6 for all sufficiently large $n \in \mathbb{N}$, all $x \in B$, and all $i \in \mathbb{N}$, we find some $w \in \mathcal{W}_{e^{h_n L} ch_n^q}$ such that

$$\varphi(ih_n, x, w) = \widetilde{\Phi}_{h_n}(ih_n, x).$$

This yields

$$\begin{aligned}
\|\widetilde{\Phi}_{h_n}(ih_n, x)\|_{A_n} &\leq \|\widetilde{\Phi}_{h_n}(ih_n, x)\|_A + d_H(A_n, A) \\
&= \|\varphi(ih_n, x, w)\|_A + d_H(A_n, A) \\
&\leq \beta(\|x\|_A, ih_n) + d_H(A_n, A) + (1+\varepsilon)\gamma(e^{h_n L} ch_n^q) \\
&\leq \beta(\|x\|_{A_n} + d_H(A_n, A), ih_n) + d_H(A_n, A) + (1+\varepsilon)\gamma(e^{h_n L} ch_n^q),
\end{aligned}$$

which shows the assertion for $\varepsilon_n = d_H(A_n, A) + (1+\varepsilon)\gamma(e^{h_n L} ch_n^q)$.

(ii) ⇒ (i): Similar to the arguments in case "(i) ⇒ (ii)," we obtain that for sufficiently large $n \in \mathbb{N}$ there exist $\gamma_n(D_n \cdot)$-robust attracting sets $\widetilde{A}_n$ for the inflated time-h_n map (3.3) of the α_n-inflated system for suitable constants $C_n \to 1$ and $\alpha_n \to \alpha_0$, such that $d_H(\widetilde{A}_n, A) \to 0$. Hence by Proposition 5.1 we obtain that A is a γ-robust attracting set for each $\tilde{\alpha}_0 \in (0, \alpha_0)$ for (2.1). It remains to show that A is an attractor. If this is not the case, then by Lemmas 2.8 and 2.9 there exists an attractor $\widetilde{A}$ for φ with $\widetilde{A} \subset A$, $\widetilde{A} \neq A$, and attracted neighborhood B. Denote $\eta := d_H(\widetilde{A}, A) > 0$. Again following the arguments from the case "(i) ⇒ (ii)," this implies that for all sufficiently large $n \in \mathbb{N}$ the attractors A_n for the numerical systems $\widetilde{\Phi}_{h_n}$ must satisfy $\operatorname{dist}(A_n, \widetilde{A}) < \eta/2$. This implies $d_H(A_n, A) \geq \eta/2$ and hence contradicts the convergence $d_H(A_n, A) \to 0$ for $n \to \infty$.

(iii) ⇒ (i): Fixing some $T > 0$ and some $\varepsilon > 0$, by Gronwall's lemma for all $n > 0$ sufficiently large and all $x \in B$ we obtain the inequality

$$\|\widetilde{\Phi}_h(i_n h_n, x) - \varphi(T, x)\| \leq \varepsilon,$$

where $i_n \in \mathbb{N}$ can be chosen such that $|T - i_n h_n| < \varepsilon$. Hence from the convergence of A_n to A and from the properties of β_n and β we obtain

$$\|\varphi(T,x)\|_A \leq \beta(\|x\|_A + \varepsilon, T + \varepsilon) + \varepsilon,$$

and since $\varepsilon > 0$ was arbitrary by continuity of β, this implies

$$\|\varphi(T,x)\|_A \leq \beta(\|x\|_A, T),$$

which implies that A is an attracting set since $T > 0$ was arbitrary. The fact that A is an attractor follows similarly to the case "(ii) ⇒ (i)," above. □

In other words, Theorem 6.2 states that a sequence of "numerical" attractors converges to a "real" attractor if and only if the elements of this sequence are either uniform robust or attracting with a uniform rate.

Remark 6.3. Note that we have used the minimality of attractors only in the proof of the implication "(i) ⇒ (ii)." Hence the equivalence "(i) ⇔ (iii)" and the implication "(ii) ⇒ (i)" remain true for general attracting sets.

In the next theorem we shift our attention to a sequence of uniformly robust numerical attractors (in the sense of Theorem 6.2 (ii)) without the a priori assumption about convergence of these sets. It turns out that this sequence of numerical attractors converges to a set if and only if this set is an attractor.

THEOREM 6.4. *Consider the continuous time system* (2.1) *and a family of numerical one-step approximation* $\widetilde{\Phi}_{h_n}$ *satisfying* (2.6) *for a sequence of time-steps* $h_n \to 0$ *as* $n \to \infty$*. Let* A_n *be attractors for the discrete time systems induced by these numerical approximations, each with the same attracted neighborhood* B*, assume that they are* γ_n*-robust for the numerical system* $\widetilde{\Phi}_{h_n}$*, for some* $\alpha_0 > 0$ *and* $\gamma_n(r) \leq \gamma(C_n r) + \rho_n$ *for some suitable* $\gamma \in \mathcal{K}$ *and sequences* $C_n \to 1$ *and* $\rho_n \to 0$*, and let* $A \subset B$ *be a compact set.*

Then the following statements are equivalent.

(i) *A is an attractor for* (2.1) *with attracted neighborhood* B*.*

(ii) $d_H(A_n, A) \to 0$ *as* $n \to \infty$*.*

In this case, A is γ-robust for (2.1) *for* γ *and each* $\tilde{\alpha}_0 \in (0, \alpha_0)$*.*

Proof. (i) ⇒ (ii): Since by Lemma 2.7 and Theorem 4.1 the attractor A is γ-robust for some suitable $\gamma \in \mathcal{K}$, by Proposition 6.1(a) and Lemma 2.8 we can conclude $\operatorname{dist}(A_n, A) \to 0$ as $n \to \infty$. For the converse "dist" estimate, by the assumption on the γ_n-robustness of the A_n, for each $\varepsilon > 0$ and all $n \in \mathbb{N}$ sufficiently large there exist attracting sets $\widetilde{A}_n$ for the time-h_n map of the continuous time system with $\operatorname{dist}(\widetilde{A}_n, A_n) \leq \varepsilon$. By Lemma 2.8 each of these sets contains an attractor for the time-h_n map (2.5) and attracted neighborhood B, which by Lemma 5.2 coincides with A. This implies $\operatorname{dist}(A, A_n) \leq \varepsilon$, and since $\varepsilon > 0$ was arbitrary we obtain $\operatorname{dist}(A, A_n) \to 0$ as $n \to \infty$. This shows the desired convergence.

(ii) ⇒ (i): This follows from the implication "(ii) ⇒ (i)" in Theorem 6.2. □

In other words, Theorem 6.4 states that a sequence of uniformly robust "numerical" attractors converges to some set A if and only if it is a "real" attractor.

Finally, we are going to investigate the rates of convergence of A_n to A under the assumptions of Theorem 6.4.

THEOREM 6.5. *Consider the continuous time system* (2.1) *and a family of numerical one-step approximation* $\widetilde{\Phi}_{h_n}$ *satisfying* (2.6) *for a sequence of time-steps* $h_n \to 0$ *as* $n \to \infty$ *and constants* c*,* $q > 0$*. Let* A_n *be attractors for the discrete time systems* *induced by these numerical approximations, each with the same attracted neighborhood B, and assume that they are γ_n-robust for the numerical system $\widetilde{\Phi}_{h_n}$ for some $\alpha_0 > 0$ and $\gamma_n(r) \le \gamma(C_n r) + \rho_n$ for some suitable $\gamma \in \mathcal{K}$ and sequences $C_n \to 1$ and $\rho_n \to 0$. Let $A \subset B$ be a compact set, and assume that one of the following two conditions is satisfied:*

(i) *A is an attractor for (2.1) with attracted neighborhood B.*

(ii) *$d_H(A_n, A) \to 0$ as $n \to \infty$.*

Then for all sufficiently large $n \in \mathbb{N}$ we obtain the estimates

$$\operatorname{dist}(A, A_n) \le \gamma(C_n e^{Lh_n} c h_n^q) + \rho_n \quad \textit{and} \quad \operatorname{dist}(A_n, A) \le \gamma(c h_n^q)$$

for the rate of convergence of A_n to A.

Proof. Under the assumptions, Theorem 6.4 implies that A is γ-robust for (2.1). Hence by Proposition 6.1 we obtain the existence of attracting sets for the numerical systems and the time-h_n maps, respectively, with the desired distances. By Lemmas 2.8, 2.9, and 5.2, the attractors A_n and A are contained in these attracting sets, and hence the "dist" estimates remain valid. □

REFERENCES

[1] F. Albertini and E. D. Sontag, *Continuous control–Lyapunov functions for asymptotically stable continuous time–varying systems*, Internat. J. Control, 72 (1999), pp. 1630–1641.

[2] J.-P. Aubin and H. Frankowska, *Set-Valued Analysis*, Birkhäuser Boston, Boston, 1990.

[3] M. Dellnitz and A. Hohmann, *A subdivision algorithm for the computation of unstable manifolds and global attractors*, Numer. Math., 75 (1997), pp. 293–317.

[4] P. Deuflhard and F. Bornemann, *Numerische Mathematik.* II: *Integration gewöhnlicher Differentialgleichungen*, de Gruyter, Berlin, 1994.

[5] B. M. Garay, *Discretization and Morse–Smale dynamical systems on planar discs*, Acta Math. Univ. Comenian. (N.S.), 63 (1994), pp. 25–38.

[6] B. M. Garay, *On structural stability of ordinary differential equations with respect to discretization methods*, Numer. Math., 72 (1996), pp. 449–479.

[7] B. M. Garay and P. E. Kloeden, *Discretization near compact invariant sets*, Random Comput. Dyn., 5 (1997), pp. 93–123.

[8] L. Grüne, *Persistence of attractors for one–step discretizations of ordinary differential equations*, IMA J. Numer. Anal., 21 (2001), pp. 751–767.

[9] L. Grüne, *Asymptotic Behavior of Dynamical and Control Systems under Perturbation and Discretization*, Lecture Notes in Math. 1783, Springer-Verlag, Berlin, 2002.

[10] W. Hahn, *Stability of Motion*, Springer-Verlag, Berlin, Heidelberg, 1967.

[11] J. K. Hale, X.-B. Lin, and G. Raugel, *Upper semicontinuity of attractors for approximations of semigroups and partial differential equations*, Math. Comp., 50 (1988), pp. 89–123.

[12] J. K. Hale and G. Raugel, *Lower semicontinuity of attractors of gradient systems and applications*, Ann. Mat. Pura Appl. (4), 154 (1989), pp. 281–326.

[13] A. Iserles, *A First Course in the Numerical Analysis of Differential Equations*, Cambridge Texts Appl. Math., Cambridge University Press, Cambridge, UK, 1996.

[14] O. Junge, *Mengenorientierte Methoden zur numerischen Analyse dynamischer Systeme*, Shaker Verlag, Aachen, 2000; Dissertation, Universität Paderborn.

[15] O. Junge, *Rigorous discretization of subdivision techniques*, in EQUADIFF 99, B. Fiedler, K. Gröger, and J. Sprekels, eds., World Scientific, Singapore, 2000, pp. 916–918.

[16] P. E. Kloeden, *Asymptotically stable attracting sets in the Navier–Stokes equations*, Bull. Austral. Math. Soc., 34 (1986), pp. 37–52.

[17] P. E. Kloeden and V. S. Kozyakin, *The inflation of attractors and their discretization: The autonomous case*, Nonlinear Anal., 40 (2000), pp. 333–343.

[18] P. E. Kloeden and J. Lorenz, *Stable attracting sets in dynamical systems and their one-step discretizations*, SIAM J. Numer. Anal., 23 (1986), pp. 986–995.

[19] Y. Lin, E. D. Sontag, and Y. Wang, *A smooth converse Lyapunov theorem for robust stability*, SIAM J. Control Optim., 34 (1996), pp. 124–160.

[20] E. D. Sontag, *The ISS philosophy as a unifying framework for stability–like behavior*, in Nonlinear Control in the Year 2000, Volume 2, A. Isidori, F. Lamnabhi-Lagarrigue, and W. Respondek, eds., Lecture Notes in Control Inform. Sci. 259, Springer-Verlag, London, 2000, pp. 443–468.

[21] J. Stoer and R. Bulirsch, *Introduction to Numerical Analysis*, Springer-Verlag, New York, 1980.

[22] A. M. Stuart and A. R. Humphries, *Dynamical Systems and Numerical Analysis*, Cambridge University Press, Cambridge, UK, 1996.

SIAM J. NUMER. ANAL.
Vol. 41, No. 6, pp. 2114–2130

ON THE VELOCITY-PRESSURE-VORTICITY LEAST-SQUARES MIXED FINITE ELEMENT METHOD FOR THE 3D STOKES EQUATIONS[*]

HUO-YUAN DUAN[†] AND GUO-PING LIANG[†]

Abstract. This paper provides a new strict mathematical analysis for the velocity-pressure-vorticity least-squares methods (i.e., the standard linear element method and the Bochev–Gunzburger method) for the 3D Stokes problem with homogeneous velocity boundary condition. The analysis shows that, in general, the divergence of the vorticity does not affect the coerciveness and the accuracy. This admits the use of the edge element for the vorticity to reduce the number of whole unknowns. Moreover, the analysis also shows that, in the standard linear element method, the L^2 error bound for the velocity is $\mathcal{O}(h^{3/2})$ generally. Numerical examples are presented.

Key words. Stokes equation, velocity-pressure-vorticity least-squares mixed finite element method, Aubin–Nitsche duality argument

AMS subject classification. 65N30

DOI. 10.1137/S0036142901399604

1. Introduction. As far as the Stokes problem is concerned, it is well known that the primal form in terms of velocity-pressure requires the approximating spaces to satisfy the Babuška–Brezzi condition in order to obtain stability and convergence. Generally, this condition excludes the use of simple equal low-order elements (e.g., the continuous linear elements). Besides, an indefinite system is often a source of trouble in practical implementation; see [13], [14] for more details.

In addition, when some important variable (e.g., the vorticity) in the Stokes problem is needed, it seems impossible to obtain accurate approximate solutions in the mixed method in terms of vector potential-vorticity, with the use of the continuous linear elements; see, e.g., [13].

During the past decade, the least-squares mixed finite element method has been demonstrated to be very effective as an alternative approach to numerically solve the Stokes problem (see [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12]), which offers many advantages such as that any combination of simple low-order approximating spaces (e.g., the linear elements) can be employed for all variables and that the weak problem is generally coercive; cf. [2].

Undoubtedly, the least-squares scheme in terms of velocity-pressure-vorticity for the Stokes problem is such a method (cf. [1], [3], [6], [9], [10]). Due to solving only seven unknowns (the three dimensional (3D) case), this scheme is the simplest and the most widely used one in all schemes of least-squares type for the Stokes problem. To our best knowledge, however, a strict mathematical analysis for this scheme has been missing for many years. (In the earlier works [9], [35], the mathematical analysis and conclusions are not correct due to the fact that the H^1 coerciveness does not hold for all variables.)

On the other hand, as is pointed out by several authors (see [1], [2], [3], [4], [5]), in three dimensions the divergence of the vorticity is indispensable to both coerciveness

[*] Received by the editors December 12, 2001; accepted for publication (in revised form) April 10, 2003; published electronically November 25, 2003.

http://www.siam.org/journals/sinum/41-6/39960.html

[†] Institute of Mathematics, Chinese Academy of Sciences, Beijing 100080, People's Republic of China (dhymath@yahoo.com.cn, guoping@math03.math.ac.cn).

and accuracy, although it is not so in two dimensions. This viewpoint is drawn from the Agmon–Doulis–Nirenberg (ADN) theory (cf. [4], [5]). However, up until now, it in fact has not been very clear whether both coerciveness and accuracy are affected if the divergence of the vorticity is not introduced. Clearly, if the divergence of the vorticity can be dropped, then the tangential continuous element (i.e., the edge element or the Nédélec element [23], [24]) can be used for this variable, and as a consequence, the number of unknowns will be greatly reduced, without loss of accuracy.

Therefore, it is of interest and importance to investigate the coerciveness and the accuracy for the velocity-pressure-vorticity least-squares method when employing the continuous elements and not including the divergence of the vorticity.

In this paper, we show that the weak problem of this scheme is indeed coercive for both pressure and vorticity over the L^2 space and for velocity over the H^1 space and that the accuracy is not affected generally, even if the divergence of the vorticity is not introduced into the weak problem.

In particular, without including the divergence of the vorticity in the Bochev–Gunzburger method (cf. [4], [2], [3]), we show that the error bounds for all variables are still optimal. This corrects the traditional viewpoint.

Moreover, for the least-squares scheme (without the divergence of the vorticity) with the use of the continuous linear elements for all variables, in general, we show that the L^2 error bound for the velocity is $\mathcal{O}(h^{3/2})$. The numerical results given in the last section support these theoretical results on error bounds.

We remark that all these L^2 error estimates are derived from the classical Aubin–Nitsche duality argument (cf. [15], [16], [18]), which usually prerequisitely requires that the domain occupied by the flow be suitably smooth to ensure the existence of the solution and some regularities.

In addition, owing to that the auxiliary variational problem, introduced in the duality argument, cannot be coercive over the whole space for both pressure and vorticity and that it is no longer a least-squares first-order system, the existence of the solution cannot be directly deduced from either the Lax–Milgram lemma (cf. [15], [18]) or the ADN theory (cf. [4], [5]). In fact, it turns out that the existence of the solution and the corresponding regularities, which are derived from a constructive approach in this paper, are not trivial.

Let us mention the recent work [34]. A completely different approach is taken in [34] to obtain the same error bounds in the energy norm, but no error estimates are given for the velocity in the L^2 norm.

This paper is outlined as follows: In section 2, we recall some Hilbert spaces and inequalities and the regularities to classical problems. In section 3, for the velocity-pressure-vorticity least-squares method, dropping the divergence of the vorticity, we obtain both coerciveness and basic error bounds. In section 4, using the Aubin–Nitsche duality argument, we derive an improved error bound $\mathcal{O}(h^{3/2})$ for the velocity in the standard linear element method. In section 5, we show that the error bounds in the Bochev–Gunzburger method are still optimal, even without introducing the divergence of vorticity. In section 6, numerical examples are given to verify the theoretical results in sections 3 and 4.

2. Inequalities and regularities.

Let $\Omega \subset \Re^3$ be an open bounded domain, with boundary $\Gamma = \partial\Omega$ and $\boldsymbol{n}$ the unit normal vector to Γ. Elementary differential operators are recalled as follows:

$$\partial^r v = \frac{\partial^{|r|} v}{\partial x_1^{r_1} \partial x_2^{r_2} \partial x_3^{r_3}}, \quad |r| = r_1 + r_2 + r_3,$$

$$\bigtriangledown v = (\partial v/\partial x_1, \partial v/\partial x_2, \partial v/\partial x_3), \quad \mathbf{curl}\, \boldsymbol{v} = \bigtriangledown \times \boldsymbol{v}, \quad \mathrm{div}\, \boldsymbol{v} = \bigtriangledown \cdot \boldsymbol{v}.$$

We introduce the Hilbert spaces

$$\begin{aligned}
L^2(\Omega) &= \{v; \textstyle\int_\Omega v^2 < \infty\},\\
L^2_0(\Omega) &= \{q \in L^2(\Omega); \textstyle\int_\Omega q = 0\},\\
H^m(\Omega) &= \{\partial^r v \in L^2(\Omega), 0 \le |r| \le m\} \quad (m \ge 1),\\
H^1_0(\Omega) &= \{v \in L^2(\Omega); \bigtriangledown v \in (L^2(\Omega))^3, v_{|\Gamma} = 0\},\\
H(\mathbf{curl}; \Omega) &= \{\boldsymbol{v} \in (L^2(\Omega))^3; \mathbf{curl}\, \boldsymbol{v} \in (L^2(\Omega))^3\},\\
H_0(\mathbf{curl}; \Omega) &= \{\boldsymbol{v} \in H(\mathbf{curl}; \Omega); \boldsymbol{v} \times \boldsymbol{n}_{|\Gamma} = \mathbf{0}\},\\
H(\mathrm{div}; \Omega) &= \{\boldsymbol{v} \in (L^2(\Omega))^3; \mathrm{div}\, \boldsymbol{v} \in L^2(\Omega)\},\\
H_0(\mathrm{div}; \Omega) &= \{\boldsymbol{v} \in H(\mathrm{div}; \Omega); \boldsymbol{v} \cdot \boldsymbol{n}_{|\Gamma} = 0\},
\end{aligned}$$

where all the above spaces are equipped with natural norms (cf. [16], [13], [27], [26]). The following norms will be encountered in this paper:

$$||\boldsymbol{v}||^2_{0,\mathrm{div}} = ||\boldsymbol{v}||^2_0 + ||\mathrm{div}\, \boldsymbol{v}||^2_0, \qquad ||\boldsymbol{v}||^2_{0,\mathbf{curl}} = ||\boldsymbol{v}||^2_0 + ||\mathbf{curl}\, \boldsymbol{v}||^2_0.$$

In addition, H^{-1} is the dual space of $H^1_0(\Omega)$.

Two of Green's formulae of integration by parts are as follows (cf. [13]):

$$\begin{aligned}
(\boldsymbol{v}, \bigtriangledown \phi) + (\mathrm{div}\, \boldsymbol{v}, \phi) &= \int_\Gamma \boldsymbol{v} \cdot \boldsymbol{n}\, \phi \quad \forall \boldsymbol{v} \in H(\mathrm{div}; \Omega), \forall \phi \in H^1(\Omega),\\
(\mathbf{curl}\, \boldsymbol{v}, \boldsymbol{\phi}) - (\boldsymbol{v}, \mathbf{curl}\, \boldsymbol{\phi}) &= \int_\Gamma \boldsymbol{v} \times \boldsymbol{n}\, \boldsymbol{\phi} \quad \forall \boldsymbol{v} \in H(\mathbf{curl}; \Omega), \forall \boldsymbol{\phi} \in (H^1(\Omega))^3.
\end{aligned}$$

PROPOSITION 2.1 (see [13, 20, 21, 30]). *Assume that $\Omega \subset \Re^3$ is a simply connected and bounded domain with a Lipschitz continuous boundary Γ. Then*

$$||\boldsymbol{v}||_0 \le C\, \{||\mathbf{curl}\, \boldsymbol{v}||_0 + ||\mathrm{div}\, \boldsymbol{v}||_0\} \quad \forall \boldsymbol{v} \in H_0(\mathrm{div}; \Omega) \cap H(\mathbf{curl}; \Omega). \tag{2.1}$$

PROPOSITION 2.2 (see [13, 31]). *Assume that $\Omega \subset \Re^3$ is a simply connected bounded domain with $C^{1,1}$ boundary Γ or is a bounded and convex polyhedron. Then*

$$|\boldsymbol{v}|_1 \le C\, \{||\mathbf{curl}\, \boldsymbol{v}||_0 + ||\mathrm{div}\, \boldsymbol{v}||_0\} \quad \forall \boldsymbol{v} \in H_0(\mathrm{div}; \Omega) \cap (H^1(\Omega))^3. \tag{2.2}$$

PROPOSITION 2.3 (see [13, 17, 22]). *Assume that $\Omega \subset \Re^3$ is a simply connected and bounded domain with a Lipschitz continuous boundary Γ. Given $\boldsymbol{\chi} \in (H^{-1}(\Omega))^3$ and $g \in L^2_0(\Omega)$, there exists a unique solution $(\boldsymbol{u}, p) \in (H^1_0(\Omega))^3 \times (H^1(\Omega) \cap L^2_0(\Omega))$ to the Stokes problem*

$$-\Delta\, \boldsymbol{u} + \bigtriangledown p = \boldsymbol{\chi}, \quad \mathrm{div}\, \boldsymbol{u} = g, \quad \boldsymbol{u}_{|\Gamma} = \mathbf{0}, \tag{2.3}$$

the solution of which satisfies

$$||\boldsymbol{u}||_1 + ||p||_0 \le C\, (||\boldsymbol{\chi}||_{-1} + ||g||_0). \tag{2.4}$$

Moreover, if $g \equiv 0$, then $p \in L^2_0(\Omega)$ satisfies

$$(p, \mathrm{div}\, \boldsymbol{v}) = (\mathbf{curl}\, \boldsymbol{u}, \mathbf{curl}\, \boldsymbol{v}) - (\boldsymbol{\chi}, \boldsymbol{v}) \quad \forall \boldsymbol{v} \in (H^1_0(\Omega))^3. \tag{2.5}$$

If additionally $\Gamma \in C^{r+2}$, $\boldsymbol{\chi} \in (H^r(\Omega))^3$, and $g \in H^{r+1}(\Omega)$ with $r = 0, 1$, we have

$$||\boldsymbol{u}||_{r+2} + ||p||_{r+1} \le C\, (||\boldsymbol{\chi}||_r + ||g||_{r+1}). \tag{2.6}$$

3. Coerciveness without the divergence of the vorticity. We consider the Stokes problem in 3D space,

$$-\Delta\, \boldsymbol{u} + \bigtriangledown p = \boldsymbol{f}, \quad \operatorname{div} \boldsymbol{u} = 0, \quad \text{in } \Omega, \quad \boldsymbol{u} = \boldsymbol{0}, \quad \text{on } \Gamma, \tag{3.1}$$

where $\boldsymbol{u} \in (H_0^1(\Omega))^3$ and $p \in L_0^2(\Omega)$ are velocity and pressure, respectively, $\boldsymbol{f} \in (L^2(\Omega))^3$ is the given function, and $\Omega \subset \Re^3$ is the domain occupied by the flow, with boundary $\Gamma = \partial\Omega$ and the unit normal vector $\boldsymbol{n}$ to Γ.

Let the vorticity

$$\boldsymbol{\omega} = \mathbf{curl}\, \boldsymbol{u} \tag{3.2}$$

be a new unknown; in the light of $\mathbf{curl}\,\mathbf{curl}\, \boldsymbol{u} = -\Delta\, \boldsymbol{u} + \bigtriangledown \operatorname{div} \boldsymbol{u}$ and $\operatorname{div} \boldsymbol{u} = 0$, we can write (3.1) in the first-order system as follows:

$$\mathbf{curl}\, \boldsymbol{\omega} + \bigtriangledown p = \boldsymbol{f}, \quad \boldsymbol{\omega} = \mathbf{curl}\, \boldsymbol{u}, \quad \operatorname{div} \boldsymbol{u} = 0, \quad \text{in } \Omega, \tag{3.3}$$

where

$$\boldsymbol{u}_{|\Gamma} = \boldsymbol{0}, \qquad \int_\Omega p = 0. \tag{3.4}$$

Remark 3.1. Clearly, there naturally hold (cf. [11], [33])

$$\boldsymbol{\omega} \cdot \boldsymbol{n}_{|\Gamma} = 0, \quad \operatorname{div} \boldsymbol{\omega} = 0, \tag{3.5}$$

which are redundant equations for the first-order system (3.3) and (3.4). This can be seen from Theorem 3.1.

Let

$$\mathcal{J}(\boldsymbol{u}, p, \boldsymbol{\omega}) = ||\mathbf{curl}\, \boldsymbol{\omega} + \bigtriangledown p||_0^2 + ||\boldsymbol{\omega} - \mathbf{curl}\, \boldsymbol{u}||_0^2 + ||\operatorname{div} \boldsymbol{u}||_0^2, \tag{3.6}$$

$$\mathcal{J}^+(\boldsymbol{u}, p, \boldsymbol{\omega}) = \mathcal{J}(\boldsymbol{u}, p, \boldsymbol{\omega}) + ||\operatorname{div} \boldsymbol{\omega}||_0^2. \tag{3.7}$$

THEOREM 3.1. *Under the same conditions as in Proposition* 2.2*, there holds*

$$\mathcal{J}(\boldsymbol{u}, p, \boldsymbol{\omega}) \geq C\, \{||\boldsymbol{u}||_1^2 + ||p||_0^2 + ||\boldsymbol{\omega}||_0^2\} \tag{3.8}$$

for all $(\boldsymbol{u}, p, \boldsymbol{\omega}) \in (H_0^1(\Omega))^3 \times (H^1(\Omega) \cap L_0^2(\Omega)) \times H(\mathbf{curl}; \Omega)$.

Proof. Let $\alpha > 0$ be a constant to be determined; in the light of $(\boldsymbol{\omega}, \mathbf{curl}\, \boldsymbol{u}) = (\mathbf{curl}\, \boldsymbol{\omega}, \boldsymbol{u})$, we have

$$\begin{aligned} ||\boldsymbol{\omega} - \mathbf{curl}\, \boldsymbol{u}||_0^2 = &\frac{1}{2}\, \{||\boldsymbol{\omega} - \mathbf{curl}\, \boldsymbol{u} + \alpha\, \mathbf{curl}\, \boldsymbol{u}||_0^2 + ||\boldsymbol{\omega} - \mathbf{curl}\, \boldsymbol{u} - \alpha\, \boldsymbol{\omega}||_0^2\} \\ &+ \alpha \left(1 - \frac{\alpha}{2}\right) \{||\mathbf{curl}\, \boldsymbol{u}||_0^2 + ||\boldsymbol{\omega}||_0^2\} - 2\, \alpha\, (\mathbf{curl}\, \boldsymbol{\omega}, \boldsymbol{u}). \end{aligned} \tag{3.9}$$

In the light of $(\bigtriangledown p, \boldsymbol{u}) = -(\operatorname{div} \boldsymbol{u}, p)$, we have

$$\begin{aligned} ||\mathbf{curl}\, \boldsymbol{\omega} + \bigtriangledown p||_0^2 = &||\mathbf{curl}\, \boldsymbol{\omega} + \bigtriangledown p - \alpha\, \boldsymbol{u}||_0^2 \\ &+ 2\, \alpha\, (\mathbf{curl}\, \boldsymbol{\omega}, \boldsymbol{u}) - 2\, \alpha\, (\operatorname{div} \boldsymbol{u}, p) - \alpha^2\, ||\boldsymbol{u}||_0^2, \end{aligned} \tag{3.10}$$

where for some constant $\varepsilon_1 > 0$ we have

$$-2\, \alpha\, (\operatorname{div} \boldsymbol{u}, p) \geq -\varepsilon_1\, ||p||_0^2 - \frac{C_1\, \alpha^2}{\varepsilon_1}\, ||\operatorname{div} \boldsymbol{u}||_0^2. \tag{3.11}$$

Owing to the inf-sup condition [13]

$$\sup_{\boldsymbol{v}\in(H_0^1(\Omega))^3}\frac{(\operatorname{div}\boldsymbol{v},q)}{||\boldsymbol{v}||_1}\geq C\,||q||_0\quad\forall q\in L_0^2(\Omega),\tag{3.12}$$

we can find $\boldsymbol{v}^*\in(H_0^1(\Omega))^3$ such that

$$(\operatorname{div}\boldsymbol{v}^*,p)=||p||_0^2,\quad ||\boldsymbol{v}^*||_1\leq C\,||p||_0.\tag{3.13}$$

Let $\gamma>0$ be a constant to be determined; in the light of $(\mathbf{curl}\,\boldsymbol{\omega},\boldsymbol{v}^*)=(\boldsymbol{\omega},\mathbf{curl}\,\boldsymbol{v}^*)$ and $(\operatorname{div}\boldsymbol{v}^*,p)=||p||_0^2$, we have

$$\begin{aligned}||\mathbf{curl}\,\boldsymbol{\omega}+\bigtriangledown p||_0^2=&||\mathbf{curl}\,\boldsymbol{\omega}+\bigtriangledown p+\gamma\,\boldsymbol{v}^*||_0^2\\&-2\,\gamma\,(\boldsymbol{\omega},\mathbf{curl}\,\boldsymbol{v}^*)+2\,\gamma\,||p||_0^2-\gamma^2\,||\boldsymbol{v}^*||_0^2,\end{aligned}\tag{3.14}$$

where for some constant $\varepsilon_2>0$, in the light of $||\boldsymbol{v}^*||_1\leq C\,||p||_0$, we have

$$-2\,\gamma\,(\boldsymbol{\omega},\mathbf{curl}\,\boldsymbol{v}^*)\geq-\varepsilon_2\,||\boldsymbol{\omega}||_0^2-\frac{C_2\,\gamma^2}{\varepsilon_2}\,||p||_0^2.\tag{3.15}$$

In the light of $||\boldsymbol{v}^*||_0^2\leq C\,||\boldsymbol{v}^*||_1^2\leq C_3\,||p||_0^3$, we then get

$$||\mathbf{curl}\,\boldsymbol{\omega}+\bigtriangledown p||_0^2\geq\gamma\,\left[2-\gamma\,\left(C_3+\frac{C_2}{\varepsilon_2}\right)\right]\,||p||_0^2-\varepsilon_2\,||\boldsymbol{\omega}||_0^2.\tag{3.16}$$

From (3.10), (3.11), and (3.16), we have

$$\begin{aligned}2\,||\mathbf{curl}\,\boldsymbol{\omega}+\bigtriangledown p||_0^2\geq&\left\{\gamma\,\left[2-\gamma\,\left(C_3+\frac{C_2}{\varepsilon_2}\right)\right]-\varepsilon_1\right\}\,||p||_0^2\\&-\varepsilon_2\,||\boldsymbol{\omega}||_0^2-\alpha^2\,||\boldsymbol{u}||_0^2\\&+2\,\alpha\,(\mathbf{curl}\,\boldsymbol{\omega},\boldsymbol{u})-\frac{C_1\,\alpha^2}{\varepsilon_1}\,||\operatorname{div}\boldsymbol{u}||_0^2,\end{aligned}\tag{3.17}$$

where, in the light of $||\boldsymbol{u}||_0\leq C\,\{||\mathbf{curl}\,\boldsymbol{u}||_0+||\operatorname{div}\boldsymbol{u}||_0\}$, we have

$$-\alpha^2\,||\boldsymbol{u}||_0^2\geq-C_4\,\alpha^2\,||\mathbf{curl}\,\boldsymbol{u}||_0^2-C_4\,\alpha^2\,||\operatorname{div}\boldsymbol{u}||_0^2.\tag{3.18}$$

Summing (3.9) and (3.17) and in the light of (3.18), we have

$$\begin{aligned}&||\boldsymbol{\omega}-\mathbf{curl}\,\boldsymbol{u}||_0^2+2\,||\mathbf{curl}\,\boldsymbol{\omega}+\bigtriangledown p||_0^2\\&\geq\alpha\,\left[1-\alpha\,\left(\frac{1}{2}+C_4\right)\right]\,||\mathbf{curl}\,\boldsymbol{u}||_0^2+\left[\alpha\,\left(1-\frac{\alpha}{2}\right)-\varepsilon_2\right]\,||\boldsymbol{\omega}||_0^2\\&\quad+\left\{\gamma\,\left[2-\gamma\,\left(C_3+\frac{C_2}{\varepsilon_2}\right)\right]-\varepsilon_1\right\}\,||p||_0^2\\&\quad-\alpha^2\,\left(\frac{C_1}{\varepsilon_1}+C_4\right)\,||\operatorname{div}\boldsymbol{u}||_0^2.\end{aligned}\tag{3.19}$$

Therefore, let $\beta > 0$ be a constant to be determined, and we have

$$\begin{aligned} &||\boldsymbol{\omega} - \mathbf{curl}\, \boldsymbol{u}||_0^2 + 2\, ||\mathbf{curl}\, \boldsymbol{\omega} + \bigtriangledown p||_0^2 + \beta\, ||\mathrm{div}\, \boldsymbol{u}||_0^2 \\ &\quad \geq \alpha \left[1 - \alpha \left(\frac{1}{2} + C_4\right)\right] ||\mathbf{curl}\, \boldsymbol{u}||_0^2 + \left[\alpha \left(1 - \frac{\alpha}{2}\right) - \varepsilon_2\right] ||\boldsymbol{\omega}||_0^2 \\ &\qquad + \left\{\gamma \left[2 - \gamma \left(C_3 + \frac{C_2}{\varepsilon_2}\right)\right] - \varepsilon_1\right\} ||p||_0^2 \\ &\qquad + \left[\beta - \alpha^2 \left(\frac{C_1}{\varepsilon_1} + C_4\right)\right] ||\mathrm{div}\, \boldsymbol{u}||_0^2. \end{aligned} \tag{3.20}$$

Taking

$$0 < \alpha < \frac{2}{1 + 2\, C_4}, \quad 0 < \varepsilon_2 < \alpha \left(1 - \frac{\alpha}{2}\right), \tag{3.21}$$

$$0 < \gamma < \frac{2\, \varepsilon_2}{\varepsilon_2\, C_3 + C_2}, \quad 0 < \varepsilon_1 < \gamma \left[2 - \gamma \left(C_3 + \frac{C_2}{\varepsilon_2}\right)\right], \tag{3.22}$$

$$\beta > \alpha^2 \left(\frac{C_1}{\varepsilon_1} + C_4\right), \tag{3.23}$$

we have

$$\begin{aligned} &||\boldsymbol{\omega} - \mathbf{curl}\, \boldsymbol{u}||_0^2 + 2\, ||\mathbf{curl}\, \boldsymbol{\omega} + \bigtriangledown p||_0^2 + \beta\, ||\mathrm{div}\, \boldsymbol{u}||_0^2 \\ &\geq C\, \{||\mathbf{curl}\, \boldsymbol{u}||_0^2 + ||\mathrm{div}\, \boldsymbol{u}||_0^2 + ||p||_0^2 + ||\boldsymbol{\omega}||_0^2\}. \end{aligned} \tag{3.24}$$

Finally, we have

$$\begin{aligned} \mathcal{J}(\boldsymbol{u}, p, \boldsymbol{\omega}) &\geq \frac{1}{\max(2, \beta)} \{||\boldsymbol{\omega} - \mathbf{curl}\, \boldsymbol{u}||_0^2 + 2\, ||\mathbf{curl}\, \boldsymbol{\omega} + \bigtriangledown p||_0^2 + \beta\, ||\mathrm{div}\, \boldsymbol{u}||_0^2\} \\ &\geq C\, \{||\mathbf{curl}\, \boldsymbol{u}||_0^2 + ||\mathrm{div}\, \boldsymbol{u}||_0^2 + ||p||_0^2 + ||\boldsymbol{\omega}||_0^2\}. \end{aligned} \tag{3.25}$$

We further have

$$\mathcal{J}(\boldsymbol{u}, p, \boldsymbol{\omega}) \geq C\, \{||\boldsymbol{u}||_1^2 + ||p||_0^2 + ||\boldsymbol{\omega}||_0^2\} \tag{3.26}$$

because of $||\boldsymbol{u}||_1 \leq C\, \{||\mathbf{curl}\, \boldsymbol{u}||_0 + ||\mathrm{div}\, \boldsymbol{u}||_0\}$. □

COROLLARY 3.1. *Under the same hypotheses as in Theorem* 3.1*, we have*

$$\mathcal{J}^+(\boldsymbol{u}, p, \boldsymbol{\omega}) \geq C\, \{||\boldsymbol{u}||_1^2 + ||p||_0^2 + ||\boldsymbol{\omega}||_{\mathrm{div}}^2\} \tag{3.27}$$

for all $(\boldsymbol{u}, p, \boldsymbol{\omega}) \in (H_0^1(\Omega))^3 \times (H^1(\Omega) \cap L_0^2(\Omega)) \times (H(\mathbf{curl}; \Omega) \cap H(\mathrm{div}; \Omega))$.

Now, we describe the finite element method.

Let $\mathcal{C}_h$ be the regular triangulation of Ω into tetrahedra (cf. [15]). Define

$$V_h = \{v \in H^1(\Omega); v_{|_K} \in \mathcal{P}_1(K)\ \forall K \in \mathcal{C}_h\}, \tag{3.28}$$

where $\mathcal{P}_1(K)$ is the space of linear polynomials. Let $\tilde{v} \in V_h$ be the standard interpolant to $v \in H^2(\Omega)$, and from the standard interpolation theory in [15], we have

$$||v - \tilde{v}||_0 + h\, ||v - \tilde{v}||_1 \leq C\, h^2\, ||v||_2. \tag{3.29}$$

Define

$$U_h = (V_h \cap H_0^1(\Omega))^3, \quad Q_h = V_h \cap L_0^2(\Omega), \quad W_h = (V_h)^3. \tag{3.30}$$

The finite element method is to find $(\boldsymbol{u}_h, p_h, \boldsymbol{\omega}_h) \in U_h \times Q_h \times W_h$ such that

$$\begin{aligned} &(\mathbf{curl}\,\boldsymbol{\omega}_h + \bigtriangledown p_h, \mathbf{curl}\,\boldsymbol{z} + \bigtriangledown q) \\ &\quad + (\boldsymbol{\omega}_h - \mathbf{curl}\,\boldsymbol{u}_h, \boldsymbol{z} - \mathbf{curl}\,\boldsymbol{v}) + (\operatorname{div}\boldsymbol{u}_h, \operatorname{div}\boldsymbol{v}) \\ &\quad = (\boldsymbol{f}, \mathbf{curl}\,\boldsymbol{z} + \bigtriangledown q) \quad \forall(\boldsymbol{v}, q, \boldsymbol{z}) \in U_h \times Q_h \times W_h. \end{aligned} \tag{3.31}$$

THEOREM 3.2. *Under the same conditions as in Theorem* 3.1*, let* $(\boldsymbol{u}, p, \boldsymbol{\omega})$ *and* $(\boldsymbol{u}_h, p_h, \boldsymbol{\omega}_h)$ *be the solutions of* (3.3) *and* (3.31)*, respectively. If* $(\boldsymbol{u}, p, \boldsymbol{\omega}) \in (H^2(\Omega))^3 \times H^2(\Omega) \times (H^2(\Omega))^3$*, then*

$$||\boldsymbol{u} - \boldsymbol{u}_h||_1 + ||p - p_h||_0 + ||\boldsymbol{\omega} - \boldsymbol{\omega}_h||_0 \le C\,h\,\{||\boldsymbol{u}||_2 + ||p||_2 + ||\boldsymbol{\omega}||_2\}. \tag{3.32}$$

Proof. Let $\mathcal{A}((\boldsymbol{u}, p, \boldsymbol{\omega}); (\boldsymbol{v}, q, \boldsymbol{z}))$ be the bilinear form on $(H_0^1(\Omega))^3 \times (H^1(\Omega) \cap L_0^2(\Omega)) \times H(\mathbf{curl}; \Omega)$, defined by

$$\begin{aligned} \mathcal{A}((\boldsymbol{u}, p, \boldsymbol{\omega}); (\boldsymbol{v}, q, \boldsymbol{z})) &= (\mathbf{curl}\,\boldsymbol{\omega} + \bigtriangledown p, \mathbf{curl}\,\boldsymbol{z} + \bigtriangledown q) \\ &\quad + (\boldsymbol{\omega} - \mathbf{curl}\,\boldsymbol{u}, \boldsymbol{z} - \mathbf{curl}\,\boldsymbol{v}) + (\operatorname{div}\boldsymbol{u}, \operatorname{div}\boldsymbol{v}). \end{aligned} \tag{3.33}$$

We have the error orthogonality

$$\mathcal{A}((\boldsymbol{u} - \boldsymbol{u}_h, p - p_h, \boldsymbol{\omega} - \boldsymbol{\omega}_h); (\boldsymbol{v}, q, \boldsymbol{z})) = 0 \quad \forall(\boldsymbol{v}, q, \boldsymbol{z}) \in U_h \times Q_h \times W_h, \tag{3.34}$$

where $(\boldsymbol{u}, p, \boldsymbol{\omega})$ and $(\boldsymbol{u}_h, p_h, \boldsymbol{\omega}_h)$ are the solutions of (3.3) and (3.31), respectively.

Moreover, let $(\tilde{\boldsymbol{u}}, \tilde{p}, \tilde{\boldsymbol{\omega}}) \in U_h \times Q_h \times W_h$ be the standard interpolants to $(\boldsymbol{u}, p, \boldsymbol{\omega}) \in (H_0^1(\Omega) \cap H^2(\Omega))^3 \times (H^2(\Omega) \cap L_0^2(\Omega)) \times (H^2(\Omega))^3$, respectively, and the interpolation error estimations similar to (3.29) hold.

In the light of (3.34) and the Schwarz inequality, we have

$$\begin{aligned} &\mathcal{A}((\boldsymbol{u} - \boldsymbol{u}_h, p - p_h, \boldsymbol{\omega} - \boldsymbol{\omega}_h); (\boldsymbol{u} - \boldsymbol{u}_h, p - p_h, \boldsymbol{\omega} - \boldsymbol{\omega}_h)) \\ &\quad = \mathcal{A}((\boldsymbol{u} - \boldsymbol{u}_h, p - p_h, \boldsymbol{\omega} - \boldsymbol{\omega}_h); (\boldsymbol{u} - \tilde{\boldsymbol{u}}, p - \tilde{p}, \boldsymbol{\omega} - \tilde{\boldsymbol{\omega}})) \\ &\quad \le \mathcal{A}((\boldsymbol{u} - \boldsymbol{u}_h, p - p_h, \boldsymbol{\omega} - \boldsymbol{\omega}_h); (\boldsymbol{u} - \boldsymbol{u}_h, p - p_h, \boldsymbol{\omega} - \boldsymbol{\omega}_h))^{1/2} \\ &\qquad \times \mathcal{A}((\boldsymbol{u} - \tilde{\boldsymbol{u}}, p - \tilde{p}, \boldsymbol{\omega} - \tilde{\boldsymbol{\omega}}); (\boldsymbol{u} - \tilde{\boldsymbol{u}}, p - \tilde{p}, \boldsymbol{\omega} - \tilde{\boldsymbol{\omega}}))^{1/2}. \end{aligned}$$

We then have

$$\begin{aligned} &\mathcal{A}((\boldsymbol{u} - \boldsymbol{u}_h, p - p_h, \boldsymbol{\omega} - \boldsymbol{\omega}_h); (\boldsymbol{u} - \boldsymbol{u}_h, p - p_h, \boldsymbol{\omega} - \boldsymbol{\omega}_h))^{1/2} \\ &\quad \le \mathcal{A}((\boldsymbol{u} - \tilde{\boldsymbol{u}}, p - \tilde{p}, \boldsymbol{\omega} - \tilde{\boldsymbol{\omega}}); (\boldsymbol{u} - \tilde{\boldsymbol{u}}, p - \tilde{p}, \boldsymbol{\omega} - \tilde{\boldsymbol{\omega}}))^{1/2}. \end{aligned} \tag{3.35}$$

Therefore, we get

$$\begin{aligned} &||\boldsymbol{u} - \boldsymbol{u}_h||_1 + ||p - p_h||_0 + ||\boldsymbol{\omega} - \boldsymbol{\omega}_h||_0 \\ &\quad \le C\,\mathcal{A}((\boldsymbol{u} - \boldsymbol{u}_h, p - p_h, \boldsymbol{\omega} - \boldsymbol{\omega}_h); (\boldsymbol{u} - \boldsymbol{u}_h, p - p_h, \boldsymbol{\omega} - \boldsymbol{\omega}_h))^{1/2} \\ &\quad \le C\,\mathcal{A}((\boldsymbol{u} - \tilde{\boldsymbol{u}}, p - \tilde{p}, \boldsymbol{\omega} - \tilde{\boldsymbol{\omega}}); (\boldsymbol{u} - \tilde{\boldsymbol{u}}, p - \tilde{p}, \boldsymbol{\omega} - \tilde{\boldsymbol{\omega}}))^{1/2} \\ &\quad \le C\,\{||\boldsymbol{\omega} - \tilde{\boldsymbol{\omega}}||_{0,\mathbf{curl}} + ||p - \tilde{p}||_1 + ||\boldsymbol{u} - \tilde{\boldsymbol{u}}||_1\} \\ &\quad \le C\,h\,\{||\boldsymbol{u}||_2 + ||p||_2 + ||\boldsymbol{\omega}||_2\}. \qquad \square \end{aligned} \tag{3.36}$$

Remark 3.2. It is obvious that one can use tangential continuous elements such as Nédélec edge elements (see [23], [24]) to approximate the vorticity, and the error bound becomes

$$||\boldsymbol{u} - \boldsymbol{u}_h||_1 + ||p - p_h||_0 + ||\boldsymbol{\omega} - \boldsymbol{\omega}_h||_0 \le C\,h\,\{||\boldsymbol{u}||_2 + ||p||_2 + ||\boldsymbol{f}||_1\} \tag{3.37}$$

since, in this case, from [19], [25] we have $||\tilde{\boldsymbol{\omega}} - \boldsymbol{\omega}||_{0,\mathbf{curl}} \leq C\, h\, \{||\boldsymbol{\omega}||_1 + ||\mathbf{curl}\, \boldsymbol{\omega}||_1\}$; here $\tilde{\boldsymbol{\omega}}$ is the interpolant in the Nédélec space.

Remark 3.3. Theorem 3.1 holds for the Navier–Stokes equations as follows:

$$-\nu\, \Delta\, \boldsymbol{u} + \boldsymbol{a} \cdot \bigtriangledown \boldsymbol{u} + \sigma\, \boldsymbol{u} + \bigtriangledown p = \boldsymbol{f}, \quad \operatorname{div} \boldsymbol{u} = 0, \quad \text{in } \Omega, \quad \boldsymbol{u}_{|\Gamma} = \mathbf{0}, \tag{3.38}$$

where $\boldsymbol{a}$ is the convection term satisfying $\operatorname{div} \boldsymbol{a} = 0$, and $\nu > 0$ is the viscosity and $\sigma \geq 0$ is a constant. Specifically, applying the same argument as in Theorem 3.1, we have

$$\begin{aligned} &||\boldsymbol{\omega} - \mathbf{curl}\, \boldsymbol{u}||_0^2 + ||\nu\, \mathbf{curl}\, \boldsymbol{\omega} + \boldsymbol{a} \cdot \bigtriangledown \boldsymbol{u} + \sigma\, \boldsymbol{u} + \bigtriangledown p||_0^2 + ||\operatorname{div} \boldsymbol{u}||_0^2 \\ &\quad \geq C\, \Bigg\{ \nu^2\, ||\mathbf{curl}\, \boldsymbol{u}||_0^2 + \nu^2\, ||\boldsymbol{\omega}||_0^2 \\ &\qquad + \frac{\nu^2}{\nu^2 + \sigma^2 + ||\boldsymbol{a}||_\infty^2}\, ||p||_0^2 + \frac{\nu^2 + \sigma^2 + ||\boldsymbol{a}||_\infty^2}{\nu^2}\, ||\operatorname{div} \boldsymbol{u}||_0^2 + \nu\, \sigma\, ||\boldsymbol{u}||_0^2 \Bigg\}, \end{aligned} \tag{3.39}$$

where $C > 0$ is independent of $\boldsymbol{a}, \sigma, \nu$.

Remark 3.4. The linear elasticity problem is as follows:

$$-\mu\, \Delta\, \boldsymbol{u} - (\lambda + \mu) \bigtriangledown \operatorname{div} \boldsymbol{u} = \boldsymbol{f}, \quad \text{in } \Omega, \quad \boldsymbol{u} = \mathbf{0}, \quad \text{on } \Gamma, \tag{3.40}$$

where λ and μ are Lamé coefficients. Introducing

$$\boldsymbol{\omega} = \mathbf{curl}\, \boldsymbol{u}, \quad p = (\lambda + 2\, \mu) \operatorname{div} \boldsymbol{u}, \tag{3.41}$$

we can rewrite (3.40) in the form

$$\mu\, \mathbf{curl}\, \boldsymbol{\omega} - \bigtriangledown p = \boldsymbol{f}, \quad \operatorname{div} \boldsymbol{u} - \frac{1}{\lambda + 2\, \mu} p = 0, \quad \boldsymbol{\omega} = \mathbf{curl}\, \boldsymbol{u}. \tag{3.42}$$

Then, we can obtain

$$\begin{aligned} &||\boldsymbol{\omega} - \mathbf{curl}\, \boldsymbol{u}||_0^2 + ||\mu\, \mathbf{curl}\, \boldsymbol{\omega} - \bigtriangledown p||_0^2 + \left|\left| \operatorname{div} \boldsymbol{u} - \frac{1}{\lambda + 2\, \mu}\, p \right|\right|_0^2 \\ &\quad \geq C\, \{\mu^2\, ||\mathbf{curl}\, \boldsymbol{u}||_0^2 + \mu^2\, ||\boldsymbol{\omega}||_0^2 + ||p||_0^2 + ||\operatorname{div} \boldsymbol{u}||_0^2\}, \end{aligned} \tag{3.43}$$

where $C > 0$ is independent of λ and μ.

Remark 3.5. So far, we have shown that the divergence of the vorticity is not essential and does not affect the coerciveness. This corrects the traditional viewpoint (cf. [1], [2], [3], [4], [5]).

4. On the L^2 error bound for velocity. In this section, with the classical Aubin–Nitsche duality argument, we show that the L^2 error bound for the velocity is in fact $\mathcal{O}(h^{3/2})$.

Let us consider the auxiliary variational problem: to find $(\boldsymbol{u}^*, p^*, \boldsymbol{\omega}^*) \in (H_0^1(\Omega))^3 \times (H^1(\Omega) \cap L_0^2(\Omega)) \times H(\mathbf{curl}; \Omega)$ such that

$$\mathcal{A}((\boldsymbol{u}^*, p^*, \boldsymbol{\omega}^*); (\boldsymbol{v}, q, \boldsymbol{z})) = (\boldsymbol{\chi}, \boldsymbol{v}) \tag{4.1}$$

for all $(\boldsymbol{v}, q, \boldsymbol{z}) \in (H_0^1(\Omega))^3 \times (H^1(\Omega) \cap L_0^2(\Omega)) \times H(\mathbf{curl}; \Omega)$, where $\boldsymbol{\chi} \in (H^1(\Omega))^3$.

THEOREM 4.1. *Under the same conditions as in Proposition* 2.3, *problem* (4.1) *has a unique solution* $(\boldsymbol{u}^*, p^*, \boldsymbol{\omega}^*)$, *which satisfies*

$$||\boldsymbol{u}^*||_2 \leq C\, ||\boldsymbol{\chi}||_0, \quad ||p^*||_2 + ||\boldsymbol{\omega}^*||_2 \leq C\, ||\boldsymbol{\chi}||_1. \tag{4.2}$$

Proof. The proof is divided into four steps, where the last step is used for verification.

Step 1. We consider the following problem: to find $\boldsymbol{v}_0 \in (H_0^1(\Omega))^3$ and $q_0 \in H^1(\Omega) \cap L_0^2(\Omega)$ such that

$$-\Delta\,\boldsymbol{v}_0 + \bigtriangledown q_0 = \boldsymbol{\chi}, \quad \mathrm{div}\boldsymbol{v}_0 = 0, \quad \boldsymbol{v}_{0|_\Gamma} = \mathbf{0}, \tag{4.3}$$

the solution of which satisfies

$$||\boldsymbol{v}_0||_2 + ||q_0||_1 \le C\,||\boldsymbol{\chi}||_0, \quad ||\boldsymbol{v}_0||_3 + ||q_0||_2 \le C\,||\boldsymbol{\chi}||_1. \tag{4.4}$$

Moreover, $q_0 \in L_0^2(\Omega)$ satisfies

$$(q_0, \mathrm{div}\,\boldsymbol{v}) = (\mathbf{curl}\,\mathbf{curl}\,\boldsymbol{v}_0 - \boldsymbol{\chi}, \boldsymbol{v}) \quad \forall \boldsymbol{v} \in (H_0^1(\Omega))^3. \tag{4.5}$$

Step 2. We consider the following problem: to find $\boldsymbol{u}_0 \in (H_0^1(\Omega))^3$ and $p_0 \in H^1(\Omega) \cap L_0^2(\Omega)$ such that

$$-\Delta\,\boldsymbol{u}_0 + \bigtriangledown p_0 = \boldsymbol{v}_0, \quad \mathrm{div}\,\boldsymbol{u}_0 = -q_0, \quad \boldsymbol{u}_{0|_\Gamma} = \mathbf{0}, \tag{4.6}$$

the solution of which satisfies

$$||\boldsymbol{u}_0||_2 + ||p_0||_1 \le C\,||\boldsymbol{\chi}||_0, \quad ||\boldsymbol{u}_0||_3 + ||p_0||_2 \le C\,||\boldsymbol{\chi}||_1. \tag{4.7}$$

Step 3. Define

$$\boldsymbol{u}^* = \boldsymbol{u}_0 + \boldsymbol{v}_0, \quad p^* = p_0 + q_0, \quad \boldsymbol{\omega}^* = \mathbf{curl}\,\boldsymbol{u}_0. \tag{4.8}$$

We have

$$||\boldsymbol{u}^*||_2 + ||p^*||_1 + ||\boldsymbol{\omega}^*||_1 \le C\,||\boldsymbol{\chi}||_0, \tag{4.9}$$
$$||\boldsymbol{u}^*||_3 + ||p^*||_2 + ||\boldsymbol{\omega}^*||_2 \le C\,||\boldsymbol{\chi}||_1. \tag{4.10}$$

Step 4. Note that

$$\mathbf{curl}\,\boldsymbol{\omega}^* + \bigtriangledown p^* = \boldsymbol{v}_0, \tag{4.11}$$
$$\boldsymbol{\omega}^* - \mathbf{curl}\,\boldsymbol{u}^* = -\mathbf{curl}\,\boldsymbol{v}_0, \quad \mathrm{div}\,\boldsymbol{u}^* = -q_0. \tag{4.12}$$

We can easily verify that $(\boldsymbol{u}^*, p^*, \boldsymbol{\omega}^*)$ satisfies (4.1). □

COROLLARY 4.1. *Let* $\boldsymbol{u}$ *and* $\boldsymbol{u}_h$ *be the exact and the approximate solutions. Under the same hypotheses as in Theorem* 4.1, *if* $(\boldsymbol{u}, p, \boldsymbol{\omega}) \in (H^2(\Omega))^3 \times H^2(\Omega) \times (H^2(\Omega))^3$, *we have*

$$||\boldsymbol{u} - \boldsymbol{u}_h||_0 \le C\,h^{3/2}\,\{||\boldsymbol{u}||_2 + ||p||_2 + ||\boldsymbol{\omega}||_2\}. \tag{4.13}$$

Proof. In (4.1), take $\boldsymbol{\chi} = \boldsymbol{u} - \boldsymbol{u}_h$.

Let $(\boldsymbol{u}, p, \boldsymbol{\omega}), (\boldsymbol{u}^*, p^*, \boldsymbol{\omega}^*)$ be the solutions to (3.3) and (4.1), respectively, and $(\tilde{\boldsymbol{u}}, \tilde{p}, \tilde{\boldsymbol{\omega}}), (\widetilde{\boldsymbol{u}^*}, \widetilde{p^*}, \widetilde{\boldsymbol{\omega}^*})$ are their corresponding interpolants in $U_h \times Q_h \times W_h$.

From the standard Aubin–Nitsche duality argument, we have

$$\begin{aligned} ||\boldsymbol{u} - \boldsymbol{u}_h||_0^2 \le\, & \mathcal{A}((\boldsymbol{u} - \tilde{\boldsymbol{u}}, p - \tilde{p}, \boldsymbol{\omega} - \tilde{\boldsymbol{\omega}}); (\boldsymbol{u} - \tilde{\boldsymbol{u}}, p - \tilde{p}, \boldsymbol{\omega} - \tilde{\boldsymbol{\omega}}))^{1/2} \\ & \times \mathcal{A}((\boldsymbol{u}^* - \widetilde{\boldsymbol{u}^*}, p^* - \widetilde{p^*}, \boldsymbol{\omega}^* - \widetilde{\boldsymbol{\omega}^*}); (\boldsymbol{u}^* - \widetilde{\boldsymbol{u}^*}, p^* - \widetilde{p^*}, \boldsymbol{\omega}^* - \widetilde{\boldsymbol{\omega}^*}))^{1/2}. \end{aligned} \tag{4.14}$$

In the light of the standard interpolation properties and (4.2) and (3.32), we conclude that (4.13) holds. □

Remark 4.1. Similarly, we can obtain

$$||\boldsymbol{u} - \boldsymbol{u}_h||_{-1} \leq C\, h^2 \,\{||\boldsymbol{u}||_2 + ||p||_2 + ||\boldsymbol{\omega}||_2\}. \tag{4.15}$$

Remark 4.2. In addition, we may investigate the L^2 error bounds for both pressure and vorticity, but, in general, there do not yield improved error estimates, even if the divergence and the homogeneous normal boundary condition of the vorticity are added. Nonetheless, for a pathological domain (cf. [28]) for which there holds $H_0(\mathbf{curl};\Omega) \cap (H^1(\Omega))^3 \equiv (H_0^1(\Omega))^3$ (this does not hold generally; cf. [29]), we can show that the L^2 error bounds for both velocity and pressure are $\mathcal{O}(h^2)$ in the standard linear element method without the divergence of the vorticity. If introducing both the divergence and the homogeneous normal boundary condition of the vorticity, we can further show that the L^2 error bound for this variable is also $\mathcal{O}(h^2)$; see [32] for details.

5. On the Bochev–Gunzburger method. In this section, following an argument similar to that of the previous section, we show that the error bounds for all variables are still optimal in the Bochev–Gunzburger method, even without including the divergence of the vorticity in the weak problem.

Let us first recall the Bochev–Gunzburger method: to find $(\boldsymbol{u}_h, p_h, \boldsymbol{\omega}_h) \in V_h \times Q_h \times W_h$ such that

$$\mathcal{A}_h((\boldsymbol{u}_h, p_h, \boldsymbol{\omega}_h); (\boldsymbol{v}, q, \boldsymbol{z})) = (\boldsymbol{f}, \mathbf{curl}\, \boldsymbol{z} + \bigtriangledown q) \quad \forall (\boldsymbol{v}, q, \boldsymbol{z}) \in V_h \times Q_h \times W_h, \tag{5.1}$$

where Q_h, W_h are still defined as in (3.30), but

$$V_h = \{\boldsymbol{v} \in (H_0^1(\Omega))^3; \boldsymbol{v}_{|_K} \in (\mathcal{P}_2(K))^3, K \in \mathcal{C}_h\} \tag{5.2}$$

with $\mathcal{P}_2(K)$ the space of quadratic polynomials, and

$$\begin{aligned}\mathcal{A}_h((\boldsymbol{u}, p, \boldsymbol{\omega}); (\boldsymbol{v}, q, \boldsymbol{z})) &= (\mathbf{curl}\, \boldsymbol{\omega} + \bigtriangledown p, \mathbf{curl}\, \boldsymbol{z} + \bigtriangledown q)\\ &\quad + h^{-2}\,(\boldsymbol{\omega} - \mathbf{curl}\, \boldsymbol{u}, \boldsymbol{z} - \mathbf{curl}\, \boldsymbol{v}) + h^{-2}\,(\mathrm{div}\, \boldsymbol{u}, \mathrm{div}\, \boldsymbol{v}).\end{aligned} \tag{5.3}$$

Remark 5.1. The original Bochev–Gunzburger method includes the divergence of the vorticity. As is proved in section 3, the introduction of the divergence of the vorticity is unnecessary for the coerciveness. In what follows, we will show that it does not affect the accuracy, either.

THEOREM 5.1. *Let $(\boldsymbol{u}, p, \boldsymbol{\omega})$ and $(\boldsymbol{u}_h, p_h, \boldsymbol{\omega}_h)$ be the exact solution to* (3.3) *and the finite element solution to* (5.1)*, respectively. Under the same conditions as in Theorem* 3.1*, if $(\boldsymbol{u}, p, \boldsymbol{\omega}) \in (H^3(\Omega))^3 \times H^2(\Omega) \times (H^2(\Omega))^3$, then*

$$||\boldsymbol{u} - \boldsymbol{u}_h||_1 + ||p - p_h||_0 + ||\boldsymbol{\omega} - \boldsymbol{\omega}_h||_0 \leq C\, h\, \{||\boldsymbol{u}||_3 + ||p||_2 + ||\boldsymbol{\omega}||_2\}, \tag{5.4}$$

$$||\boldsymbol{\omega} - \boldsymbol{\omega}_h - \mathbf{curl}\,(\boldsymbol{u} - \boldsymbol{u}_h)||_0 + ||\mathrm{div}\,(\boldsymbol{u} - \boldsymbol{u}_h)||_0 \leq C\, h^2\, \{||\boldsymbol{u}||_3 + ||p||_2 + ||\boldsymbol{\omega}||_2\}. \tag{5.5}$$

Proof. It is obvious that

$$\begin{aligned}&\mathcal{A}_h((\boldsymbol{v}, q, \boldsymbol{z}); (\boldsymbol{v}, q, \boldsymbol{z}))\\ &\quad \geq C\, \{\mathcal{J}(\boldsymbol{v}, q, \boldsymbol{z}) + h^{-2}\,(||\boldsymbol{z} - \mathbf{curl}\, \boldsymbol{v}||_0^2 + ||\mathrm{div}\, \boldsymbol{v}||_0^2)\}\\ &\quad \geq C\, \{||\boldsymbol{v}||_1^2 + ||\boldsymbol{z}||_0^2 + ||q||_0^2 + h^{-2}\,(||\boldsymbol{z} - \mathbf{curl}\, \boldsymbol{v}||_0^2 + ||\mathrm{div}\, \boldsymbol{v}||_0^2)\},\end{aligned} \tag{5.6}$$

where we have used Theorem 3.1. Applying the standard interpolation estimation, from (5.6) we can easily obtain (5.4) and (5.5). □

Now we turn to the L^2 error bound for the vorticity.

Let us consider the auxiliary variational problem: to find $(\boldsymbol{u}^*, p^*, \boldsymbol{\omega}^*) \in (H_0^1(\Omega))^3 \times (H^1(\Omega) \cap L_0^2(\Omega)) \times H(\mathbf{curl}; \Omega)$ such that

$$\mathcal{A}_h((\boldsymbol{u}^*, p^*, \boldsymbol{\omega}^*); (\boldsymbol{v}, q, \boldsymbol{z})) = (\boldsymbol{\chi}, \boldsymbol{z}) \tag{5.7}$$

holds for all $(\boldsymbol{v}, q, \boldsymbol{z}) \in (H_0^1(\Omega))^3 \times (H^1(\Omega) \cap L_0^2(\Omega)) \times H(\mathbf{curl}; \Omega)$, with $\boldsymbol{\chi} \in H(\mathbf{curl}; \Omega)$.

THEOREM 5.2. *Under the same conditions as in Proposition* 2.3*, problem* (5.7) *has a unique solution* $(\boldsymbol{u}^*, p^*, \boldsymbol{\omega}^*)$*, which satisfies*

$$||\boldsymbol{u}^*||_1 + ||\boldsymbol{\omega}^*||_0 + ||p^*||_0 \leq C\, ||\boldsymbol{\chi}||_0. \tag{5.8}$$

Moreover, there exists $(\boldsymbol{u}_0, p_0, \boldsymbol{\omega}_0) \in (H_0^1(\Omega))^3 \times (H^1(\Omega) \cap L_0^2(\Omega)) \times H(\mathbf{curl}; \Omega)$ *such that*

$$||\boldsymbol{u}_0||_3 + ||\boldsymbol{\omega}_0||_2 + ||p_0||_2 \leq C\, ||\boldsymbol{\chi}||_0, \tag{5.9}$$

$$\mathbf{curl}\,(\boldsymbol{\omega}^* - \boldsymbol{\omega}_0) + \bigtriangledown (p^* - p_0) = \mathbf{0}, \tag{5.10}$$

$$||\boldsymbol{\omega}^* - \boldsymbol{\omega}_0 - \mathbf{curl}\,(\boldsymbol{u}^* - \boldsymbol{u}_0)||_0 \leq C\, h^2\, ||\boldsymbol{\chi}||_0, \tag{5.11}$$

$$||\mathrm{div}\,(\boldsymbol{u}^* - \boldsymbol{u}_0)||_0 \leq C\, h^2\, ||\boldsymbol{\chi}||_0. \tag{5.12}$$

Proof. The proof is divided into five steps, where the last step is used for verification.

Step 1. We consider the following problem: to find $\boldsymbol{v}_0 \in (H_0^1(\Omega))^3$ and $q_0 \in H^1(\Omega) \cap L_0^2(\Omega)$ such that

$$-\Delta\, \boldsymbol{v}_0 + \bigtriangledown q_0 = \mathbf{curl}\, \boldsymbol{\chi}, \quad \mathrm{div}\, \boldsymbol{v}_0 = 0, \quad \boldsymbol{v}_{0|_\Gamma} = \mathbf{0}, \tag{5.13}$$

the solution of which satisfies

$$||\boldsymbol{v}_0||_1 + ||q_0||_0 \leq C\, ||\boldsymbol{\chi}||_0, \quad ||\boldsymbol{v}_0||_2 + ||q_0||_1 \leq C\, ||\mathbf{curl}\, \boldsymbol{\chi}||_0. \tag{5.14}$$

Moreover, $q_0 \in L_0^2(\Omega)$ satisfies

$$(q_0, \mathrm{div}\, \boldsymbol{v}) = (\mathbf{curl}\, \boldsymbol{v}_0 - \boldsymbol{\chi}, \mathbf{curl}\, \boldsymbol{v}) \quad \forall \boldsymbol{v} \in (H_0^1(\Omega))^3. \tag{5.15}$$

Step 2. We consider the following problem: to find $\boldsymbol{u}_0 \in (H_0^1(\Omega))^3$ and $p_0 \in H^1(\Omega) \cap L_0^2(\Omega)$ such that

$$-\Delta\, \boldsymbol{u}_0 + \bigtriangledown p_0 = \boldsymbol{v}_0, \quad \mathrm{div}\, \boldsymbol{u}_0 = 0, \quad \boldsymbol{u}_{0|_\gamma} = \mathbf{0}, \tag{5.16}$$

the solution of which satisfies

$$||\boldsymbol{u}_0||_3 + ||p_0||_2 \leq C\, ||\boldsymbol{\chi}||_0. \tag{5.17}$$

Step 3. We consider the following problem: to find $\boldsymbol{u}^+ \in (H_0^1(\Omega))^3$ and $p^+ \in H^1(\Omega) \cap L_0^2(\Omega)$ such that

$$-\Delta\, \boldsymbol{u}^+ + \bigtriangledown p^+ = -h^2\, \mathbf{curl}\, \boldsymbol{\chi} + h^2\, \mathbf{curl}\, \mathbf{curl}\, \boldsymbol{v}_0, \quad \mathrm{div}\, \boldsymbol{u}^+ = h^2\, q_0, \quad {\boldsymbol{u}^+}_{|_\Gamma} = \mathbf{0}, \tag{5.18}$$

the solution of which satisfies

$$||\boldsymbol{u}^+||_1 + ||p^+||_0 \leq C\, h^2\, ||\boldsymbol{\chi}||_0, \quad ||\boldsymbol{u}^+||_2 + ||p^+||_1 \leq C\, h^2\, ||\mathbf{curl}\, \boldsymbol{\chi}||_0. \tag{5.19}$$

Step 4. Define

$$\boldsymbol{u}^* = \boldsymbol{u}_0 + \boldsymbol{u}^+, \quad p^* = p_0 + p^+ - h^2\, q_0, \quad \boldsymbol{\omega}^* = \mathbf{curl}\,\boldsymbol{u}^* + h^2\,(\boldsymbol{\chi} - \mathbf{curl}\,\boldsymbol{v}_0), \tag{5.20}$$

$$\boldsymbol{\omega}_0 = \mathbf{curl}\,\boldsymbol{u}_0. \tag{5.21}$$

We have

$$\mathbf{curl}\,(\boldsymbol{\omega}^* - \boldsymbol{\omega}_0) + \bigtriangledown(p^* - p_0) = \mathbf{0}, \tag{5.22}$$

$$||\boldsymbol{\omega}^* - \boldsymbol{\omega}_0 - \mathbf{curl}\,(\boldsymbol{u}^* - \boldsymbol{u}_0)||_0 \le C\,h^2\,||\boldsymbol{\chi}||_0, \tag{5.23}$$

$$||\mathrm{div}\,(\boldsymbol{u}^* - \boldsymbol{u}_0)||_0 \le C\,h^2\,||\boldsymbol{\chi}||_0. \tag{5.24}$$

Step 5. Note that

$$\mathbf{curl}\,\boldsymbol{\omega}^* + \bigtriangledown p^* = \boldsymbol{v}_0, \tag{5.25}$$

$$\boldsymbol{\omega}^* - \mathbf{curl}\,\boldsymbol{u}^* = h^2\,(\boldsymbol{\chi} - \mathbf{curl}\,\boldsymbol{v}_0), \quad \mathrm{div}\,\boldsymbol{u}^* = h^2\,q_0. \tag{5.26}$$

We can easily verify that $(\boldsymbol{u}^*, p^*, \boldsymbol{\omega}^*)$ satisfies (5.7). □

THEOREM 5.3. *Let $\boldsymbol{\omega}$ and $\boldsymbol{\omega}_h$ be the exact and the approximate solutions. Under the same hypotheses as in Theorem* 5.2 *and if $(\boldsymbol{u}, p, \boldsymbol{\omega}) \in (H^3(\Omega))^3 \times H^2(\Omega) \times (H^2(\Omega))^3$, then there holds*

$$||\boldsymbol{\omega} - \boldsymbol{\omega}_h||_0 + ||\boldsymbol{u} - \boldsymbol{u}_h||_1 \le C\,h^2\,\{||\boldsymbol{u}||_3 + ||p||_2 + ||\boldsymbol{\omega}||_2\}. \tag{5.27}$$

Proof. In (5.7), take $\boldsymbol{\chi} = \boldsymbol{\omega} - \boldsymbol{\omega}_h$.

Let $(\tilde{\boldsymbol{u}}, \tilde{p}, \tilde{\boldsymbol{\omega}})$ and $(\tilde{\boldsymbol{u}_0}, \tilde{p_0}, \tilde{\boldsymbol{\omega}_0})$ be the interpolants in $V_h \times Q_h \times W_h$ to $(\boldsymbol{u}, p, \boldsymbol{\omega})$ and $(\boldsymbol{u}_0, p_0, \boldsymbol{\omega}_0)$, respectively.

By the duality argument, we have

$$\begin{aligned} &||\boldsymbol{\omega} - \boldsymbol{\omega}_h||_0^2 \\ &\quad \le \mathcal{A}_h((\boldsymbol{u} - \tilde{\boldsymbol{u}}, p - \tilde{p}, \boldsymbol{\omega} - \tilde{\boldsymbol{\omega}}); (\boldsymbol{u} - \tilde{\boldsymbol{u}}, p - \tilde{p}, \boldsymbol{\omega} - \tilde{\boldsymbol{\omega}}))^{1/2} \\ &\qquad \times \mathcal{A}_h((\boldsymbol{u}^* - \tilde{\boldsymbol{u}_0}, p^* - \tilde{p_0}, \boldsymbol{\omega}^* - \tilde{\boldsymbol{\omega}_0}); (\boldsymbol{u}^* - \tilde{\boldsymbol{u}_0}, p^* - \tilde{p_0}, \boldsymbol{\omega}^* - \tilde{\boldsymbol{\omega}_0}))^{1/2}, \end{aligned} \tag{5.28}$$

where

$$\begin{aligned} &\mathcal{A}_h((\boldsymbol{u} - \tilde{\boldsymbol{u}}, p - \tilde{p}, \boldsymbol{\omega} - \tilde{\boldsymbol{\omega}}); (\boldsymbol{u} - \tilde{\boldsymbol{u}}, p - \tilde{p}, \boldsymbol{\omega} - \tilde{\boldsymbol{\omega}}))^{1/2} \\ &\qquad \le C\,h\,\{||\boldsymbol{u}||_3 + ||p||_2 + ||\boldsymbol{\omega}||_2\}, \end{aligned} \tag{5.29}$$

$$\begin{aligned} &\mathcal{A}_h((\boldsymbol{u}^* - \tilde{\boldsymbol{u}_0}, p^* - \tilde{p_0}, \boldsymbol{\omega}^* - \tilde{\boldsymbol{\omega}_0}); (\boldsymbol{u}^* - \tilde{\boldsymbol{u}_0}, p^* - \tilde{p_0}, \boldsymbol{\omega}^* - \tilde{\boldsymbol{\omega}_0}))^{1/2} \\ &\quad \le C\,\mathcal{A}_h((\boldsymbol{u}_0 - \tilde{\boldsymbol{u}_0}, p_0 - \tilde{p_0}, \boldsymbol{\omega}_0 - \tilde{\boldsymbol{\omega}_0}); (\boldsymbol{u}_0 - \tilde{\boldsymbol{u}_0}, p_0 - \tilde{p_0}, \boldsymbol{\omega}_0 - \tilde{\boldsymbol{\omega}_0}))^{1/2} \\ &\qquad + C\,h^{-1}\,(||\boldsymbol{\omega}^* - \boldsymbol{\omega}_0 - \mathbf{curl}\,(\boldsymbol{u}^* - \boldsymbol{u}_0)||_0 + ||\mathrm{div}\,(\boldsymbol{u}^* - \boldsymbol{u}_0)||_0) \\ &\quad \le C\,h\,\{||\boldsymbol{u}_0||_3 + ||p_0||_2 + ||\boldsymbol{\omega}_0||_2 + ||\boldsymbol{\omega} - \boldsymbol{\omega}_h||_0\} \\ &\quad \le C\,h\,||\boldsymbol{\omega} - \boldsymbol{\omega}_h||_0. \end{aligned} \tag{5.30}$$

Hence we get

$$||\boldsymbol{\omega} - \boldsymbol{\omega}_h||_0 \le C\,h^2\,\{||\boldsymbol{u}||_3 + ||p||_2 + ||\boldsymbol{\omega}||_2\}. \tag{5.31}$$

From (5.5) and (5.27), we then immediately know that

$$\begin{aligned} ||\boldsymbol{u} - \boldsymbol{u}_h||_1 &\le C\,\{||\boldsymbol{\omega} - \boldsymbol{\omega}_h - \mathbf{curl}\,(\boldsymbol{u} - \boldsymbol{u}_h)||_0 + ||\mathrm{div}\,(\boldsymbol{u} - \boldsymbol{u}_h)||_0 + ||\boldsymbol{\omega} - \boldsymbol{\omega}_h||_0\} \\ &\le C\,h^2\,\{||\boldsymbol{u}||_3 + ||p||_2 + ||\boldsymbol{\omega}||_2\}, \end{aligned} \tag{5.32}$$

where we have used the inequality $||\boldsymbol{v}||_1 \leq C\{||\mathbf{curl}\,\boldsymbol{v}||_0 + ||\mathrm{div}\,\boldsymbol{v}||_0\}$ on $(H_0^1(\Omega))^3$. □

Remark 5.2. Regarding the pressure, following a similar argument, we can establish the L^2 error bound $\mathcal{O}(h^2)$ for this variable.

Remark 5.3. When Ω is convex polyhedron, both methods (3.31) and (5.1) are coercive. When additionally the exact solution $(\boldsymbol{u}, p, \boldsymbol{\omega})$ is in $(H^3(\Omega))^3 \times H^2(\Omega) \times (H^2(\Omega))^3$, we obtain basic error bounds (3.32), (5.4), and (5.5).

If the solution of the Stokes problem (2.3) is in the same spaces as above (the right-hand side functions $\boldsymbol{\chi}$ and g may be required to be in $(H^1(\Omega))^3$ and $H^2(\Omega)$, respectively), then with the help of the Aubin–Nitsche duality argument we obtain the L^2 error bound (4.13) for the linear element method (3.31) and recover the optimal error bounds for the Bochev–Gunzburger method (5.1).

To our best knowledge, from [13, p. 88] we know that if the boundary Γ is in C^3 (the right-hand side functions need corresponding regularities), then it holds that $(\boldsymbol{u}, p, \boldsymbol{\omega})$ is in $(H^3(\Omega))^3 \times H^2(\Omega) \times (H^2(\Omega))^3$.

6. Numerical examples.

In this section we report the results of numerical examples to illustrate the theoretical error bounds for the least-squares linear element method (3.31).

We take the domain as $\Omega = [0,1]^3$ and consider a 3D Stokes problem

$$-\Delta\,\boldsymbol{u} + \bigtriangledown p = \boldsymbol{f}, \quad \mathrm{div}\,\boldsymbol{u} = 0, \quad \boldsymbol{u}_{|\Gamma} = \mathbf{0}, \tag{6.1}$$

the exact solution of which is known: let

$$\boldsymbol{\Phi} = \begin{pmatrix} x(1-x)y^2(1-y)^2z^2(1-z)^2 \\ x^2(1-x)^2y(1-y)z^2(1-z)^2 \\ x^2(1-x)^2y^2(1-y)^2z(1-z) \end{pmatrix}$$

and

$$p = -2xyz + x^2 + y^2 + z^2 + xy + xz + yz - x - y - z.$$

We take

$$\boldsymbol{u} = \mathbf{curl}\,\boldsymbol{\Phi}, \quad \boldsymbol{\omega} = \mathbf{curl}\,\boldsymbol{u}, \quad \boldsymbol{f} = \mathbf{curl}\,\boldsymbol{\omega} + \bigtriangledown p,$$

and we can easily verify that such a $(\boldsymbol{u}, p, \boldsymbol{\omega}, \boldsymbol{f})$ satisfies (6.1).

We partition Ω into a set of h^{-3} cubic subdomains with side-length h and use piecewise trilinear elements $(\mathcal{P}_1(K))^3$ to approximate all variables. We also set $p_h(0,0,0) = 0$, instead of $\int_\Omega p_h = 0$, to ensure the uniqueness.

In our computer codes, we use the double precision conjugate gradient method (CGM) to solve the associated linear system with an initial guess $(0,0,0,0,0,0,0)$, and the stopping criterion is the l^2 norm of the residual vector less than 10^{-11}. Computational results are collected in Tables 1– 7. The rates of convergence in Tables 5, 6, and 7 are computed the following intuitive way: for any two consecutive sets of data with respect to the mesh sizes $h1$ and $h2$, the rate of conv:=$\ln(\|e1\|/\|e2\|)/\ln(h1/h2)$.

We find that the numerical results in Tables 5– 7 support the conclusions obtained in this paper.

Table 1

The dimension of the matrix from the least-squares method.

h	Number of unknowns	Number of nontrivial entries
1/4	430	12100
1/6	1452	51654
1/8	3458	137696
1/10	6784	288514
1/12	11766	522396
1/14	18740	857630
1/16	28042	1312504
1/18	40008	1905306

Table 2

Relative errors for velocity $\boldsymbol{u}_h = (u_{1,h}, u_{2,h}, u_{3,h})^T$.

	$u_{1,h}$	$u_{1,h}$	$u_{2,h}$	$u_{2,h}$	$u_{3,h}$	$u_{3,h}$
h	L^2-Rel	H^1-Rel	L^2-Rel	H^1-Rel	L^2-Rel	H^1-Rel
1/4	4.300E-1	7.146E-1	3.528E-1	7.030E-1	3.964E-1	7.072E-1
1/6	2.089E-1	4.778E-1	1.555E-1	4.737E-1	1.854E-1	4.744E-1
1/8	1.232E-1	3.579E-1	8.772E-2	3.567E-1	1.072E-1	3.562E-1
1/10	8.147E-2	2.861E-1	5.746E-2	2.861E-1	7.011E-2	2.851E-1
1/12	5.812E-2	2.383E-1	4.168E-2	2.389E-1	4.970E-2	2.377E-1
1/14	4.378E-2	2.043E-1	3.249E-2	2.052E-1	3.729E-2	2.038E-1
1/16	3.431E-2	1.788E-1	2.667E-2	1.798E-1	2.921E-2	1.784E-1
1/18	2.774E-2	1.589E-1	2.271E-2	1.599E-1	2.369E-2	1.587E-1

Table 3

Relative errors for vorticity $\boldsymbol{\omega}_h = (\omega_{1,h}, \omega_{2,h}, \omega_{3,h})^T$.

	$\omega_{1,h}$	$\omega_{1,h}$	$\omega_{2,h}$	$\omega_{2,h}$	$\omega_{3,h}$	$\omega_{3,h}$
h	L^2-Rel	H^1-Rel	L^2-Rel	H^1-Rel	L^2-Rel	H^1-Rel
1/4	1.533E-1	4.190E-1	1.333E-1	4.116E-1	1.963E-1	4.538E-1
1/6	7.369E-2	2.833E-1	6.700E-2	2.777E-1	9.370E-2	2.982E-1
1/8	4.780E-2	2.150E-1	4.094E-2	2.099E-1	5.999E-2	2.235E-1
1/10	3.658E-2	1.741E-1	2.855E-2	1.696E-1	4.535E-2	1.804E-1
1/12	3.053E-2	1.470E-1	2.198E-2	1.430E-1	3.732E-2	1.523E-1
1/14	2.670E-2	1.277E-1	1.821E-2	1.242E-1	3.213E-2	1.323E-1
1/16	2.398E-2	1.132E-1	1.592E-2	1.103E-1	2.837E-2	1.173E-1
1/18	2.189E-2	1.020E-1	1.444E-2	9.957E-2	2.546E-2	1.056E-1

TABLE 4
Relative errors for pressure p_h.

	p_h	p_h
h	L^2-Rel	H^1-Rel
1/4	2.004E-1	4.968E-1
1/6	9.470E-2	3.296E-1
1/8	5.498E-2	2.467E-1
1/10	3.599E-2	1.973E-1
1/12	2.547E-2	1.644E-1
1/14	1.902E-2	1.410E-1
1/16	1.473E-2	1.234E-1
1/18	1.170E-2	1.098E-1

TABLE 5
Rates of convergence for velocity $\boldsymbol{u}_h = (u_{1,h}, u_{2,h}, u_{3,h})^T$.

	$u_{1,h}$	$u_{1,h}$	$u_{2,h}$	$u_{2,h}$	$u_{3,h}$	$u_{3,h}$
h	L^2-rate	H^1-rate	L^2-rate	H^1-rate	L^2-rate	H^1-rate
1/4	–	–	–	–	–	-
1/6	1.780	0.993	2.021	0.974	1.874	0.985
1/8	1.836	1.003	1.990	0.986	1.904	0.996
1/10	1.853	1.004	1.896	0.988	1.903	0.998
1/12	1.852	1.003	1.761	0.989	1.887	0.997
1/14	1.838	0.999	1.616	0.986	1.864	0.998
1/16	1.825	0.998	1.478	0.990	1.829	0.997
1/18	1.805	1.002	1.365	0.996	1.778	0.993

TABLE 6
Rates of convergence for vorticity $\boldsymbol{\omega}_h = (\omega_{1,h}, \omega_{2,h}, \omega_{3,h})^T$.

	$\omega_{1,h}$	$\omega_{1,h}$	$\omega_{2,h}$	$\omega_{2,h}$	$\omega_{3,h}$	$\omega_{3,h}$
h	L^2-rate	H^1-rate	L^2-rate	H^1-rate	L^2-rate	H^1-rate
1/4	–	–	–	–	–	–
1/6	1.839	0.965	1.697	0.971	1.824	1.036
1/8	1.505	0.959	1.712	0.973	1.550	1.002
1/10	1.199	0.946	1.615	0.955	1.254	0.960
1/12	0.992	0.928	1.434	0.936	1.069	0.929
1/14	0.870	0.913	1.221	0.914	0.971	0.913
1/16	0.805	0.903	1.006	0.889	0.932	0.901
1/18	0.774	0.885	0.828	0.869	0.919	0.892

Table 7
Rates of convergence for pressure p_h.

	p_h	p_h
h	L^2-rate	H^1-rate
1/4	–	–
1/6	1.849	1.012
1/8	1.890	1.007
1/10	1.899	1.001
1/12	1.896	1.001
1/14	1.894	0.996
1/16	1.914	0.998
1/18	1.955	0.991

Acknowledgments. The authors wish to thank the anonymous referees for valuable comments that improved the paper greatly. In particular, the authors wish to thank one of the two referees for pointing out the mistakes in the first version of this paper.

Special thanks go to Professor C.-L. Chang (Cleveland State University) and Professor S.-Y. Yang (National Central University, Taiwan) for their help with the computations reported in this paper. Special thanks also go to the associate editor Professor G. Starke for his helpful suggestions and advice.

REFERENCES

[1] B.-N. Jiang, *On least squares finite element method*, Comput. Methods Appl. Mech. Engrg., 152 (1998), pp. 239–257.

[2] P. B. Bochev and M. D. Gunzburger, *Finite element methods of least-squares type*, SIAM Rev., 40 (1998), pp. 789–837.

[3] J. M. Deang and M. D. Gunzburger, *Issues related to least-squares finite element methods for the Stokes equations*, SIAM J. Sci. Comput., 20 (1998), pp. 878–906.

[4] P. B. Bochev and M. D. Gunzburger, *Analysis of least-squares finite element methods for the Stokes equations*, Math. Comp., 63 (1994), pp. 479–506.

[5] C.-L. Chang and M. D. Gunzburger, *A finite element method for first order systems in three dimensions*, Appl. Math. Comput., 23 (1987), pp. 171–184.

[6] P. B. Bochev, *Analysis of least-squares finite element methods for the Navier–Stokes equations*, SIAM J. Numer. Anal., 34 (1997), pp. 1817–1844.

[7] Z.-Q. Cai, T. A. Manteuffel, and S. F. Mccormick, *First-order system least-squares for velocity-vorticity-pressure form of the Stokes equations, with application to linear elasticity*, Electron. Trans. Numer. Anal., 3 (1995), pp. 150–159.

[8] Z.-Q. Cai, T. A. Manteuffel, and S. F. Mccormick, *First-order system least squares for the Stokes equations, with application to linear elasticity*, SIAM J. Numer. Anal., 34 (1997), pp. 1727–1741.

[9] C.-L. Chang, *An error estimate of the least squares finite element method for the Stokes problem in three dimensions*, Math. Comp., 63 (1994), pp. 41–50.

[10] B.-N. Jiang and C.-L. Chang, *Least-squares finite elements for the Stokes problem*, Comput. Methods Appl. Mech. Engrg., 78 (1990), pp. 297–311.

[11] B.-N. Jiang, *The Least Squares Finite Element Method*, Springer-Verlag, Berlin, 1998.

[12] J. Bramble and J. Pasciak, *Least squares method for Stokes equations based on a discrete minus one inner product*, J. Comput. Appl. Math., 74 (1996), pp. 155–173.

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

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

[15] P.-G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North–Holland, Amsterdam, 1978.

[16] J. T. Oden and J. N. Reddy, *An Introduction to the Mathematical Theory of Finite Elements*, Wiley, New York, 1974.

[17] O. A. Ladyzhenskaya, *The Mathematical Theory of Viscous Incompressible Flow*, Gordon and Breach, New York, 1963.

[18] S. C. Brenner and L. R. Scott, *The Mathematical Theory of Finite Element Methods*, Springer-Verlag, Berlin, 1996.

[19] C. Amrouche, C. Bernardi, M. Dauge, and V. Girault, *Vector potentials in three-dimensional non-smooth domains*, Math. Methods Appl. Sci., 21 (1998), pp. 823–864.

[20] M. Krizek and P. Neittaanmaki, *On the validity of Friedrichs' inequality*, Math. Scand., 54 (1984), pp. 17–26.

[21] J. Saranen, *On an inequality of Friedrichs*, Math. Scand., 51 (1982), pp. 310–322.

[22] G. P. Galdi, *An Introduction to the Mathematical Theory of the Navier-Stokes Equations*, *Vol.* I, Springer-Verlag, New York, 1994.

[23] J. C. Nédélec, *Mixed finite elements in* $\Re^3$, Numer. Math., 35 (1980), pp. 315–341.

[24] J. C. Nédélec, *A new family of mixed finite elements in* $\Re^3$, Numer. Math., 50 (1986), pp. 57–81.

[25] D. Boffi, *Fortin operator and discrete compactness for edge elements*, Numer. Math., 87 (2000), pp. 229–246.

[26] P. Grisvard, *Elliptic Problems in Non-Smooth Domains*, Pitman, London, 1985.

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

[28] A. Buffa, M. Costabel, and D. Sheen, *On traces for* $H(\mathbf{curl};\Omega)$ *in Lipschitz domains*, J. Math. Anal. Appl., 276 (2002), pp. 845–876.

[29] A. Buffa and P.-G. Ciarlet, *On traces for functional spaces related to Maxwell's equations*, Math. Methods Appl. Sci., 24 (2001), pp. 9–30.

[30] A. Alonso and A. Valli, *Some remarks on the characterization of the space of tangential traces of* $H(\mathbf{curl};\Omega)$ *and the construction of an extension operator*, Manuscripta Math., 89 (1996), pp. 159–178.

[31] F. E. Dabaghi and O. Pironneau, *Stream vectors in three dimensional aerodynamics*, Numer. Math., 48 (1986), pp. 561–589.

[32] H.-Y. Duan, *Studies on Mixed Finite Element Methods*, Ph.D. thesis, Institute of Mathematics, Academy of Mathematics and System Sciences, Chinese Academy of Sciences, Beijing, People's Republic of China, 2002.

[33] A. Bendali, J. M. Dominguez, and S. Gallic, *A variational approach for the vector potential formulation of the Stokes and Navier-Stokes problems in three dimensional domains*, J. Math. Anal. Appl., 107 (1985), pp. 537–560.

[34] C.-L. Chang and S.-Y. Yang, *Analysis of the* L^2 *least-squares finite element method for the velocity-vorticity-pressure Stokes equations with velocity boundary conditions*, Appl. Math. Comput., 130 (2002), pp. 121–144.

[35] C.-L. Chang and B.-N. Jiang, *An error analysis of least-squares finite element method of the velocity-pressure-vorticity formulation for Stokes problem*, Comput. Methods Appl. Mech. Engrg., 84 (1990), pp. 247–255.

SIAM J. Numer. Anal.
Vol. 41, No. 6, pp. 2131–2148

ELIMINATION OF FIRST ORDER ERRORS IN TIME DEPENDENT SHOCK CALCULATIONS*

MALIN SIKLOSI[†] AND GUNILLA KREISS[†]

Abstract. First order errors downstream of shocks have been detected in computations with higher order shock-capturing schemes in one and two dimensions. We use matched asymptotic expansions to analyze the phenomenon for one dimensional time dependent hyperbolic systems and show how to design the artificial viscosity term in order to avoid the first order error. Numerical computations verify that second order accurate solutions are obtained.

Key words. hyperbolic conservation laws, shock wave, artificial viscosity, asymptotic analysis

AMS subject classifications. 35L65, 35L67, 65M06, 65M02

DOI. 10.1137/S0036142902400457

1. Introduction.

In many cases, solutions of conservation laws obtained by formally higher order methods are only first order accurate downstream of shocks; see, e.g., [2], [5], and [4]. Basically, errors from the shock region follow outgoing characteristics and pollute the solution downstream. Examples in one space dimension in which this effect can be seen are steady-state calculations for systems with a source term and time dependent calculations for systems with nonconstant solution. The effect cannot be seen in one dimensional Riemann problems, because the exact global conservation determines the postshock states.

This degeneration in accuracy is troublesome, even though the first order term for reasonable mesh-sizes seems to be small in many cases. In some applications, e.g., aeroacoustics, where waves with small amplitude need to be computed accurately, it is particularly important to achieve very high accuracy. It is also important to understand the phenomenon more deeply in order to be able to design new methods which do not suffer from this deficiency.

The aim of this paper is to show that the first order error can be understood by matched asymptotic analysis of the modified equation and that the analysis can be used to construct methods that yield second order accurate solutions.

We consider the case of systems with time dependent solutions. We assume that the numerical solution can be modeled by a slightly viscous equation, a so-called modified equation. In the shock layer, the coefficient of the viscous term is $\mathcal{O}(h)$, where h is the grid size. We analyze the solution of the modified equation using matched asymptotic expansions. It is assumed that an inner solution is valid in the shock region, and an outer solution is valid elsewhere. The two solutions are matched in a so-called matching zone. From the analysis, we see that generally the outer solution contains a term of $\mathcal{O}(h)$ downstream of the shock. We also see that if the inner solution satisfied a certain condition, the $\mathcal{O}(h)$ term would be eliminated. Based on this observation, we design a matrix valued viscosity coefficient, which gives the inner solution the right shape to eliminate the $\mathcal{O}(h)$ downstream term. We construct a numerical scheme, using this matrix valued viscosity coefficient, and show in numerical

*Received by the editors January 2, 2002; accepted for publication (in revised form) April 4, 2003; published electronically November 25, 2003.
http://www.siam.org/journals/sinum/41-6/40045.html

[†]Department of Numerical Analysis and Computer Science, Royal Institute of Technology, SE-100 44 Stockholm, Sweden (malins@nada.kth.se, gunillak@nada.kth.se).

experiments that the first order downstream error really is eliminated. However, we do not claim to have constructed an efficient and robust numerical method which can be used in realistic computations.

Similar analysis and construction of a matrix valued viscosity coefficient is done in [8] for the case of a steady-state solution of a system with a source term. In [3], matched asymptotic expansions for a problem that is very similar to the problem studied in this paper are analyzed for other purposes. The phenomenon has also been studied by other methods in [5] and [2]. In [5], analytic examples are constructed where the numerical solution is only first order accurate downstream of a shock, although the numerical scheme is formally second order. It is also shown that a converging numerical method will yield solutions having the formal order of accuracy in domains where no characteristics have passed through a shock. In [2], the first order downstream error is numerically detected in solutions of a shock-sound interaction problem solved by a fourth order ENO method. A scalar, linear equation is used to model the problem. It can be seen that the solution of the model problem computed with the fourth order ENO method behaves qualitatively differently depending on whether the discontinuity is located on a cell interface or in the interior of a cell. In the first case, the solution is fourth order in all of the domain, but in the second case the solution is only first order downstream of the discontinuity. Based on this observation, the numerical method is modified such that the shock position will always be on a cell interface, and fourth order accuracy of the solution of the shock-sound interaction problem is obtained both upstream and downstream. Also in [1], shock wave solutions are analyzed, and it is concluded that the structure in the shock region is of crucial importance for the solution outside the shock region. However, the analysis in [1] concerns another numerical phenomenon and considers methods where the shock is so narrow that it is not well modeled by the solution of a slightly viscous equation.

This paper is organized as follows. In section 2, we use asymptotic analysis to explain the first order downstream error and derive a matrix valued viscosity coefficient that eliminates it. In section 3, we implement a numerical method using the matrix valued viscosity coefficient and show in computations that the first order downstream error is eliminated.

2. Analysis.

2.1. The inviscid problem.

Consider the inviscid problem

$$\mathbf{u}_t + \mathbf{f}(\mathbf{u})_x = 0, \quad 0 \le x \le x_{\text{end}}, \tag{1}$$

$$\mathbf{u}(x,0) = \mathbf{g}(x), \tag{2}$$

where $\mathbf{u}(x,t), \mathbf{g}(x) \in \mathbf{R}^n$, $\mathbf{f} : \mathbf{R}^n \to \mathbf{R}^n$, and $\mathbf{g}$ is a piecewise smooth function. We denote the Jacobian of the flux function $\mathbf{f}'(\mathbf{u})$ by $J(\mathbf{u})$. We assume that the eigenvalues of $J(\mathbf{u})$, denoted $\lambda_i(\mathbf{u}), i = 1, 2, \dots, n$, are real and ordered in increasing order and that the eigenvectors span $\mathbf{R}^n$.

The initial and boundary conditions are chosen such that a shock forms at some inner point $s(t)$. At the shock, the solution satisfies the Rankine–Hugoniot condition

$$\dot{s}[\mathbf{u}] = [\mathbf{f}(\mathbf{u})].$$

Here $[\mathbf{u}] = \mathbf{u}^+ - \mathbf{u}^-$, where $\mathbf{u}^\pm = \lim_{\delta \to 0^+} \mathbf{u}(s(t) \pm \delta, t)$. Corresponding notation for other quantities will be used frequently.

We assume that the shock is a classical Lax 1-shock, i.e.,

$$\begin{aligned} \dot{s} &< \lambda_1^-, \\ \lambda_1^+ < \dot{s} &< \lambda_2^+, \end{aligned}$$

and that the matrix

$$D = \begin{pmatrix} S_{II}^+ & [\mathbf{u}] \end{pmatrix} \tag{3}$$

is nonsingular. Here the columns of S_{II}^+ are the eigenvectors of J^+ corresponding to the eigenvalues $\lambda_2^+, \lambda_3^+, \dots, \lambda_n^+$.

To complete the problem we also need boundary conditions. At each boundary we need as many boundary conditions as there are ingoing characteristics. We consider pointwise boundary conditions, i.e., boundary conditions where the quantities involved are prescribed pointwise at the boundary to some function of time. One example of such boundary conditions is when the ingoing characteristic variables are prescribed as a function of time. We call these boundary conditions mathematical boundary conditions to distinguish them from numerical boundary conditions. For more details concerning mathematical boundary conditions for hyperbolic equations, we refer to [9].

Remark. For 1-shocks and n-shocks there is just one downstream side. Hence, the first order error appears on only one side of the shock. For other Lax shocks, both sides of the shock are downstream sides, and first order errors appear on both sides. The phenomenon can be analyzed by the same method in both cases, but the analysis becomes less involved when only one side must be considered. Hence, here we analyze a 1-shock.

2.2. The slightly viscous model. We intend to study the behavior of numerical solutions of (1); i.e., we want to study the behavior of discrete functions that are the solutions of difference equations. A useful technique for studying the behavior of solutions to difference equations is to model the difference equation by a differential equation. Such a differential equation is often called a modified equation; see, e.g., [11], [6]. Many numerical solutions of (1) can be viewed as higher order accurate solutions of the modified equation

$$\mathbf{u}_t^\varepsilon + \mathbf{f}(\mathbf{u}^\varepsilon)_x = (\Gamma \mathbf{u}_x^\varepsilon)_x, \quad 0 \le x \le x_{\text{end}}.$$

In the shock region, the modified equation can be shown to be valid only for weak shocks; see, e.g., [7]. However, our computations indicate that it applies also for strong shocks. In the neighborhood of a shock layer we must have $\Gamma = \mathcal{O}(h)$, where h is the grid size, in order to avoid oscillations in the solution. Outside the shock region, Γ can be smaller. In this paper we consider methods which can be modeled by

$$\mathbf{u}_t^\varepsilon + \mathbf{f}(\mathbf{u}^\varepsilon)_x = \varepsilon(\phi \mathbf{u}_x^\varepsilon)_x + c_2 \varepsilon^2 \mathbf{u}_{xx}^\varepsilon, \tag{4}$$

where $\varepsilon = c_1 h$ and c_1 and c_2 a scalar constants. Here ϕ is a smooth function of $(x - s(t))/\varepsilon$ satisfying

$$\phi\left(\frac{x - s(t)}{\varepsilon}\right) = \begin{cases} 1 & \text{for } |\frac{x-s(t)}{\varepsilon}| \le K_0, \\ 0 & \text{for } |\frac{x-s(t)}{\varepsilon}| \ge K_1, \end{cases}$$

where $K_0 < K_1$ are constants with K_0 sufficiently large.

We must also model the initial data. In computations, the shape of the shock profile will depend on the method. If the initial data does not have exactly the right shape, the profile will after a short time adjust and obtain the right shape. In this process, small diffusion waves appear and flow out of the shock region, following the outgoing characteristics; see [12] and the references therein. We are not interested in studying this initial effect, and consequently we assume that the initial profile is exactly the right profile for the method that is modeled. We specify the initial profile in (12) and (13).

We consider the same mathematical boundary conditions for $\mathbf{u}^\varepsilon$ as for $\mathbf{u}$. When (1) is solved numerically, the mathematical boundary conditions must be augmented by numerical boundary conditions. Correspondingly, additional boundary conditions that model the numerical boundary conditions are needed for $\mathbf{u}^\varepsilon$. Numerical boundary conditions can introduce boundary layers in the solution. We consider numerical boundary conditions where such effects are $\mathcal{O}(h^2)$ or smaller, e.g., extrapolation of outgoing characteristic variables.

We define the position of the viscous shock layer as the smallest x-value such that $\mathbf{u}^{\varepsilon(1)}(x,t) = (\mathbf{u}^{-(1)} + \mathbf{u}^{+(1)})/2$, and denote this point by x_ε; i.e., the viscous shock position is defined as the point where the first component of the viscous solution $\mathbf{u}^\varepsilon$ is halfway between the right and left states in the corresponding inviscid shock.

2.3. Asymptotic expansions. We assume the following: The solution of (4) can be described by an inner solution, valid in the shock layer, and an outer solution, valid elsewhere. These solutions can be expanded in powers of ε and matched in a region of overlap. Also, the position of the shock layer can be expanded in ε. To leading order, the outer solution is equal to the solution of the corresponding inviscid problem.

We will now show that the outer solution downstream of the shock contains an $\mathcal{O}(h)$ term; i.e., downstream, the solution of (4) is just a first order approximation of the solution of the corresponding inviscid problem (1). There is no $\mathcal{O}(h)$ term upstream.

The inner solution is expressed using the variables $(\tilde{x}, \tilde{t})$, where

$$\tilde{x} = \frac{x - s(t)}{\varepsilon},$$
$$\tilde{t} = t.$$

Thus we have expansions of the form

$$\text{(5)} \qquad \text{Outer:} \quad \mathbf{u}^\varepsilon \sim \mathbf{u}(x,t) + \varepsilon \mathbf{u}_1(x,t) + \varepsilon^2 \mathbf{u}_2(x,t) + \cdots,$$
$$\text{Inner:} \quad \mathbf{u}^\varepsilon \sim \mathbf{U}_0(\tilde{x},\tilde{t}) + \varepsilon \mathbf{U}_1(\tilde{x},\tilde{t}) + \varepsilon^2 \mathbf{U}_2(\tilde{x},\tilde{t}) + \cdots,$$
$$\text{(6)} \qquad \text{Position:} \; x_\varepsilon \sim s(t) + \varepsilon x_1(t) + \varepsilon^2 x_2(t) + \cdots.$$

In [3], analysis of the asymptotic expansions for a very similar problem is presented, and also the existence of an asymptotic expansion is treated. For a detailed presentation of matched asymptotic expansions, we refer to [10].

We will match the inner and the outer solutions at an upstream and a downstream matching point, $x_m^-(t)$ and $x_m^+(t)$. The matching points must satisfy $\lim_{\varepsilon \to 0} |x_m^\pm - s| = 0$. We will also need $e^{\mp \tilde{x}_m^\pm} = \mathbf{o}(1)$. Choosing $x_m^\pm = s \mp \varepsilon \log(\varepsilon)$, we have $e^{\mp \tilde{x}_m^\pm} = \mathcal{O}(\varepsilon)$, and both requirements are satisfied.

The viscous problem (4) models a method which is a second order accurate approximation of (1) away from the shock region. We claim that the solution will be second order accurate upstream of the shock, but in general only first order downstream. Hence we must show that $\mathbf{u}_1 = 0$ upstream and $\mathbf{u}_1 \neq 0$ downstream. To do this we need equations, initial data, and boundary conditions for $\mathbf{u}_1$. Via the boundary conditions in the shock region, the outer solution will be coupled to the inner solution. Specifically, to derive boundary conditions for $\mathbf{u}_1$ we need information about $\mathbf{U}_0$. Hence, we derive equations and boundary conditions also for $\mathbf{U}_0$.

To obtain equations for the terms in the outer and inner expansions we substitute the expansions into (4), Taylor expand, and collect terms multiplying the same power of ε. The equation for $\mathbf{U}_0$ is

$$(\phi \mathbf{U}_{0\tilde{x}})_{\tilde{x}} + \dot{s}\mathbf{U}_{0\tilde{x}} - \mathbf{f}(\mathbf{U}_0)_{\tilde{x}} = 0, \quad -\infty < \tilde{x} < \infty, \tag{7}$$

where we have used that the relations between derivatives in x and t and derivatives in $\tilde{x}$ and $\tilde{t}$ are

$$\begin{aligned} \frac{\partial}{\partial x} &= \frac{1}{\varepsilon}\frac{\partial}{\partial \tilde{x}}, \\ \frac{\partial}{\partial t} &= -\frac{\dot{s}}{\varepsilon}\frac{\partial}{\partial \tilde{x}} + \frac{\partial}{\partial \tilde{t}}. \end{aligned}$$

The inner and outer expansions of $\mathbf{u}^\varepsilon$ are assumed to be valid in a region of overlap containing the matching points $x_m^\pm$. Hence, in the matching points, we must have $\lim_{\varepsilon \to 0} |\mathbf{U}_0 - \mathbf{u}| = 0$, i.e.,

$$\lim_{\varepsilon \to 0} |\mathbf{U}_0(\mp \log(\varepsilon), \tilde{t}) - \mathbf{u}(s \mp \varepsilon \log(\varepsilon), t)| = 0,$$

where we have used that $x_m^\pm = s \mp \varepsilon \log(\varepsilon)$. Evaluating the limit, we arrive at the matching conditions

$$\mathbf{U}_0(\pm\infty, \tilde{t}) = \mathbf{u}^\pm(t), \tag{8}$$

where $\mathbf{U}_0(\pm\infty, \tilde{t}) = \lim_{\tilde{x} \to \pm\infty} \mathbf{U}_0(\tilde{x}, \tilde{t})$. Note that (7) and (8) determine the shape of $\mathbf{U}_0$ but not the exact position of the shock layer.

We define $\hat{\mathbf{U}}(\tilde{x}, \tilde{t})$ by

$$\hat{\mathbf{U}}_{\tilde{x}\tilde{x}} + \dot{s}\hat{\mathbf{U}}_{\tilde{x}} - \mathbf{f}(\hat{\mathbf{U}})_{\tilde{x}} = 0, \quad -\infty < \tilde{x} < \infty, \tag{9}$$

$$\hat{\mathbf{U}}(\tilde{x}, \tilde{t}) = \mathbf{u}(s \pm 0, \tilde{t}) \quad \text{as } \tilde{x} \to \pm\infty, \tag{10}$$

$$\hat{\mathbf{U}}^{(1)}(0, \tilde{t}) = (\mathbf{u}^{-(1)} + \mathbf{u}^{+(1)})/2. \tag{11}$$

We see that $\hat{\mathbf{U}}$ differs from $\mathbf{U}_0$ in two ways. First, $\hat{\mathbf{U}}$ is independent of ϕ, which makes the equation for $\hat{\mathbf{U}}$ much easier to analyze. Second, the position of $\hat{\mathbf{U}}$ is fixed at $\tilde{x} = 0$. We note that both problems are independent of ε.

Let us first compare the shape of solutions of (7) and (8) with the shape of solutions of (9) and (10), disregarding the difference in shock position. It is easy to show that $\hat{\mathbf{U}}$ approaches its limit values exponentially fast as $\tilde{x} \to \pm\infty$. If one constructs the equations for the difference $\hat{\mathbf{U}} - \mathbf{U}_0$ and uses the exponential behavior of $\hat{\mathbf{U}}$, one can conclude that $\mathbf{U}_0$ also approaches its limit values exponentially fast and that $|\hat{\mathbf{U}} - \mathbf{U}_0|_\infty < e^{-K}$, where K is a large constant.

Since the position of the shock layer has the expansion (6), we have, except for exponentially small terms, to leading order

$$\mathbf{U}_0(\tilde{x}, \tilde{t}) = \hat{\mathbf{U}}(\tilde{x} - x_1(\tilde{t}), \tilde{t}).$$

Below we will derive an ordinary differential equation for $x_1(\tilde{t})$. The initial value of $x_1(\tilde{t})$ is determined by the initial condition $\mathbf{g}^\varepsilon$, which we now specify:

$$\text{Outer region: } \mathbf{g}^\varepsilon(x) = \mathbf{g}(x), \tag{12}$$

$$\text{Inner region: } \mathbf{g}^\varepsilon(\tilde{x}) = \hat{\mathbf{U}}(\tilde{x}, 0). \tag{13}$$

This is sufficient for our purposes. However, if one considers more terms in the inner expansion, one would have to add the corresponding terms to (13). Note that (13) means that $x_1(0) = 0$.

The equation for $\mathbf{u}_1$ is

$$\mathbf{u}_{1t} + (\mathbf{f}'(\mathbf{u})\mathbf{u}_1)_x = 0, \quad x \in \text{outer region}, \tag{14}$$

where we have used that $\phi = 0$ in the outer region. We also need initial data and boundary conditions for $\mathbf{u}_1$. The initial conditions for $\mathbf{u}^\varepsilon$, (12), gives $\mathbf{u}_1(x, 0) = 0$. Since $\mathbf{u}^\varepsilon$ and $\mathbf{u}$ satisfy the same mathematical boundary conditions and since boundary layer effects, due to numerical boundary conditions, are assumed to be $\mathcal{O}(h^2)$ or smaller, we conclude that all boundary conditions for $\mathbf{u}_1$ at $x = 0$ and $x = x_{\text{end}}$ are homogeneous. At the upstream side of the shock, no further boundary conditions are needed since all characteristics of (14) are going into the shock. Since $\mathbf{u}_1$ in the upstream region is the solution of a homogeneous equation with homogeneous initial data and homogeneous boundary conditions, we have $\mathbf{u}_1 \equiv 0$ in the upstream region. To determine $\mathbf{u}_1$ in the downstream region, we also need boundary conditions at $x = s^+$. We derive such boundary conditions in the next section.

2.4. Downstream boundary condition for the first order outer term.

Integration of the viscous equation (4) over the shock layer, from matching point x_m^- to matching point x_m^+, gives

$$\int_{x_m^-}^{x_m^+} \mathbf{u}_t^\varepsilon \, dx + [\mathbf{f}(\mathbf{u}^\varepsilon)]_{x_m^-}^{x_m^+} = \mathcal{O}(\varepsilon^2), \tag{15}$$

where we have used that ϕ vanishes in the matching regions. Using the outer expansion of $\mathbf{u}^\varepsilon$, we obtain

$$[\mathbf{f}(\mathbf{u}^\varepsilon)]_{x_m^-}^{x_m^+} = [\mathbf{f}(\mathbf{u})]_{x_m^-}^{x_m^+} + \varepsilon[J(\mathbf{u})\mathbf{u}_1]_{x_m^-}^{x_m^+} + \mathcal{O}(\varepsilon^2). \tag{16}$$

By integrating the inviscid (1) over the same interval, we obtain

$$[\mathbf{f}(\mathbf{u})]_{x_m^-}^{x_m^+} = \dot{s}[\mathbf{u}] - \int_{x_m^-}^{s^-} \mathbf{u}_t \, dx - \int_{s^+}^{x_m^+} \mathbf{u}_t \, dx. \tag{17}$$

Note that $\mathbf{u}$ is discontinuous at $x = s(t)$ and the Rankine–Hugoniot condition applies across the discontinuity. After taking into account that $\mathbf{u}_1 \equiv 0$ to the left of the shock layer and introducing (16) and (17) into (15), we arrive at

$$\dot{s}[\mathbf{u}] + \varepsilon J(\mathbf{u}(x_m^+, t))\mathbf{u}_1(x_m^+, t) + I_1 = \mathcal{O}(\varepsilon^2), \tag{18}$$

where we have introduced the notation

$$I_1 = \int_{x_m^-}^{s^-} (\mathbf{u}_t^\varepsilon - \mathbf{u}_t) \, dx + \int_{s^+}^{x_m^+} (\mathbf{u}_t^\varepsilon - \mathbf{u}_t) \, dx.$$

After Taylor expansion of $\mathbf{u}$ and $\mathbf{u}_1$ around $x = s^+$, (18) can be rewritten as

$$\dot{s}[\mathbf{u}] + \varepsilon J(\mathbf{u}(s^+, t))\mathbf{u}_1(s^+, t) + I_1 = \mathbf{o}(\varepsilon). \tag{19}$$

In the coordinate system $(\tilde{x}, \tilde{t})$ we have

$$I_1 = -\dot{s}A + \varepsilon I_2,$$

where

$$A = \int_{\tilde{x}_m^-}^{0^-} (\mathbf{u}^\varepsilon - \mathbf{u})_{\tilde{x}}\, d\tilde{x} + \int_{0^+}^{\tilde{x}_m^+} (\mathbf{u}^\varepsilon - \mathbf{u})_{\tilde{x}}\, d\tilde{x},$$
$$I_2 = \int_{\tilde{x}_m^-}^{0^-} (\mathbf{u}^\varepsilon - \mathbf{u})_{\tilde{t}}\, d\tilde{x} + \int_{0^+}^{\tilde{x}_m^+} (\mathbf{u}^\varepsilon - \mathbf{u})_{\tilde{t}}\, d\tilde{x}.$$

Evaluating the integral yields

$$A = [\mathbf{u}] + [\mathbf{u}^\varepsilon - \mathbf{u}]_{\tilde{x}_m^-}^{\tilde{x}_m^+}.$$

By using the outer expansion of $\mathbf{u}^\varepsilon$, taking into account that $\mathbf{u}_1$ is zero upstream and Taylor expanding $\mathbf{u}_1$ around $x = s^+$, we obtain

$$A = [\mathbf{u}] + \varepsilon \mathbf{u}_1^+ + \mathbf{o}(\varepsilon).$$

Next, consider I_2. Using the inner expansion of $\mathbf{u}^\varepsilon$, the Taylor expansion of $\mathbf{u}$ around $x = s \pm 0$, and $\mathbf{U}_0(\tilde{x}, \tilde{t}) = \hat{\mathbf{U}}(\tilde{x} - x_1, \tilde{t})$, we obtain

$$I_2 = \int_{\tilde{x}_m^-}^{0} (\hat{\mathbf{U}}(\tilde{x} - x_1, \tilde{t}) - \mathbf{u}^-)_{\tilde{t}}\, d\tilde{x} + \int_{0}^{\tilde{x}_m^+} (\hat{\mathbf{U}}(\tilde{x} - x_1, \tilde{t}) - \mathbf{u}^+)_{\tilde{t}}\, d\tilde{x} + \mathbf{o}(1).$$

We rewrite I_2 in two steps. First we make the substitution $\hat{x} = \tilde{x} - x_1$. Next, we use the fact that $\hat{\mathbf{U}}$ approaches the limit values exponentially fast, and the matching points are chosen such that $e^{\mp\tilde{x}_m^\pm} = \mathcal{O}(\varepsilon)$. Hence we can extend the integration interval to infinity, still keeping the remainder term $\mathbf{o}(1)$. We obtain

$$I_2 = I_{3\tilde{t}} - (x_1[\mathbf{u}])_{\tilde{t}} + \mathbf{o}(1),$$

where

$$I_3(\tilde{t}) = \int_{-\infty}^{0} (\hat{\mathbf{U}}(\tilde{x}, \tilde{t}) - \mathbf{u}^-)\, d\tilde{x} + \int_{0}^{\infty} (\hat{\mathbf{U}}(\tilde{x}, \tilde{t}) - \mathbf{u}^+)\, d\tilde{x}.$$

Since I_2, I_3, x_1, and $[\mathbf{u}]$ are functions of $\tilde{t}$ only, and since $\tilde{t} = t$, this can be written as

$$I_2(t) = \frac{\partial}{\partial t}\left(I_3(t) - x_1(t)[\mathbf{u}](t)\right) + \mathbf{o}(1).$$

Hence we have

$$I_1 = -\dot{s}[\mathbf{u}] + \varepsilon(-\dot{s}\mathbf{u}_1^+ + I_{3t} - (x_1[\mathbf{u}])_t) + \mathbf{o}(\varepsilon).$$

Substituting this into (19) and rearranging, we obtain

$$(J^+ - \dot{s}I)\mathbf{u}_1^+ - (x_1[\mathbf{u}])_t + I_{3t} = \mathbf{o}(1).$$

Hence the equations for $\mathbf{u}_1^+$ and $x_1(t)$ are

$$(J^+ - \dot{s}I)\mathbf{u}_1^+(t) - (x_1(t)[\mathbf{u}])_t = -I_{3t}, \tag{20}$$

$$x_1(0) = 0. \tag{21}$$

The two equations (20) and (21) constitute the boundary conditions for $\mathbf{u}_1$ at $x = s^+$. To make (20) and (21) easier to understand, we rewrite them using the characteristic variables of $\mathbf{u}_1$. Let w_I be the characteristic variable of $\mathbf{u}_1$ going into the shock, and $\mathbf{w}_{II}$ be the characteristic variables going out of the shock. We then have

$$\mathbf{u}_1^+ = (S_I^+ S_{II}^+) \begin{pmatrix} \mathrm{w}_I^+ \\ \mathbf{w}_{II}^+ \end{pmatrix},$$

where S_I^+ is the eigenvector of J^+ corresponding to the eigenvalue λ_1^+ and the columns of S_{II}^+ are the eigenvectors of J^+ corresponding to the eigenvalues $\lambda_2^+, \lambda_3^+, \ldots, \lambda_n^+$. Expressed in the characteristic variables, the boundary condition is

$$\begin{pmatrix} \mathbf{w}_{II}^+ \\ \dot{x}_1 \end{pmatrix} = \begin{pmatrix} \Lambda_{II}^+ - \dot{s}I & 0 \\ 0 & -1 \end{pmatrix}^{-1} D^{-1} \left(-I_{3t} + x_1[\mathbf{u}]_t - S_I^+(\lambda_1^+ - \dot{s})\mathrm{w}_I^+\right), \tag{22}$$

$$x_1(0) = 0, \tag{23}$$

where $\Lambda_{II}^+ = \mathrm{diag}(\lambda_2^+, \lambda_3^+, \ldots, \lambda_n^+)$ and D is defined by (3).

By solving (22) and (23) for $x_1(t)$ and then substituting the solution into (22) again, we can express $\mathbf{w}_{II}^+$ in w_I^+ and known functions of time. The energy method (see [9]) shows that the equation, boundary conditions, and initial data for $\mathbf{w}$ constitute a well-posed problem. Well-posedness implies that for any I_{3t} there exists a unique solution. The boundary condition for $\mathbf{w}$ at $x = s^+$ is homogeneous if $I_{3t} \equiv 0$, and nonhomogeneous otherwise. Since $\mathbf{w}$ is a transformation of $\mathbf{u}_1$, the same applies for $\mathbf{u}_1$.

It is now clear that I_{3t} is crucial for the order of accuracy of $\mathbf{u}^\varepsilon$. In the special case $I_{3t} \equiv 0$ we have $\mathbf{u}_1 \equiv 0$ in the downstream region, since $\mathbf{u}_1$ is the solution of a homogeneous equation with homogeneous initial data and boundary conditions. From (5) it then follows that $\mathbf{u}^\varepsilon$ is a second order accurate approximation of $\mathbf{u}$. However, in the general case, we have $\mathbf{u}_1(x,t) \neq 0$ for $x > s$, and $\mathbf{u}^\varepsilon$ will be a first order accurate approximation of $\mathbf{u}$.

2.5. A matrix valued viscosity coefficient eliminating the $\mathcal{O}(h)$ error.

We will now investigate whether it is possible to design the viscosity term such that the first order downstream error is eliminated and second order accurate solutions are obtained. We consider a method which has the modified equation

$$\mathbf{u}_t^\varepsilon + \mathbf{f}(\mathbf{u}^\varepsilon)_x = \varepsilon(\phi(x)E(\mathbf{u}^\varepsilon)\mathbf{u}_x^\varepsilon)_x + c_2\varepsilon^2\mathbf{u}_{xx}^\varepsilon, \tag{24}$$

where $E(\mathbf{u}^\varepsilon)$ is a matrix valued function. The solutions given by such a method can be analyzed in the same way as in the previous sections. The only point which will change in the analysis is the equation for $\hat{\mathbf{U}}$. The new equation for $\hat{\mathbf{U}}$ is

$$(E(\hat{\mathbf{U}})\hat{\mathbf{U}}_{\tilde{x}})_{\tilde{x}} + \dot{s}\hat{\mathbf{U}}_{\tilde{x}} - \mathbf{f}(\hat{\mathbf{U}})_{\tilde{x}} = 0, \tag{25}$$

together with the conditions (10) and (11). The boundary condition for $\mathbf{u}_1$ at $x = s^+$ is still given by (22) and (23). If $E(\hat{\mathbf{U}})$ can be chosen such that $I_{3t} \equiv 0$, we will have $\mathbf{u}_1(x,t) \equiv 0$ also in the downstream region.

We note that if $\hat{\mathbf{U}} = \hat{\mathbf{U}}^*$ with

$$\hat{\mathbf{U}}^* = \mathbf{u}^- + \gamma(\tilde{x})[\mathbf{u}], \tag{26}$$

where γ is a scalar smooth function, we obtain

$$I_3 = c_\gamma[\mathbf{u}],$$

where

$$c_\gamma = \int_{-\infty}^{0} \gamma(\tilde{x})\, d\tilde{x} + \int_{0}^{\infty} (\gamma(\tilde{x}) - 1)\, d\tilde{x}.$$

If γ is antisymmetric around $(0, 0.5)$ with

$$\gamma(-\infty) = 0, \quad \gamma'(-\infty) = 0, \quad \gamma(\infty) = 1, \quad \gamma'(\infty) = 0,$$

then $c_\gamma = 0$ and the boundary conditions (10) and (11) are satisfied.

It now remains to investigate whether it is possible to choose the matrix valued function $E(\hat{\mathbf{U}})$ such that $\hat{\mathbf{U}}^*$ satisfies (25). Integrating (25) from $-\infty$ to $\tilde{x}$ and substituting $\hat{\mathbf{U}}^*$ gives

$$\gamma'(\tilde{x})E(\hat{\mathbf{U}}^*)[\mathbf{u}] = \mathbf{q}(\hat{\mathbf{U}}^*), \tag{27}$$

where

$$\mathbf{q}(\mathbf{U}) = \mathbf{f}(\mathbf{U}) - \mathbf{f}(\mathbf{u}^-) - \dot{s}(\mathbf{U} - \mathbf{u}^-).$$

Note that $E(\hat{\mathbf{U}})$ is a function of $\hat{\mathbf{U}}$ only, with no explicit $\tilde{x}$ dependence. Hence, in order to solve (27) for $E(\hat{\mathbf{U}})$, we must be able to express γ' as a function of $\hat{\mathbf{U}}$. This is the case if we can express γ' in terms of γ, and if γ is monotone. Now solving (27) for $E(\hat{\mathbf{U}}^*)$ gives

$$E(\hat{\mathbf{U}}^*) = \frac{1}{\gamma'} \frac{\mathbf{q}(\hat{\mathbf{U}}^*)\mathbf{q}^T(\hat{\mathbf{U}}^*)}{\mathbf{q}^T(\hat{\mathbf{U}}^*)[\mathbf{u}]}. \tag{28}$$

To ensure that $E(\mathbf{u}^\varepsilon)$ is bounded as $\tilde{x} \to \pm\infty$ we must also require

$$\lim_{\tilde{x}\to-\infty} \frac{\gamma}{\gamma'} = M^-, \quad \lim_{\tilde{x}\to\infty} \frac{\gamma - 1}{\gamma'} = M^+,$$

where $|M^\pm| < \infty$.

Note that in order to evaluate $E(\hat{\mathbf{U}})$ the quantities $\dot{s}$, $\mathbf{u}^-$, and $\mathbf{u}^+$ must be known or estimated.

Remark. Prescribing the viscous profile as above means that the solution follows a straight line in phase space between the upstream and the downstream states. Many other shapes of the solution, and hence, paths in phase space, would also be possible. The properties of $E(\mathbf{u}^\varepsilon)$ will depend on which path is chosen. In order to obtain a stable method, it is necessary that the total viscosity coefficient of the method be positive definite. Since the term $c_2\varepsilon^2\mathbf{u}^\varepsilon_{xx}$ is also present, it is sufficient that $E(\mathbf{u}^\varepsilon)$ be positive semidefinite. We have found that the choice (28) is not positive semidefinite for all problems. In order to design a robust numerical method, we must further investigate what paths in phase space to use. Probably, this will differ depending on the equation that is to be solved. However, we are interested only in showing that it is possible to obtain second order accuracy also downstream, and for this purpose it is good enough to use $E(\mathbf{u}^\varepsilon)$ defined by (28).

3. Numerical experiments. In this section we test how the matrix valued viscosity coefficient derived in the previous section behaves in computations, and compare the results to corresponding computations with a scalar viscosity coefficient.

3.1. The test problems. We consider two test problems. In both problems, the equations, domain, initial data, and boundary condition at $x = x_{\text{end}}$ are the same, while the boundary condition at $x = 0$ differs.

We consider the Euler equations with Riemann initial data connected by a 1-shock. That is, we consider (1) and (2) with

$$\mathbf{u} = \begin{pmatrix} \rho \\ \rho u \\ E \end{pmatrix}, \quad \mathbf{f}(\mathbf{u}) = \begin{pmatrix} \rho u \\ \rho u^2 + p \\ u(E+p) \end{pmatrix}, \quad x_{\text{end}} = 6,$$

$$\mathbf{u}(x,0) = \begin{cases} \mathbf{u}_L & \text{for } x \le s_0, \\ \mathbf{u}_R & \text{for } x > s_0, \end{cases} \tag{29}$$

where E and p are connected by the equation of state for a polytropic gas

$$E = \frac{p}{\gamma - 1} + \frac{1}{2}\rho u^2, \quad \gamma = 1.4.$$

Since $\mathbf{u}_L$ and $\mathbf{u}_R$ are connected by a 1-shock, they are fully determined if ρ_L, u_L, and p_L—the initial density, velocity, and pressure at $x \le s_0$—and p_R, the initial pressure at $x > s_0$, are specified. We have used $\rho_L = 3$, $u_L = 1.2$, $p_L = 2$, and $p_R = 5$. This gives a 1-shock with speed ≈ -0.26. We have not specified the initial shock position s_0 explicitly. In the computations, which will be further described below, we started the computation at $t = -1$, with the profile (31) and the shock located at $x = 1.75$. We computed for one time unit using $\mathbf{u}(0,t) = \mathbf{u}_L$. In this way we obtained a good initial profile. We do this to avoid pollution of the numerical solution by disturbances due to nonperfect initial data. We have also used a rather large x_{end} in order to avoid reflection of such disturbances at the boundary $x = x_{\text{end}}$.

At $x = x_{\text{end}}$ we have the boundary condition

$$R_1(x_{\text{end}}, t) = R_1(x_{\text{end}}, 0),$$

where $R_1 = u - 2c/(\gamma - 1)$ is the Riemann invariant connected to λ_1, and $c = \sqrt{\gamma p/\rho}$ is the local speed of sound.

At $x = 0$ the boundary condition is specified by

$$p(0,t) = p_L(1 + \alpha d(t)),$$
$$\rho(0,t) = \rho_L \left(\frac{p(0,t)}{p_L}\right)^{1/\gamma},$$
$$u(0,t) = u_L + \frac{2}{\gamma - 1}(c(0,t) - c_L)$$

(see [2]); i.e., a disturbance with amplitude α is introduced into the Riemann invariant R_1, while the two other Riemann invariants are held constant. If α is small, this models an acoustic disturbance. We have considered the following two test problems:

Test problem 1: $\alpha = -0.2, \quad d(t) = (1 - \mathrm{e}^{-5t}) \sin 2t;$

Test problem 2: $\alpha = -0.1, \quad d(t) = \mathrm{e}^{-10(t-0.7)^2}.$

In the test problems, α is rather large, in order to make the first order effect more visible.

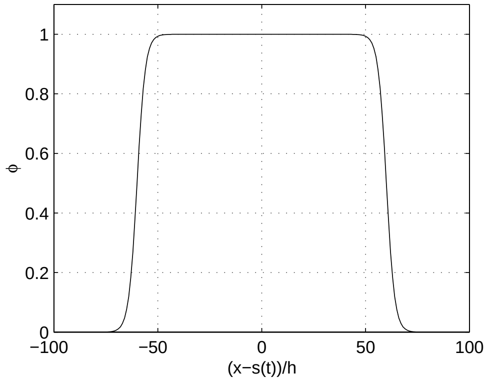


FIG. 1. *The function ϕ, with $s_1 = 60$ and $s_2 = 4$.*

3.2. The standard method. A common way to solve (1) is to discretize in space using central differences and add artificial viscosity. To avoid oscillations in the solution, the viscosity must be $\mathcal{O}(h)$ in the shock layer. Outside the shock region, the viscosity can be smaller. We obtain a formally second order method, whose solutions can be modeled by (4), using the semidiscrete scheme

$$(\mathbf{u}_j)_t + D_0\mathbf{f}(\mathbf{u}_j) = \kappa_1 h D_+ \phi_j D_- \mathbf{u}_j + \zeta h^2 D_+ D_- \mathbf{u}_j. \tag{30}$$

For test problem 1, we used $\kappa_2 = 1$ and $\zeta = 20$, and for test problem 2, we used $\kappa_1 = 0.5$ and $\zeta = 40$. We discretized in space by introducing $x_j = jh$, $h = 1/N$, $j = 0, 1, \ldots, N$, where $\mathbf{u}_j(t)$ is a grid function with $\mathbf{u}_j(t) \approx \mathbf{u}^\varepsilon(x_j, t)$. The system of ODEs (30) was solved with the classical fourth order Runge–Kutta method. The time step was chosen as $k = 0.5h$, i.e., CFL-number 0.5.

The switch ϕ was

$$\phi(x) = \begin{cases} 0.5\tanh((x - s(t) + s_1 h)/s_2 h) + 0.5, & x \le s(t), \\ 0.5\tanh((x - s(t) - s_1 h)/s_2 h) + 0.5, & x > s(t), \end{cases}$$

with $s_1 = 60$ and $s_2 = 4$; see Figure 1. Generally, there will be approximately $2s_1$ points where $\phi > 0.5$, and hence we have used a very wide switch. The parameter s_2 determines how steep the gradient of ϕ is in the transition area. The shock position $s(t)$ was numerically determined.

At $x = 6$ we used the mathematical boundary condition

$$R_1(6, t) = R_1(6, 0)$$

and the numerical boundary conditions

$$R_i(6, t) = 2R_i(6 - h, t) - R_i(6 - 2h, t), \quad i = 2, 3,$$

where the Riemann invariants R_2 and R_3 are

$$R_2 = \frac{p}{\rho^\gamma}, \quad R_3 = u + \frac{2c}{\gamma - 1}.$$

The initial data was obtained in the following way. We started the computations at $t = -1$ with the profile (31) and the shock located at $x = 1.75$. We computed for one time unit using $\mathbf{u}(0, t) = \mathbf{u}_L$.

We will refer to this method as the standard method.

3.3. The matrix viscosity method. We will now introduce a method which can be modeled by (24), and we will refer to it as the matrix viscosity method. The matrix viscosity method is the same as the standard method, except that (30) is replaced by

$$(\mathbf{u}_j)_t + D_0\mathbf{f}(\mathbf{u}_j) = \kappa_2 h D_+ \phi_j E_j D_- \mathbf{u}_j + \zeta h^2 D_+ D_- \mathbf{u}_j.$$

Here $E_j \approx E(\mathbf{u}^\varepsilon(x_j, t))$. Our implementation is described below. When solving test problem 1, we used $\kappa_2 = 15$, $\zeta = 20$, and CFL-number 0.05. For test problem 2, we used $\kappa_2 = 7$, $\zeta = 40$, and CFL-number 0.1.

To implement E_j in a robust and accurate way is difficult. The expression (28) is not suited for computations. The solution changes rapidly in the shock layer from being close to $\mathbf{u}^-$ to being close to $\mathbf{u}^+$. The quantities $\dot{s}$, $\mathbf{u}^-$, and $\mathbf{u}^+$ must be numerically determined; hence it is difficult to compute $\mathbf{q}$ with high accuracy. Also, both $\mathbf{q}$ and γ' tend rapidly to zero as $\tilde{x} \to \pm\infty$. However, for large $\tilde{x}$ we can linearize the expression for $\mathbf{q}$ and find

$$\mathbf{q} = \begin{cases} \gamma(J^- - \dot{s}I)[\mathbf{u}] & \text{as } \tilde{x} \to -\infty, \\ (\gamma - 1)(J^+ - \dot{s}I)[\mathbf{u}] & \text{as } \tilde{x} \to \infty. \end{cases}$$

By the assumptions on γ we find

$$E = \begin{cases} E^- & \text{as } \tilde{x} \to -\infty, \\ E^+ & \text{as } \tilde{x} \to \infty, \end{cases}$$

where

$$E^\pm = M^\pm \frac{(J^\pm - \dot{s}I)[\mathbf{u}][\mathbf{u}]^T (J^\pm - \dot{s}I)^T}{[\mathbf{u}]^T (J^\pm - \dot{s}I)^T [\mathbf{u}]}.$$

We have used

$$\gamma(\tilde{x}) = \frac{1}{2}(\tanh(\tilde{x}) + 1),$$

and hence we have $M^- = 1/2$ and $M^+ = -1/2$.

In the computations we have used

$$E_j = (1 - \gamma(\tilde{x}_j))E^- + \gamma(\tilde{x}_j)E^+.$$

The quantities $\dot{s}$, $\mathbf{u}^-$, and $\mathbf{u}^+$ were numerically determined. First, we approximated $\mathbf{u}^\pm$ by simply taking the value of the numerical solution 20 points upstream and downstream, respectively, of the approximated shock position. The shock speed $\dot{s}$ was approximated by

$$\dot{s}_{\text{approx}} = \sum_{k=1}^{3} [\mathbf{f}^{(k)}(\mathbf{u})] \Big/ \sum_{k=1}^{3} [\mathbf{u}^{(k)}].$$

By adding the jumps in the different components of $\mathbf{f}(\mathbf{u})$ and dividing by the sum of the jumps in the different components of $\mathbf{u}$, we avoid introducing large errors in $\dot{s}_{\text{approx}}$ due to rounding effects. This method for approximating $\mathbf{u}^\pm$ and $\dot{s}$ was used when test problem 1 was solved, and we obtain second order accuracy both upstream and downstream. The results are further presented in section 3.4.

The approximation of $\mathbf{u}^{\pm}$ mentioned above has an error which is small, but independent of h. Hence, there will be a small $\mathcal{O}(1)$ error in our approximation of $E(\mathbf{u}^{\varepsilon})$, which will cause a small $\mathcal{O}(h)$ error in the solution. This first order effect became evident as we tried to solve test problem 2. In order to eliminate it, we needed to use an estimate of $\mathbf{u}^{\pm}$ with an error which is $\mathcal{O}(h)$. We obtain this if we make use of the fact that the solution in the shock region follows a straight line in phase space between $\mathbf{u}^{-}$ and $\mathbf{u}^{+}$, and that the function γ determines how fast the solution is approaching the limit values. We pick the value of the solution $2\kappa_2$ points upstream and downstream of the approximate shock position. Since γ is known, we know how far from the end states these values are and correct for this. The shock speed $\dot{s}$ is still computed as above. Test problem 2 was solved using this improved approximation of $\mathbf{u}^{\pm}$, and we obtained second order accuracy both upstream and downstream.

The implementation of $E(\mathbf{u}^{\varepsilon})$ described above requires a fixed number of computations, independent of h.

As initial profile at $t = -1$, we used

$$\mathbf{u}_j = \mathbf{u}_L + \gamma(\tilde{x}_j)(\mathbf{u}_R - \mathbf{u}_L); \tag{31}$$

i.e., in the shock region the profile satisfies (26). The initial profile (31) is used to avoid diffusion waves in the solution (see section 2.2). Ideally, using this initial profile, no such waves should appear. However, in our numerical computations we see diffusion waves, but they are very small.

In computations not reported here, we have also tried to use

$$\mathbf{u}(x,-1) = \begin{cases} \mathbf{u}_L & \text{for } x \leq 1.75, \\ \mathbf{u}_R & \text{for } x > 1.75. \end{cases} \tag{32}$$

As expected, the profile rapidly adjusts, and diffusion waves appear and move out of the shock region following outgoing characteristics. Also as expected, for the matrix method, the diffusion waves are much larger with (32) as initial data than if (31) is used. However, the order of accuracy of the solution behind the diffusion waves is the same.

3.4. Results. We have numerically investigated the rate of convergence of the standard method and of the matrix viscosity method by solving the two test problems described in section 3.1 with successively refined space step.

First, consider test problem 1. In Figure 2 we see the solution at $t = 1.25$. We have solved test problem 1 with successively halved space step h with both methods. We started with $h = 0.02$. In all, we computed six solutions with the matrix viscosity method and eight solutions with the standard method.

The computational order of accuracy, r_h, was estimated in the standard way,

$$r_h = \log\left(\frac{||\rho u_{4h} - \rho u_{2h}||}{||\rho u_{2h} - \rho u_h||}\right) \Big/ \log 2,$$

where ρu_h denotes the discrete approximation of ρu with space step h, and the norm used was the discrete L_2-norm on the interval $(0, 0.7)$ in the upstream region and $(1.4, 2.2)$ in the downstream region. In Table 1 we see that the standard method is second order accurate upstream, but only first order accurate downstream of the

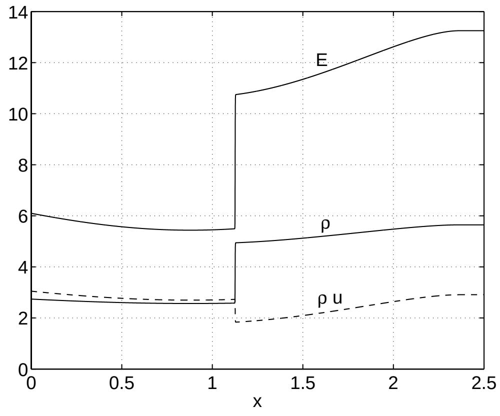


Fig. 2. *The solution of test problem* 1 *at* $t = 1.25$*. The solution is computed numerically using the standard method with* $h = 1.5625 \cdot 10^{-4}$.

Table 1

Estimated order of accuracy (r_h) *and absolute error* ($||e_h||$) *for the* ρu*-component of the solution of test problem* 1*, computed by the standard method.*

	Upstream		Downstream	
h	r_h	$\lVert e_h\rVert$	r_h	$\lVert e_h\rVert$
$1 \cdot 10^{-2}$		$1.5 \cdot 10^{-3}$		$3.1 \cdot 10^{-2}$
$5 \cdot 10^{-3}$	3.92	$1.0 \cdot 10^{-4}$	2.17	$6.9 \cdot 10^{-3}$
$2.5 \cdot 10^{-3}$	2.38	$1.9 \cdot 10^{-5}$	2.15	$1.6 \cdot 10^{-3}$
$1.25 \cdot 10^{-3}$	2.00	$4.6 \cdot 10^{-6}$	1.46	$5.7 \cdot 10^{-4}$
$6.25 \cdot 10^{-4}$	2.00	$1.2 \cdot 10^{-6}$	1.20	$2.5 \cdot 10^{-4}$
$3.125 \cdot 10^{-4}$	2.00	$2.9 \cdot 10^{-7}$	1.08	$1.2 \cdot 10^{-4}$
$1.5625 \cdot 10^{-4}$	2.00	$7.2 \cdot 10^{-8}$	1.03	$5.7 \cdot 10^{-5}$

Table 2

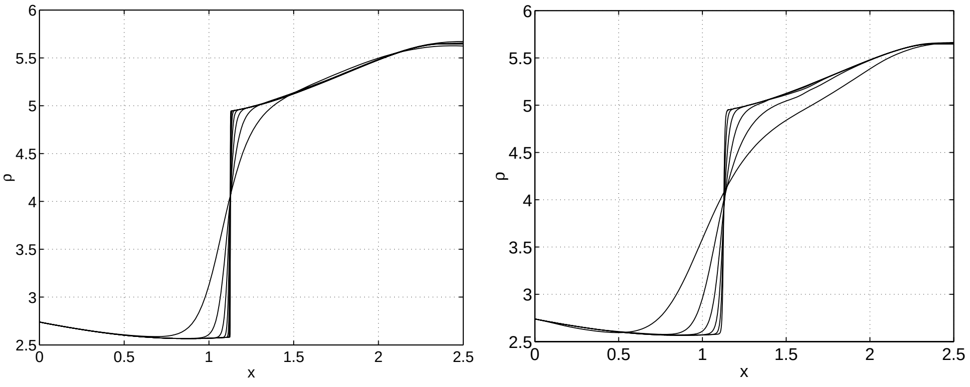
Estimated order of accuracy (r_h) *and absolute error* ($||e_h||$) *for the* ρu*-component of the solution of test problem* 1*, computed by the matrix viscosity method.*

	Upstream		Downstream	
h	r_h	$\lVert e_h\rVert$	r_h	$\lVert e_h\rVert$
$1 \cdot 10^{-2}$		$6.3 \cdot 10^{-3}$		$1.8 \cdot 10^{-2}$
$5 \cdot 10^{-3}$	4.71	$2.4 \cdot 10^{-4}$	1.93	$4.8 \cdot 10^{-3}$
$2.5 \cdot 10^{-3}$	3.68	$1.9 \cdot 10^{-5}$	1.96	$1.2 \cdot 10^{-3}$
$1.25 \cdot 10^{-3}$	2.01	$4.6 \cdot 10^{-6}$	2.00	$3.1 \cdot 10^{-4}$
$6.25 \cdot 10^{-4}$	2.00	$1.2 \cdot 10^{-6}$	1.99	$7.7 \cdot 10^{-5}$

shock. The matrix viscosity method is second order accurate, both upstream and downstream of the shock; see Table 2.

In Figure 3 we see an overview of how the ρ-component of the solution converges, and in Figure 4 we see a close-up. Note that the aim when designing the matrix viscosity method was to avoid the first order error outside the shock region. Hence, the matrix viscosity method performs better than the standard method for fine grids, where the first order downstream error destroys the convergence rate of the standard method. For coarse grids, however, the matrix viscosity method is not better.

If the order of accuracy is r, then the error in the ρu-component of the solution



(a) Standard method

(b) Matrix viscosity method

FIG. 3. *Overview of the convergence of test problem* 1. *In the plots we see the* ρ*-component of the solution. In both cases, the most viscous solution is computed using* $h = 0.02$. *Additional solutions are computed using successively halved space step. For the standard method we see eight different solutions, and for the matrix viscosity method six solutions.*

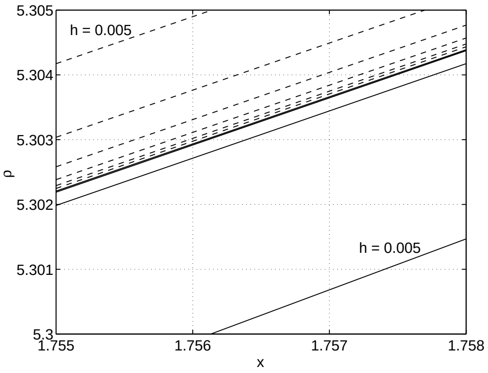


FIG. 4. *Close-up of the convergence of the* ρ*-component for test problem* 1. *Solid lines: the matrix viscosity method; dashed lines: the standard method. For both methods, the two coarsest solutions are not seen in the close-up. As* h *is successively halved, the solutions from the matrix viscosity method increase and the solutions from the standard method decrease.*

can be estimated:

$$||e_h|| = \frac{1}{2^r - 1}||u_{2h} - u_h||.$$

Hence, $||e_h||$ is computed with $r = 2$ in the upstream region for the standard method, and both upstream and downstream for the matrix method. For the standard method downstream of the shock, we have used $r = 1$. Again we use the discrete L_2-norm, on the same interval as above.

Corresponding computations for the ρ- and E-components of the solution give qualitatively the same result.

By plotting the shock profile in phase space (see Figure 5), we see that the shock

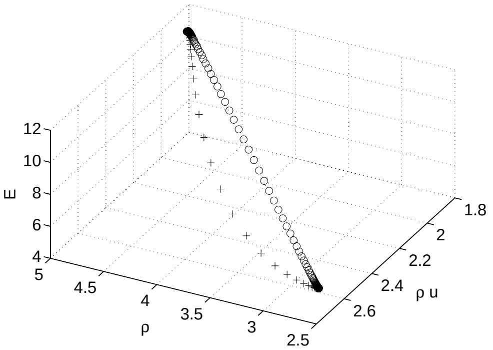


FIG. 5. *Numerical phase diagram of the shock profile computed by the matrix viscosity method* (o) *and the standard method* (+). *Both solutions are computed using* $h = 6.25 \cdot 10^{-4}$. *The shock profile computed by the matrix viscosity method follows a straight line in phase space quite closely.*

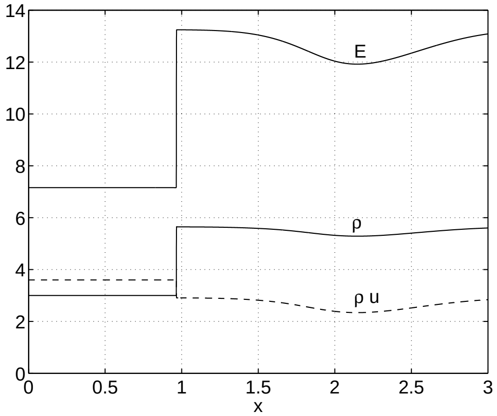


FIG. 6. *The solution of test problem* 2 *at* $t = 1.9$. *The solution is computed numerically using the standard method with* $h = 1.5625 \cdot 10^{-4}$.

profile obtained by the matrix viscosity method approximately follows a straight line between the shock states. The shock profile of the standard method clearly has another shape.

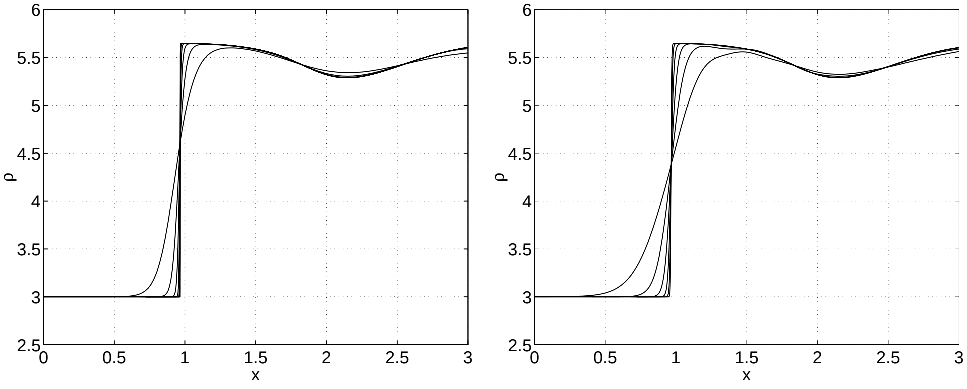
Next, consider test problem 2. In Figure 6, we see the solution of test problem 2 at $t = 1.9$. Test problem 2 was also solved with successively halved space step h with both methods, starting with $h = 0.02$. Again, in all we computed eight solutions with the standard method and six solutions with the matrix viscosity method. In Table 3, we see the estimated order of accuracy for the standard method. Again, we have used the interval $(0, 0.7)$ in the upstream region. In the downstream region the discrete L_2-norm was computed on the interval $(1.1, 3)$. Upstream, the solution is second order. Downstream the convergence is slower, and the order of accuracy is slowly approaching one. In Table 4 we see that the matrix viscosity method is second order accurate both upstream and downstream. In Figure 7 we see an overview of how the ρ-component converges, and in Figure 8 we see a close-up. In phase space, the shock profiles of the solutions of test problem 2 are qualitatively the same as in Figure 5.

TABLE 3
Estimated order of accuracy (r_h) and absolute error ($||e_h||$) for the ρu-component of the solution of test problem 2, computed by the standard method.

	Upstream		Downstream	
h	r_h	$\|\|e_h\|\|$	r_h	$\|\|e_h\|\|$
$1 \cdot 10^{-2}$		$6.2 \cdot 10^{-4}$		$5.0 \cdot 10^{-2}$
$5 \cdot 10^{-3}$	7.93	$2.6 \cdot 10^{-6}$	1.95	$1.3 \cdot 10^{-2}$
$2.5 \cdot 10^{-3}$	2.58	$4.3 \cdot 10^{-7}$	1.96	$3.3 \cdot 10^{-3}$
$1.25 \cdot 10^{-3}$	2.33	$8.5 \cdot 10^{-8}$	1.91	$8.8 \cdot 10^{-4}$
$6.25 \cdot 10^{-4}$	2.01	$2.1 \cdot 10^{-8}$	1.88	$2.4 \cdot 10^{-4}$
$3.125 \cdot 10^{-4}$	2.00	$5.3 \cdot 10^{-9}$	1.69	$7.4 \cdot 10^{-5}$
$1.5625 \cdot 10^{-4}$	2.00	$1.3 \cdot 10^{-9}$	1.39	$2.8 \cdot 10^{-5}$

TABLE 4
Estimated order of accuracy (r_h) and absolute error ($||e_h||$) for the ρu-component of the solution of test problem 2, computed by the matrix viscosity method.

	Upstream		Downstream	
h	r_h	$\|\|e_h\|\|$	r_h	$\|\|e_h\|\|$
$1 \cdot 10^{-2}$		$1.5 \cdot 10^{-3}$		$2.1 \cdot 10^{-2}$
$5 \cdot 10^{-3}$	4.24	$7.9 \cdot 10^{-5}$	2.06	$4.9 \cdot 10^{-3}$
$2.5 \cdot 10^{-3}$	8.01	$3.1 \cdot 10^{-7}$	2.33	$9.8 \cdot 10^{-4}$
$1.25 \cdot 10^{-3}$	1.86	$8.5 \cdot 10^{-8}$	2.23	$2.1 \cdot 10^{-4}$
$6.25 \cdot 10^{-4}$	2.01	$2.1 \cdot 10^{-8}$	2.00	$5.2 \cdot 10^{-5}$



(a) Standard method

(b) Matrix viscosity method

FIG. 7. *Overview of the convergence of test problem 2. We see the ρ-component of the solution. In both cases, the most viscous solution is computed using $h = 0.02$. Additional solutions are computed using successively halved space step. For the standard method we see eight different solutions, and for the matrix viscosity method six solutions.*

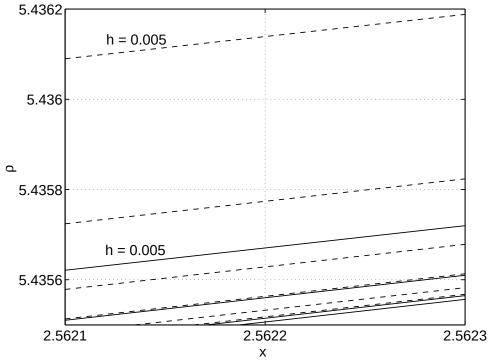


FIG. 8. *Close-up of the convergence of the ρ-component for test problem* 2. *Solid lines: the matrix viscosity method; dashed lines: the standard method. For both methods, the two coarsest solutions are not seen in the close-up. As h is successively halved, the solutions from both methods decrease.*

REFERENCES

[1] M. Arora and P. L. Roe, *On postshock oscillations due to capturing schemes in unsteady flows*, J. Comput. Phys., 130 (1997), pp. 25–40.

[2] J. Casper and M. H. Carpenter, *Computational considerations for the simulation of shock-induced sound*, SIAM J. Sci. Comput., 19 (1998), pp. 813–828.

[3] G. Efraimsson and G. Kreiss, *Approximate solutions to slightly viscous conservation laws*, Quart. Appl. Math., to appear.

[4] G. Efraimsson, J. Nordström, and G. Kreiss, *Artificial Dissipation and Accuracy Downstream of Slightly Viscous Shocks*, Paper 2001-2608, American Institute of Aeronautics and Astronautics, Reston, VA, 2001.

[5] B. Engquist and B. Sjögreen, *The convergence rate of finite difference schemes in the presence of shocks*, SIAM J. Numer. Anal., 35 (1998), pp. 2464–2485.

[6] J. Goodman and A. Majda, *The validity of the modified equation for nonlinear shock waves*, J. Comput. Phys., 58 (1985), pp. 336–348.

[7] S. Karni and S. Čanić, *Computations of slowly moving shocks*, J. Comput. Phys., 136 (1997), pp. 132–139.

[8] G. Kreiss, G. Efraimsson, and J. Nordström, *Elimination of first order errors in shock calculations*, SIAM J. Numer. Anal., 38 (2001), pp. 1986–1998.

[9] H.-O. Kreiss and J. Lorenz, *Initial-Boundary Value Problems and the Navier–Stokes Equations*, Academic Press, New York, 1989.

[10] P. A. Lagerstrom, *Matched Asymptotic Expansions*, Springer-Verlag, New York, 1988.

[11] R. J. LeVeque, *Numerical Methods for Conservation Laws*, Birkhäuser Verlag, Basel, Switzerland, 1992.

[12] A. Szepessy and Z. Xin, *Nonlinear stability of viscous shock waves*, Arch. Ration. Mech. Anal., 122 (1993), pp. 53–103.

SIAM J. Numer. Anal.
Vol. 41, No. 6, pp. 2149–2160

POINTWISE ERROR ESTIMATES FOR DIFFERENCES IN PIECEWISE LINEAR FINITE ELEMENT APPROXIMATIONS*

ALFRED H. SCHATZ[†] AND LARS B. WAHLBIN[†]

Abstract. We consider piecewise linear finite element approximations u_h to u the solution of an elliptic boundary value problem. New estimates for the differences $|e(x_1) - e(x_2)|$ (where $e(x) = u(x) - u_h(x)$ is the error and x_1 and x_2 are any two points in the domain) are obtained in terms of weighted L_∞ norms. As a consequence, so-called asymptotic expansion inequalities are derived that have been applied to obtain asymptotically exact a posteriori estimators for the gradient on each element.

Key words. finite element methods, pointwise estimates, boundary value problems

AMS subject classifications. 65N30, 65N15

DOI. 10.1137/S0036142902405928

1. Introduction and statement of results.

The aim of this paper is to provide both local and global estimates for differences in the error at any two points in a domain when the finite element method is used with continuous piecewise linear functions to approximate solutions of second order elliptic boundary value problems. The results of this paper rely on, and can be viewed as extensions of, the results given in Schatz [6] and [7], which play an essential role in their derivation. An important feature of all of these pointwise estimates for the error is that they are bounded in terms of weighted L_∞ norms that sharply localize the dependence of the error on the solution. Using this feature, the results of [6] were applied in Hoffmann et al. [4] to obtain local asymptotically exact a posteriori estimators for quadratic or higher order finite elements. The results of this paper have been applied in [9] to obtain local asymptotically exact a posteriori estimators in the case of linear finite elements.

An outline of this paper is as follows: In section 1.1 we consider a global Neumann problem. The main result for this case is stated in Theorem 1. An important corollary is then given, namely, Corollary 1, which contains a so-called asymptotic error expansion inequality. It is this result that is used in [9]. Section 1.2 is concerned with local estimates. The main result is given in Theorem 2, and the corresponding local asymptotic error expansion inequality is given in Corollary 2. Section 2 contains some preliminaries needed for the proofs of Theorems 1 and 2. Section 3 contains a proof of Theorem 1 and section 4 a proof of Theorem 2.

1.1. A global Neumann problem.

Let Ω be a bounded domain in $\mathbb{R}^N$, $N \geq 2$, with smooth boundary $\partial\Omega$, and consider the Neumann problem with a homogeneous conormal boundary condition,

$$\mathcal{L}u = -\sum_{i,j=1}^{N} \frac{\partial}{\partial x_i}\Big(a_{ij}(x)\frac{\partial u}{\partial x_j}\Big) + \sum_{i=1}^{N} b_i(x)\frac{\partial u}{\partial x_i} + c(x)u = f(x) \quad \text{in } \Omega, \tag{1.1}$$

[*] Received by the editors April 17, 2002; accepted for publication (in revised form) April 22, 2003; published electronically November 25, 2003. This research was supported by the National Science Foundation, USA, grant DMS-0071412.

http://www.siam.org/journals/sinum/41-6/40592.html

[†] Department of Mathematics, Cornell University, Ithaca, NY 14853 (schatz@math.cornell.edu, wahlbin@math.cornell.edu).

$$\frac{\partial u}{\partial n_{\mathcal{L}}} = \sum_{i,j=1}^{N} \Big(a_{ij} \frac{\partial u}{\partial x_j} \Big) n_i = 0 \quad \text{on } \partial\Omega. \tag{1.2}$$

We shall assume that the coefficients are smooth and that $\mathcal{L}$ is uniformly elliptic in Ω; i.e., there exists a constant $c_{\text{ell}} > 0$ such that

$$c_{\text{ell}} |\zeta|^2 \le \sum_{i,j=1}^{N} a_{ij} \zeta_i \zeta_j \quad \text{for all } \zeta \in \mathbb{R}^N. \tag{1.3}$$

Let $u \in W_2^1(\Omega)$ be a weak solution of (1.1), (1.2) satisfying

$$\begin{aligned} A(v,v) &= \int_\Omega \left(\sum_{i,j=1}^{N} a_{ij}(x) \frac{\partial u}{\partial x_i} \, \frac{\partial v}{\partial x_j} + \sum_{i=1}^{N} b_i(x) \frac{\partial u}{\partial x_i} v + c(x) uv dx \right) \\ &= \int_\Omega fv dx = (f,v) \quad \text{for all } v \in W_2^1(\Omega). \end{aligned} \tag{1.4}$$

Here for simplicity it will be assumed that $A(\cdot,\cdot)$ is coercive; i.e., for some $c_{co} > 0$

$$c_{co} \|u\|^2_{W_2^1(\Omega)} \le A(u,u). \tag{1.5}$$

Now consider the approximation of u using the finite element method. For $0 < h < 1$, let S_h be a family of subspaces of $W_\infty^1(\Omega)$, each S_h consisting of continuous piecewise linear functions defined on a quasi-uniform triangulation by simplexes of roughly size h that fit $\partial\Omega$ exactly. Thus curved faces are allowed at the boundary. By a quasi-uniform triangulation of roughly size h we mean that there exist constants $0 < c_* < c^*$ such that each simplex τ contains a ball of radius $c_* h$ and is contained in a ball of radius $c^* h$, where c_* and c^* are independent of τ and h. Now let $u_h \in S_h$ be the finite element approximation of u defined by

$$A(u_h, \varphi) = (f, \varphi) \quad \text{for all } \varphi \in S_h \tag{1.6}$$

so that

$$A(u - u_h, \varphi) = 0 \quad \text{for all } \varphi \in S_h. \tag{1.7}$$

We are interested in estimating the differences $e(x_2) - e(x_1)$, where x_1 and x_2 are arbitrary points in $\overline{\Omega}$, and $e(x) = u(x) - u_h(x)$. Our estimates will be given in terms of weighted L_∞ norms that we will now describe. Set $\rho = |x_2 - x_1|$, $\overline{x} = \frac{x_1 + x_2}{2}$, and for any real s

$$\sigma_{\overline{x}}^s(y) = \left(\frac{\max(h,\rho)}{|\overline{x} - y| + \max(h,\rho)} \right)^s. \tag{1.8}$$

For $p = 1$ or ∞, define the weighted norms

$$\|u\|_{W_p^1(\Omega),\overline{x},s} = \|\sigma_{\overline{x}}^s(y) u(y)\|_{L_p(\Omega)} + \|\sigma_{\overline{x}}^s(y) \nabla u(y)\|_{L_p(\Omega)}. \tag{1.9}$$

Our first result is as follows.

Theorem 1. *Let u and u_h satisfy* (1.7). *There exist positive constants C and k such that if x_1 and x_2 are any two points in $\overline{\Omega}$, then the following hold:*

(i) *If* $\rho \leq kh$ *and* $0 \leq s \leq 1$,

$$|e(x_2) - e(x_1)| \leq C\rho\Big(\ln\frac{1}{h}\Big)^{\overline{s}}\Big(\min_{\chi \in S^h}\|u-\chi\|_{W^1_\infty(\Omega),\overline{x},s}\Big). \tag{1.10}$$

Here, $\overline{s} = 1$ *if* $s = 1$, *and* $\overline{s} = 0$ *otherwise.*

(ii) *If* $\rho \geq kh$ *and* $0 \leq s < 1$,

$$|e(x_2) - e(x_1)| \leq Ch\Big(\ln\frac{\rho}{h}\Big)\Big(\inf_{\chi \in S^h}\|u-\chi\|_{W^1_\infty(\Omega),\overline{x},s}\Big). \tag{1.11}$$

In (1.10) and (1.11), C depends on c_{ell}, c_{c_0}, s, Ω, and the maximum norm of the coefficients of $\mathcal{L}$ and sufficiently many of their derivatives. It is independent of u, u_h, h, x_1, and x_2 (and hence also of ρ).

Remark 1.1. The case $s = 1$ has been excluded from the estimate (1.11). An estimate for this case easily follows from the proof given in section 3 but involves an additional logarithmic factor. We have not included this estimate since it is not the sharpest possible. However, the proof of a sharper result would require a much longer paper: essentially, we would not *use* the results of [6] but *redo* them with appropriate modifications.

An immediate consequence of Theorem 1 are so-called asymptotic error expansion inequalities.

Corollary 1. *Suppose that Theorem* 1 *holds, and in addition, let* $\varepsilon > 0$ *be arbitrary but fixed. Then the following hold:*

(i) *If* $\rho \leq kh$,

$$|e(x_2) - e(x_1)| \leq C_\varepsilon \rho h\left(\sum_{|\alpha|=2}|D^\alpha u(\overline{x})| + h^{1-\varepsilon}\|u\|_{W^3_\infty(\Omega)}\right). \tag{1.12}$$

(ii) *If* $\rho \geq kh$,

$$|e(x_2) - e(x_1)| \leq C_\varepsilon h^2\Big(\ln\frac{\rho}{h}\Big)\left(\sum_{|\alpha|=2}|D^\alpha u(\overline{x})| + \rho^{1-\varepsilon}\|u\|_{W^3_\infty(\Omega)}\right). \tag{1.13}$$

Corollary 1 has been applied to the problem of a posteriori estimates in [9]. The inequalities (1.12) and (1.13) follow from (1.10) and (1.11), respectively, by using standard approximation properties of piecewise linear functions and Taylor's theorem.

Remark 1.2. In the case of quadratic or higher elements, $r = 3, 4, \dots$, we have from [6] that

$$|e(x)| \leq C_\varepsilon h^r\left(\sum_{|\alpha|=r}|D^\alpha u(x)| + h^{1-\varepsilon}\|u\|_{W^{r+1}_\infty(\Omega)}\right).$$

However, Demlow [2] has shown that such an estimate is impossible in the piecewise linear case, $r = 2$ (even allowing for logarithmic factors). In [9], Corollary 1 above is a substitute estimate, strong enough to give the desired result.

1.2. Local estimates. We shall next describe some results that are local analogues of the estimates (1.12) and (1.13). Toward this end, for $d > 0$, let $B_{2d}(x_0)$ denote the ball of radius $2d$ centered at x_0. Furthermore, let $S^h(B_{2d}(x_0))$, $0 < h < 1$, be a family of continuous piecewise linear elements defined on quasi-uniform triangulations of roughly size h that cover $B_{2d}(x_0)$. It will be assumed that $kh < d$ for some fixed k sufficiently large. We wish to estimate the error $|e(x_2) - e(x_1)|$ for any two points $x_1, x_2 \in B_d(x_0)$. Here $u_h \in S^h(B_{2d}(x_0))$, and $e(x) = u(x) - u_h(x)$ satisfies the local equations

$$\begin{aligned} A(e,\varphi) = \int_{B_{2d}(x_0)} \Bigg(\sum_{i,j=1}^{N} a_{ij}(x) \frac{\partial e}{\partial x_i}\, \frac{\partial \varphi}{\partial x_j} \\ + \sum_{i=1}^{N} b_i(x) \frac{\partial e}{\partial x_i} \varphi + c(x) e \varphi \Bigg) dx = 0 \\ \text{for all } \varphi \in \overset{\circ}{S}{}^h(B_{2d}(x_0)), \end{aligned} \tag{1.14}$$

where $\overset{\circ}{S}{}^h(B_{2d}(x_0))$ denotes the subspace of $S^h(B_{2d}(x_0))$ of functions whose support is contained in $B_{2d}(x_0)$. Our main local result is as follows.

THEOREM 2. *Let u and u_h satisfy* (1.14), *let $1 \le p \le \infty$, and let t be a nonnegative integer. There exist positive constants C and k such that if x_1 and x_2 are any two points in $\overline{B}_d(x_0)$, $\rho = |x_2 - x_1|$, $\overline{x} = \frac{x_1+x_2}{2}$, $d \ge \max(\rho, kh)$, then the following hold:*

(i) *If $\rho \le kh$ and $0 \le s \le 1$,*

$$\begin{aligned} |e(x_2) - e(x_1)| \le C\rho \Bigg(\Big(\ln \frac{1}{h} \Big)^{\overline{s}} \min_{\chi \in S^h} \|u - \chi\|_{W^1_\infty(B_{2d}(x_0)), \overline{x}, s} \\ + d^{-t-1-N/\rho} \|e\|_{W_p^{-t}(B_{2d}(x_0))} \Bigg), \end{aligned} \tag{1.15}$$

where $\overline{s} = 1$ if $s = 1$, and $\overline{s} = 0$ otherwise.

(ii) *If $\rho \ge kh$ and $0 \le s < 1$,*

$$\begin{aligned} |e(x_2) - e(x_1)| \le & Ch \ln \Big(\frac{\rho}{h} \Big) \Big(\min_{\chi \in S^h} \|u - \chi\|_{W^1_\infty(B_{2d}(x_0)), \overline{x}, s} \Big) \\ & + C\rho d^{-t-1-N/p} \|e\|_{W_p^{-t}(B_{2d}(x_0))}. \end{aligned} \tag{1.16}$$

In (1.15) and (1.16), C depends on c_{ell}, c_{co}, s, t, p, and the maximum norms of the coefficients and sufficiently many derivatives. It is independent of u, u_h, h, x_1, x_2 (hence also of ρ), and d. Furthermore, for any open set G and $\frac{1}{p} + \frac{1}{q} = 1$,

$$\|u\|_{W_p^{-t}(G)} = \sup \int_G u \eta dx. \tag{1.17}$$

As a consequence of Theorem 2 we have the following local "asymptotic expansion inequalities."

COROLLARY 2. *Suppose that Theorem 2 holds, and in addition, let $\varepsilon > 0$ be arbitrary but fixed. Then, for all x_1 and x_2 in $B_d(x_0)$, the following hold:*

(i) *If $\rho \le kh$,*

$$\begin{aligned} |e(x_2) - e(x_1)| \le C_\varepsilon \rho h \Bigg(\sum_{|\alpha|=2} |D^\alpha u(\overline{x})| + h^{1-\varepsilon} \|u\|_{W^3_\infty(B_{2d}(x_0))} \Bigg) \\ + C d^{-t-1N/p} \|e\|_{W_p^{-t}(B_{2d}(x_0))}. \end{aligned} \tag{1.18}$$

(ii) *If $\rho \geq kh$,*

$$|e(x_2) - e(x_1)| \leq C_\varepsilon h^2 \ln\left(\frac{\rho}{h}\right)\left[\left(\sum_{|\alpha|=2} |D^\alpha u(\overline{x})| + \rho^{1-\varepsilon}\|u\|_{W^3_\infty(B_{2d}(x_0))}\right)\right] + C\rho d^{-t-1-N/p}\|e\|_{W_p^{-t}(B_{2d}(x_0))}. \tag{1.19}$$

2. Preliminaries. Here we shall collect some known results that will be essential in our proofs of Theorems 1 and 2. Lemma 2.1 is concerned with a special case of estimates derived in [6] for the gradient at a point for the global Neumann problem. Lemma 2.2 deals with analogous local estimates given in [7]. Lemma 2.3 is concerned with some pointwise estimates given in Krasovskii [5] for the Green's function for the problem (1.1), (1.2).

LEMMA 2.1. *Let Ω be a bounded domain in $\mathbb{R}^N$ with a smooth boundary $\partial\Omega$, and let $S^h(\Omega)$ be the continuous piecewise linear functions defined on a globally quasi-uniform mesh of roughly size h, $0 < h < 1$, that fits the boundary exactly. Let $u \in W^1_\infty(\Omega)$ and $u_h \in S^h(\Omega)$ satisfy* (1.7), *and suppose* $0 \leq s \leq 1$. *Then there exists a constant C such that if $x \in \overline{\Omega}$ is arbitrary,*

$$|e(x)| + |\nabla e(x)| \leq C\Big(\ln\frac{1}{h}\Big)^{\overline{s}}\Big(\min_{\chi\in S^h}\|u-\chi\|_{W^1_\infty(\Omega),x,s}\Big). \tag{2.1}$$

Here $\overline{s} = 1$ if $s = 1$, and $\overline{s} = 0$ otherwise. C depends on c_{ell}, c_{co}, s, Ω, and the maximum norm of the coefficients of $\mathcal{L}$ and sufficiently many of their derivatives. It is independent of u, u_h, h, and x.

Remark 2.1. At its points of discontinuity, $\nabla u_h(x)$ is to be interpreted as the limit from inside any simplex for which x is a boundary point.

We shall need a local analogue of Lemma 2.1. In Lemma 2.2 we shall use the notation of section 1.2.

LEMMA 2.2. *Let $u \in W^1_\infty(B_{2d}(x_0))$ and $u_h \in S^h(B_{2d}(x_0))$ satisfy* (1.14). *There exists a constant c such that if $x \in \overline{B_d(x_0)}$ is arbitrary, then if* $0 \leq s \leq 1$,

$$|e(x)| + |\nabla e(x)| \leq C\Big(\Big(\ln\frac{1}{h}\Big)^{\overline{s}}\min_{\chi\in S^h(B_{2d}(x_0))}\|u-\chi\|_{W^1_\infty(B_{2d}(x_0)),x,s} + d^{-t-1-N/p}\|e\|_{W_p^{-t}(B_{2d}(x_0))}\Big). \tag{2.2}$$

Here $\overline{s} = 1$ if $s = 1$, and $\overline{s} = 0$ otherwise. C depends on c_{ell}, c_{co}, s, and the maximum norm of the coefficients of $\mathcal{L}$ and sufficiently many of their derivatives. It is independent of u, u_h, h, d, and x.

Let $G^x(y)$ denote the Green's function for the problem (1.1), (1.2) with singularity at x.

LEMMA 2.3. *There exists a constant C such that if $x, y \in \overline{\Omega}$,*

$$|D_x^\alpha D_y^\beta G^x(y)| \leq C|x-y|^{2-N-|\alpha+\beta|} \text{ for } |\alpha+\beta| > 0. \tag{2.3}$$

Here C depends on c_{co}, c_{ell}, and various norms of the coefficients of $\mathcal{L}$.

3. A proof of Theorem 1, global estimates. For simplicity of proof, we will assume that Ω is convex. We start by proving (1.10).

In this case $\rho \le kh$, and by the fundamental theorem of calculus, there exists a point $\widehat{x}$ on the line joining x_1 and x_2 such that, in view of (2.1),

$$|e(x_2) - e(x_1)| \le \rho |\nabla e(\widehat{x})| \le c\rho \Big(\ln \frac{1}{h} \Big)^{\overline{s}} \min_{\chi \in S^h} \|u - \chi\|_{W^1_\infty(\Omega), \widehat{x}, s}. \tag{3.1}$$

We next note that since the function $\lambda \to \frac{\lambda}{A+\lambda}$, $A > 0$, is an increasing function for $\lambda > 0$, we have the following string of elementary inequalities:

$$\begin{aligned} \sigma_{\widehat{x}}(y) = \frac{\max(\rho, h)}{|\widehat{x} - y| + \max(\rho, h)} &\le \frac{2\max(\rho, h)}{|\overline{x} - y| - |\widehat{x} - \overline{x}| + 2\max(\rho, h)} \\ &\le \frac{2\max(\rho, h)}{|\overline{x} - y| + \max(\rho, h)} = 2\sigma_{\overline{x}}(y). \end{aligned} \tag{3.2}$$

In the next to last step we have used that $|\widehat{x} - \overline{x}| \le \rho$. It follows from (3.2) and the definition (1.9) that

$$\|v\|_{W^1_\infty(\Omega), \widehat{x}, s} \le 2\|v\|_{W^1_\infty(\Omega), \overline{x}, s} \text{ for all } v \in W^1_\infty.$$

Using this in (3.1), the inequality (1.10) follows (the case $\rho \le kh$).

We now turn to a proof of (1.11) (the case $\rho \ge kh$). We can then write

$$e(x_2) - e(x_1) = A(e(y), G^{x_2} - G^{x_1}), \tag{3.3}$$

where $G^x(y)$ is the Green's function for the adjoint problem with singularity at x,

$$\mathcal{L}^* G^x(y) = \delta^x(y) \text{ in } \Omega, \ \frac{\partial G^x}{\partial n_{\mathcal{L}^*}} = 0 \text{ on } \partial\Omega. \tag{3.4}$$

Hence for any $\psi \in S^h(\Omega)$

$$\begin{aligned} |e(x_2) - e(x_1)| &\le |A(e, G^{x_2} - G^{x_1} - \psi)| \\ &\le C\Big(\|e\|_{W^1_\infty(\Omega), \overline{x}, s} \|G^{x_2} - G^{x_1} - \psi\|_{W^1_1(\Omega), \overline{x}, -s} \Big). \end{aligned} \tag{3.5}$$

The remainder of this section will be devoted to showing that, for $0 \le s < 1$ and $\rho \ge kh$,

$$\|e\|_{W^1_\infty(\Omega), \overline{x}, s} \le C \inf_{\chi \in S^h} \|u - \chi\|_{W^1_\infty(\Omega), \overline{x}, s} \tag{3.6}$$

and that there exists a $\psi \in S^h$ so that

$$\|G^{x_2} - G^{x_1} - \psi\|_{W^1_1(\Omega), \overline{x}, -s} \le Ch \Big(\ln \frac{\rho}{h} \Big), \tag{3.7}$$

where C is as in Theorem 1. Granting the last two inequalities for a moment, the proof of (1.11) follows from (3.5), (3.6), and (3.7), which would complete the proof of Theorem 1.

We now turn to the proof of (3.6). In view of (2.1), we have for any $z \in \overline{\Omega}$

$$|e(z)| + |\nabla e(z)| \le C \inf_{\chi \in S^h} \Big(\|\sigma^s_z(y)(u - \chi)(y)\|_{L_\infty(\Omega)} + \|\sigma^s_z(y)(\nabla(u - \chi))(y)\|_{L_\infty(\Omega)} \Big).$$

Now multiplying both sides by $\sigma^s_{\overline{x}}(z)$, we obtain

$$\begin{aligned}\sigma^s_{\overline{x}}(z)|e(z)| &+ \sigma^s_{\overline{x}}(z)|\nabla e(z)| \\ &\le C \inf_{\chi\in S^h}\Big(\|\sigma^s_{\overline{x}}(z)\sigma^s_z(y)(u-\chi)(y)\|_{L_\infty(\Omega)} \\ &\qquad + \|\sigma^s_{\overline{x}}(z)\sigma^s_z(y)(\nabla(u-\chi))(y)\|_{L_\infty(\Omega)}\Big).\end{aligned} \tag{3.8}$$

For the product of weights in the right-hand side, we have

$$\begin{aligned}\sigma_{\overline{x}}(z)\sigma_z(y) &\equiv \Big(\frac{\rho}{|\overline{x}-z|+\rho}\Big)\Big(\frac{\rho}{|z-y|+\rho}\Big) \le \Big(\frac{\rho}{|\overline{x}-z|+\rho}\Big)\Big(\frac{|\overline{x}-z|+\rho}{|z-y|+|\overline{x}-z|+\rho}\Big) \\ &= \frac{\rho}{|\overline{x}-z|+|z-y|+\rho} \le \frac{\rho}{|\overline{x}-y|+\rho} \equiv \sigma_{\overline{x}}(y).\end{aligned}$$

Using this in the right-hand side of (3.8), which then becomes independent of z, and taking the maximum norm of the left-hand side in z, we arrive at (3.6).

It remains to prove (3.7). For $d > 0$ and $x \in \Omega$, let

$$M_d(x) = B_d(x) \cap \Omega = \{y \in \Omega : |y-x| < d\}.$$

Next, for $i = 1, 2$, define $\psi_i \in S^h$ by

$$\psi_i(y) = (G^{x_i}(y))_{\mathrm{Int}}, \tag{3.9}$$

where the interpolation operator is of any standard type involving local averaging; see, e.g., Clément [1], Hilbert [3], or Scott and Zhang [10]. Set

$$\psi = \psi_1 - \psi_2.$$

Using the triangle inequality (recall that $\rho = |x_1 - x_2|$ and $\overline{x} = (x_1 + x_2)/2$), we have

$$\begin{aligned}&\|G^{x_2} - G^{x_1} - \psi\|_{W^1_1(\Omega),\overline{x},-s} \\ &\quad\le 4^s \sum_{i=1}^{2} \|G^{x_i} - \psi_i\|_{W^1_1(M_{4\rho}(x_i))} + \|G^{x_2} - G^{x_1} - \psi\|_{W^1_1(\Omega/M_{3\rho}(\overline{x})),\overline{x},-s} \\ &\quad= I_1 + I_2.\end{aligned} \tag{3.10}$$

We begin by estimating each term in I_1. Now

$$\begin{aligned}&\|G^{x_i} - \psi_i\|_{W^1_1(M_{4\rho}(x_i))} \\ &\quad\le \|G^{x_i} - \psi_i\|_{W^1_1(M_{2kh}(x_i))} + \|G^{x_i} - \psi_i\|_{W^1_1(M_{4\rho}(x_i)/M_{2kh}(x_i))}.\end{aligned} \tag{3.11}$$

Using stability properties of the ψ_i for the first term on the right, we have

$$\begin{aligned}\|G^{x_i} - \psi_i\|_{W^1_1(M_{2kh}(x_i))} &\le C\|G^{x_i}\|_{W^1_1(M_{3kh}(x_i))} \\ &\le C\int_{0\le r\le 3kh} \frac{r^{N-1}}{r^{N-1}}dr \le Ch.\end{aligned} \tag{3.12}$$

Here we have set $r = |x - x_i|$ and used the bounds (2.3) for the Green's function. Now using approximation properties of the ψ_i, and (3.12) in (3.11),

$$\begin{aligned}&\|G^{x_i} - \psi_i\|_{W^1_1(M_{4\rho}(x_i)/M_{2kh}(x_i))} \\ &\quad\le Ch\|G^{x_i}\|_{W^2_1(M_{5\rho}(x_i)/M_{kh}(x_i))} \\ &\quad\le Ch\int_{kh\le r\le 5\rho} \frac{r^{N-1}}{r^N}dr \le Ch\ln\Big(\frac{\rho}{h}\Big).\end{aligned} \tag{3.13}$$

Combining (3.13), (3.12), and (3.11), we arrive at

$$I_1 \leq Ch\Big(\ln\frac{\rho}{h}\Big). \tag{3.14}$$

It remains to estimate I_2. For this purpose, let

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

and introduce the "annuli"

$$\begin{aligned}
&\Omega_j = \{y \in \Omega : d_{j+1} < |y - \overline{x}| < d_j\} \text{ for } j = 0, 1, 2, \ldots,\\
&\Omega_j' = \Omega_{j-1} \cup \overline{\Omega}_j \cup \Omega_{j+1} \text{ for } j = 1, 2, \ldots,\\
&\Omega_0' = \overline{\Omega}_0 \cup \Omega_1.
\end{aligned}$$

Also set

$$J = \left[\ln_2\left(\frac{1}{3\rho}\right)\right] + 1.$$

Using the triangle inequality and assuming for convenience that $\Omega \subset M_1(\overline{x})$,

$$\begin{aligned}
I_2 &\leq \sum_{j=0}^{J}\Big(\frac{d_j}{\rho}\Big)^s \|G^{x_2} - G^{x_1} - \psi\|_{W_1^1(\Omega_j)}\\
&\leq C\sum_{j=0}^{J}\frac{d_j^{N+s}}{\rho^s}\|G^{x_2} - G^{x_1} - \psi\|_{W_\infty^1(\Omega_j)}\\
&\leq Ch\sum_{j=0}^{J}\frac{d_j^{N+s}}{\rho^s}|G^{x_2} - G^{x_1}|_{W_\infty^2(\Omega_j')},
\end{aligned} \tag{3.15}$$

where we have used the standard approximation properties of ψ. Next, by the fundamental theorem of calculus and (2.3), we have for any multi-index β with $|\beta| = 2$

$$|D_y^\beta(G^{x_2}(y) - G^{x_1}(y))| \leq \sum_{|\alpha|=1}\int_\Gamma |D_x^\alpha D_y^\beta G^x|dx \leq C\rho d_j^{-N-1}, \tag{3.16}$$

where Γ is the line joining x_1 and x_2. Using (3.16) in (3.15), we arrive at

$$I_2 \leq Ch\sum_{j=0}^{J}\frac{d_j^{N+s}\rho d_j^{-N-1}}{\rho^s} = Ch\sum_{j=0}^{J}\Big(\frac{\rho}{d_j}\Big)^{1-s},$$

or

$$I_2 \leq C(s)h \quad \text{for } 0 \leq s < 1. \tag{3.17}$$

Combining (3.16) and (3.14) into (3.10) leads to (1.11), which completes the proof of Theorem 1.

Remark 3.1. If $s = 1$, then $I_2 \leq Ch\ln\frac{1}{\rho}$. Hence the estimate (1.11) may be extended to the case $s = 1$ but with an additional $\ln\frac{1}{\rho}$ factor; cf. Remark 1.1.

4. A proof of Theorem 2, interior estimates. The proof of interior estimates involves some additional technical difficulties when compared with the proof of Theorem 1. These difficulties can be overcome by well-established techniques developed in [8]. There are some parts where the proof relies on techniques used in the proof of Theorem 1. In order to avoid tedious and repetitive details, we shall restrict ourselves to giving an outline of the proof, except at essentially new points.

Proof. We begin with a proof of (1.15). Proceeding as in the proof of (1.10), we have

$$|e(x_2) - e(x_1)| \leq \rho |\nabla e(\widehat{x})|,$$

where $\widehat{x}$ is a point on the line joining x_1 and x_2. Applying (2.2), we obtain for $0 \leq s \leq 1$

$$\begin{aligned} |e(x_2) - e(x_1)| &\leq C\rho\Big[\Big(\ln\frac{1}{h}\Big)^{\overline{s}}\Big(\min_{\chi\in S^h}\|u-\chi\|_{W^1_\infty(B_d(\widehat{x})),\widehat{x},s}\Big) + d^{-1-t-N/p}\|e\|_{W_p^{-t}(B_d(\widehat{x}))}\Big] \\ &\leq C\rho\Big[\Big(\ln\frac{1}{h}\Big)^{\overline{s}}\Big(\min_{\chi\in S^h}\|u-\chi\|_{W^1_\infty(B_{2d}(x_0)),\overline{x},s}\Big) + d^{-1-t-N/p}\|e\|_{W_p^{-t}(B_{2d}(x_0))}\Big]. \end{aligned}$$

In the last step we used that $|\widehat{x} - \overline{x}| \leq p/2$. This is (1.15).

The proof of (1.16) will be separated into two steps, following the general procedure established in [8], with a slight modification given in [7]. In the first step the case $d = 1$ will be treated, and, in the second step, the result for any $d < 1$ will be reduced to the first case via a standard scaling argument.

Step 1. Let $\widehat{d} > 0$ be any number independent of u and h. Assume we could show, for $x_1, x_2 \in B_{\widehat{d}}(\widehat{x}_0)$, any $\widehat{x}_0 \in B_1(x_0)$, that

$$|e(x_2) - e(x_1)| \leq Ch\ln(\rho/h)\min_{\chi\in S^h}\|u-\chi\|_{W^1_\infty(B_1(\widehat{x}_0)),\overline{x},s} + C\rho\|e\|_{W_p^{-t}(B_1(\widehat{x}_0))}. \tag{4.1}$$

Then clearly (1.16) for $d = 1$ would follow by a covering argument (and inserting points between x_1 and x_2 if $p = |x_2 - x_1| > \widehat{d}$, i.e., when the weight is of no consequence).

For notational simplicity, we shall now let $\widehat{x}_0 = 0$ and write B_d for $B_d(0)$.

It would be convenient if $A(\cdot,\cdot)$ were coercive on B_1, i.e., if

$$c_{\mathrm{co}}\|u\|^2_{W^1_2(B_1)} \leq A(u,u) \qquad \text{for all} \quad u \in W^1_2(B_1).$$

This is not assumed. However, this difficulty may be overcome by modifying the form in the following manner. For $\mu > 0$ and $\widehat{d} > 0$, set

$$A_\mu(u,v) = A(u,v) + \mu\int_{B_1}\widehat{w}(x)u(x)v(x)dx,$$

where $\widehat{w} \in C^\infty$, $\widehat{w} \geq 0$, $\widehat{w} = 0$ for $|x| < 2\widehat{d}$, $\widehat{w} = 1$ for $|x| \geq 3\widehat{d}$. Let $Q = \sum_{i=1}^N \|b_i\|_{L_\infty(B_1)}| + \|c\|_{L_\infty(B_1)}$. From Lemma 2.1 of [7] we know that there do exist $\widehat{d}$ and μ depending only on Q and c_{ell}, such that

$$\frac{c_{\mathrm{ell}}}{4}\|u\|^2_{W^1_2(B_1)} \leq A_\mu(u,u) \text{ for } \quad u \in W^1_2(B_1). \tag{4.2}$$

Note that, by construction,

$$A_\mu(u,v) = A(u,v) \text{ for all } \quad v \in \overset{\circ}{W}{}^1_2(B_{2\widehat{d}}). \tag{4.3}$$

We now begin the proof of (1.16) by slightly modifying the proof of Theorem 1. Let $G^x(y)$ be the Green's function for the Neumann problem

$$\mathcal{L}^* G^x(y) = \delta^x(y) \text{ in } B_1, \ \frac{\partial G^x(y)}{\partial n_{\mathcal{L}^*}} = 0 \text{ on } \partial B_1,$$

where $\mathcal{L}^*$ is the adjoint of $\mathcal{L}_\lambda = \mathcal{L} + \mu \widehat{w} I$. Let $w(y) \in C_0^\infty$ be a cut-off function,

$$w(y) = \begin{cases} 1 & \text{for } |y| \leq \widehat{d}, \\ 0 & \text{for } |y| \geq 3\widehat{d}/2. \end{cases}$$

Then, for $x_1, x_2 \in B_{\widehat{d}}$,

$$e(x_2) - e(x_1) = A_\mu(we, G^{x_2} - G^{x_1}). \tag{4.4}$$

A straightforward but tedious calculation yields

$$A_\mu(we, G^{x_2} - G^{x_1}) = A_\mu(e, w(G^{x_2} - G^{x_1})) + R, \tag{4.5}$$

where, for any $1 \leq p \leq \infty$ and integer $t \geq 0$,

$$|R| \leq C\|e\|_{W_p^{-t}(B_1)} \left(\sum_{|\alpha|=1}^{2} \sum_{|\beta|=0}^{1} \|(D^\alpha w) D^\beta (G^{x_1} - G^{x_2})\|_{W_q^t(B_1 \setminus B_{\widehat{d}})} \right). \tag{4.6}$$

Here we have used the fact that $D^\alpha w$ vanishes on $B_{\widehat{d}}$ for $|\alpha| \geq 1$. Estimating $D_y^\beta(G^{x_2} - G^{x_1})$ as in (3.16), we arrive at

$$|R| \leq C\rho \|e\|_{W_p^{-t}(B_1)}. \tag{4.7}$$

According to (4.5), it remains to estimate $I = A_\mu(e, w(G^{x_2} - G^{x_1})) = A(e, w(G^{x_2} - G^{x_1}))$, by (4.3). Hence, for any $\psi \in \overset{\circ}{S}{}^h(B_{2\widehat{d}})$,

$$I = A(e, w(G^{x_2} - G^{x_1}) - \psi).$$

Analogously to (3.9), let $\psi_i = (wG^{x_i})_{\mathrm{Int}}$, $i = 1, 2$, be a suitable (locally averaged) interpolant of wG^{x_i}, and set $\psi = \psi_2 - \psi_1$. For h small enough, $\psi \in \overset{\circ}{S}{}^h(B_{2\widehat{d}})$ since w is supported in $B_{3\widehat{d}/2}$. Thus

$$|I| \leq C\|e\|_{W_\infty^1(B_{2\widehat{d}}), \overline{x}, s} \|w(G^{x_2} - G^{x_1}) - \psi\|_{W_1^1(B_{2\widehat{d}}), \overline{x}, -s}. \tag{4.8}$$

The last term on the right of (4.8) can be estimated by following the same procedure used in estimating the terms in (3.10) to obtain

$$\|w(G^{x_2} - G^{x_1}) - \psi\|_{W_1^1(B_{2\widehat{d}}), \overline{x}, -s} \leq Ch \ln\left(\frac{\rho}{h}\right) \quad \text{for } 0 \leq s < 1. \tag{4.9}$$

Combining (4.9) with (4.8) and using Lemma 2.2 with $d = 2\widehat{d}$,

$$|I| \leq Ch \ln(\rho/h) \Big(\min_{\chi \in S^h} \|u - \chi\|_{W_1^1(B_1), \overline{x}, s} + \|e\|_{W_p^{-1}(B_1)} \Big).$$

In view of (4.4), (4.5), and (4.7), this gives

$$|e(x_2) - e(x_1)| \le Ch \ln\left(\frac{\rho}{h}\right) \min_{\chi \in S^h} \|u - \chi\|_{W^1_\infty(B_1),\bar{x},s} \\ + C\rho\Big(\frac{h}{\rho}\ln\left(\frac{\rho}{h}\right) + 1\Big)\|e\|_{W_p^{-t}(B_1)}.$$

Since $\rho \ge kh$, where we may assume without loss of generality that $k \ge 2$, we have $|h/\rho \ln(\rho/h)| \le 1/e$. Hence this gives (4.1) and, as already noted, proves (1.16) for $d = 1$.

Step 2. We now turn to a proof of (1.16) in the case $d < 1$. As remarked before, this will be reduced to the case $d = 1$ by a standard scaling argument (see [7], [8]).

For simplicity we shall take x_0 to be the origin. Then the transformation $z = \frac{x}{d}$ maps $B_{2d}(x_0)$ to $B_2(x_0)$. Set $\widehat{u}(z) = u(dz)$, $\widehat{u}_h(z) = u_h(dz)$, and $\widehat{e}(z) = \widehat{u}(z) - \widehat{u}_h(z)$. By a change in variables we have

$$A(e, \chi) \equiv d^{N-2}\widehat{A}(\widehat{e}, \widehat{\chi}) = 0 \text{ for all } \chi \in \overset{\circ}{S}{}^h(B_{2d})$$

or

$$\widehat{A}(\widehat{e}, \widehat{\chi}) = 0 \text{ for all } \widehat{\chi} \in \overset{\circ}{S}{}^{h/d}(B_2), \tag{4.10}$$

where

$$\widehat{A}(\widehat{e}, \widehat{\chi}) = \int_{B_2} \left(\sum_{i,j=1}^N a_{ij}(dz) \frac{\partial \widehat{e}}{\partial z_i} \frac{\partial \widehat{\chi}}{\partial z_j} + \sum_{i=1}^N d b_i(dz) \frac{\partial \widehat{e}}{\partial z_i} \widehat{\chi} + d^2 c(dz) \widehat{e}\widehat{\chi} \right) dz.$$

Thus the mesh size for (4.10) for $\widehat{e}$ is h/d. Notice that the size of c_{ell} and the coefficients and derivatives of $\widehat{A}$ are no larger than those for A. Hence Theorem 2 in the case that $\rho \ge kh$ and $d = 1$ may be applied with appropriate bounds for c_{ell}, $\sum_{i=1}^N d\|b_i(dz)\|_{L_\infty(B_2)}$, and $d^2\|c(dz)\|_{L_\infty(B_2)}$ (and sufficiently many of their derivatives) that can be chosen independent of $d \le 1$. We obtain

$$|e(z_2) - e(z_1)| \le C\Big(\frac{h}{d} \ln \frac{\rho/d}{h/d}\Big)\Big[\min_{\chi \in S^{h/d}} \|\sigma^s_{\bar{x}}(\nabla_z(u - \chi))\|_{L_\infty(B_2)} \\ + \|\sigma^s_{\bar{x}}(u - \chi)\|_{L_\infty(B_2)})\Big] \\ + C\Big(\frac{\rho}{d}\Big)\|e\|_{W_p^{-t}(B_2)}, \tag{4.11}$$

where in (4.11)

$$\sigma^s_{\bar{x}} = \Big(\frac{\rho/d}{|z| + \rho/d}\Big).$$

The inequality (1.16) now follows from (4.11) by scaling B_2 back to the ball B_{2d}, via $z = \frac{x}{d}$, and taking into account how each norm transforms. This completes the outline of the proof.

Acknowledgment. The authors thank a referee for detailed comments which helped improve the presentation.

REFERENCES

[1] P. Clément, *Approximation by finite element functions using local regularization*, Rev. Française Automat. Informat. Recherche Opérationnelle Sér. Rouge Anal. Numér., 9 (1975), pp. 77–84.

[2] A. Demlow, *Piecewise linear finite elements are not localized*, Math. Comp., to appear.

[3] S. Hilbert, *A mollifier useful for approximations in Sobolev spaces and some applications to approximating solutions of differential equations*, Math. Comp., 27 (1973), pp. 81–89.

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

[5] Ju. P. Krasovskii, *Properties of Green's function and generalized solutions of elliptic boundary value problems*, Soviet Math., 10 (1969), pp. 54–120.

[6] A. H. Schatz, *Pointwise error estimates and asymptotic error expansion inequalities for the finite element method on irregular grids: Part* I. *Global estimates*, Math. Comp., 67 (1998), pp. 877–899.

[7] A. H. Schatz, *Pointwise error estimates and asymptotic error expansion inequalities for the finite element method on irregular grids: Part* II. *Interior estimates*, SIAM J. Numer. Anal., 38 (2000), pp. 1269–1293.

[8] A. H. Schatz and L. B. Wahlbin, *Interior maximum norm estimates for finite element methods*, Math. Comp., 31 (1977), pp. 414–442.

[9] A. H. Schatz and L. B. Wahlbin, *Asymptotically exact a posteriori estimators for the pointwise gradient error on each element in irregular meshes. Part* II: *The piecewise linear case*, Math. Comp., to appear.

[10] L. R. Scott and S. Zhang, *Finite element interpolation of nonsmooth functions satisfying boundary conditions*, Math. Comp., 54 (1990), pp. 483–493.

SIAM J. NUMER. ANAL.
Vol. 41, No. 6, pp. 2161–2179

ERROR ANALYSIS OF A SEMIDISCRETE NUMERICAL SCHEME FOR DIFFUSION IN AXIALLY SYMMETRIC SURFACES[*]

KLAUS DECKELNICK[†], GERHARD DZIUK[‡], AND CHARLES M. ELLIOTT[§]

Abstract. We analyze a semidiscrete numerical scheme for approximating the evolution of axially symmetric surfaces by surface diffusion. The fourth order equation is split into two coupled second order problems, which are approximated by linear finite elements. We prove error bounds for the resulting scheme and present numerical test calculations that confirm our analysis.

Key words. surface diffusion, finite elements, error estimates, fourth order parabolic equation

AMS subject classifications. 65N30, 35K55

DOI. 10.1137/S0036142902405382

1. Introduction.

In recent years motion by mean curvature has been extensively studied from the computational point of view. However, the related curvature flow of motion by the surface Laplacian has received far less attention in the numerical analysis literature. The geometrical problem is to find a time-dependent surface $\Gamma(t)$ evolving according to the law of motion

$$V = \Delta_{\Gamma(t)}\kappa \qquad \text{on } \Gamma(t), \tag{1.1}$$

where V and κ denote, respectively, the normal velocity and the mean curvature of the surface. Our sign convention is that κ with respect to the outer normal is positive for spheres. The Laplace–Beltrami or surface Laplacian operator for Γ is denoted by Δ_Γ. This evolution has interesting geometrical properties: if $\Gamma(t)$ is a closed surface bounding a domain $\Omega(t)$, then the volume of $\Omega(t)$ is preserved and the surface area of $\Gamma(t)$ decreases. It is known that for closed curves in the plane or closed surfaces in $\mathbb{R}^3$ balls are asymptotically stable subject to small perturbations; see [9], [10]. However, it is also known that topological changes such as pinch-off are possible [11], [13].

Equation (1.1) is referred to as a surface diffusion equation because it models the diffusion of mass within the bounding surface of a solid body. At the atomistic level atoms on the surface move along the surface due to a driving force consisting of a chemical potential difference. For a surface with constant surface energy density the appropriate chemical potential in this setting is the mean curvature κ. This leads to the flux law

$$\rho V = -\mathrm{div}_\Gamma \mathbf{j},$$

where ρ is the mass density and $\mathbf{j}$ is the mass flux in the surface, with the constitutive flux law [12], [14]

$$\mathbf{j} = -D\nabla_\Gamma \kappa.$$

[*] Received by the editors April 9, 2002; accepted for publication (in revised form) April 28, 2003; published electronically November 25, 2003.

http://www.siam.org/journals/sinum/41-6/40538.html

[†] Institut für Analysis und Numerik, Universitatsplatz 2, 39106 Magdeburg, Germany (klaus.deckelnick@mathematik.uni-magdeburg.de).

[‡] Institut für Angewandte Mathematik, Albert Ludwigs Universität Freiburg, Hermann Herder Str. 10, 79104 Freiburg, Germany (gerd@mathematik.uni-freiburg.de).

[§] Center for Mathematical Analysis and Applications, School of Mathematical Sciences, University of Sussex, Falmer, Brighton BN1 9QH, UK (c.m.elliott@sussex.ac.uk).

Here, D is the diffusion constant. From these equations we obtain the law (1.1) after an appropriate nondimensionalization. The notion of surface diffusion is due to Mullins [14] and for a review we refer to [2].

In applications one is interested in the stability of so-called whiskers, which are axially symmetric cylindrical bodies of small diameter with respect to their length; see [15], [3], [1], and [16]. We shall be concerned with an axially symmetric cylindrical body, whose boundary

$$\Gamma(t) = \{\mathbf{x} \in \mathbb{R}^3 \,|\, \mathbf{x} = (x, r(x,t)\cos\phi, r(x,t)\sin\phi), x \in [0,L], \phi \in [0,2\pi]\}$$

evolves by surface diffusion. We assume that the radius r is a smooth positive function, which is periodic in x, so that $r(0,t) = r(L,t)$. In these coordinates the mean curvature of $\Gamma(t)$ is

$$\kappa = \frac{1}{r\sqrt{1+r_x^2}} - \frac{r_{xx}}{\sqrt{1+r_x^2}^3} = \frac{1}{r\sqrt{1+r_x^2}} - \left(\frac{r_x}{\sqrt{1+r_x^2}}\right)_x, \tag{1.2}$$

while the normal velocity and surface Laplacian of the mean curvature of the surface, respectively, are given by

$$V = \frac{r_t}{\sqrt{1+r_x^2}}, \qquad \Delta_\Gamma \kappa = \frac{1}{r\sqrt{1+r_x^2}}\left(\frac{r\kappa_x}{\sqrt{1+r_x^2}}\right)_x.$$

It follows from these two equations that r satisfies the quasi-linear fourth order parabolic problem

$$r_t = \frac{1}{r}\left(\frac{r\kappa_x}{\sqrt{1+r_x^2}}\right)_x \quad \text{in } I \times (0,T], \tag{1.3}$$

$$r(0,t) = r(L,t) \quad \text{in } (0,T], \tag{1.4}$$

$$\kappa(0,t) = \kappa(L,t) \quad \text{in } (0,T], \tag{1.5}$$

$$r(\cdot,0) = r_0 \quad \text{in } I, \tag{1.6}$$

where $I = (0,L)$ and κ is given by (1.2). The initial function r_0 is assumed to be periodic and positive.

Our concern in this paper is the analysis of a finite element discretization based on the above natural splitting of the fourth order problem into two coupled second order equations for the radial variable r and the mean curvature κ. We note that [4] proposed a similar second order splitting scheme and used $R = r^2$ and κ as the variables. Our principal result is an error estimate for the spatial discretization, which is actually attained in numerical experiments.

The paper is organized as follows: in section 2 we introduce the numerical scheme, prove the local existence and uniqueness of the discrete solution, and formulate our main error estimate. This result is proved in section 3, while section 4 contains numerical tests.

2. The discrete problem.

As already mentioned in the introduction, our discretization of (1.3) is based on the idea of splitting the elliptic part, which is of fourth order, into two second order operators. This is similar in spirit to the second order splitting techniques proposed for the numerical approximation of the Cahn–Hilliard equation in [8]. To begin, we deduce from (1.2)

$$r\kappa = \frac{1}{\sqrt{1+r_x^2}} - r\left(\frac{r_x}{\sqrt{1+r_x^2}}\right)_x = \sqrt{1+r_x^2} - \left(\frac{rr_x}{\sqrt{1+r_x^2}}\right)_x . \tag{2.1}$$

Thus (1.3) and (2.1) allow the variational formulation

$$\int_I rr_t\eta dx = -\int_I \frac{r\kappa_x\eta_x}{\sqrt{1+r_x^2}}dx \qquad \forall\eta \in H^1_{per}(I), \tag{2.2}$$

$$\int_I r\kappa\zeta dx = \int_I \sqrt{1+r_x^2}\,\zeta dx + \int_I \frac{rr_x\zeta_x}{\sqrt{1+r_x^2}}dx \qquad \forall\zeta \in H^1_{per}(I), \tag{2.3}$$

where $H^1_{per}(I) = \{\eta \in H^1(I)\,|\,\eta(0) = \eta(L)\}$. We employ (2.2), (2.3) in order to define a semidiscrete scheme using linear finite elements to approximate r and κ. Let $0 = x_0 < x_1 < \cdots < x_N = L$, $h_j := x_j - x_{j-1}$, and $h := \max_{1\le j\le N} h_j$. We shall make an inverse assumption of the form

$$h \le \rho h_j \qquad \forall j = 1, \dots, N, \tag{2.4}$$

where $\rho > 0$ is independent of h. The space of linear finite elements is defined by

$$X_h := \{\phi_h \in C^0(\bar{I})\,|\,\phi_{h|[x_{j-1},x_j]} \in P^1, 1 \le j \le N, \phi_h(0) = \phi_h(L)\}.$$

Our discrete problem now reads as follows: find $r_h, \kappa_h : [0, T] \to X_h$ such that

$$\int_I r_h r_{h,t}\eta_h dx = -\int_I \frac{r_h\kappa_{h,x}\eta_{h,x}}{\sqrt{1+r_{h,x}^2}}dx \qquad \forall\eta_h \in X_h, t \in (0, T], \tag{2.5}$$

$$\int_I r_h\kappa_h\zeta_h dx = \int_I \sqrt{1+r_{h,x}^2}\,\zeta_h dx + \int_I \frac{r_h r_{h,x}\zeta_{h,x}}{\sqrt{1+r_{h,x}^2}}dx \qquad \forall\zeta_h \in X_h, t \in [0, T], \tag{2.6}$$

$$r_h(0) = I_h r_0, \tag{2.7}$$

where I_h denotes the Lagrange interpolation operator.

LEMMA 2.1. *There exists $T_h > 0$ such that* (2.5)–(2.7) *has a unique solution $(r_h, \kappa_h) \in C^1([0, T_h]; X_h \times X_h)$ satisfying $\frac{1}{2}\min_{[0,L]} r_0 \le r_h \le 2\max_{[0,L]} r_0$ in $[0, L] \times [0, T_h]$.*

Proof. Choose a smooth globally Lipschitz-continuous function $\beta : \mathbb{R} \to \mathbb{R}$ with the properties $\beta(s) = s$ for $\frac{1}{2}\min_{[0,L]} r_0 \le s \le 2\max_{[0,L]} r_0$, $\frac{1}{4}\min_{[0,L]} r_0 \le \beta(s) \le 4\max_{[0,L]} r_0$ for all $s \in \mathbb{R}$. We first consider the following modified problem: find $r_h, \kappa_h : [0, T] \to X_h$ such that

$$\int_I \beta(r_h) r_{h,t}\eta_h dx = -\int_I \frac{r_h\kappa_{h,x}\eta_{h,x}}{\sqrt{1+r_{h,x}^2}}dx \qquad \forall\eta_h \in X_h, t \in [0, T], \tag{2.8}$$

$$\int_I \beta(r_h)\kappa_h\zeta_h dx = \int_I \sqrt{1+r_{h,x}^2}\,\zeta_h dx + \int_I \frac{r_h r_{h,x}\zeta_{h,x}}{\sqrt{1+r_{h,x}^2}}dx \qquad \forall\zeta_h \in X_h, t \in [0, T], \tag{2.9}$$

$$r_h(0) = I_h r_0. \tag{2.10}$$

Denoting by $\psi_1, \dots, \psi_N$ the usual nodal basis of X_h, we can represent (r_h, κ_h) as

$$r_h(\cdot, t) = \sum_{j=1}^N r_j(t)\psi_j, \qquad \kappa_h(\cdot, t) = \sum_{j=1}^N \kappa_j(t)\psi_j \tag{2.11}$$

and write $\underline{r}(t) = (r_1(t), \ldots, r_N(t))^T$, $\underline{\kappa}(t) = (\kappa_1(t), \ldots, \kappa_N(t))^T$. In view of the properties of β we may rewrite (2.9) in the form $\underline{\kappa}(t) = G(\underline{r}(t))$ with a Lipschitz-continuous mapping $G : \mathbb{R}^N \to \mathbb{R}^N$. Inserting this into (2.8) and using again the properties of β, we may write this relation as

$$\underline{r}'(t) = F(\underline{r}(t)), \quad \underline{r}(0) = (r_0(x_1), \ldots, r_0(x_N))^T,$$

with a Lipschitz-continuous $F : \mathbb{R}^N \to \mathbb{R}^N$. The existence and uniqueness of $\underline{r}$ on some interval $[0, T_h]$ follows directly from the theory of ODEs. The corresponding functions r_h and κ_h given by (2.11) will then solve (2.8)–(2.10). Since $r_h(0) = I_h r_0$ and by making T_h smaller if necessary, we may assume that $\frac{1}{2}\min_{[0,L]} r_0 \leq r_h \leq 2\max_{[0,L]} r_0$ in $[0, L] \times [0, T_h]$ so that, in view of the properties of β, (r_h, κ_h) also solves (2.5)–(2.7). □

Using $\eta_h = \kappa_h$ in (2.5) and $\zeta_h = r_{h,t}$ in (2.6) and taking the difference of the resulting equations, we obtain

$$\begin{aligned} 0 &= \int_I \sqrt{1 + r_{h,x}^2}\, r_{h,t} dx + \int_I \frac{r_h r_{h,x} r_{h,tx}}{\sqrt{1 + r_{h,x}^2}} dx + \int_I \frac{r_h \kappa_{h,x}^2}{\sqrt{1 + r_{h,x}^2}} dx \\ &= \frac{d}{dt} \int_I r_h \sqrt{1 + r_{h,x}^2} dx + \int_I \frac{r_h \kappa_{h,x}^2}{\sqrt{1 + r_{h,x}^2}} dx. \end{aligned}$$

Thus

$$\sup_{0 \leq t \leq T_h} \int_I r_h \sqrt{1 + r_{h,x}^2} dx + \int_0^{T_h} \int_I \frac{r_h \kappa_{h,x}^2}{\sqrt{1 + r_{h,x}^2}} dx dt \leq C(r_0). \tag{2.12}$$

Before we formulate an error estimate for the scheme (2.5)–(2.7), we state a local existence and uniqueness result for the continuous problem.

THEOREM 2.2. *Suppose that $r_0 \in H^4_{per}(I)$ is strictly positive. Then there exists $T_0 > 0$ such that (1.3)–(1.6) has a unique solution (r, κ), which satisfies $r \in L^\infty\big(0, T_0; H^4_{per}(I)\big)$, $r_t \in L^2\big(0, T_0; H^2_{per}(I)\big)$, and $r(x, t) > 0$ for all $(x, t) \in I \times [0, T_0]$.*

Proof. A similar result was proved in [11] for a formulation of (1.1) in terms of the distance function to a fixed reference curve. Since the resulting equation has the same structure as (1.3)–(1.6), the methods employed in [11] can be applied to our situation. □

We denote by $[0, T_{\max})$, $T_{\max} \in (0, \infty]$ the maximal time interval on which the solution from Theorem 2.2 exists and fix $T < T_{\max}$. Then there exist constants $0 < c_0 \leq C_0$ and $M \geq 0$ (depending on T) such that

$$c_0 \leq r \leq C_0, \quad |r_x| \leq C_0 \quad \text{on } [0, L] \times [0, T], \tag{2.13}$$

$$\sup_{t \in (0,T)} \|r(., t)\|^2_{H^4(I)} + \int_0^T \|r_t\|^2_{H^2(I)} dt \leq M^2. \tag{2.14}$$

Combining these bounds with (1.2), (1.3), and the inequality $\|f\|_{L^\infty(I)} \leq C\|f\|_{H^1(I)}$, we note for later use

$$\|\kappa(.,t)\|_{H^{1,\infty}(I)}+\|\kappa(.,t)\|_{H^2(I)}+\|r_t(.,t)\|_{L^2(I)} \le C \qquad \text{uniformly in } t \in [0,T], \tag{2.15}$$

where C depends on L, c_0, C_0, and M.

Our main result is the following error estimate, the proof of which will be given in the next section.

THEOREM 2.3. *There exists an $h_0 > 0$ such that for all $0 < h \le h_0$ the discrete solution (r_h, κ_h) exists on $[0,T]$ and*

$$\sup_{0\le t\le T} \|(r-r_h)(t)\|^2_{H^1(I)} + \int_0^T \|\kappa-\kappa_h\|^2_{H^1(I)} dt \le Ch^2. \tag{2.16}$$

The constant C depends on L, T, c_0, C_0, M, and ρ.

3. Proof of Theorem 2.3.

Let us define

$$\hat{T}_h := \sup\Big\{ t \in [0,T] \,|\, (r_h,\kappa_h) \text{ solves (2.5)–(2.7) on } [0,t] \text{ and } \frac{1}{2}c_0 \le r_h \le 2C_0,\ |r_{h,x}| \le 2C_0 \text{ on } [0,t]\Big\}.$$

By choosing T_h smaller if necessary (in order to satisfy the bound on $r_{h,x}$), we may deduce from Lemma 2.1 that $\hat{T}_h > 0$. Our aim is to show that $\hat{T}_h = T$ for small h. This will be achieved by proving the bounds (2.16) on $[0,\hat{T}_h]$, which will enable us to continue the discrete solution. By the definition of $\hat{T}_h$ we have

$$\frac{1}{2}c_0 \le r_h \le 2C_0, \quad |r_{h,x}| \le 2C_0 \text{ on } [0,L]\times[0,\hat{T}_h). \tag{3.1}$$

In what follows, we shall denote by C a constant which may depend on L, T, c_0, C_0, M, and ρ. Additional dependencies of C will be stated explicitly. We start with a useful auxiliary lemma.

LEMMA 3.1. *Let $v \in H^1_{per}(I), t \in [0,\hat{T}_h)$. Then we have for $\epsilon > 0$*

$$\Big|\int_I \frac{r_h}{r} v\, r\, r_t dx - \int_I v\, r_h r_{h,t} dx\Big| \le \epsilon\|\kappa_x - \kappa_{h,x}\|^2_{L^2(I)} + C_\epsilon\|v\|^2_{H^1(I)} + Ch^2 + C\|r-r_h\|^2_{H^1(I)}.$$

Proof. Fix $t \in [0,\hat{T}_h)$ and denote by $Q_h : L^2(I) \to X_h$ the following weighted projection: for a given $u \in L^2(I)$ let $Q_h u \in X_h$ be defined by

$$\int_I r_h u\, \zeta_h dx = \int_I r_h Q_h u\, \zeta_h dx \qquad \forall \zeta_h \in X_h. \tag{3.2}$$

We claim that

$$\|u - Q_h u\|_{L^2(I)} + h\|u_x - (Q_h u)_x\|_{L^2(I)} \le Ch\|u_x\|_{L^2(I)} \qquad \forall u \in H^1_{per}(I). \tag{3.3}$$

To see this, we first note that (3.1), (3.2), and an interpolation inequality imply

$$\frac{c_0}{2}\int_I |u-Q_h u|^2 \le \int_I r_h(u-Q_h u)(u-Q_h u) = \int_I r_h(u-Q_h u)(u-I_h u) \le 2C_0\|u-Q_h u\|_{L^2(I)}\, h\|u_x\|_{L^2(I)},$$

which yields the first part of (3.3). In view of (2.4) we have that $\|v_{h,x}\|_{L^2(I)} \leq Ch^{-1}\|v_h\|_{L^2(I)}$ for $v_h \in X_h$, and therefore

$$\begin{aligned}\|u_x - (Q_h u)_x\|_{L^2(I)} &\leq \|u_x - (I_h u)_x\|_{L^2(I)} + \|(I_h u)_x - (Q_h u)_x\|_{L^2(I)} \\ &\leq 2\|u_x\|_{L^2(I)} + Ch^{-1}\|I_h u - Q_h u\|_{L^2(I)} \\ &\leq 2\|u_x\|_{L^2(I)} + Ch^{-1}\big(\|u - I_h u\|_{L^2(I)} + \|u - Q_h u\|_{L^2(I)}\big) \\ &\leq C\|u_x\|_{L^2(I)},\end{aligned}$$

where we used the bound on $\|u - Q_u\|_{L^2(I)}$. This proves (3.3).

Next we infer from (3.2) and (2.5) that

$$\int_I v\, r_h r_{h,t} dx = \int_I Q_h v\, r_h r_{h,t} dx = -\int_I \frac{r_h \kappa_{h,x} (Q_h v)_x}{\sqrt{1 + r_{h,x}^2}} dx \qquad \forall v \in H^1_{per}(I).$$

If we combine this relation with (2.2), we may continue with

$$\begin{aligned}&\int_I \frac{r_h}{r} v\, r\, r_t dx - \int_I v\, r_h r_{h,t} dx \\ &= -\int_I \frac{r\kappa_x (\frac{r_h}{r} v)_x}{\sqrt{1 + r_x^2}} dx + \int_I \frac{r_h \kappa_{h,x} (Q_h v)_x}{\sqrt{1 + r_{h,x}^2}} dx \\ &= -\int_I \frac{r\kappa_x v}{\sqrt{1 + r_x^2}} \frac{r_{h,x} r - r_x r_h}{r^2} dx + \int_I \frac{r - r_h}{\sqrt{1 + r_x^2}} \kappa_x v_x \, dx \\ &\quad + \int_I \left(\frac{r_h}{\sqrt{1 + r_{h,x}^2}} - \frac{r}{\sqrt{1 + r_x^2}} \right) \kappa_{h,x} (Q_h v)_x dx \\ &\quad + \int_I \frac{r}{\sqrt{1 + r_x^2}} (\kappa_{h,x} - \kappa_x)(Q_h v)_x dx + \int_I \frac{r\, \kappa_x}{\sqrt{1 + r_x^2}} (Q_h v - v)_x \, dx \\ &\equiv \sum_{i=1}^5 S_i.\end{aligned}$$

In view of (2.13), (2.15), and (3.1), we then have

$$\begin{aligned}|S_1| &\leq C \int_I |v| \big(|r - r_h| + |r_x - r_{h,x}| \big) dx \leq \|v\|^2_{L^2(I)} + C\|r - r_h\|^2_{H^1(I)}, \\ |S_2| &\leq \|v_x\|^2_{L^2(I)} + C\|r - r_h\|^2_{L^2(I)}.\end{aligned}$$

Next, (2.15) and (3.3) imply

$$\begin{aligned}|S_3| &\leq \int_I \left| \frac{r_h}{\sqrt{1 + r_{h,x}^2}} - \frac{r}{\sqrt{1 + r_x^2}} \right| \big(|\kappa_x| + |\kappa_{h,x} - \kappa_x| \big) |(Q_h v)_x| \, dx \\ &\leq C \int_I \big(|r - r_h| + |r_x - r_{h,x}| \big) |(Q_h v)_x| \, dx + C \int_I |\kappa_x - \kappa_{h,x}| \, |(Q_h v)_x| \, dx \\ &\leq \epsilon \|\kappa_x - \kappa_{h,x}\|^2_{L^2(I)} + C_\epsilon \|v_x\|^2_{L^2(I)} + C\|r - r_h\|^2_{H^1(I)},\end{aligned}$$

and similarly,

$$|S_4| \leq \epsilon \|\kappa_x - \kappa_{h,x}\|^2_{L^2(I)} + C_\epsilon \|v_x\|^2_{L^2(I)}.$$

Finally, integration by parts, (1.3), (2.15), and (3.3) yield

$$\begin{aligned}|S_5| &= \left| -\int_I \left(\frac{r\,\kappa_x}{\sqrt{1+r_x^2}} \right)_x (Q_h v - v)\,dx \right| \\ &\leq Ch\|v_x\|_{L^2(I)}\|r_t\|_{L^2(I)} \leq Ch^2 + C\|v_x\|^2_{L^2(I)}.\end{aligned}$$

Collecting the above estimates concludes the proof of the lemma. □

As a first application of the above result we derive a differential inequality for the L^2-error.

LEMMA 3.2.

$$\frac{1}{2}\frac{d}{dt}\|r - r_h\|^2_{L^2(I)} \leq \epsilon\|\kappa_x - \kappa_{h,x}\|^2_{L^2(I)} + C_\epsilon\|r - r_h\|^2_{H^1(I)} + Ch^2.$$

Proof. Clearly,

$$\begin{aligned}\frac{1}{2}\frac{d}{dt}\|r - r_h\|^2_{L^2(I)} &= \int_I (r - r_h)(r_t - r_{h,t})dx \\ &= \int_I \frac{1}{r}(r - r_h) r r_t dx - \int_I \frac{1}{r_h}(r - r_h) r_h r_{h,t} dx.\end{aligned} \tag{3.4}$$

If we apply Lemma 3.1 to the function

$$v := \frac{1}{r_h}(r - r_h)(\cdot, t) \qquad \text{for } t \in (0, \hat{T}_h),$$

the result follows. □

The main part of the proof of Theorem 2.3 consists in controlling the H^1-seminorms of $r - r_h$ and $\kappa - \kappa_h$. The idea is to mimic the argument which led to the a priori estimate (2.12) in such a way that it can be applied to the difference between exact and discrete solution. This suggests using $\eta_h = I_h\kappa - \kappa_h$, $\zeta_h = I_h r_t - r_{h,t}$ in the error relations satisfied by $r - r_h$, $\kappa - \kappa_h$. In order to derive these relations we use $\eta = \eta_h \in X_h$ in (2.2) and $\zeta = \zeta_h \in X_h$ in (2.3) and take the difference with (2.5), (2.6), respectively. This leads to

$$\int_I \big(r\,r_t - r_h r_{h,t}\big)\eta_h dx = -\int_I \left(\frac{r\kappa_x}{\sqrt{1+r_x^2}} - \frac{r_h\kappa_{h,x}}{\sqrt{1+r_{h,x}^2}} \right)\eta_{h,x} dx \quad \forall \eta_h \in X_h, \tag{3.5}$$

$$\begin{aligned}\int_I (r\,\kappa - r_h\kappa_h)\zeta_h dx &= \int_I \big(\sqrt{1+r_x^2} - \sqrt{1+r_{h,x}^2}\big)\zeta_h dx \\ &\quad + \int_I \left(\frac{r r_x}{\sqrt{1+r_x^2}} - \frac{r_h r_{h,x}}{\sqrt{1+r_{h,x}^2}} \right)\zeta_{h,x} dx \quad \forall \zeta_h \in X_h.\end{aligned} \tag{3.6}$$

LEMMA 3.3. *We have for all $\epsilon > 0$*

$$\begin{aligned}&\frac{d}{dt}\int_I r_h \left(\sqrt{1+r_{h,x}^2} - \frac{r_{h,x} r_x + 1}{\sqrt{1+r_x^2}} \right) dx + \int_I \frac{r_h}{\sqrt{1+r_{h,x}^2}}(\kappa_x - \kappa_{h,x})^2 dx \\ &\leq C\epsilon\|\kappa - \kappa_h\|^2_{H^1(I)} + C_\epsilon(1 + \|r_t\|_{H^2(I)})\|r - r_h\|^2_{H^1(I)} + C_\epsilon h^2(1 + \|r_t\|^2_{H^2(I)}).\end{aligned} \tag{3.7}$$

Proof. Using $\zeta_h = I_h r_t - r_{h,t}$ in (3.6), we obtain

$$\begin{aligned}\int_I (r\kappa - r_h\kappa_h)(I_h r_t - r_{h,t})dx &= \int_I \big(\sqrt{1+r_x^2} - \sqrt{1+r_{h,x}^2}\big)(I_h r_t - r_{h,t})dx \\ &\quad + \int_I r_h \left(\frac{r_x}{\sqrt{1+r_x^2}} - \frac{r_{h,x}}{\sqrt{1+r_{h,x}^2}}\right)(r_{tx} - r_{h,tx})dx \\ &\quad + \int_I (r - r_h)\frac{r_x}{\sqrt{1+r_x^2}}(r_{tx} - r_{h,tx})dx \\ &\quad + \int_I \left(\frac{r r_x}{\sqrt{1+r_x^2}} - \frac{r_h r_{h,x}}{\sqrt{1+r_{h,x}^2}}\right)\big((I_h r_t)_x - r_{tx}\big)dx.\end{aligned}\tag{3.8}$$

Note first that the second integral can be written as

$$\begin{aligned}&\int_I r_h \left(\frac{r_x}{\sqrt{1+r_x^2}} - \frac{r_{h,x}}{\sqrt{1+r_{h,x}^2}}\right)(r_{tx} - r_{h,tx})dx \\ &= \int_I r_h \frac{\partial}{\partial t}\left(\sqrt{1+r_{h,x}^2} - \frac{r_{h,x}r_x + 1}{\sqrt{1+r_x^2}}\right)dx \\ &\quad + \int_I r_h r_{t,x}\left(\frac{r_{h,x}}{\sqrt{1+r_x^2}} - \frac{r_{h,x}}{\sqrt{1+r_{h,x}^2}} + \frac{r_x}{\sqrt{1+r_x^2}} - \frac{1 + r_x r_{h,x}}{1+r_x^2}\frac{r_x}{\sqrt{1+r_x^2}}\right)dx \\ &= \frac{d}{dt}\int_I r_h\left(\sqrt{1+r_{h,x}^2} - \frac{r_{h,x}r_x + 1}{\sqrt{1+r_x^2}}\right)dx - \int_I r_{h,t}\left(\sqrt{1+r_{h,x}^2} - \frac{r_{h,x}r_x + 1}{\sqrt{1+r_x^2}}\right)dx \\ &\quad + \int_I r_h r_{t,x}\left(\frac{r_{h,x}}{\sqrt{1+r_x^2}} - \frac{r_{h,x}}{\sqrt{1+r_{h,x}^2}} + \frac{r_x}{\sqrt{1+r_x^2}} - \frac{1 + r_x r_{h,x}}{1+r_x^2}\frac{r_x}{\sqrt{1+r_x^2}}\right)dx.\end{aligned}$$

Integration by parts together with (1.2) implies for the third term in (3.8)

$$\begin{aligned}&\int_I (r - r_h)\frac{r_x}{\sqrt{1+r_x^2}}(r_{tx} - r_{h,tx})dx \\ &= -\int_I (r_x - r_{h,x})\frac{r_x}{\sqrt{1+r_x^2}}(r_t - r_{h,t})dx - \int_I (r - r_h)\left(\frac{r_x}{\sqrt{1+r_x^2}}\right)_x (r_t - r_{h,t})dx \\ &= -\int_I r_t(r_x - r_{h,x})\frac{r_x}{\sqrt{1+r_x^2}}dx + \int_I r_{h,t}\sqrt{1+r_x^2}\,dx - \int_I r_{h,t}\frac{r_{h,x}r_x + 1}{\sqrt{1+r_x^2}}dx \\ &\quad - \int_I (r - r_h)\frac{1}{r\sqrt{1+r_x^2}}(r_t - r_{h,t})dx + \int_I (r - r_h)\kappa(r_t - r_{h,t})dx.\end{aligned}$$

Inserting the above equations into (3.8), we derive

$$\begin{aligned}\int_I (r\kappa - r_h\kappa_h)(I_h r_t - r_{h,t})dx &= \frac{d}{dt}\int_I r_h\left(\sqrt{1+r_{h,x}^2} - \frac{r_{h,x}r_x + 1}{\sqrt{1+r_x^2}}\right)dx \\ &\quad + \int_I \big(\sqrt{1+r_x^2} - \sqrt{1+r_{h,x}^2}\big)I_h r_t dx - \int_I r_t(r_x - r_{h,x})\frac{r_x}{\sqrt{1+r_x^2}}\,dx\end{aligned}\tag{3.9}$$

$$+\int_I r_h r_{tx}\left(\frac{r_{h,x}}{\sqrt{1+r_x^2}}-\frac{r_{h,x}}{\sqrt{1+r_{h,x}^2}}+\frac{r_x}{\sqrt{1+r_x^2}}-\frac{1+r_x r_{h,x}}{1+r_x^2}\frac{r_x}{\sqrt{1+r_x^2}}\right)dx$$
$$-\int_I (r-r_h)\frac{1}{r\sqrt{1+r_x^2}}(r_t-r_{h,t})dx+\int_I (r-r_h)\kappa(r_t-r_{h,t})dx$$
$$+\int_I\left(\frac{rr_x}{\sqrt{1+r_x^2}}-\frac{r_h r_{h,x}}{\sqrt{1+r_{h,x}^2}}\right)\big((I_h r_t)_x-r_{tx}\big)dx.$$

Let us next insert $\eta_h = I_h\kappa - \kappa_h$ into (3.5):

$$\begin{aligned}(3.10)\qquad &\int_I (rr_t - r_h r_{h,t})(I_h\kappa-\kappa_h)dx\\ &=-\int_I\left(\frac{r\kappa_x}{\sqrt{1+r_x^2}}-\frac{r_h\kappa_{h,x}}{\sqrt{1+r_{h,x}^2}}\right)\big((I_h\kappa)_x-\kappa_{h,x}\big)dx\\ &=\int_I\left(\frac{r\kappa_x}{\sqrt{1+r_x^2}}-\frac{r_h\kappa_{h,x}}{\sqrt{1+r_{h,x}^2}}\right)\big(\kappa_x-(I_h\kappa)_x\big)dx\\ &\quad-\int_I\frac{r_h}{\sqrt{1+r_{h,x}^2}}(\kappa_x-\kappa_{h,x})^2dx\\ &\quad-\int_I\left(\frac{r}{\sqrt{1+r_x^2}}-\frac{r_h}{\sqrt{1+r_{h,x}^2}}\right)\kappa_x(\kappa_x-\kappa_{h,x})dx.\end{aligned}$$

Combining (3.9) and (3.10), we obtain

$$(3.11)\quad \frac{d}{dt}\int_I r_h\left(\sqrt{1+r_{h,x}^2}-\frac{r_{h,x}r_x+1}{\sqrt{1+r_x^2}}\right)dx+\int_I\frac{r_h}{\sqrt{1+r_{h,x}^2}}(\kappa_x-\kappa_{h,x})^2dx=\sum_{i=1}^{8}\tilde{S}_i,$$

where

$$\tilde{S}_1=\int_I(r\kappa-r_h\kappa_h)(I_h r_t-r_{h,t})dx-\int_I(rr_t-r_h r_{h,t})(I_h\kappa-\kappa_h)dx$$
$$-\int_I(r-r_h)\kappa(r_t-r_{h,t})dx,$$
$$\tilde{S}_2=-\int_I r_t\left(\sqrt{1+r_x^2}-\sqrt{1+r_{h,x}^2}-(r_x-r_{h,x})\frac{r_x}{\sqrt{1+r_x^2}}\right)dx,$$
$$\tilde{S}_3=-\int_I r_h r_{tx}\left(\frac{r_{h,x}}{\sqrt{1+r_x^2}}-\frac{r_{h,x}}{\sqrt{1+r_{h,x}^2}}+\frac{r_x}{\sqrt{1+r_x^2}}-\frac{1+r_x r_{h,x}}{1+r_x^2}\frac{r_x}{\sqrt{1+r_x^2}}\right)dx,$$
$$\tilde{S}_4=\int_I(r-r_h)\frac{1}{r\sqrt{1+r_x^2}}(r_t-r_{h,t})dx,$$
$$\tilde{S}_5=\int_I\big(\sqrt{1+r_x^2}-\sqrt{1+r_{h,x}^2}\big)(r_t-I_h r_t)dx,$$

$$\tilde{S}_6 = -\int_I \left(\frac{r r_x}{\sqrt{1+r_x^2}} - \frac{r_h r_{h,x}}{\sqrt{1+r_{h,x}^2}} \right) \big((I_h r_t)_x - r_{tx} \big) dx,$$

$$\tilde{S}_7 = \int_I \left(\frac{r \kappa_x}{\sqrt{1+r_x^2}} - \frac{r_h \kappa_{h,x}}{\sqrt{1+r_{h,x}^2}} \right) \big(\kappa_x - (I_h \kappa)_x \big) dx,$$

$$\tilde{S}_8 = -\int_I \left(\frac{r}{\sqrt{1+r_x^2}} - \frac{r_h}{\sqrt{1+r_{h,x}^2}} \right) \kappa_x (\kappa_x - \kappa_{h,x}) dx.$$

The terms $\tilde{S}_1, \ldots, \tilde{S}_8$ have been organized in such a way that each of them is quadratic in an appropriate difference. To see this, let us examine them in more detail. First,

$$\begin{aligned}\tilde{S}_1 =& \int_I (r\kappa - r_h \kappa_h)(r_t - r_{h,t}) dx - \int_I (r r_t - r_h r_{h,t})(\kappa - \kappa_h) dx - \int_I (r - r_h) \kappa (r_t - r_{h,t}) dx \\ &+ \int_I (r\kappa - r_h\kappa_h)(I_h r_t - r_t) dx - \int_I (r r_t - r_h r_{h,t})(I_h \kappa - \kappa) dx \\ =& -\int_I r_t (\kappa - \kappa_h)(r - r_h) dx + \int_I (r\kappa - r_h \kappa_h)(I_h r_t - r_t) dx - \int_I (r r_t - r_h r_{h,t})(I_h\kappa - \kappa) dx \\ \equiv& A_1 + A_2 + A_3.\end{aligned}$$

Using an interpolation estimate, (2.15), and the continuous embedding $H^1(I) \hookrightarrow L^\infty(I)$, we obtain

$$\begin{aligned}|A_1 + A_2| &\le C\|r_t\|_{L^2(I)} \|\kappa - \kappa_h\|_{L^2(I)} \|r - r_h\|_{L^\infty(I)} + Ch\|r_{tx}\|_{L^2(I)} \|r\kappa - r_h\kappa_h\|_{L^2(I)} \\ &\le \epsilon \|\kappa - \kappa_h\|^2_{L^2(I)} + C_\epsilon \|r - r_h\|^2_{H^1(I)} + C_\epsilon h^2 \|r_{tx}\|^2_{L^2(I)},\end{aligned}$$

while

$$A_3 = \int_I (I_h\kappa - \kappa) r_h r_{h,t} dx - \int_I \frac{r_h}{r} (I_h \kappa - \kappa) r r_t dx + \int_I (I_h \kappa - \kappa) \Big(\frac{r_h}{r} - 1 \Big) r r_t dx.$$

We infer from Lemma 3.1 with $v = \kappa - I_h\kappa$ and well-known interpolation estimates that

$$\begin{aligned}|A_3| \le& \epsilon \|\kappa_x - \kappa_{h,x}\|^2_{L^2(I)} + C_\epsilon \|\kappa - I_h \kappa\|^2_{H^1(I)} + Ch^2 + C\|r - r_h\|^2_{H^1(I)} \\ &+ C\|r_t\|_{L^2(I)} \|r - r_h\|_{L^\infty(I)} \|\kappa - I_h \kappa\|_{L^2(I)} \\ \le& \epsilon \|\kappa_x - \kappa_{h,x}\|^2_{L^2(I)} + C_\epsilon h^2 \|\kappa\|^2_{H^2(I)} + C\|r - r_h\|^2_{H^1(I)}.\end{aligned}$$

Recalling (2.15), we conclude

$$|\tilde{S}_1| \le \epsilon \|\kappa - \kappa_h\|^2_{H^1(I)} + C_\epsilon \|r - r_h\|^2_{H^1(I)} + C_\epsilon (1 + \|r_{tx}\|^2_{L^2(I)}) h^2.$$

Next, observing that

$$\left| \sqrt{1+q^2} - \sqrt{1+p^2} - (q-p) \frac{q}{\sqrt{1+q^2}} \right| \le C(q-p)^2 \qquad \forall q, p \in \mathbb{R}, \tag{3.12}$$

we obtain

$$|\tilde{S}_2| \le C\|r_t\|_{L^\infty(I)} \|r_x - r_{h,x}\|^2_{L^2(I)} \le C\|r_t\|_{H^1(I)} \|r_x - r_{h,x}\|^2_{L^2(I)}.$$

Let us now examine $\tilde{S}_3$. A short calculation shows

$$\begin{aligned}
&\frac{p}{\sqrt{1+q^2}} - \frac{p}{\sqrt{1+p^2}} + \frac{q}{\sqrt{1+q^2}} - \frac{1+pq}{1+q^2}\frac{q}{\sqrt{1+q^2}} \\
&= \frac{p(1+q^2)\big(\sqrt{1+p^2}-\sqrt{1+q^2}\big) - q^2\sqrt{1+p^2}(p-q)}{\sqrt{1+q^2}^3\sqrt{1+p^2}} \\
&= \frac{p}{\sqrt{1+q^2}\sqrt{1+p^2}}\left(\sqrt{1+p^2}-\sqrt{1+q^2}-(p-q)\frac{p}{\sqrt{1+p^2}}\right) \\
&\qquad + \frac{p-q}{\sqrt{1+q^2}^3(1+p^2)}\big(p^2(1+q^2)-q^2(1+p^2)\big),
\end{aligned}$$

which implies in view of (3.12)

$$\left|\frac{p}{\sqrt{1+q^2}} - \frac{p}{\sqrt{1+p^2}} + \frac{q}{\sqrt{1+q^2}} - \frac{1+pq}{1+q^2}\frac{q}{\sqrt{1+q^2}}\right| \quad \leq \quad C(p-q)^2$$

for all $p, q \in \mathbb{R}$. Therefore,

$$|\tilde{S}_3| \leq C\|r_{tx}\|_{L^\infty(I)}\|r_x - r_{h,x}\|^2_{L^2(I)} \leq C\|r_{tx}\|_{H^1(I)}\|r_x - r_{h,x}\|^2_{L^2(I)}.$$

If we write

$$\tilde{S}_4 = \int_I (r-r_h)\frac{1}{r\sqrt{1+r_x^2}}(r_t - r_{h,t})dx = \int_I \frac{r_h}{r} v r r_t dx - \int_I v r_h r_{h,t} dx$$

with $v = \frac{r-r_h}{r_h r\sqrt{1+r_x^2}}$ and apply Lemma 3.1, we deduce

$$\begin{aligned}
|\tilde{S}_4| &\leq \epsilon\|\kappa_x - \kappa_{h,x}\|^2_{L^2(I)} + C_\epsilon\left\|\frac{r-r_h}{r_h r\sqrt{1+r_x^2}}\right\|^2_{H^1(I)} + Ch^2 + C\|r-r_h\|^2_{H^1(I)} \\
&\leq \epsilon\|\kappa_x - \kappa_{h,x}\|^2_{L^2(I)} + C_\epsilon\|r-r_h\|^2_{H^1(I)} + Ch^2.
\end{aligned}$$

In view of interpolation estimates, Young's inequality, and (2.15),

$$\begin{aligned}
|\tilde{S}_5| &\leq Ch\|r_{tx}\|_{L^2(I)}\|r_x - r_{h,x}\|_{L^2(I)} \leq Ch^2\|r_{tx}\|^2_{L^2(I)} + C\|r_x - r_{h,x}\|^2_{L^2(I)}, \\
|\tilde{S}_6| &\leq Ch\|r_{txx}\|_{L^2(I)}\|r - r_h\|_{H^1(I)} \leq Ch^2\|r_{txx}\|^2_{L^2(I)} + C\|r - r_h\|^2_{H^1(I)}, \\
|\tilde{S}_7| &\leq Ch\|\kappa_{xx}\|_{L^2(I)}\big(\|\kappa_x - \kappa_{h,x}\|_{L^2(I)} + \|r - r_h\|_{H^1(I)}\big) \\
&\leq \epsilon\|\kappa_x - \kappa_{h,x}\|^2_{L^2(I)} + C_\epsilon h^2 + C\|r - r_h\|^2_{H^1(I)}.
\end{aligned}$$

Finally,

$$|\tilde{S}_8| \leq \epsilon\|\kappa_x - \kappa_{h,x}\|^2_{L^2(I)} + C_\epsilon\|r - r_h\|^2_{H^1(I)}.$$

If we insert the above estimates for $\tilde{S}_1, \ldots, \tilde{S}_8$ into (3.11), the result is

$$\begin{aligned}
&\frac{d}{dt}\int_I r_h\left(\sqrt{1+r_{h,x}^2} - \frac{r_{h,x}r_x + 1}{\sqrt{1+r_x^2}}\right)dx + \int_I \frac{r_h}{\sqrt{1+r_{h,x}^2}}(\kappa_x - \kappa_{h,x})^2 dx \\
&\qquad \leq C\epsilon\|\kappa - \kappa_h\|^2_{H^1(I)} + C_\epsilon(1 + \|r_t\|_{H^2(I)})\|r - r_h\|^2_{H^1(I)} + C_\epsilon(1 + \|r_t\|^2_{H^2(I)})h^2,
\end{aligned}$$

which completes the proof of the lemma. □

Remark 3.4. (a) In order to interpret the integral

$$\int_I r_h \left(\sqrt{1+r_{h,x}^2} - \frac{r_{h,x} r_x + 1}{\sqrt{1+r_x^2}} \right) dx \tag{3.13}$$

occurring in (3.7), we note that

$$\nu = \frac{1}{\sqrt{1+r_x^2}} \big(-r_x, \cos\phi, \sin\phi \big), \quad \nu_h = \frac{1}{\sqrt{1+r_{h,x}^2}} \big(-r_{h,x}, \cos\phi, \sin\phi \big)$$

are the unit outward normals to

$$\Gamma(t) = \{ \mathbf{x} \in \mathbb{R}^3 \,|\, \mathbf{x} = (x, r(x,t)\cos\phi, r(x,t)\sin\phi), x \in [0,L], \phi \in [0, 2\pi] \},$$
$$\Gamma_h(t) = \{ \mathbf{x} \in \mathbb{R}^3 \,|\, \mathbf{x} = (x, r_h(x,t)\cos\phi, r_h(x,t)\sin\phi), x \in [0,L], \phi \in [0, 2\pi] \},$$

respectively. Observing that $dS = r_h \sqrt{1+r_{h,x}^2}\, dx d\phi$ is the surface element on Γ_h, a short calculation shows that

$$\int_I r_h \left(\sqrt{1+r_{h,x}^2} - \frac{r_{h,x} r_x + 1}{\sqrt{1+r_x^2}} \right) dx = \frac{1}{2\pi} \int_{\Gamma_h} |\nu - \nu_h|^2 dS.$$

A similar relation was used in [5], [6] in an error analysis for the mean curvature flow of graphs.

(b) Under the conditions (2.13) and (3.1), the expression (3.13) is equivalent to $\|r_x - r_{h,x}\|^2_{H^1(I)}$. To see this, note that

$$\begin{aligned}
&\sqrt{1+r_{h,x}^2} - \frac{r_{h,x} r_x + 1}{\sqrt{1+r_x^2}} \\
&= \frac{\big(\sqrt{1+r_{h,x}^2}\sqrt{1+r_x^2} - (r_{h,x} r_x + 1) \big) \big(\sqrt{1+r_{h,x}^2}\sqrt{1+r_x^2} + (r_{h,x} r_x + 1) \big)}{\sqrt{1+r_x^2} \big(\sqrt{1+r_{h,x}^2}\sqrt{1+r_x^2} + (r_{h,x} r_x + 1) \big)} \\
&= \frac{(r_x - r_{h,x})^2}{\sqrt{1+r_x^2} \big(\sqrt{1+r_{h,x}^2}\sqrt{1+r_x^2} + (r_{h,x} r_x + 1) \big)},
\end{aligned}$$

which implies

$$\begin{aligned}
\frac{c_0}{4(1+C_0^2)\sqrt{1+4C_0^2}} \|r_x - r_{h,x}\|^2_{H^1(I)} &\le \int_I r_h \left(\sqrt{1+r_{h,x}^2} - \frac{r_{h,x} r_x + 1}{\sqrt{1+r_x^2}} \right) dx \\
&\le C_0 \|r_x - r_{h,x}\|^2_{H^1(I)},
\end{aligned} \tag{3.14}$$

since

$$\begin{aligned}
1 &\le \sqrt{1+r_x^2} \big(\sqrt{1+r_{h,x}^2}\sqrt{1+r_x^2} + (r_{h,x} r_x + 1) \big) \\
&\le \sqrt{1+r_x^2} \big(\sqrt{1+r_{h,x}^2}\sqrt{1+r_x^2} + \sqrt{1+r_{h,x}^2}\sqrt{1+r_x^2} \big) \;\le\; 2(1+C_0^2)\sqrt{1+4C_0^2}.
\end{aligned}$$

It remains to derive an estimate for $\|\kappa - \kappa_h\|_{L^2(I)}$.

Lemma 3.5.

$$\|\kappa - \kappa_h\|_{L^2(I)} \le C\big(\|r - r_h\|_{H^1(I)} + \|\kappa_x - \kappa_{h,x}\|_{L^2(I)} + h\big).$$

Proof. Clearly,

$$\begin{aligned}&\int_I r_h(\kappa-\kappa_h)^2 dx\\ &= -\int_I (r-r_h)\kappa(\kappa-\kappa_h)dx + \int_I (r\kappa - r_h\kappa_h)(\kappa - I_h\kappa)dx + \int_I (r\kappa - r_h\kappa_h)(I_h\kappa-\kappa_h)dx.\end{aligned}$$

Using (3.6) in order to rewrite the third integral, we deduce

$$\begin{aligned}\int_I r_h(\kappa-\kappa_h)^2 dx &= -\int_I (r-r_h)\kappa(\kappa-\kappa_h)dx + \int_I (r\kappa - r_h\kappa_h)(\kappa - I_h\kappa)dx\\ &\quad + \int_I (I_h\kappa - \kappa_h)\big(\sqrt{1+r_x^2} - \sqrt{1+r_{h,x}^2}\big)dx + \int_I \left(\frac{r r_x}{\sqrt{1+r_x^2}} - \frac{r_h r_{h,x}}{\sqrt{1+r_{h,x}^2}}\right)\big(I_h\kappa - \kappa_h\big)_x dx\\ &\le C\|r-r_h\|_{L^2(I)}\|\kappa-\kappa_h\|_{L^2(I)} + C\big(\|r-r_h\|_{L^2(I)} + \|\kappa-\kappa_h\|_{L^2(I)}\big)\|\kappa - I_h\kappa\|_{L^2(I)}\\ &\quad + C\|I_h\kappa-\kappa_h\|_{L^2(I)}\|r_x - r_{h,x}\|_{L^2(I)} + C\|(I_h\kappa)_x - \kappa_{h,x}\|_{L^2(I)}\|r-r_h\|_{H^1(I)}\\ &\le \epsilon\|\kappa-\kappa_h\|^2_{L^2(I)} + C_\epsilon\|r-r_h\|^2_{H^1(I)} + C_\epsilon h^2 + C\|\kappa_x - \kappa_{h,x}\|^2_{L^2(I)}.\end{aligned}$$

Here we have again used (2.15). Choosing $\epsilon = \frac{c_0}{4}$ and recalling (3.1), we complete the proof of the lemma. □

We are now in position to complete the proof of Theorem 2.3. Combining Lemmas 3.2, 3.3, and 3.5 and (3.1), we obtain with $\lambda = \frac{c_0}{2\sqrt{1+4C_0^2}}$

$$\begin{aligned}&\frac{1}{2}\frac{d}{dt}\|r-r_h\|^2_{L^2(I)} + \frac{d}{dt}\int_I r_h\left(\sqrt{1+r_{h,x}^2} - \frac{r_{h,x}r_x+1}{\sqrt{1+r_x^2}}\right)dx + \lambda\|\kappa_x-\kappa_{h,x}\|^2_{L^2(I)}\\ &\le C\epsilon\|\kappa_x - \kappa_{h,x}\|^2_{L^2(I)} + C_\epsilon(1+\|r_t\|_{H^2(I)})\|r-r_h\|^2_{H^1(I)} + C_\epsilon(1+\|r_t\|^2_{H^2(I)})h^2.\end{aligned}$$

Choosing ϵ sufficiently small and recalling (3.14), the function

$$\phi(t) := \frac{1}{2}\|(r-r_h)(t)\|^2_{L^2(I)} + \int_I r_h\left(\sqrt{1+r_{h,x}^2} - \frac{r_{h,x}r_x+1}{\sqrt{1+r_x^2}}\right)(t)dx$$

satisfies

$$\phi'(t) + \frac{\lambda}{2}\|\kappa_x - \kappa_{h,x}\|^2_{L^2(I)} \le C(1+\|r_t\|^2_{H^2(I)})h^2 + C(1+\|r_t\|_{H^2(I)})\phi(t), \qquad 0 \le t \le \hat{T}_h. \tag{3.15}$$

Now, (2.7) and (3.14) yield $\phi(0) \le Ch^2$, so that Gronwall's lemma implies

$$\phi(t) \le C\,h^2\left(1 + \int_0^T \|r_t\|^2_{H^2(I)}dt\right)\exp\left(\int_0^T C(1+\|r_t\|_{H^2(I)})dt\right), \qquad 0 \le t \le \hat{T}_h.$$

Therefore,

$$\sup_{0<t<\hat{T}_h} \|(r-r_h)(t)\|_{H^1(I)} \le Ch, \tag{3.16}$$

and, using (3.15) together with Lemma 3.5,

$$\int_0^{\hat{T}_h} \|\kappa - \kappa_h\|^2_{H^1(I)} dt \le Ch^2. \tag{3.17}$$

We can now prove that $\hat{T}_h = T$. If not, we would have $\hat{T}_h < T$; the smoothness of r, (3.16), and an inverse estimate then would imply that

$$\|(r - r_h)(t)\|_{H^{1,\infty}(I)} \le C\sqrt{h}, \qquad 0 \le t \le \hat{T}_h,$$

which combined with (2.13) would give

$$\frac{3}{4}c_0 \le r_h \le \frac{3}{2}C_0,\ |r_{h,x}| \le \frac{3}{2}C_0 \quad \text{in } I \times [0, \hat{T}_h]$$

provided that $h \le h_0$ and h_0 is sufficiently small. However, then we could extend the discrete solution to an interval $[0, \hat{T}_h + \delta]$ for some $\delta > 0$ with

$$\frac{1}{2}c_0 \le r_h \le 2C_0,\ |r_{h,x}| \le 2C_0 \quad \text{in } I \times [0, \hat{T}_h + \delta],$$

which contradicts the definition of $\hat{T}_h$. Thus $\hat{T}_h = T$ for $h \le h_0$ and (3.16), (3.17) imply our result.

4. Numerical results.

We use the notation

$$r_j(t) = r_h(x_j, t),\ \ \kappa_j(t) = \kappa_h(x_j, t),\ \ j = 0, \ldots, N,$$
$$q_j(t) = \sqrt{h_j^2 + \big(r_j(t) - r_{j-1}(t)\big)^2},\ \ j = 1, \ldots, N.$$

The spatially discrete problem (2.5), (2.6) then is translated into the following system of ODEs. By a dot we denote the time derivative. For numerical tests we shall use an additional right-hand side f which we include in the equations here.

$$\begin{aligned}
&\frac{h_j}{6}(r_{j-1} + r_j)\dot{r}_{j-1} + \left(\frac{h_j}{6}r_{j-1} + \frac{1}{2}(h_j + h_{j+1})r_j + \frac{h_{j+1}}{6}r_{j+1}\right)\dot{r}_j \\
&\qquad + \frac{h_{j+1}}{6}(r_j + r_{j+1})\dot{r}_{j+1} - \frac{r_{j-1} + r_j}{q_j}\kappa_{j-1} + \left(\frac{r_{j-1} + r_j}{q_j} + \frac{r_j + r_{j+1}}{q_{j+1}}\right)\kappa_j \\
&\qquad - \frac{r_j + r_{j+1}}{q_{j+1}}\kappa_{j+1} = \frac{1}{2}\big(q_j(r_{j-1} + r_j)f_{j-\frac{1}{2}} + q_{j+1}(r_j + r_{j+1})f_{j+\frac{1}{2}}\big), \\
&\frac{h_j}{6}(r_{j-1} + r_j)\kappa_{j-1} + \left(\frac{h_j}{6}r_{j-1} + \frac{1}{2}(h_j + h_{j+1})r_j + \frac{h_{j+1}}{6}r_{j+1}\right)\kappa_j \\
&\qquad + \frac{h_{j+1}}{6}(r_j + r_{j+1})\kappa_{j+1} + \frac{r_{j-1} + r_j}{q_j}r_{j-1} - \left(\frac{r_{j-1} + r_j}{q_j} + \frac{r_j + r_{j+1}}{q_{j+1}}\right)r_j \\
&\qquad + \frac{r_j + r_{j+1}}{q_{j+1}}r_{j+1} = q_j + q_{j+1}
\end{aligned} \tag{4.1}$$

for $j = 1, \ldots, N$, $t \in (0, T]$, with periodic boundary conditions and initial condition $r_j(0) = r_0(x_j)$, $j = 0, \ldots, N$. For the right-hand side term involving f we have used a simple integration formula and the notation $f_{j\pm\frac{1}{2}} = f((x_j + x_{j\pm 1})/2)$.

The time discretization is done via a semi-implicit scheme which also linearizes the problem. Furthermore we use mass lumping at suitable positions. Let $\tau > 0$ be the time step size and $M = [T/\tau]$. For a generic function w we denote by w^m ($0 \leq m \leq M$) the evaluation on the mth time level: $w^m = w(\cdot, m\tau)$. The fully discrete scheme then reads as follows.

Algorithm 4.1. Let $r_j^0 = r_0(x_j)$, $j = 0, \dots, N$. For $m = 1, \dots, M$ solve

$$\begin{aligned}
&\frac{1}{\tau}(h_j + h_{j+1}) r_j^{m-1}(r_j^m - r_j^{m-1}) \\
&\quad - \frac{r_{j-1}^{m-1} + r_j^{m-1}}{q_j^{m-1}} \kappa_{j-1}^m + \left(\frac{r_{j-1}^{m-1} + r_j^{m-1}}{q_j^{m-1}} + \frac{r_j^{m-1} + r_{j+1}^{m-1}}{q_{j+1}^{m-1}} \right) \kappa_j^m - \frac{r_j^{m-1} + r_{j+1}^{m-1}}{q_{j+1}^{m-1}} \kappa_{j+1}^m \\
&\qquad = \frac{1}{2}\big(q_j^{m-1}(r_{j-1}^{m-1} + r_j^{m-1}) f_{j-\frac{1}{2}}^m + q_{j+1}^{m-1}(r_j^{m-1} + r_{j+1}^{m-1}) f_{j+\frac{1}{2}}^m \big), \\
&(h_j + h_{j+1}) r_j^{m-1} \kappa_j^m \\
&\quad + \frac{r_{j-1}^{m-1} + r_j^{m-1}}{q_j^{m-1}} r_{j-1}^m - \left(\frac{r_{j-1}^{m-1} + r_j^{m-1}}{q_j^{m-1}} + \frac{r_j^{m-1} + r_{j+1}^{m-1}}{q_{j+1}^{m-1}} \right) r_j^m + \frac{r_j^{m-1} + r_{j+1}^{m-1}}{q_{j+1}^{m-1}} r_{j+1}^m \\
&\qquad = q_j^{m-1} + q_{j+1}^{m-1}
\end{aligned}$$

for $j = 1, \dots, N$, $m = 1, \dots, M$.

In every time step a linear system for $\underline{r}^m = (r_1^m, \dots, r_N^m)$ and $\underline{\kappa}^m = (\kappa_1^m, \dots, \kappa_N^m)$ of the form

$$\frac{1}{\tau} M^{m-1} \underline{r}^m + S^{m-1} \underline{\kappa}^m = \underline{c}^{m-1}, \tag{4.2}$$

$$M^{m-1} \underline{\kappa}^m - S^{m-1} \underline{r}^m = \underline{d}^{m-1} \tag{4.3}$$

has to be solved. Here M^{m-1} is a suitable mass matrix, S^{m-1} is a stiffness matrix, and $\underline{c}^{m-1}$, $\underline{d}^{m-1}$ are right-hand sides depending on the quantities of the $(m-1)$st time step with built-in periodic boundary conditions. Note that the time discretization is semi-implicit with respect to the position r but is fully implicit with respect to curvature κ. The linear system (4.2), (4.3) was solved by inserting the second equation into the first one, which leads to the following linear system for $\underline{r}^m$:

$$\left(\frac{1}{\tau} M^{m-1} + S^{m-1}(M^{m-1})^{-1} S^{m-1} \right) \underline{r}^m = \underline{c}^{m-1} - S^{m-1}(M^{m-1})^{-1} \underline{d}^{m-1}. \tag{4.4}$$

Note that the matrix M^{m-1} is a diagonal matrix. The system (4.4) was solved by a conjugate gradient method.

For all computations we have used uniform spatial grids $h_j = h$ with h as indicated.

We test the scheme with a known continuous solution. We choose

$$r(x,t) = (1 + 0.25 \sin \pi(x-1))(1 + 0.125 \cos t)$$

on the interval $I = [0, 2]$ for $T = 1$ and calculate the corresponding right-hand side f from (1.3) and (1.2). Now we are able to compute the error between continuous solution r, κ and discrete solution r_h^m, κ_h^m and calculate the experimental order of convergence from the errors for two grids. As time step size we have chosen $\tau = 0.1h^2$.

Table 4.1

Absolute errors in various norms and experimental orders of convergence (in brackets) for the test problem for the choice $\tau = 0.1h^2$.

N	h	$\|r - r_h\|_{L^\infty(H^1)}$	$\|\kappa - \kappa_h\|_{L^2(H^1)}$
20	0.1	0.3010	2.2669
40	0.05	0.1544 (0.96)	1.1693 (0.96)
80	0.025	0.07784 (0.99)	0.5892 (0.99)
160	0.0125	0.03903 (1.00)	0.2952 (1.00)
320	0.00625	0.01953 (1.00)	0.1477 (1.00)

Table 4.2

Absolute errors in various norms and experimental orders of convergence (in brackets) for the test problem for the choice $\tau = 0.1h$.

N	h	$\|r - r_h\|_{L^\infty(H^1)}$	$\|\kappa - \kappa_h\|_{L^2(H^1)}$
20	0.1	0.2575	2.2597
40	0.05	0.1399 (0.88)	1.1672 (0.95)
80	0.025	0.07363 (0.93)	0.5886 (0.99)
160	0.0125	0.03790 (0.96)	0.2950 (1.00)
320	0.00625	0.01922 (0.98)	0.1476 (1.00)

The results are shown in Table 4.1. We measured the errors

$$\|r - r_h\|_{L^\infty((0,T),H^1(I))} \quad \text{and} \quad \|\kappa - \kappa_h\|_{L^2((0,T),H^1(I))}.$$

The results confirm the error estimates in Theorem 2.3 precisely. A quite astonishing result is that these convergence results experimentally also hold in the case of linear coupling of time step size and spatial grid size (see Table 4.2), in particular, that no stability problems arise even though the scheme is only semi-implicit. This is in some sense similar to the case of mean curvature flow, for which in [7] stability of a semi-implicit scheme was proved without any time step restriction.

In [3] it was shown that solutions of axially symmetric surface diffusion may exhibit the following dynamical behavior: After an initial rapid decay, some perturbations slowly grow in amplitude and finally lead to pinch-off. We recomputed an example from [3], for which the initial surface is given by

$$r_0(x) = 1 + 0.05\left(\sin\left(\frac{m+1}{2}x\right) + \sin\left(\frac{m}{2}x\right)\right), \; x \in (0, n\pi). \tag{4.5}$$

Figure 4.1 shows the rapid decay of perturbations for $m = 10$. For better visibility we scaled the graphics vertically by 100.

For $m = 14$ we show the long time behavior of the solution $r = r(x,t)$. In order to make the dynamical behavior more transparent we plot the solution in Figure 4.2 for $t \in [0, 10]$ and in Figure 4.3 for $t \in [20, 27.861]$. We have used 400 nodes and a time step size $\tau = 0.1\, h^2$. Note that our error analysis is only valid as long as r is bounded away from zero. For calculations near the pinch-off singularity we adapted the time step according to $\tau = 0.1\, h^2 \min_{[0,4\pi]} r_h^3$, a criterion which was found experimentally. Finally, we computed the solution of axisymmetric surface diffusion for the initial surface given by

$$r_0(x) = 1 - 0.95\, |x| \sin\frac{\pi}{x}, \; x \in (-1, 1). \tag{4.6}$$

Here we have used 500 spatial nodes and a time discretization as in the previous example. The results are shown in Figure 4.4.

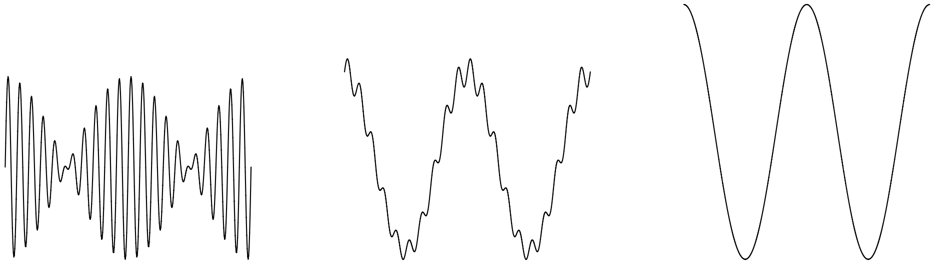


Fig. 4.1. *Evolution of the initial surface given by* (4.5) *with* $m = 10$, $n = 8$ *for* $t = 0.0$, 0.01, 0.1, *vertically scaled by* 100.

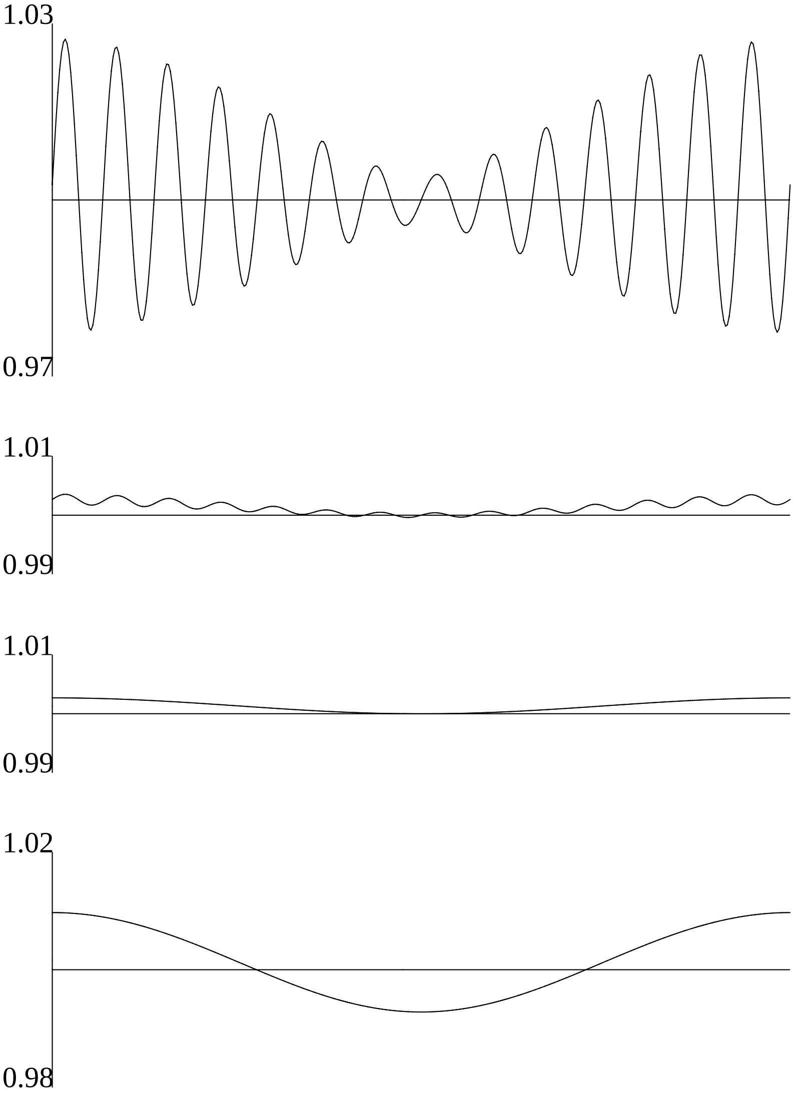


Fig. 4.2. *Evolution of the axially symmetric initial surface given by* (4.5) *with* $m = 14$, $n = 4$ *under surface diffusion. The horizontal axis runs from* 0 *to* 4π, *and the vertical axis is scaled by* 100. *Time steps* $t = 0.00$, 0.0014, 0.10, *and* 10.0.

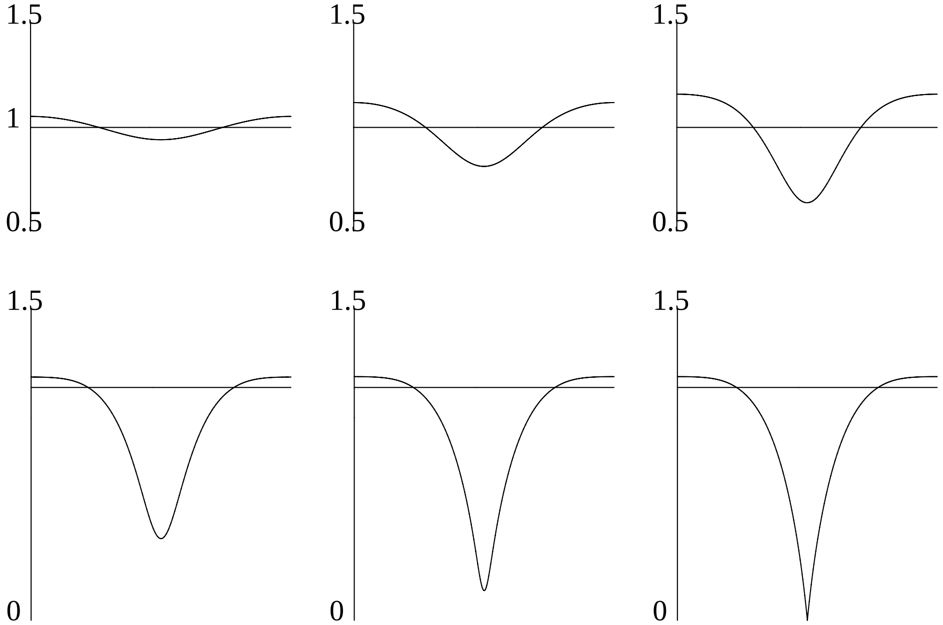


Fig. 4.3. *Evolution of the axially symmetric initial surface given by* (4.5) *with* $m = 14$, $n = 4$ *under surface diffusion. The horizontal axis runs from* 0 *to* 4π, *and the vertical axis is scaled by* 10. *Time steps* $t = 20.0$, 25.0, 27.0, 27.75, 27.86, *and* 27.861.

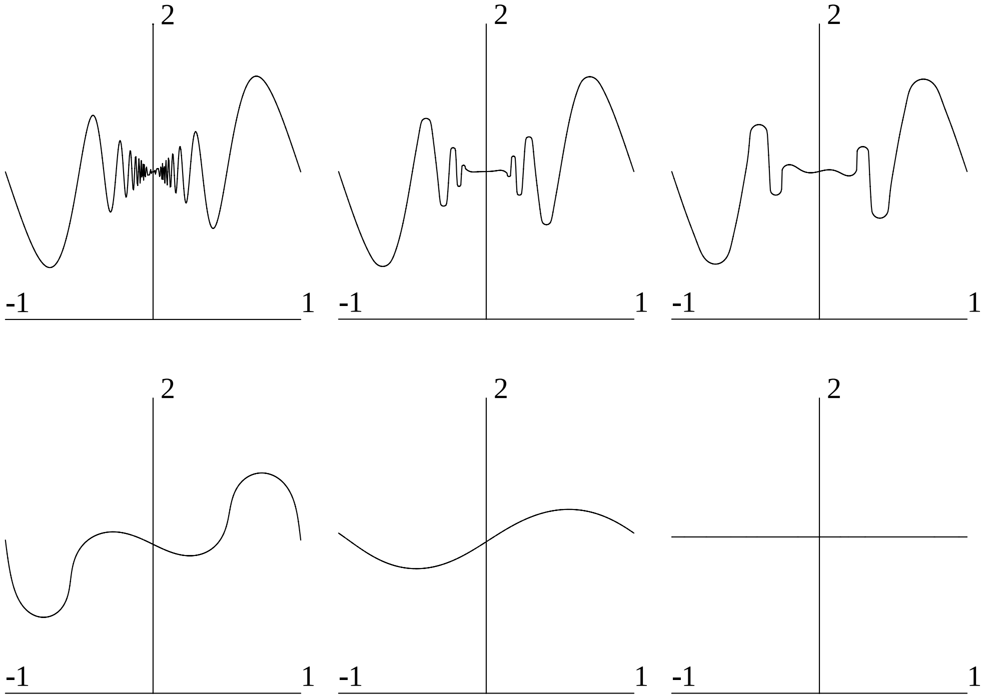


Fig. 4.4. *Evolution of the axially symmetric initial surface given by* (4.6) *under surface diffusion. Time steps* $t = 0.00$, $6.26 \cdot 10^{-7}$, $7.59 \cdot 10^{-6}$, $6.97 \cdot 10^{-4}$, $6.45 \cdot 10^{-3}$, *and* $9.82 \cdot 10^{-2}$.

REFERENCES

[1] A. J. Bernoff, A. L. Bertozzi, and T. P. Witelski, *Axisymmetric surface diffusion: Dynamics and stability of self-similar pinchoff*, J. Statist. Phys., 93 (1998), pp. 725–776.

[2] J. W. Cahn and J. E. Taylor, *Surface motion by surface diffusion*, Acta Metall. Mater., 42 (1994), pp. 1045–1063.

[3] B. D. Coleman, R. S. Falk, and M. Moakher, *Stability of cylindrical bodies in the theory of surface diffusion*, Phys. D, 89 (1995), pp. 123–135.

[4] B. D. Coleman, R. S. Falk, and M. Moakher, *Space–time finite element methods for surface diffusion with applications to the theory of the stability of cylinders*, SIAM J. Sci. Comput., 17 (1996), pp. 1434–1448.

[5] K. Deckelnick and G. Dziuk, *Discrete anisotropic curvature flow of graphs*, M2AN Math. Model. Numer. Anal., 33 (1999), pp. 1203–1222.

[6] K. Deckelnick and G. Dziuk, *Error estimates for a semi implicit fully discrete finite element scheme for the mean curvature flow of graphs*, Interfaces Free Bound., 2 (2000), pp. 341–359.

[7] G. Dziuk, *Numerical schemes for the mean curvature flow of graphs*, in Proceedings of the IUTAM Symposium on Variations of Domains and Free-Boundary Problems in Solid Mechanics, P. Argoul, M. Frémond, and Q. S. Nguyen, eds., Kluwer Academic Publishers, Dordrecht, Boston, London, 1999, pp. 63–70.

[8] C. M. Elliott, D. A. French, and F. A. Milner, *A second order splitting method for the Cahn-Hilliard equation*, Numer. Math., 54 (1989), pp. 575–590.

[9] C. M. Elliott and H. Garcke, *Existence results for diffusive surface motion laws*, Adv. Math. Sci. Appl., 7 (1997), pp. 467–490.

[10] J. Escher, U. F. Mayer, and G. Simonett, *The surface diffusion flow for immersed hypersurfaces*, SIAM J. Math. Anal., 29 (1998), pp. 1419–1433.

[11] Y. Giga and K. Ito, *On pinching of curves moved by surface diffusion*, Commun. Appl. Anal., 2 (1998), pp. 393–405.

[12] C. Herring, *Surface diffusion as a motivation for sintering*, in The Physics of Powder Metallurgy, W. E. Kingston, ed., McGraw–Hill, New York, 1951, pp. 143–179.

[13] U. F. Mayer and G. Simonett, *Self–intersections for the surface diffusion and the volume-preserving mean curvature flow*, Differential Integral Equations, 13 (2000), pp. 1189–1199.

[14] W. W. Mullins, *Theory of thermal grooving*, J. Appl. Phys., 28 (1957), pp. 333–339.

[15] F. A. Nichols and W. W. Mullins, *Surface–(interface–) and volume–diffusion contributions to morphological changes driven by capillarity*, Trans. Metall. Soc., AIME, 233 (1965), pp. 1840–1847.

[16] H. Wong, M. J. Miksis, P. W. Voorhees, and S. H. Davis, *Universal pinch off of rods by capillarity-driven surface diffusion*, Scripta Mat., 39 (1998), pp. 55–60.

SIAM J. Numer. Anal.
Vol. 41, No. 6, pp. 2180–2196

PERFORMING INTERPOLATION AND ANTERPOLATION ENTIRELY BY FAST FOURIER TRANSFORM IN THE 3-D MULTILEVEL FAST MULTIPOLE ALGORITHM[*]

JUKKA SARVAS[†]

Abstract. The fast multipole methods are used for solving a scalar acoustic or vector electromagnetic wave equation by integral equation methods with a large number of unknowns. In this paper a new method is presented for performing interpolation and anterpolation in both spherical coordinates θ and ϕ by FFT in the three-dimensional (3-D) multilevel fast multipole algorithm (MLFMA). The key idea is to approximate functions on the unit sphere by truncated Fourier series in two variables rather than by the usual spherical harmonics.

The proposed method is exact in interpolating and anterpolating and has the high numerical efficiency of FFT. The method is numerically compared to the method of performing interpolation and anterpolation using Lagrangian interpolation, which presently is probably the fastest method for those operations, and the results suggest that the proposed new method is equally or more efficient, depending on the desired accuracy.

Key words. fast multipole method, 3-D multilevel fast multipole algorithm, interpolation, anterpolation, fast Fourier transform

AMS subject classifications. 65T40, 78A40

DOI. 10.1137/S0036142902405655

1. Introduction.

The fast multipole methods, with two levels [1] or multiple levels [2], [3], are usually applied to solving a scalar acoustic or vector electromagnetic wave equation as an integral equation with a large number of unknowns. They are designed to speed up the matrix-vector multiply in a numerical solution using an iterative method. The multilevel fast multipole algorithm (MLFMA) leads to a computational cost of order $N \log N$, with N being the number of unknowns [4], [3], compared with the usual cost of order N^2 for the direct matrix-vector multiply. In this paper, we consider MLFMA for the scalar wave equation. Our results can easily be extended to the vector case in the usual way; see, e.g., [5], [3].

The key issue in MLFMA is to control the computational accuracy and cost of the scattered fields, both globally and locally, by a tree-like data structure. The global field representations are iteratively constructed by shifts and interpolations, and the local ones iteratively by translations, shifts, and anterpolations. We assume that the reader is familiar with the data structure of the MLFMA; see, e.g., [2], [5], [6], [7], [3]. Here we only outline the procedure for the needs of the present paper.

We consider a scatterer in 3-space; either its surface is triangulated for the surface integral method, or its volume is subdivided into a three-dimensional (3-D) grid for a volume integral method. In both cases, the unknown scalar source density is presented in terms of appropriate local basis functions. The scatterer is enclosed in a large cube, which is partitioned into eight subcubes. Every subcube is then, successively, subdivided until the finest level is reached, with the side length of a cube equal to about half of the wave length. The levels are indexed $M = 0, 1, \ldots, M_{\max}$. On the

[*] Received by the editors April 12, 2002; accepted for publication (in revised form) March 27, 2003; published electronically November 25, 2003DATEDATE. This work was financially supported by the Academy of Finland.

http://www.siam.org/journals/sinum/41-6/40565.html

[†] Electromagnetics Laboratory, Helsinki University of Technology, P.O. Box 3000, FIN-02015 HUT Helsinki, Finland (jukka.sarvas@hut.fi).

finest level $M = M_{\max}$, those cubes containing basis functions are indexed, and subsequently, on each level the nonempty parental cubes are indexed. The data structure is constructed in uptree, or *aggregation*, steps and in downtree, or *disaggregation*, steps.

In the aggregation steps, for every level $M = 1, 2, \ldots, M_{\max} - 1$ and for each nonempty cube Q on that level, the far field, due to sources in Q, is formed from the far fields of the subcubes of Q on the level $M + 1$.

The far fields, which are smooth functions on the unit sphere, are in this paper presented by truncated Fourier series, i.e., by *trigonometric polynomials*, in the spherical coordinates θ and ϕ, rather than in terms of the spherical harmonics. For the trigonometric polynomial presentation, we extend the functions on the unit sphere in the θ coordinate from $0 \le \theta \le \pi$ to $-\pi \le \theta \le \pi$. This extension makes them smooth 2π-periodic functions in both θ and ϕ, and consequently, a numerically very efficient approximation of these functions by trigonometric polynomials can be performed using the fast Fourier transform (FFT). The redundancy due to the extension into $-\pi \le \theta \le 0$ can be completely removed in the data storage and in the computations with FFT.

In forming the level M far fields from the level $M + 1$ fields, shifting and interpolation are used. The interpolation can be carried out for trigonometric polynomials accurately with FFT without any approximation error, which makes this operation numerically very efficient and straightforward for trigonometric polynomials.

In the disaggregation steps local presentations for the scattered fields F are formed in terms of plane waves by the integral

$$F(x) = \int_{|z|=1} v(z) e^{ikx\cdot z} dz, \quad x \in \mathbb{R}^3, \tag{1.1}$$

where the integral is a surface integral over the unit sphere $\{z \in \mathbb{R}^3 : |z| = 1\}$, $\mathbb{R}^3$ is the 3-space, and the function v in (1.1) is called the *amplitude field* of F. In a disaggregation step from the level $M - 1$ to M, the local field for each level M cube Q, due to sources outside of Q and its immediate neighbor cubes on level M, is formed in terms of plane waves. This is done by translating the far fields of cubes on the level M to a local field in Q and shifting and anterpolating the level $M + 1$ local field, corresponding to the parental cube of Q, into Q.

In this paper, we also present the amplitude fields in (1.1) in terms of trigonometric polynomials. Our choice makes it possible to carry out the anterpolation accurately and effectively by FFT.

In the disaggregation step from the level $M_{\max} - 1$ to the level $M_{\max}$, we eventually need to compute the scattered fields from the local amplitude fields v as in (1.1). After having estimated the amplitude field v and the function $e^{ikx\cdot z}$ in (1.1) by trigonometric polynomials, the resulting field integral can be computed efficiently in a closed form without any extra integration error.

In the last section of this paper, our method using trigonometric polynomials and FFT in the interpolation and anterpolation in three dimensions is compared to the Lagrangian interpolation method of performing those operations [2], [6], [3], which presently is probably the fastest method in this area. The comparison shows that, for a lower accuracy demand, both methods numerically are about equally efficient, but for a higher accuracy, the method of this paper is more efficient.

Furthermore, with no approximation errors in interpolation and anterpolation, the present method also makes accuracy control in MLFMA easier than in using other methods. Though the method of this paper was intended to be used primarily with the time-harmonic MLFMA, it is also well suited for the time-dependent MLFMA [8].

To finish the introduction, we fix some notation. We denote by $\mathbb{R}$ and $\mathbb{C}$ the real and complex numbers, respectively. The imaginary unit is $i = \sqrt{-1}$. The complex conjugate of $z \in \mathbb{C}$ is denoted by $\overline{z}$. For a real number t, $\mathit{floor}(t)$ is the largest integer $\leq t$, and $\mathit{ceil}(t)$ is the smallest integer $\geq t$. For a sequence $a = (a_1, \dots, a_n)$ of numbers, we denote by $\mathit{length}(a)$ the length n of the sequence. The element-by-element product $c = \{c_n\} = \{a_n b_n\}$ of two sequences $a = \{a_n\}$ and $b = \{b_n\}$, of the same length, is denoted by $c = ab$. For a number a_n in a sequence $a = \{a_n\}$ we also use the notation $a_n = a(n)$.

2. Discrete Fourier transform. We use a centralized discrete Fourier transform (DFT). For a sequence $u = \{u_n\}$ of length N, we define the centralized DFT $v = \mathcal{F}_N u$ of u, with the period N, by the formula

$$v_n = v(n) = (\mathcal{F}_N u)(n) = \sum_{m=-N_1}^{N_2} u(m) e^{-i \frac{2\pi}{N} nm}, \tag{2.1}$$

$-N_1 \leq n \leq N_2$, where $N_1 = \mathit{floor}(N/2)$ and $N_2 = N - N_1 - 1$. We also write $\mathcal{F}$ for $\mathcal{F}_N$.

In the context of DFT, we always consider the sequences u and v in (2.1) to be extended to be N-periodic for all n, i.e., $u(n + N) = u(n)$ for all integers n.

The inverse transform (IDFT) $\mathcal{F}_N^{-1}$ is defined by

$$u(m) = (\mathcal{F}_N^{-1} v)(m) = \frac{1}{N} \sum_{n=-N_1}^{N_2} v(n) e^{i \frac{2\pi}{N} mn} \quad \text{for all } m. \tag{2.2}$$

For a (P, Q)-matrix $u = \{u_{mn};\, -P_1 \leq m \leq P_2,\, -Q_1 \leq n \leq Q_2\}$, with $P_1 = \mathit{floor}(P/2), P_2 = P - P_1 - 1$, and $Q_1 = \mathit{floor}(Q/2)$, $Q_2 = Q - Q_1 - 1$, we denote the nth column by $u(:, n)$, and the mth row by $u(m, :)$, respectively. We also write $u(m, n)$ for u_{mn}. Again, we tacitly assume that each column $u(:, n)$ is extended to be a P-periodic sequence, and likewise, each row is extended to be a Q-periodic sequence.

We define our two-dimensional (2-D) discrete centralized Fourier transform $\mathcal{F}_{P,Q}$, with the periods P and Q, for a (P, Q)-matrix u by

$$(\mathcal{F}_{P,Q} u)(m, n) = \sum_{p=-P_1}^{P_2} \sum_{q=-Q_1}^{Q_2} u(p, q) e^{-i(\frac{2\pi}{P} mp + \frac{2\pi}{Q} nq)} \tag{2.3}$$

for all m, n. We also denote $\mathcal{F}_{P,Q}$ by $\mathcal{F}$. The inverse transform $\mathcal{F}_{P,Q}^{-1}$ is defined by

$$(\mathcal{F}_{P,Q}^{-1} v)(p, q) = \frac{1}{PQ} \sum_{m=-P_1}^{P_2} \sum_{n=-Q_1}^{Q_2} v(m, n) e^{i(\frac{2\pi}{P} pm + \frac{2\pi}{Q} qn)} \tag{2.4}$$

for all p, q. The 2-D DFT and its inverse transform can also be given, and usually are computed, in the terms of one-dimensional (1-D) DFT's as follows:

$$(\mathcal{F}_{P,Q} u)(m, n) = (\mathcal{F}_Q v(m, :))(n), \quad \text{where} \quad v(p, q) = (\mathcal{F}_P u(:, q))(p), \tag{2.5}$$

for all m, n, p, and q, or the order is reversed: first take $\mathcal{F}_Q$ rowwise and then $\mathcal{F}_P$ columnwise. A similar equation holds for $\mathcal{F}_{P,Q}^{-1}$ in terms of $\mathcal{F}_P^{-1}$ and $\mathcal{F}_Q^{-1}$.

The DFT is closely related to truncated Fourier series, i.e., *trigonometric polynomials.* Consider a trigonometric polynomial

$$U(t) = \sum_{n=-N}^{N'} a_n e^{int}, \quad t \in \mathbb{R}, \tag{2.6}$$

with $N' = N - 1$ or N, depending on whether the length of the coefficient sequence $\{a_n\}$ is, respectively, even or odd, and we say that the *kind* of U, denoted by $Kind(U)$, is then even or odd, respectively. The integer N is the *degree* of U, denoted by *Degree*(U), and it also is the half bandwidth of U.

As is well known, the coefficient sequence $a = \{a_n\}$ is related to U by equations

$$a = \frac{1}{M}\mathcal{F}_M(u) \quad \text{and} \quad u = M\mathcal{F}_M^{-1}(a), \tag{2.7}$$

where $u(n) = U(n\frac{2\pi}{M})$, $-N \le n \le N'$, $M = N + N' + 1$ is the period of $\mathcal{F}_M$, and u is the sample sequence of U, denoted by $u = \mathit{Sample}(U)$. The equations (2.7) show that we can identify a trigonometric polynomial U either with the sample sequence u or with its coefficient sequence a, which we denote by $a = \mathit{Coef}(U)$.

It is also well known that a smooth 2π-periodic function $V(t) = V(t + 2\pi)$, $t\epsilon\mathbb{R}$, can be approximated by an interpolating trigonometric polynomial (2.6) so that

$$u(n) = U(t_n) = V(t_n), \quad t_n = n\frac{2\pi}{M}, \quad -N \le n \le N',$$

and the approximation $U \simeq V$ improves as the *sampling rate* $N = \mathit{Degree}(U)$ increases.

We also consider trigonometric polynomials, i.e., truncated Fourier series, in two variables, but only of the following form:

$$U(\theta, \phi) = \sum_{m=-M}^{M} \sum_{n=-N}^{N-1} a(m,n) e^{i(m\theta + n\phi)}, \tag{2.8}$$

where $M, N > 0$ are integers. The coefficient matrix a is a $(2M+1, 2N)$-matrix, and

$$a = \frac{1}{(2M+1)2N}\mathcal{F}_{2M+1,2N}(u) \quad \text{with} \tag{2.9}$$

$$u(m,n) = U\left(m\frac{2\pi}{2M+1}, n\frac{\pi}{N}\right) \tag{2.10}$$

for $-M \le m \le M$, $-N \le n \le N-1$. We denote that $u = \mathit{Sample}(U)$ and $(M,N) = \mathit{Degree}(U)$, and also call (M,N) the *sampling rate* for U. As in the case of one variable, we can approximate a smooth function $V(\theta, \phi)$, 2π-periodic both in θ and ϕ, by the trigonometric polynomial $U(\theta, \phi)$ in (2.8), by choosing

$$a = \frac{1}{(2M+1)2N}\mathcal{F}v \quad \text{with} \quad v(m,n) = V\left(m\frac{2\pi}{2M+1}, n\frac{\pi}{N}\right)$$

for all $-M \le m \le M$ and $-N \le n \le N-1$.

3. Smooth functions on a sphere. We want to approximate far fields and amplitude fields on a unit sphere by trigonometric polynomials using DFT. For that purpose we must extend them to be 2π-periodic in both the spherical coordinates θ and ϕ.

Let $S = \{x\epsilon\mathbb{R}^3 : |x| = 1\}$ be the unit sphere in $\mathbb{R}^3$, and let $W : S \longrightarrow \mathbb{C}$ be a complex valued function on S, i.e., $W(x) \in \mathbb{C}$ for $x = (x_1, x_2, x_3) \in S$. If we write $x = (x_1, x_2, x_3)$ in the spherical coordinates (θ, ϕ), we can identify W with the function $V(\theta, \phi)$, for $0 \leq \theta \leq \pi$, $-\pi \leq \phi \leq \pi$, given by

$$V(\theta, \phi) = W(\sin\theta\cos\phi, \sin\theta\sin\phi, \cos\theta) \tag{3.1}$$

for $0 \leq \theta \leq \pi$, $-\pi \leq \phi \leq \pi$. By (3.1) we now can extend $V(\theta, \phi)$ for all $\theta, \phi \in \mathbb{R}$. This makes the extension $V(\theta, \phi)$ be 2π-periodic both in θ and ϕ and satisfy the equation

$$V(-\theta, \phi) = V(\theta, \phi + \pi) \quad \text{for all } \theta, \phi \in \mathbb{R}. \tag{3.2}$$

In general, we here call a function $V(\theta, \phi)$, $\theta, \phi \in \mathbb{R}$, *spherical* if it is 2π-periodic in θ and ϕ and satisfies (3.2).

We approximate a smooth spherical function V with a trigonometric polynomial of the form

$$U(\theta, \phi) = \sum_{m=-M}^{M} \sum_{n=-N}^{N-1} a(m, n) e^{i(m\theta + n\phi)}, \tag{3.3}$$

where

$$a = \frac{1}{(2M+1)2N} \mathcal{F}_{2M+1,2N}(v), \tag{3.4}$$

$$v(m, n) = V\left(m\frac{2\pi}{2M+1}, n\frac{\pi}{N}\right) \tag{3.5}$$

for $-M \leq m \leq M$, $-N \leq n \leq N - 1$. Because in (3.4) the first period $2M + 1$ is odd and the second period $2N$ is even, it is easily seen that (3.5) and the sphericality of V imply that

$$a(-m, n) = (-1)^n a(m, n) \quad \text{for all } m, n, \tag{3.6}$$

a condition which also makes $U(\theta, \phi)$ a spherical function. This implies that $u = \mathit{Sample}(U) = v$ satisfies the condition

$$u(-m, n) = u(m, n + N) \quad \text{for all } m, n. \tag{3.7}$$

We here call a $(2M + 1, 2N)$-matrix u *spherical* if it satisfies (3.7), and *transspherical* if it satisfies (3.6).

In the later interpolation and anterpolation operations we will frequently apply DFT on spherical matrices and IDFT on transspherical ones. The conditions (3.6) and (3.7) can be effectively utilized and the resulting computation reduced by half by using the following lemma, which is easily proved.

Lemma 3.1. *Consider a $(2M + 1, 2N)$-matrix u. If u is spherical and v_1 and v_2 are obtained from u by taking DFT columnwise and rowwise, i.e.,*

$$v_1(m, n) = (\mathcal{F}_{2M+1} u(:, n))(m) \quad \textit{and} \tag{3.8}$$

$$v_2(m, n) = (\mathcal{F}_{2N} u(m, :))(n) \tag{3.9}$$

for all m, n, then v_1 is spherical and v_2 is transspherical. Conversely, if u is transspherical and v_1 and v_2 are as in (3.8) *and* (3.9)*, then v_1 is transspherical and v_2 is spherical. Furthermore, the lemma also holds if $\mathcal{F}$ is replaced by $\mathcal{F}^{-1}$.*

The above lemma with the (2.5) yields the following corollary.

Corollary 3.2. *If u is a spherical $(2M+1, 2N)$-matrix, then $v = \mathcal{F}_{2M+1,2N}(u)$ is transspherical. If v is a transspherical $(2M+1, 2N)$-matrix, then $u = \mathcal{F}^{-1}_{2M+1,2N}(v)$ is spherical.*

We usually use FFT when computing the DFT and IDFT. The above lemma and its corollary can be used to reduce this computing of spherical and transspherical matrices, as follows. For a given spherical $(2M+1, 2N)$-matrix u, compute $w = \mathcal{F}_{2M+1,2N}u$ as follows. Let

$$\begin{aligned} v(m,n) &= (\mathcal{F}u(:,n))(m) \text{ for } -M \le m \le M, -N \le n \le -1, \\ v(m,n) &= v(-m, n-N) \text{ for } -M \le m \le M, 0 \le n \le N-1, \\ w(m,n) &= (\mathcal{F}v(m,:))(n) \text{ for } -M \le m \le 0, -N \le n \le N-1, \\ w(m,n) &= (-1)^n w(-m,n) \text{ for } 1 \le m \le M, -N \le n \le N-1. \end{aligned}$$

Similarly, we can, with a reduced workload, compute $\mathcal{F}^{-1}_{(2M+1,2N)}v$ for a transspherical matrix v.

Note also that only one half of a spherical, or a transspherical, matrix must be stored, and also that the DFT and IDFT can be performed using that half storage space. These reductions in computing the DFT and IDFT can be used in the interpolation and anterpolation operations, because transsphericality is preserved in zero-padding and truncation of coefficient sequences in DFT.

In addition to the possibility of performing interpolation and anterpolation by FFT, the use of trigonometric polynomials in MLFMA has the advantage that, after estimating functions by trigonometric polynomials, the field integrals can be computed efficiently in a closed form without any further integration error. The needed integral formula is given in the following lemma.

Lemma 3.3. *Let U and V be trigonometric polynomials in two variables and of degree (M, N). If U and V also are spherical functions, then*

$$\int_{-\pi}^{\pi}\int_{-\pi}^{\pi} U(\theta,\phi)\overline{V(\theta,\phi)}d\theta d\phi$$

$$= \frac{4\pi^2}{(2M+1)N}\left[\sum_{n=0}^{N-1} u(0,n)\overline{v(0,n)} + \sum_{m=1}^{M}\sum_{n=-N}^{N-1} u(m,n)\overline{v(m,n)}\right],$$

where $u = Sample(U)$ and $v = Sample(V)$.

Proof. Let $a = Coef(U)$ and $b = Coef(V)$. Start by using the orthogonality of terms $e^{i(m\theta+n\phi)}$ in integrating over $-\pi \le \theta, \phi \le \pi$, and then use (2.9), the Parseval's identity for 2-D DFT, and the sphericality property of u and v to get

$$\int_{-\pi}^{\pi}\int_{-\pi}^{\pi} U(\theta,\phi)\overline{V(\theta,\phi)}d\theta d\phi$$

$$= \int_{-\pi}^{\pi}\int_{-\pi}^{\pi}\left[\sum_{m=-M}^{M}\sum_{n=-N}^{N-1} a(m,n)e^{i(m\phi+n\phi)}\right]\left[\sum_{m=-M}^{M}\sum_{n=-N}^{N-1} \overline{b(m,n)}e^{-i(m\phi+n\phi)}\right]d\theta d\phi$$

$$= 4\pi^2 \sum_{m=-M}^{M} \sum_{n=-N}^{N-1} a(m,n)\overline{b(m,n)} = \frac{4\pi^2}{(2M+1)2N} \sum_{m=-M}^{M} \sum_{n=-N}^{N-1} u(m,n)\overline{v(m,n)}$$

$$= \frac{4\pi^2}{(2M+1)N} \left[\sum_{n=0}^{N-1} u(0,n)\overline{v(0,n)} + \sum_{m=1}^{M} \sum_{n=-N}^{N-1} u(m,n)\overline{v(m,n)} \right],$$

where in the last step we have used the equation $u(-m,n)\overline{v(-m,n)} = u(m,n+N)\overline{v(m,n+N)} = u(m,n-N)\overline{v(m,n-N)}$ for all m,n, due to the sphericality of u and v and the periodicity in n with the period $2N$. This completes the proof. □

4. Interpolation and anterpolation. Interpolation and anterpolation are here applied to trigonometric polynomials of one and two variables. These operations are essentially zero-padding and truncation, applied to coefficient sequences or matrices.

We say that a sequence c is obtained from a sequence b by *zero-padding* by M, and denote $c = \mathit{Pad}(b, M)$, if c is obtained from b by adding M zeros at both ends of b.

If U and V are trigonometric polynomials of one variable and

$$\mathit{Coef}(V) = \mathit{Pad}(\mathit{Coef}(U), M),$$

we say that $v = \mathit{Sample}(V)$ is obtained from $u = \mathit{Sample}(U)$ by *interpolating* u from the sampling rate $N = \mathit{Degree}(U)$ to the sampling rate $N + M$, and we denote that by

$$v = \mathit{Interp}(u, N+M). \tag{4.1}$$

Due to (2.7), the interpolation can be computed by FFT using the equation

$$\mathit{Interp}(u, N+M) = (L+2M)\mathcal{F}^{-1}_{L+2M}(\mathit{Pad}(a, M)), \tag{4.2}$$

where $a = L^{-1}\mathcal{F}_L(u)$, $L = \mathit{Length}(u)$, and $N = \mathit{floor}(L/2)$.

For a trigonometric polynomial U of two variables the interpolation of $u = \mathit{Sample}(U)$ from the sampling rate $(N_1, N_2) = \mathit{Degree}(U)$ to the sampling rate $(N_1 + M_1, N_2 + M_2)$ is defined in an analogous way, and the resulting matrix is denoted by $\mathit{Interp}(u, N_1 + M_1, N_2 + M_2)$. It is computed by FFT by applying (4.2) to the coefficient matrix

$$a = (2N_1 + 1)^{-1}(2N_2 + 1)^{-1}\mathcal{F}_{2N_1+1,2N_2}(u)$$

column- and rowwise. If u is spherical, the computational cost is lowered in the way discussed at the end of the last section.

We say that a sequence c is obtained from a sequence $b = (b_{-N}, \dots, b_{N'})$ by *truncating* b by M elements from each end if $c = (b_{-N+M}, \dots, b_{N'-M})$, and we denote $c = \mathit{Trunc}(b, M)$.

If U and V are trigonometric polynomials of one variable and

$$\mathit{Coef}(V) = \mathit{Trunc}(\mathit{Coef}(U), M),$$

we say that V and $v = \mathit{Sample}(V)$ are obtained from U and $u = \mathit{Sample}(U)$, respectively, by *anterpolating* U and u from the degree $N = \mathit{Degree}(U)$ to the degree $N - M$, and we denote

$$V = \mathit{Anterp}(U, N-M) \quad \text{and} \quad v = \mathit{Anterp}(u, N-M).$$

The sample sequence $v = \mathit{Anterp}(u, N - M)$ can be computed by FFT using the equation

$$\mathit{Anterp}(u, N - M) = (L - 2M)\mathcal{F}_{L-2M}^{-1}(\mathit{Trunc}(a, M)), \tag{4.3}$$

where $a = L^{-1}\mathcal{F}_L(u)$, $L = \mathit{Length}(u)$, and $N = \mathit{floor}(L/2)$.

The anterpolation can be, in an obvious way, extended to functions U given by a Fourier series, i.e., $\mathit{Degree}(U) = \infty$.

The anterpolation of a trigonometric polynomial of two variables is defined in an analogous way and is computed by (4.3) column- and rowwise. For a spherical u, the resulting reductions of computational cost can be utilized.

In MLFMA the following theorem is fundamental for implementing the anterpolation in an optimal manner. The theorem is essentially due to the fact that if in DFT we undersample a trigonometric polynomial appropriately, the resulting aliasing corrupts only coefficients of higher order, leaving the lower ones unchanged.

Theorem 4.1 (anterpolation of a product). *Let U and V be trigonometric polynomials of one variable and of the odd kind. Let*

$$\mathit{Degree}(U) = N \quad \text{and} \quad \mathit{Degree}(V) \geq N + M$$

for an integer $M \geq 0$. Then

$$\mathit{Anterp}(UV, M) = \mathit{Anterp}(W, M), \tag{4.4}$$

where W is a trigonometric polynomial of degree $N + M$, of the odd kind, and given by its sample sequence $\mathit{Sample}(W) = w$ so that

$$w(n) = u_1(n)v_1(n) \quad \text{for} \quad -N - M \leq n \leq N + M, \quad \text{where}$$
$$u_1 = \mathit{Interp}(\mathit{Sample}(U), N + M), \quad v_1 = \mathit{Sample}(\mathit{Anterp}(V, N + M)).$$

The theorem also holds for V with $\mathit{Degree}(V) = \infty$; i.e., V is a function given by a Fourier series.

Proof. Let $a = \mathit{Coef}(U)$, $b = \mathit{Coef}(V)$, and $c = \mathit{Coef}(UV)$. If $-M \leq m \leq M$, then

$$c(m) = \sum_{n=-N}^{N} a(n)b(m - n) = \sum_{n=-N-M}^{N+M} \tilde{a}(n)\tilde{b}(m - n) = (\tilde{a} * \tilde{b})(m), \tag{4.5}$$

where the sequences $\tilde{a}$ and $\tilde{b}$ are the P-periodic extensions of the sequences $\mathit{Pad}(a, M)$ and $\{b(n);\ -N-M \leq n \leq N+M\}$, and the sequence $\tilde{a}*\tilde{b}$ is the P-periodic convolution of $\tilde{a}$ and $\tilde{b}$ with $P = 2(N + M) + 1$. By the convolution theorem for the inverse DFT and by the coefficient rule (2.7), we get

$$\begin{aligned}\tilde{a} * \tilde{b} &= \mathcal{F}_P\big(\mathcal{F}_P^{-1}(\tilde{a} * \tilde{b})\big) = \mathcal{F}_P\big(P(\mathcal{F}_P^{-1}\tilde{a})(\mathcal{F}_P^{-1}\tilde{b})\big)\\ &= \frac{1}{P}\mathcal{F}_P(u_1v_1) = \frac{1}{P}\mathcal{F}_P(w) = \mathit{Coef}(W).\end{aligned}$$

This with (4.5) implies (4.4) and proves the theorem. □

For anterpolation with respect to the ϕ variable, we also need the following lemma.

LEMMA 4.2. *Let U_1, U_2, and U_3 be trigonometric polynomials of one variable with*

$$Degree(U_1) = N \geq Degree(U_3) = M.$$

Assume that $\overline{U_2}U_3$ can be approximated with sufficient accuracy by a trigonometric polynomial V, i.e., $\overline{U_2}U_3 \simeq V$, with $Degree(V) = N$, and U_1, V, and U_3 are all of the even kind. Then

$$\int_{-\pi}^{\pi} U_1(t)U_2(t)\overline{U_3(t)}dt \simeq \int_{-\pi}^{\pi} W(t)\overline{U_3(t)}dt,$$

where $W = Anterp(S, M)$ and S is the trigonometric polynomial with $Sample(S)(n) = U_1(t_n)U_2(t_n), t_n = n\pi/N$ for $-N \leq n \leq N-1$.

Proof. Let $a = Coef(U_3)$, $b = Coef(S)$, and $t_n = n\pi/N$ for $-N \leq n \leq N-1$. Reasoning as in Lemma 3.3 with the trigonometric polynomials U and V of one variable, we get

$$\begin{aligned}
&\int_{-\pi}^{\pi} U_1(t)U_2(t)\overline{U_3(t)}dt \simeq \int_{-\pi}^{\pi} U_1(t)\overline{V(t)}dt \\
&= \frac{2\pi}{2N}\sum_{n=-N}^{N-1} U_1(t_n)\overline{V(t_n)} \simeq \frac{\pi}{N}\sum_{n=-N}^{N-1} S(t_n)\overline{U_3(t_n)} = 2\pi\sum_{n=-N}^{N-1} b(n)\overline{a(n)} \\
&= 2\pi\sum_{m=-M}^{M-1} b(n)\overline{a(n)} = \int_{-\pi}^{\pi} W(t)\overline{U_3(t)}dt,
\end{aligned}$$

which completes the proof of the lemma. □

5. Translation from a far field to a local field.

We begin with a short review of the derivation of the fundamental translation formula from a far field to a local field; see, e.g., [1], [3]. This formula is the key step in the two-level and multilevel (nonstatic) fast multipole methods.

We start with the well-known expansion for the 3-D Helmholtz kernel $e^{ik|x|}/(4\pi|x|)$ in terms of incoming multipoles (e.g., see [3, p. 80]),

$$\frac{e^{ik|p+x|}}{4\pi|p+x|} = \sum_{n=0}^{\infty} a_n j_n(k|x|)P_n\left(\frac{p}{|p|}\cdot\frac{x}{|x|}\right), \tag{5.1}$$

where

$$a_n = \frac{ik}{4\pi}(-1)^n(2n+1)h_n^{(1)}(k|p|),$$

$h_n^{(1)}$ is a spherical Hankel function of the first kind, j_n is a spherical Bessel function, and P_n is the Legendre polynomial of order n. The series converges absolutely and uniformly for $|x| \leq r$ with any fixed $r < |p|$. If we truncate the series at $n = L$ and use the identity (e.g., see [9, p. 31])

$$\int_{|z|=1} e^{ikz\cdot x}P_n\left(\frac{p}{|p|}\cdot z\right)dz = 4\pi i^n j_n(k|x|)P_n\left(\frac{p}{|p|}\cdot\frac{x}{|x|}\right),$$

we get

$$\sum_{n=0}^{L} a_n j_n(k|x|) P_n\left(\frac{p}{|p|}\cdot\frac{x}{|x|}\right) = \int_{|z|=1} \frac{1}{4\pi} T_L\left(k|p|, \frac{p}{|p|}\cdot z\right) e^{ikx\cdot z} dz, \tag{5.2}$$

where the surface integral is over the unit sphere and

$$T_L\left(k|p|, \frac{p}{|p|}\cdot z\right) = \frac{k}{4\pi}\sum_{n=0}^{L}(2n+1) i^{n+1} h_n^{(1)}(k|p|) P_n\left(\frac{p}{|p|}\cdot z\right), \tag{5.3}$$

where and T_L is the Rokhlin translation function (see [1]).

Now, consider a source distribution $\rho(y)$, $y \in \mathbb{R}^3$, with $\rho(y) = 0$ outside of a bounded domain $D \subset \mathbb{R}^3$. Let

$$F(x) = \frac{1}{4\pi}\int_D \frac{e^{ik|x-y|}}{|x-y|}\rho(y)dy, \quad x \in \mathbb{R}^3, \tag{5.4}$$

be the scalar field due to this source, where k is the wave number. Suppose $D \subset \{z \in \mathbb{R}^3;\ |z| < R\}$ and $p \in \mathbb{R}^3$ with $R < |p|$.

The expansion (5.1), when applied to $x - y$ instead of x, allows us to expand $F(p + x)$ locally at the point p for all $x, |x| < r$, with any $r < |p| - R$:

$$F(p+x) = \int_D \left[\sum_{n=0}^{\infty} a_n j_n(k|x-y|) P_n\left(\frac{p}{|p|}\cdot\frac{(x-y)}{|x-y|}\right)\right]\rho(y)dy. \tag{5.5}$$

If we truncate the series in (5.5) at $n = L$, $F(p + x)$ will be approximated, the more accurately the larger L is, by the integral

$$\begin{aligned} F(p+x) &\simeq \int_D \left[\sum_{n=0}^{L} a_n j_n(k|x-y|) P_n\left(\frac{p}{|p|}\cdot\frac{(x-y)}{|x-y|}\right)\right]\rho(y)dy \\ &= \int_{|z|=1}\left[\frac{1}{4\pi}\int_D e^{-iky\cdot z}\rho(y)dy\right] T_L\left(k|p|, \frac{p}{|p|}\cdot z\right) e^{ikx\cdot z}dz, \end{aligned} \tag{5.6}$$

where we have applied (5.2) to $x-y$ instead of x and changed the order of integration. The obtained result is the wanted translation formula,

$$F(p+x) \simeq \int_{|z|=1} U_\infty(z) T_L\left(k|p|, \frac{p}{|p|}\cdot z\right) e^{ikx\cdot z}dz, \tag{5.7}$$

where

$$U_\infty(z) = \frac{1}{4\pi}\int_D e^{-iky\cdot z}\rho(y)dy, \quad |z| = 1, \tag{5.8}$$

is the far field due to the source distribution ρ.

Usually for the interpolation and anterpolation in MLFMA, the integral in (5.7) is thought to be evaluated by expanding $U_\infty(z)e^{ikx\cdot z}$ in spherical harmonics. It can be shown that an L, the degree of T_L, which makes (5.7) accurate is also a sufficient number of terms needed in the truncated series expansion of $U_\infty(z)e^{ikx\cdot z}$ to make it accurately estimate $U_\infty(z)e^{ikx\cdot z}$ for all x with $|x| \le r$. In practice, this estimation is done by simply representing $U_\infty(z)$ and $e^{ikx\cdot z}$ by their sample matrices. Because the spherical harmonics are orthogonal with respect to the scalar product induced by the integral over the unit sphere, the usual strategy for interpolation and anterpolation follows.

Because we have expanded U_∞ and $e^{ikx\cdot z}$ in trigonometric polynomials, we need to change the integration in (5.7) over the parameter square $-\pi \le \theta \le \pi, -\pi \le \phi \le \pi$ for the spherical coordinates θ and ϕ, because the terms $e^{i(m\theta+n\phi)}$ are orthogonal functions only with respect to scalar product induced by the integral over that square.

For making the needed change in the integral in (5.7), note that all functions involved are smooth on the unit sphere, and in the spherical coordinates they can be extended to be 2π-periodic in both θ and ϕ for all $\theta, \phi \in \mathbb{R}$. We obtain

$$(5.9) \quad \int_{|z|=1} U_\infty(z)T_L(z)e^{ikx\cdot z}dz = \int_{-\pi}^{\pi}\left[\int_0^{\pi} U_\infty(\theta,\phi)T_L(\theta,\phi)e^{ikx\cdot z(\theta,\phi)}\sin\theta\,d\theta\right]d\phi$$
$$= \frac{1}{2}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi} U_\infty(\theta,\phi)T_L(\theta,\phi)e^{ikx\cdot z(\theta,\phi)}|\sin\theta|d\theta d\phi,$$

where $z(\theta,\phi) = (\sin\theta\cos\phi, \sin\theta\sin\phi, \cos\theta)$, $U_\infty(\theta,\phi) = U_\infty(z(\theta,\phi))$, and $T_L(\theta,\phi) = T_L(z(\theta,\phi))$. Above, we have also used the fact that for a spherical function $v(\theta,\phi)$ we get

$$\int_{-\pi}^{\pi}\left[\int_{-\pi}^{0} v(\theta,\phi)|\sin\theta|d\theta\right]d\phi = \int_{-\pi}^{\pi}\left[\int_0^{\pi} v(\theta,\phi+\pi)|\sin\theta|d\theta\right]d\phi$$
$$= \int_{-\pi}^{\pi}\left[\int_0^{\pi} v(\theta,\phi)|\sin\theta|d\theta\right]d\phi,$$

because $v(-\theta,\phi) = v(\theta,\phi+\pi)$ and $v(\theta,\phi)$ is 2π-periodic in ϕ.

Note that in the right-hand side of (5.9) in the integral there is $|\sin\theta|$, which is not a smooth function at $\theta = 0$. This prevents us from approximating it in a numerically efficient way by sampling and using a trigonometric polynomial. In fact, this problem arises only in the variable θ and in the anterpolation in the disaggregation steps. However, the Fourier series of $|\sin\theta|$ is well known, and it can be directly used with the anterpolation Theorem 4.1 for implementing an efficient anterpolation strategy in the variable θ.

6. Aggregation and disaggregation steps in MLFMA. In this section, we consider the tree-like data structure of MLFMA in three dimensions by using FFT in the interpolation and anterpolation operations. The data structure is constructed in aggregation and disaggregation iteration steps.

In the aggregation iteration steps we construct the far field representations for each level. At every level M we choose an appropriate sampling rate (N, N') so that the far field is represented by a $(2N+1, 2N')$-sample matrix with a desired accuracy. Because the matrix is spherical, only a $(2N+1, N'+1)$-matrix must be stored.

Also we choose the order L for the translation function T_L for the level M. The order L can be chosen, for instance, as suggested in [3]. The sampling rate (N, N') could be taken so that $N = N' = L+1$, as is the usual choice with the spherical harmonics, or N and N' could, possibly, be chosen more economically. The optimal, accuracy-dependent values for N and N' can be numerically searched and pretabulated; see the numerical example in the next section.

After these choices, the aggregation step from level $M+1$ to level M will be performed. Let Q be a cube on level M, and let (N_1, N_1') and (N_2, N_2') be the sampling rates for levels $M+1$ and M, respectively. Let $Q_1, \ldots, Q_r$ be the level $M+1$ nonempty subcubes of Q, q_j the vector from the center of Q to that of Q_j, and U_j the trigonometric polynomial approximating the far field due to sources in Q_j and represented by the $(2N_1+1, 2N_1')$-matrix $u_j = \mathit{Sample}(U_j)$. The far field corresponding to Q is approximated by

$$U(z) = \sum_{j=1}^{r} e^{-ikq_j \cdot z} U_j(z), \quad |z| = 1,$$

and that is approximated by a trigonometric polynomial U_∞ with a sample matrix $u_\infty = \mathit{Sample}(U_\infty)$. The matrix u_∞ is formed by interpolating and shifting as follows,

$$u_\infty(m,n) = \sum_{j=1}^{r} v_j(m,n) w_j(m,n), \tag{6.1}$$

where $w_j = \mathit{Interp}(\mathit{Sample}(U_j), N_2, N_2')$, and v_j is the sample matrix

$$v_j(m,n) = e^{-ikq_j \cdot z(\theta_m, \phi_n)},$$

where $\theta_m = m\frac{2\pi}{(2N_2+1)}$, $-N_2 \le m \le N_2$, and $\phi_n = n\frac{\pi}{N_2'}$, $-N_2' \le n \le N_2' - 1$. The matrix u_∞ is saved, and the aggregation iteration step is completed.

For the disaggregation step from level $M-1$ to M, we consider a cube Q on level M. Let $M < M_{\max}$. Our aim is to form the amplitude field V of the incoming field due to sources outside both Q and its immediate level M neighbor cubes.

The field V is divided into two parts: $V = V_1 + V_2$.

The *interaction list* of Q is defined to be the list of those level M subcubes that are neither Q itself nor its immediate neighbors but are contained in the immediate level $M-1$ neighbor cubes of the parental cube P of Q. The field V_1 is due to the sources in the subcubes of the interaction list of Q. The field V_2 is due to sources outside P and its immediate $M-1$ level neighbor cubes.

First consider V_1. Let Q_j be a cube in the interaction list of Q. The field F due to sources in Q_j is represented as a far field U_∞, with origin at the center c_j of Q_j. Let U_∞ have been stored as a sample matrix u_∞ sampled with the level M sampling rate (N_1, N_1'). We need to translate F to be an incoming field in Q and to form its amplitude field. Let $T_L(z) = T_L(k|p|, \frac{p}{|p|} \cdot z)$ be the translation function from Q_j to Q with p being the vector from the center c_j of Q_j to that of Q. For convenience, we may assume that $c_j = 0$. Due to (5.7) and (5.9),

$$F(p+x) \simeq \frac{1}{2} \int_{-\pi}^{\pi} \left(\int_{-\pi}^{\pi} U_\infty(\theta,\phi) T_L(\theta,\phi) e^{ikx\cdot z(\theta,\phi)} d\phi \right) |\sin\theta| d\theta. \tag{6.2}$$

We first treat the integration with respect to ϕ. Due to the proper choice of L, the function $U_\infty(\theta,\phi)e^{ikx\cdot z(\theta,\phi)}$, with $p+x$ in Q, can be approximated by a spherical harmonics expansion of degree L which is also a trigonometric polynomial of degree $(L, L+1)$, and the function $e^{-ikx\cdot z(\theta,\phi)}$ can be approximated by a trigonometric polynomial of degree (N_1, N_1'). By applying Lemma 4.2 to the ϕ-integration with $T_L = U_1$, $U_\infty = U_2$, and $e^{-ikx\cdot z} \simeq U_3$, we get, after changing the order of integration,

$$F(p+x) \simeq \frac{1}{2} \int_{-\pi}^{\pi} \left(\int_{-\pi}^{\pi} W(\theta,\phi) |\sin\theta| e^{ikx\cdot z(\theta,\phi)} d\theta \right) d\phi, \tag{6.3}$$

where $W(\theta,\phi)$ is the trigonometric polynomial formed from the sample matrix $s(n) = U_\infty(\theta,\phi_n)T_L(\theta,\phi_n)$, $\phi_n = n\pi/(L+1)$, $-L-1 \leq n \leq L$, by anterpolating s in the ϕ-variable from degree $L+1$ to N_1', as Lemma 4.2 states. Accordingly, $\mathit{Degree}(W) = (N_1+L, N_1')$, because $\mathit{Degree}(U_\infty T_L) = (N_1+L, N_1'+L+1)$.

In practice, we form $w = \mathit{Sample}(W)$ as follows. Start with sample matrices u_∞ and $\mathit{Sample}(T_L)$ and interpolate both of them to sampling rate $(N_1+L, L+1)$. Thereafter, multiply them with each other, element by element, to get a matrix s, and anterpolate that in the ϕ-variable, i.e., rowwise, to degree (N_1+L, N_1'), to get w.

Next we treat the integration with respect to θ in (6.3). Because $e^{ikx\cdot z(\theta,\phi_n)}$ can be approximated in θ by a trigonometric polynomial of degree N_1, we can anterpolate $\frac{1}{2}W(\theta,\phi_n)|\sin\theta|$ down to degree N_1 without changing the value of the integral within the desired accuracy. This anterpolation in θ is carried out in four steps as Theorem 4.1 describes. First, interpolate $\frac{1}{2}\mathit{Sample}(W) = \frac{1}{2}w$ in θ, i.e., columnwise, from sampling rate (N_1+L, N_1) to $(2N_1+L, N_1)$ and get the sample matrix u_1. Then anterpolate $|\sin\theta|$ down to degree $2N_1+L$ as follows: truncate the series expansion

$$(6.4) \qquad |\sin\theta| = \sum_{m=-\infty}^{\infty} a_m e^{im\theta}, \qquad a_m = \begin{cases} \frac{2}{\pi}\frac{1}{1-m^2} & \text{if } m \text{ is even,} \\ 0 & \text{if } m \text{ is odd,} \end{cases}$$

by summing m only from $-2N_1-L$ to $2N_1+L$; get a trigonometric polynomial S of order $2N_1+L$; and let $u_2 = \mathit{Sample}(S)$. Next multiply the columns of u_1 by u_2, element by element, and get u_3, i.e., $u_3(m,n) = u_1(m,n)u_2(m)$, for $-2N_1-L \leq m \leq 2N_1+L$ and $-N_1' \leq n \leq N_1'-1$. Thereafter, anterpolate u_3 in θ, i.e., columnwise, from degree $(2N_1+L, N_1')$ to (N_1, N_1'), and get the sample matrix v_{Q_j}. The trigonometric polynomial V_{Q_j}, with $\mathit{Sample}(V_{Q_j}) = v_{Q_j}$, is the wanted amplitude field, i.e.,

$$(6.5) \qquad F(p+x) \simeq \int_{-\pi}^{\pi}\int_{-\pi}^{\pi} V_{Q_j}(\theta,\phi)e^{ikx\cdot z(\theta,\phi)}d\theta d\phi$$

for $p+x$ in Q. Finally, sum up v_{Q_j} over the subcubes Q_j in the interaction list of Q, and get v_1 so that the trigonometric polynomial V_1, with $\mathit{Sample}(V_1) = v_1$, is the wanted joint amplitude field.

Note that, in practice, there is a more economical order of operations for forming the needed v_1. Namely, form the above sample matrices s of the products $U_\infty T_L$ for each Q_j and sum them up; anterpolate the sum matrix u_1 rowwise and, thereafter, columnwise interpolate it to get u_3; and multiply it by u_2 and anterpolate the product as explained above. Also in practice, we can take the θ-direction sizes of the matrices s, u_3 and the length of vector u_2 to be smaller than the above theoretical upper bounds are, as the example in the next section shows. This seems to be due to the fact that, in practice, the anterpolated $|\sin\theta|$ increases the degree of $U_\infty T_L|\sin\theta|$ less than Theorem 4.1 suggests.

Next we treat the amplitude field V_2 due to sources outside P and its immediate level $M-1$ neighbor cubes. In the previous iteration step this amplitude field is already formed with origin at the center c' of P and stored as a sample matrix, say w, sampled with the level $M-1$ sampling rate (N_2, N_2'). We shift the origin from c' to the center c of Q, and obtain an incoming field in Q of the form

$$(6.6) \qquad \begin{aligned} F(x) &= \int_{-\pi}^{\pi}\int_{-\pi}^{\pi} W(\theta,\phi)e^{ik(p+x)\cdot z(\theta,\phi)}d\theta\, d\phi \\ &= \int_{-\pi}^{\pi}\int_{-\pi}^{\pi} W(\theta,\phi)e^{ikp\cdot z(\theta,\phi)}e^{ikx\cdot z(\theta,\phi)}d\theta\, d\phi \end{aligned}$$

for $c + x$ in Q with $p = c - c'$ and W being the trigonometric polynomial with $\mathit{Sample}(W) = w$. Next apply an obvious two variable version of Lemma 4.2. Because $e^{-ik(p+x)\cdot z}$ and $e^{-ikx\cdot z}$, with $c+x$ in Q, can be approximated with trigonometric polynomials of degrees (N_2, N_2') and (N_1, N_1'), respectively, we can reason that, without changing the value of the integral (6.6) within the desired accuracy, $We^{ikp\cdot z}$ in the right-hand side of (6.6) can be replaced by a trigonometric polynomial V_2, which is obtained by anterpolating S from degree (N_2, N_2') to (N_1, N_1') with

$$\mathit{Sample}(S)(m,n) = s(m,n) = w(m,n)e^{ikp\cdot z(\theta_m,\phi_n)}, \tag{6.7}$$

$\theta_m = m2\pi/(2N_2+1)$, $-N_2 \le m \le N_2$, $\phi_n = n\pi/N_2'$, $-N_2' \le n \le N_2' - 1$. This V_2 is the wanted amplitude field. Thus, in practice, we only form the sample matrix s in (6.7), anterpolate that from degree (N_2, N_2') to (N_1, N_1'), and get the wanted $v_2 = \mathit{Sample}(V_2)$.

Finally, let $V = V_1 + V_2$, and store $v = v_1 + v_2 = \mathit{Sample}(V)$. The disaggregation iteration step from level $M - 1$ to level M is completed.

After having completed the last disaggregation step from level $M_{\max} - 1$ to $M_{\max}$, we want to compute the final incoming field F for any level $M_{\max}$ cube Q due to sources outside Q and its immediate level $M_{\max}$ neighbors. This field we get from the amplitude field V, corresponding to Q, by Lemma 3.3,

$$\begin{aligned} F(c+x) &\simeq \int_{-\pi}^{\pi}\int_{-\pi}^{\pi} V(\theta,\phi)e^{ikx\cdot z(\theta,\phi)}d\theta d\phi \\ &\simeq \frac{4\pi^2}{(2N+1)N'}\left[\sum_{n=0}^{N'-1} v(0,n)e^{ikx\cdot z(0,\phi_n)} + \sum_{m=1}^{N}\sum_{n=-N'}^{N'-1} v(m,n)e^{ikx\cdot z(\theta_m,\phi_n)}\right], \end{aligned} \tag{6.8}$$

where c is the center of Q, $v = \mathit{Sample}(V)$, and $\theta_m = m2\pi/(2N+1)$, $-N \le m \le N$, $\phi_n = n\pi/N'$, $-N' \le n \le N' - 1$, because $e^{-ikx\cdot z(\theta,\phi)}$ can be approximated by a trigonometric polynomial of degree (N, N'), which is the level $M_{\max}$ sampling rate.

7. A numerical example and comparisons. We present a numerical example where the method of this paper for interpolation, anterpolation, and integrating the fields, here called the *DFT method*, is tested and compared with the usual method of using polynomial interpolation, spherical harmonics, and Gaussian integration for the same operations, here called the *filter method*; see [6], [3].

In the example we simulate the main operations of the aggregation and disaggregation steps by setting a distributed source into a cube Q corresponding to a division cube on the level $M = M_{\max} - m$ for $m = 1, 2, \ldots, 5$. We interpolate and shift the associated far fields from the level $M+1$ subcubes of Q into Q, translate the resulting field into a local field in a nonimmediate neighbor P of Q, anterpolate that local field into a level $M+1$ subcube of P, and compare the obtained field to its exact value. This procedure is carried out with both the DFT and filter methods, using in the latter case both 4 × 4 and 6 × 6 Lagrangian interpolation stencils. Finally, we compare the computational costs of the three simulations at given accuracy levels. Our numerical example illustrates the numerical efficiencies of the compared methods in MLFMA with six or fewer levels.

We normalize the wave number $k = 1$. We consider a cube Q on the level $M = M_{\max} - m, m = 1, 2, \ldots, 5$, with the center at the origin, with vertices

$$R_{r_1,r_2,r_3} = 2^m\pi(r_1,r_2,r_3), \quad r_j = \pm\frac{1}{2}, j = 1,2,3, \tag{7.1}$$

and with the side length $s_m = 2^m\pi$, and its nonimmediate neighbor cube P on the same level M with vertices $2s_m(1,0,0) + R_{r_1,r_2,r_3}$, $r_j = \pm\frac{1}{2}, j = 1,2,3$. Accordingly, $s_m = 2^m\lambda/2$ with the wave length $\lambda = 2\pi$ for $k = 1$.

The source is chosen to be the constant planar source density $\rho = 4\pi$ on the diagonal square A,

$$A = \left\{(x_1, x_2, x_3)\epsilon\mathbb{R}^3;\ -\frac{s_m}{2} \le x_1, x_2 \le \frac{s_m}{2},\ x_3 = x_1\right\}. \tag{7.2}$$

The cube Q is divided into eight level $M+1$ subcubes, and those four subcubes containing the source are enumerated $Q_1, \dots, Q_4$. The far field U_j due to the source distribution in Q_j is computed by (5.8) using a 5×5 Gaussian integration grid over the subsquares of A. In fact, this numerical integral is the actual far field U_j in our example. It is accurately sampled with $N_1 + 1$ points in the θ-direction and $2N_1$ points in the ϕ-direction, both uniformly spaced, for the DFT method, and with $L+1$ Gaussian points in the θ-direction and $2(L+1)$ uniformly spaced points in the ϕ-direction for the filter method. The far field U_∞ due to the source ρ in Q is computed by the formula

$$U_\infty = \sum_{j=1}^{4} e^{-ikq_j\cdot z} U_j, \tag{7.3}$$

where q_j is the vector from the center of Q to that of Q_j. The far fields U_j, by using the sample matrices u_j, are approximated by trigonometric polynomials in the DFT method, and by spherical harmonics in the filter method. The far field U_∞ is approximated by a sample matrix u_∞, by sampling U_∞ in a denser level M grid so that the shift function $e^{-ikq_j\cdot z}$ is directly sampled in that denser grid and each u_j is interpolated into that grid.

For the DFT method our example proceeds as follows. On each level M, we choose the degree $L \ge N_1$ for the translation function T_L and take the denser grid to be an $(N_2+1) \times 2(L+1)$ grid with $N_2 \ge L$. Thereafter, T_L is sampled in that grid. Additionally, u_∞ is formed in that grid by DFT interpolation and shifting. The element-by-element product of the two sample matrices is formed, and its columns are multiplied by the sample vector of $\frac{1}{2}Anterp(|\sin\theta|, N_2)$. The obtained local field in the cube P is shifted and DFT anterpolated down to degree (N_1, N_1) into a level $M+1$ subcube P_1 of P with center $c+q$, where $c = 2s_m(1,0,0)$ is the center of P and $q = s_m/4(-1,1,1)$. Finally, the local field in P_1 is integrated using (6.8), the obtained field is compared to the exact field, and an average relative error E_{DFT} in P_1 is computed.

For illustrating the efficiency of the DFT method, for each accuracy level $10^{-p}, p = 1,2,3$, we choose the smallest L, N_1, and N_2 so that $E_{DFT} \le 10^{-p}$. We also estimate the computational cost of the entire procedure in the example with the DFT method for each MLFMA level M and accuracy level p. The cost is estimated by counting the number of scalar multiplications and using the cost $n\log_2 n$ for the FFT of a vector with length n. The results are presented in Table 7.1.

The filter method is treated in our example as follows. For each level $M = M_{\max} - m$, $m = 1, 2, \dots, 5$, an optimal degree L_M for the translation function T_L is chosen in a way explained later. On the level M the sample matrix u_∞ is formed in the grid $(L_M + 1) \times 2L_M$ from the level $M+1$ sample matrices u_j by the polynomial interpolation and shifting using both the 4-points and 6-points Lagrangian interpolation in the ϕ- and θ-directions. The function T_L with $L = L_M$ is sampled in the same grid and multiplied,

Table 7.1

Comparison of the computational costs of the DFT method and the filter method with stencil sizes 4×4 and 6×6; here $p = 1, 2, 3$ refers to accuracy level 10^{-p}.

		DFT				4×4 stencil		6×6 stencil	
M	p	L	N_1	N_2	cost/10^3	L	cost/10^3	L	cost/10^3
$M_{\max} - 1$	1	8	5	8	7.1	11	12.1	8	10.0
	2	9	5	11	9.8			14	28.6
	3	13	7	19	24.6				
$M_{\max} - 2$	1	15	7	14	22.1	23	44.9	16	33.5
	2	18	9	21	40.5			27	92.6
	3	20	10	28	58.7				
$M_{\max} - 3$	1	27	12	27	82.1	31	89.2	28	96.3
	2	34	15	35	138			42	215
	3	37	17	49	206				
$M_{\max} - 4$	1	46	20	46	261	61	288	60	386
	2	60	24	60	444			79	684
	3	65	27	76	607				
$M_{\max} - 5$	1	105	37	104	1394	110	932	120	1522
	2	121	44	119	1917			137	2063
	3	130	48	135	2351				

element by element, by u_∞. The local field in P is shifted and anterpolated into the subcube P_1 of P; the anterpolation is carried out as an adjoint operation of the polynomial interpolation from the $(L_{M+1} + 1) \times 2L_{M+1}$ grid to the $(L_M + 1) \times 2L_M$ grid. Finally, the resulting local field in P_1 is computed, and the relative errors $E_{4\times4}$ and $E_{6\times6}$ in P_1 are computed for both stencil sizes.

For illustrating the efficiency of the filter method, for each accuracy level 10^{-p}, $p = 1, 2, 3$, and for each $n \times n$ stencil, $n = 4, 6$, we choose the smallest degrees L_M for $M = M_{\max} - m, m = 1, 2, \ldots, 5$, so that the relative error $E_{n\times n} \leq 10^{-p}$ on each level M. It turns out that this is possible for the 4×4 stencil only for $p = 1$, and for the 6×6 stencil only for $p = 1, 2$; higher accuracy levels for these stencil sizes cannot be reached. Finally, the computational cost for the entire procedure with the filter method is estimated. The cost on the level M for the interpolation, as well as for anterpolation, is about $2n(L_{M+1} + L_M)(L_M + 1)$ for an $n \times n$ stencil when the interpolation is performed first columnwise and then rowwise. The results for both stencil sizes are presented in Table 7.1.

Table 7.1 shows that the computational cost is lowest for the DFT method but roughly of the same order for all three methods. Higher accuracy is reached in the DFT method by only increasing the sample rate, while in the filter method increasing the stencil size is also needed.

The lower cost with the DFT method is due to the fact that the same accuracy level is reached in the DFT method with smaller sampling rates than in the filter method. This mainly follows from the fact that in the filter method for interpolation and anterpolation an oversampling by a factor of 2 is needed [10]; also the filter method, because of the Lagrangian interpolation, is more sensitive to sampling noise than the DFT method.

There is also another point of view on smaller sampling rates. Usually in MLFMA more than half of the computational cost arises from translating outgoing fields into local fields in the disaggregation steps. The cost of this operation for a cube Q depends on the size of the sample matrices for $u_\infty T_L$ and on the length of the interaction list of Q, which may vary from about 20 to 189. In Table 7.1 these matrix sizes are $(N_2+1)\times 2(L+1)$ for the DFT method, and $(L+1)\times 2(L+1)$, with a different L, for the filter method. We see that those sizes are about 2 times larger for the filter methods, and accordingly our example suggests that the DFT method is, altogether, about 1.5 to 2 times more efficient than the filter method, at least for MLFMA with six or fewer levels.

8. Conclusions. A new method is presented for using FFT in interpolation and anterpolation in a 3-D MLFMA. The associated far fields and the amplitude fields of the plane wave expansions of the local fields on the 3-D unit sphere are extended to be 2π-periodic in both the spherical coordinates θ and ϕ. This enables as effective a field presentation in terms of trigonometric polynomials as the usual presentation in the terms of spherical harmonics. Furthermore, with the trigonometric polynomials and DFT, an effective and exact procedure for interpolation and anterpolation, together with field computing, can be facilitated, whereas these operations can be performed only approximately when using spherical harmonics and polynomial interpolation. In a numerical example, the proposed new method is compared to the usual method employing spherical harmonics and polynomial interpolation in 3-D MLFMA. The comparison shows that the proposed new method is about from 1.5 to 2 times more efficient, at least for MLFMA with six or fewer levels.

Acknowledgments. The main part of this paper was written during the author's sabbatical visit in 1999–2000 to the Center for Computational Electromagnetics at the University of Illinois at Urbana-Champaign. The author greatly thanks Prof. Weng Chew and Dr. Jiming Song for valuable discussions and comments. Warm thanks are also due to Prof. Eric Michielssen and Dr. Sanjay Velamparambil for helpful discussions.

REFERENCES

[1] R. Coifman, V. Rokhlin, and S. Wanzura, *The fast multipole method for the wave equation: A pedestrian prescription*, IEEE Antennas and Propagation Magazine, 35 (1993), pp. 7–12.

[2] J. M. Song, C. C. Lu, and W. C. Chew, *Multilevel fast-multipole algorithm for solving combined field integral equations of electromagnetic scattering*, Microwave and Optical Technology Letters, 10 (1995), pp. 14–19.

[3] W. C. Chew, J.-M. Jin, E. Michielssen, and J. Song, *Fast and Efficient Algorithms in Computational Electromagnetics*, Artech House, Boston, London, 2001.

[4] W. C. Chew, J. Jin, C. Lu, E. Michielssen, and J. M. Song, *Fast solution methods in electromagnetics*, IEEE Trans. Antennas and Propagation, 45 (1997), pp. 533–543.

[5] J. M. Song, C. C. Lu, and W. C. Chew, *MLFMA for electromagnetic scattering from large complex objects*, IEEE Trans. Antennas and Propagation, 45 (1997), pp. 1488–1493.

[6] S. Koc and W. C. Chew, *Calculation of acoustical scattering from a cluster of scatterers*, J. Acoust. Soc. Amer., 103 (1998), pp. 721–734.

[7] M. F. Gyure and M. A. Stalzer, *A prescription for the multilevel Helmholtz FMM*, IEEE Comput. Sci. Engrg., 5 (1998), pp. 39–47.

[8] M. Lu, J. Sarvas, and E. Michielssen, *A simplified 3D plane wave time domain (PWTD) algorithm*, in Digest of the 2001 IEEE Antennas and Propagation Society International Symposium, Boston, MA, 2001, pp. 188–191.

[9] D. Colton and R. Kress, *Inverse Acoustic and Electromagnetic Scattering Theory*, Springer-Verlag, Berlin, Heidelberg, New York, 1992.

[10] S. Koc, J. Song, and W. C. Chew, *Error analysis for the numerical evaluation of the diagonal forms of the scalar spherical addition theorem*, SIAM J. Numer. Anal., 36 (1999), pp. 906–921.

SIAM J. Numer. Anal.
Vol. 41, No. 6, pp. 2197–2209

MULTILEVEL FIRST-ORDER SYSTEM LEAST SQUARES FOR NONLINEAR ELLIPTIC PARTIAL DIFFERENTIAL EQUATIONS[*]

A. L. CODD[†], T. A. MANTEUFFEL[‡], AND S. F. MCCORMICK[‡]

Abstract. A fully variational approach is developed for solving nonlinear elliptic equations that enables accurate discretization and fast solution methods. The equations are converted to a first-order system that is then linearized via Newton's method. First-order system least squares (FOSLS) is used to formulate and discretize the Newton step, and the resulting matrix equation is solved using algebraic multigrid (AMG). The approach is coupled with nested iteration to provide an accurate initial guess for finer levels using coarse-level computation. A general theory is developed that confirms the usual full multigrid efficiency: accuracy comparable to the finest-level discretization is achieved at a cost proportional to the number of finest-level degrees of freedom. In a companion paper, the theory is applied to elliptic grid generation (EGG) and supported by numerical results.

Key words. least-squares discretization, multigrid, nonlinear elliptic boundary value problems

AMS subject classifications. 35J65, 65N15, 65N30, 65N50, 65F10

DOI. 10.1137/S0036142902404406

1. Introduction.

We develop a theoretical foundation for a method for solving nonlinear elliptic equations that combines first-order system least squares (FOSLS) with Newton's method, algebraic multigrid (AMG), and nested iteration (NI). The algorithm achieves accuracy comparable to the finest-level discretization at a cost proportional to the number of finest-level degrees of freedom. In a companion paper [17], we apply this theory to the elliptic grid generation (EGG) equations and numerically validate the theory established below.

Our development assumes that the target problem is a first-order system whose associated least-squares functional applied to functions in $H^{1+\delta}$, $\delta \in (0,1)$, has quadratic part that is equivalent to the product H^1 norm. Higher-order differential systems can be recast in the standard way as a first-order system, but care must be taken to ensure such product ellipticity when it is feasible (cf. [15, 14]). Our particular interest is in quasilinear first-order systems where the nonlinearity is a product of variables, of which at most one is a derivative term. For example, if u and v are the variables in a two-dimensional problem, then we admit terms like u, v, u^2u_x, and uvv_y, but not uu_x^2. (Admitting product derivative terms while retaining $H^{1+\delta}$ spaces for the variables would prevent the use of L^2 norms for the equations and inhibit analysis of the linearized equations.)

Our algorithm applies to the first-order system in three separate stages. The outermost stage is NI, which starts on the coarsest level where the discrete nonlinear problem is solved by any appropriate method. The result is then interpolated to the

[*] Received by the editors March 21, 2002; accepted for publication (in revised form) April 16, 2003; published electronically December 5, 2003.
http://www.siam.org/journals/sinum/41-6/40440.html

[†] Centre for Mathematics and Its Applications, School of Mathematical Sciences, Australian National University, Canberra, ACT 0200, Australia (andrea.codd@anu.edu.au.).

[‡] Department of Applied Mathematics, Campus Box 526, University of Colorado at Boulder, Boulder, CO 80309–0526 (tmanteuf@boulder.colorado.edu, stevem@boulder.colorado.edu). This work was sponsored by the National Institute of Health under grant number 1–R01–EY12291–01, the National Science Foundation under grant number DMS–9706866, and the Department of Energy under grant number DE–FG03–93ER25165.

next finer level, where it is used as an initial guess for *one* Newton linearization (the middle stage) of the nonlinear problem. A functional is created on that level using a least-squares principle, and the resulting matrix equation is solved using *one, two, or three* V-cycles of AMG (the innermost stage). The result is then interpolated up to the next finer level, with the steps repeated until the finest level is processed. Our theoretical results confirm that this *direct* NI-Newton-FOSLS-AMG scheme converges in *one* overall step to an approximation on the finest level that is accurate to *the level of discretization error*. Numerical experiments for the EGG equations, described in a companion paper [17], confirm this result.

One advantage that the FOSLS system has over standard minimization techniques is that the minimum value of the functional is zero at the exact solution of the differential equation. This property has implications for adaptive refinement that are to be explored in future work. Another advantage is that FOSLS used with finite element discretization and Newton linearization of the elliptic equations results in self-adjoint positive-definite matrix problems that themselves correspond to a well-posed elliptic system, so the discrete problems can be efficiently solved by multigrid. We demonstrate this attribute qualitatively in the theory here and numerically in the companion paper.

The idea of using Newton iterations coupled with a multilevel scheme is not new. In [20], for example, a multilevel nested iteration Newton scheme was applied to differential eigenproblems. An abstract theory and numerical results confirmed the need for only one Newton step on the finest level. In [5], optimal parameters were calculated for the damped approximate Newton's method to ensure quadratic convergence of a particular finite element approximation of nonlinear elliptic partial differential equations. This was later combined with multilevel techniques [6] to obtain a convergence result that asymptotically required just one Newton linearization per level. This last result is similar to ours but does not include the derivative terms in the nonlinearity that are present in our target application, EGG. The NI approach is also used in recent work on cascadic multigrid [22], although again their form of the nonlinearity does not include the more complicated case needed here.

The "mesh-independence" theory developed for Newton's method in [21, 4, 3, 2] addresses the same property of NI that we exploit here. Unfortunately, this theory cannot easily be applied to our setting because it requires more smoothness of the infinite-dimensional iterates than ours appear to possess. We are also unable to apply the mesh-independence-based theory developed in [18, 19] because the nonlinearity for the Navier–Stokes equations treated there appears only in the lower-order terms.

This paper is organized as follows. Section 2 introduces the equations and function spaces. In section 3, we describe the NI-Newton-FOSLS method for solving the nonlinear equations. Section 4 contains theory on convergence of the Newton iterates in H^1 and on accuracy estimates for the NI-Newton-FOSLS-AMG scheme. We conclude with some remarks in the final section.

2. Setup. We use standard notation for the associated spaces. Restricting ourselves to two dimensions, consider a generic open domain, $\Omega \in R^2$, with Lipschitz boundary Γ. Suppose $m \geq 0$ and $n \geq 1$ are given integers. Let $(\cdot,\cdot)_{0,\Omega}$ denote the inner product on $L^2(\Omega)^n$, $\|\cdot\|_{0,\Omega}$ its induced norm, and $H^m(\Omega)^n$ the standard Sobolev space with norm $\|\cdot\|_{m,\Omega}$ and seminorms $|\cdot|_{i,\Omega}$ $(0 \leq i \leq m)$. (We suppress superscript n, because dependence of the vector norms on dimension is clear by context.) For $\delta \in (0,1)$, let $H^{m+\delta}(\Omega)$ (cf. [11]) denote the Sobolev space associated with the norm defined by

$$\|u\|^2_{m+\delta,\Omega} \equiv \|u\|^2_{m,\Omega} + \sum_{|\alpha|=m} \int_\Omega \int_\Omega \frac{|\partial_\alpha u(x) - \partial_\alpha u(y)|^2}{|x-y|^{2(1+\delta)}} dxdy.$$

(This definition allows the use of the "real interpolation" method [1, 11].) Also, let $H^{\frac{1}{2}}(\Gamma)$ denote the trace Sobolev space associated with the norm

$$\|u\|_{\frac{1}{2},\Gamma} \equiv \inf\{\|v\|_{1,\Omega} : v \in H^1(\Omega), \text{ trace } v = u \text{ on } \Gamma\}.$$

Finally, let $C^m(\Omega)$ denote the space of functions with continuous derivatives in Ω of up to order $m \geq 0$, and define the $C^0(\Omega)$ norm of $f \in C^0(\Omega)$ by

$$\|f\|_{\infty,\Omega} \equiv \sup_{x \in \Omega} |f(x)|.$$

From Sobolev's lemma [1], there exists a constant C, depending only on Ω and δ, such that

$$\|f\|_{\infty,\Omega} \leq C\|f\|_{1+\delta,\Omega} \qquad \forall\ f \in H^{1+\delta}(\Omega).$$

The method we develop applies to elliptic quasi-linear partial differential equations, where the highest-order derivative terms appear linearly, with the exception that their coefficients may include lower-order terms. In fact, we assume that the equations have been formulated as a $p \times q$ first-order system with appropriate boundary conditions. It is straightforward to rewrite higher-order equations in first-order form, although care must be taken to ensure that the resulting system is elliptic in the H^1 product norm (assuming that this is even feasible; cf. [15, 14]). This process is exemplified by the reformulation of the EGG equations in the companion paper [17].

Let the first-order system be represented abstractly by the p-vector equation

$$\mathbf{p}(\mathbf{J}) = \mathbf{0} \qquad \text{in} \quad \Omega, \tag{2.1}$$

with boundary conditions

$$\mathbf{BJ} = \mathbf{g} \qquad \text{on} \quad \Gamma, \tag{2.2}$$

where $\mathbf{J}$ is the q-vector of unknowns. (Here p and q are positive integers, generally with $p \geq q$.) To ensure that we can apply least squares to this system, we must have $\mathbf{p}(\mathbf{J}) \in L^2(\Omega)^p$. We want to allow product terms involving a combination of elements of $\mathbf{J}$, one of which may involve a partial derivative. Therefore, we cannot allow $\mathbf{J}$ to roam freely in $H^1(\Omega)^q$, because such products would not necessarily be in $L^2(\Omega)$ and we would thus be prevented from using $L^2(\Omega)$ norms for the functional. However, from Sobolev's lemma, the $C^0(\Omega)$ norm is bounded by the $H^{1+\delta}(\Omega)$ norm in R^2 when $\delta \in (0,1)$. Thus, everything in $H^{1+\delta}(\Omega)$ is continuous, and our product terms are in $L^2(\Omega)$. We therefore choose the space for $\mathbf{J}$ to be $H^{1+\delta}(\Omega)^q$, ensuring $\mathbf{p}(\mathbf{J}) \in L^2(\Omega)^p$. This obviously places restrictions on the allowable boundary functions $\mathbf{g}$. In fact, we assume that the solution $\mathbf{J}^*$ of (2.1)–(2.2) is in $H^{2+\delta}(\Omega)^q$, which places even further restrictions on $\mathbf{g}$. In addition, the coercivity requirement on the first Fréchet derivative of our system influences the allowable spaces for both the boundary and boundary conditions, which in turn influences the solution space for $\mathbf{J}^*$. This issue is addressed implicitly in the abstract theory of section 4 and in detail for the EGG application in the companion paper [17].

In other FOSLS applications [13, 15, 7, 8, 14], both H^{-1} and L^2 norms are used for the domain, and $H^{\frac{1}{2}}$ norms for the boundary. With appropriate smoothness [15], FOSLS functionals for general second-order elliptic partial differential equations exhibit H^1 equivalence for the functionals based on L^2 norms for the domain and $H^{\frac{1}{2}}$ norms for the boundary. In practice, while L^2 norms are used for the domain, it is common either to use L^2 norms scaled by $\frac{1}{h}$ for the boundary norms or to impose the boundary conditions. We focus here for simplicity on imposing boundary conditions, although some of our numerical results in [17] take the scaled L^2 norm approach for illustration.

It is more convenient in the analysis to consider homogeneous boundary conditions. To this end, we extend $\mathbf{g}$ smoothly into Ω: assume that we are given a q-vector function $\mathbf{E}$, defined on Ω, that satisfies

$$\mathbf{BE} = \mathbf{g} \qquad \text{on } \Gamma.$$

Now, writing $\mathbf{J} = \mathbf{D} + \mathbf{E}$ and $\mathbf{P}(\mathbf{D}) = \mathbf{p}(\mathbf{D} + \mathbf{E})$, our target problem becomes

$$\mathbf{P}(\mathbf{D}) = \mathbf{0} \qquad \text{in } \Omega, \tag{2.3}$$

with homogeneous boundary conditions

$$\mathbf{BD} = \mathbf{0} \qquad \text{on } \Gamma. \tag{2.4}$$

Generally, $\mathbf{E}$ needs to be as smooth as we require $\mathbf{J}^* = \mathbf{D}^* + \mathbf{E}$ to be, but this requirement is implicit in the following assumptions that we make on (2.3)–(2.4).

We start by defining the space on which this system is posed. For any $\nu > 0$, define

$$\mathcal{H}_\nu = \{\mathbf{D} \in H^\nu(\Omega)^q : \mathbf{BD} = \mathbf{0} \text{ on } \Gamma\}. \tag{2.5}$$

We assume that our solution $\mathbf{D}^*$ resides in $\mathcal{H}_{2+\delta}$, and we look for it in $\mathcal{H}_{1+\delta}$. Note that $\mathbf{D} \in \mathcal{H}_{1+\delta}$ implies $\mathbf{P}(\mathbf{D}) \in L^2(\Omega)^p$.

3. Method. There are several decisions to be made about how (2.3) is solved. Our basic choice is to use Newton's method and FOSLS to obtain a quadratic minimization problem, finite elements for the discretization, and then AMG with NI to solve the resulting matrix equations. Within this basic framework, we need to choose how Newton's method and FOSLS are related. A FOSLS-Newton method would involve forming the least-squares functional, setting its gradient to zero, and then solving this nonlinear problem with Newton's method. A Newton-FOSLS method would involve linearizing the equations first, and then forming the least-squares functional and setting its gradient to zero. The gradient equations that result from these two approaches differ only by a term coming from the second Fréchet derivative of the system operator that, near the solution, is dominated by the other operator terms. Because nested iteration guarantees proximity to the solution on each level, the performance of these two approaches tends to be much the same. We therefore focus on the Newton-FOSLS approach because of its theoretical and numerical simplicity.

Our method involves first applying Newton's method to nonlinear system (2.3) on the coarsest finite element level. We then form an L^2 functional from the linearized equations and minimize a coarsest-level discretization of it by AMG (or perhaps a direct matrix solver). The resulting approximate Newton iterate computed at the

$$\mathbf{D}_0 \overset{N}{\to} \mathbf{D}_1 \overset{N}{\to} \mathbf{D}_2 \overset{N}{\to} \mathbf{D}_3 \overset{N}{\to} \mathbf{D}_4 \;\cdots\; \mathbf{D}_n \overset{N}{\to} \mathbf{D}_{n+1} \overset{N}{\to} \;\cdots\; \mathbf{D}^*$$

FIG. 3.1. *The Newton-FOSLS infinite-dimensional algorithm.*

coarsest-level scale is then used on the next finer level as an initial guess for an analogous discrete Newton step there: system (2.3) is linearized about this initial guess, an L^2 minimization principle is applied, and then AMG is used to approximate the minimizer on this finer scale. This process continues, with the iterates approximated on successively finer levels, until a desired accuracy is reached.

Applying Newton's method to system (2.3) gives us the following linearized problem: given $\mathbf{D}_n \in \mathcal{H}_{1+\delta}$, find $\mathbf{D}_{n+1} \in \mathcal{H}_{1+\delta}$ such that

$$\mathbf{P}'(\mathbf{D}_n)[\mathbf{D}_{n+1} - \mathbf{D}_n] = -\mathbf{P}(\mathbf{D}_n), \tag{3.1}$$

where $\mathbf{P}'(\mathbf{D}_n)[\mathbf{D}_{n+1} - \mathbf{D}_n]$ denotes the first Fréchet derivative of $\mathbf{P}(\mathbf{D}_n)$ with respect to $\mathbf{D}_n$ in direction $\mathbf{D}_{n+1} - \mathbf{D}_n$.

To solve (3.1) for $\mathbf{D}_{n+1}$, consider the least-squares functional

$$\begin{aligned}\mathbf{G}_0(\mathbf{D}_{n+1}) &\equiv \|\mathbf{P}'(\mathbf{D}_n)[\mathbf{D}_{n+1} - \mathbf{D}_n] + \mathbf{P}(\mathbf{D}_n)\|^2_{0,\Omega} \\ &= \left(\mathbf{P}'(\mathbf{D}_n)[\mathbf{D}_{n+1} - \mathbf{D}_n] + \mathbf{P}(\mathbf{D}_n),\;\; \mathbf{P}'(\mathbf{D}_n)[\mathbf{D}_{n+1} - \mathbf{D}_n] + \mathbf{P}(\mathbf{D}_n)\right).\end{aligned}$$

Note that $\mathbf{G}_0$ depends on $\mathbf{D}_n$ and $\mathbf{E}$. To minimize $\mathbf{G}_0$, we set to zero its first Fréchet derivative, taken with respect to $\mathbf{D}_{n+1}$ (cancelling the factor 2 for convenience): given $\mathbf{D}_n \in \mathcal{H}_{1+\delta}$, find $\mathbf{D}_{n+1} \in \mathcal{H}_{1+\delta}$ such that

$$(\mathbf{P}'(\mathbf{D}_n)[\mathbf{K}],\;\; \mathbf{P}'(\mathbf{D}_n)[\mathbf{D}_{n+1} - \mathbf{D}_n] + \mathbf{P}(\mathbf{D}_n)) = \mathbf{0} \quad \forall\, \mathbf{K} \in \mathcal{H}_{1+\delta}. \tag{3.2}$$

We illustrate this infinite-dimensional Newton process in Figure 3.1, where $\overset{N}{\to}$ indicates one Newton step.

In practice, we need to discretize (3.2) on some given finite element space H^h. However, we then need an approximation for $\mathbf{D}_n$. The main point to keep in mind in this approximation is that early iterates are relatively crude approximations to $\mathbf{D}^*$, and thus they can be approximated on relatively coarse grids. In general, if final iterate $\mathbf{D}_{n+1}$ is approximated by a best approximation $\mathbf{U}_{n+1}$ in H^{h_n}, then $\mathbf{D}_n$ need only be approximated on a grid with mesh size $\mathcal{O}(h_n^{\frac{1}{2}})$. This is a natural consequence of quadratic convergence, and this premise is served well by a coarse grid of mesh size $2h$. This view gives rise to our nested iteration approach that supplies the initial guess for one Newton iterate on grid h by first iterating on grid $2h$. In particular, consider a nested sequence of $m+1$ finite-dimensional subspaces of $\mathcal{H}_{1+\delta}$ denoted by $H^{h_0} \subset H^{h_1} \subset \cdots \subset H^{h_m} \subset \mathcal{H}_{1+\delta}$, where $h_n = 2^{-n}h_0$, $0 \le n \le m$. Note that piecewise bilinears on rectangles are in $H^{1+\delta}(\Omega)$ for $\delta \in [0, \frac{1}{2})$ (see [16]). Let $\mathbf{U}_0$ denote the initial guess in H^{h_0}. For $n = 0, 1, \ldots, m$ in turn, define the grid H^{h_n} problem as follows: given $\mathbf{U}_n \in H^{h_n}$, find $\mathbf{U}_{n+1} \in H^{h_n} \subset H^{h_{n+1}}$ such that

$$\big(\mathbf{P}'(\mathbf{U}_n)[\mathbf{K}^{h_n}],\;\; \mathbf{P}'(\mathbf{U}_n)[\mathbf{U}_{n+1} - \mathbf{U}_n] + \mathbf{P}(\mathbf{U}_n)\big) = \mathbf{0} \quad \forall\, \mathbf{K}^{h_n} \in H^{h_n}. \tag{3.3}$$

3.1. NI-Newton-FOSLS. The discretization in (3.3) amounts to approximating the finest-level solution $\mathbf{U}^*_m$ by a nested iteration on subspaces H^{h_n}, $n = 0, 1, \ldots, m$. This NI approach involves first solving problem (3.3) on the coarsest subspace, H^{h_0}.

$$\begin{tikzcd}
\mathbf{U}_0 \arrow[r, "Nh_0"] \arrow[dr, "N"] & \mathbf{U}_1 \arrow[r, "Nh_1"] \arrow[dr, "N"] & \mathbf{U}_2 \arrow[r, "Nh_2"] \arrow[dr, "N"] & \mathbf{U}_3 \arrow[r, "Nh_3"] \arrow[dr, "N"] & \mathbf{U}_4 \cdots \arrow[dr, phantom, "\ddots"] & \\
& \widetilde{\mathbf{D}}_1 & \widetilde{\mathbf{D}}_2 & \widetilde{\mathbf{D}}_3 & \widetilde{\mathbf{D}}_4 & \cdots
\end{tikzcd}$$

FIG. 3.2. *The NI-Newton-FOSLS algorithm.*

In practice, we can use any sensible solution process here because this space is presumably of very low dimension. We can simply iterate with a (possibly damped) discrete Newton iteration until the error in the approximation is below discretization error. However, because our theory assumes that we are sufficiently close to $\mathbf{D}^*$, we have assumed, for convenience, that this coarsest approximation is computed by only one discrete Newton iteration applied to a sufficiently close approximation $\mathbf{U}_0 \in H^{h_0}$. Now, the resulting iterate $\mathbf{U}_1$ is interpolated to the next finer level, where it is used as an initial guess for one discrete Newton step. The resulting approximation $\mathbf{U}_2$ on subspace H^{h_1} is then used as an initial guess for the next finer level. In general, the initial guess for Newton on level h_n comes from the final Newton step on level h_{n-1}: $\mathbf{U}_n$. The process is repeated until the finest subspace is reached, where one final Newton step is then applied. Note that $\mathbf{U}_{n+1}$ can be interpreted as a discrete approximation to the result $\widetilde{\mathbf{D}}_{n+1}$ of one infinite-dimensional Newton step applied to $\mathbf{U}_n$. This NI procedure is illustrated in Figure 3.2, where $\overset{Nh_n}{\to}$ indicates one Newton step on H^{h_n} and $\overset{N}{\searrow}$ indicates one infinite-dimensional Newton step applied to discrete initial guess $\mathbf{U}_n$.

One of our main objectives in this paper is to prove that this nested iteration process, involving only one discrete Newton step on each level, produces a result on the finest level that is within discretization error of the infinite-dimensional solution.

3.2. AMG. Our theory assumes a standard V-cycle multigrid algorithm because of its superior theoretical basis. However, because of its enhanced robustness, we use AMG in practice as the matrix solver for approximating $\mathbf{U}_{n+1}$. See [12] for basic descriptions of multigrid and AMG.

AMG starts on the coarsest level with initial guess $\mathbf{V}_0 = \mathbf{U}_0$. We apply ν_0 cycles of AMG to the matrix problem arising from (3.3) with $n = 0$. The result, $\mathbf{V}_1$, becomes the initial guess for level h_1, where the process continues. In general, the initial guess for AMG on level h_n comes from the final AMG approximation on level h_{n-1}: $\mathbf{V}_n$. In Figure 3.3, we illustrate the NI-Newton-FOSLS-AMG algorithm, with $\overset{\mathcal{M}h_n}{\to}$ denoting one approximate multigrid-Newton step on H^{h_n}, and $\overset{Nh_n}{\searrow}$ denoting the exact discrete Newton step with initial guess $\mathbf{V}_n$ (with result $\widetilde{\mathbf{U}}_{n+1}$).

$$\begin{tikzcd}
\mathbf{V}_0 \arrow[r, "\mathcal{M}h_0"] \arrow[dr, "Nh_0"] & \mathbf{V}_1 \arrow[r, "\mathcal{M}h_1"] \arrow[dr, "Nh_1"] & \mathbf{V}_2 \arrow[r, "\mathcal{M}h_2"] \arrow[dr, "Nh_2"] & \mathbf{V}_3 \arrow[r, "\mathcal{M}h_3"] \arrow[dr, "Nh_3"] & \mathbf{V}_4 \cdots \arrow[dr, phantom, "\ddots"] & \\
& \widetilde{\mathbf{U}}_1 & \widetilde{\mathbf{U}}_2 & \widetilde{\mathbf{U}}_3 & \widetilde{\mathbf{U}}_4 & \cdots
\end{tikzcd}$$

FIG. 3.3. *The NI-Newton-FOSLS-AMG algorithm.*

4. Abstract theory.

Bounds on various quantities used in the theory developed here involve many different constants. To avoid proliferation, we use uppercase C to denote a generic constant that, unless otherwise specified, can change meaning with each occurrence. When it is important to track the origin of these constants, we instead use lowercase c with unique subscripts. In every occurrence, these constants are independent of h and n, but they may depend on the value of the $H^{1+\delta}(\Omega)^q$ norm of the approximation. (Here and in what follows, $\delta \in (0,1)$ is a fixed constant.) To make sure that these values are properly bounded, we start with an initial guess in a small $H^{1+\delta}(\Omega)^q$ ball about $\mathbf{D}^*$. We show that the approximations remain in this ball and, in fact, attain order h accuracy in the $H^1(\Omega)^q$ norm. This also controls the values of the $H^{1+\delta}(\Omega)^q$ norm (see Lemma 4.5 below).

Assume that there exists a solution $\mathbf{D}^*$ of (2.3) in $\mathcal{H}_{2+\delta}$. (Recall that $\mathcal{H}_\nu$ is defined in (2.5) for any $\nu > 0$.) Denote the open $H^{1+\delta}(\Omega)^q$ ball centered at $\mathbf{D}^*$ of radius $r > 0$ by $B_r \equiv \{\mathbf{D} \in \mathcal{H}_{1+\delta} : \|\mathbf{D}^* - \mathbf{D}\|_{1+\delta,\Omega} < r\}$. Several of our norm assumptions and estimates involve both integer and fractional norms. So that our statements apply to both cases, we let $\epsilon = 0$ or δ. Assume now that $\mathbf{P}[\mathbf{D}] \in H^\epsilon(\Omega)^p$ for every $\mathbf{D} \in B_r$: there exists a constant C, depending only on $\mathbf{D}^*$, $\mathbf{E}$, r, Ω, and δ, such that

$$\|\mathbf{P}(\mathbf{D})\|_{\epsilon,\Omega} \le C \quad \forall\ \mathbf{D} \in B_r.$$

Further assume uniform coercivity and continuity of $\mathbf{P}'(\mathbf{D})[\,\cdot\,]$ as a mapping from $\mathcal{H}_{1+\epsilon}(\Omega)$ to $H^\epsilon(\Omega)^p$: for every $\mathbf{D} \in B_r$, there exist constants c_c and c_b, depending only on $\mathbf{D}^*, \mathbf{E}, r, \Omega$, and δ, such that

$$\frac{1}{c_c}\|\mathbf{K}\|_{1+\epsilon,\Omega} \le \|\mathbf{P}'(\mathbf{D})[\mathbf{K}]\|_{\epsilon,\Omega} \le c_b\|\mathbf{K}\|_{1+\epsilon,\Omega} \quad \forall\ \mathbf{K} \in \mathcal{H}_{1+\epsilon}(\Omega). \tag{4.1}$$

Note that coercivity implies that $\mathbf{P}'(\mathbf{D})[\,\cdot\,]$ is one-to-one on $\mathcal{H}_{1+\epsilon}(\Omega)$ for every $\mathbf{D} \in B_r$, including, of course, $\mathbf{D} = \mathbf{D}^*$. We also assume boundedness of the second Fréchet derivative of $\mathbf{P}(\mathbf{D})$ for all $\mathbf{D} \in B_r$: for every $\mathbf{D} \in B_r$ there exists a constant c_2, depending only on $\mathbf{D}^*$, $\mathbf{E}$, r, Ω, and δ, such that

$$\|\mathbf{P}''(\mathbf{D})[\mathbf{K},\mathbf{K}]\|_{\epsilon,\Omega} \le c_2\|\mathbf{K}\|_{1+\delta,\Omega}\|\mathbf{K}\|_{1+\epsilon,\Omega} \quad \forall\ \mathbf{K} \in \mathcal{H}_{1+\epsilon}(\Omega). \tag{4.2}$$

Here, $\mathbf{P}''(\mathbf{D})[\mathbf{K},\mathbf{K}]$ denotes the second Fréchet derivative of $\mathbf{P}(\mathbf{D}_n)$ with respect to $\mathbf{D}_n$ in directions $\mathbf{K}$ and $\mathbf{K}$.

Let $\mathcal{P}^h$ and $\mathcal{Q}^h$ denote the respective $H^{1+\delta}$ and H^1 projections of $\mathcal{H}_{1+\delta}$ onto H^h. Note that

$$\|\mathcal{P}^h\mathbf{D}\|_{1+\delta,\Omega} \le \|\mathbf{D}\|_{1+\delta,\Omega} \quad \forall\ \mathbf{D} \in \mathcal{H}_{1+\delta}$$

and

$$\|\mathcal{Q}^h\mathbf{D}\|_{1,\Omega} \le \|\mathbf{D}\|_{1,\Omega} \quad \forall\ \mathbf{D} \in \mathcal{H}_{1+\delta}. \tag{4.3}$$

Assume that our finite element spaces satisfy the usual *approximation properties* (cf. [9]):

$$\|\mathbf{D}^* - \mathcal{P}^h\mathbf{D}^*\|_{\gamma,\Omega} \le c_d h^{2+\delta-\gamma}\|\mathbf{D}^*\|_{2+\delta,\Omega} \quad \forall\ \gamma \in [0, 1+\delta] \tag{4.4}$$

and

$$\|\mathbf{D}^* - \mathcal{Q}^h\mathbf{D}^*\|_{1,\Omega} \le c_d h^{1+\delta}\|\mathbf{D}^*\|_{2+\delta,\Omega}. \tag{4.5}$$

Assume that they also satisfy the *inverse estimate* (cf. [9, 11]):

$$\|\mathbf{U}\|_{\beta,\Omega} \le \frac{c_i}{h^{\beta-\gamma}}\|\mathbf{U}\|_{\gamma,\Omega} \quad \forall\ \mathbf{U} \in H^h,\ \beta \in [0, 1+\delta],\ \gamma \in [0, \beta]. \tag{4.6}$$

Assume finally that h_0 is so small that $B_r \cap H^{h_0} \neq \phi$ and that initial guess $\mathbf{U}_0$ is in $B_r \cap H^{h_0}$.

The following theory shows that $\mathbf{U}_n$ is in an $H^1(\Omega)$ ball about $\mathbf{D}^*$ of radius $(1+\eta)c_d h_n$, where η is any predetermined positive constant and c_d is the constant in (4.4) and (4.5). For simplicity, we choose $\eta = 1$ and thus define

$$\mathcal{S}_n = \{\mathbf{U} \in H^{h_n} : \|\mathbf{D}^* - \mathbf{U}\|_{1,\Omega} \le 2c_d h_n\} \tag{4.7}$$

and

$$\mathcal{S} = \cup_{n=0}^{m} \mathcal{S}_n.$$

Lemma 4.5 shows that $\mathcal{S}$ is bounded in $H^{1+\delta}(\Omega)^q$ and, hence, compact in $H^1(\Omega)^q$.

We first state our three central theorems. Their proofs follow from a series of results developed in the next subsection.

For all three theorems, we assume that $r > 0$ is sufficiently small. For Theorem 4.2, with r fixed, we assume further that $h_0 > 0$ is sufficiently small, especially so that $\mathcal{S}_0 \subset B_r$. For Theorem 4.1, with r fixed, we assume that $h_0 > 0$ is possibly smaller still. We do this so that, in addition to $\mathcal{S}_0 \subset B_r$, we are sure that the exact discrete iterate $\mathbf{U}_{n+1}$ is even closer to $\mathbf{D}^*$ than to $2c_d h_n$, which in turn allows us to deduce that the multigrid approximation $\mathbf{V}_{n+1}$ is within $2c_d h_n$ of $\mathbf{D}^*$. Finally, Theorem 4.3 also assumes that, on each level, the discrete Newton problem is approximately solved with a sufficient but fixed number ν_0 of multigrid V-cycles.

THEOREM 4.1 (Newton). *With $\mathbf{D}_n \in B_r$ given, let $\mathbf{D}_{n+1}$ be the exact infinite-dimensional Newton step defined by* (3.2). *Then $\mathbf{D}_{n+1} \in B_r$ and there exists a constant c_q, depending only on $\mathbf{D}^*$, $\mathbf{E}$, Ω, and δ, such that*

$$\|\mathbf{D}^* - \mathbf{D}_{n+1}\|_{1+\epsilon,\Omega} \le c_q\|\mathbf{D}^* - \mathbf{D}_n\|_{1+\epsilon,\Omega}\|\mathbf{D}^* - \mathbf{D}_n\|_{1+\delta,\Omega}, \quad \epsilon = 0, \delta. \tag{4.8}$$

THEOREM 4.2 (discrete Newton). *Assume that $\mathbf{U}_0 \in S_0$. Then $\mathbf{U}_{n+1} \in \mathcal{S}_{n+1}$: the Newton approximation on level h_n based on initial guess $\mathbf{U}_n$ satisfies the error bound*

$$\|\mathbf{D}^* - \mathbf{U}_{n+1}\|_{1,\Omega} \le 2c_d h_{n+1} = c_d h_n. \tag{4.9}$$

THEOREM 4.3 (inexact discrete Newton). *Assume that $\mathbf{V}_0 \in S_0$. Then $\mathbf{V}_{n+1} \in \mathcal{S}_{n+1}$: the multigrid approximation on level h_n based on initial guess $\mathbf{V}_n$ satisfies the error bound*

$$\|\mathbf{D}^* - \mathbf{V}_{n+1}\|_{1,\Omega} \le 2c_d h_n. \tag{4.10}$$

4.1. Preliminaries.

Although we pose problem (2.3) on $\mathcal{H}_{1+\delta} \subset H^{1+\delta}(\Omega)^q$, we prove convergence in the weaker $H^1(\Omega)^q$ norm. Since $\mathcal{H}_{1+\delta}$ is not complete in the $H^1(\Omega)^q$ norm, we cannot appeal to standard Newton convergence theory. Fortunately, the result we need (Theorem 4.1) is weaker.

LEMMA 4.4. *Let $\mathbf{D} \in B_r$ and $\widetilde{\mathbf{D}} = \theta\mathbf{D} + (1-\theta)\mathbf{D}^*$, with $\theta \in [0, 1]$. Then*

$$\|\widetilde{\mathbf{D}}\|_{1+\delta,\Omega} \le r + \|\mathbf{D}^*\|_{1+\delta,\Omega}.$$

Proof. The result follows directly from the triangle inequality and is thus omitted. □

LEMMA 4.5. *Suppose that* $\mathbf{U}_n \in \mathcal{S}_n$. *Then*

$$\|\mathbf{D}^* - \mathbf{U}_n\|_{1+\delta,\Omega} \le c_\delta h_n^{1-\delta}, \tag{4.11}$$

where $c_\delta = 2c_ic_d + (1+c_i)c_dh_0^\delta\|\mathbf{D}^*\|_{2+\delta,\Omega}$. *Thus,* $\mathcal{S} = \cup_{n=0}^m \mathcal{S}_n \subset B_r$, *provided that* h_0 *is so small that* $c_\delta h_0^{1-\delta} \le r$.

Proof. The bound follows the triangle inequality, (4.6), the triangle inequality again, (4.4), the definition of $\mathcal{S}_n$ in (4.7), and (4.4) again:

$$\begin{aligned}
\|\mathbf{D}^* - \mathbf{U}_n\|_{1+\delta,\Omega} &\le \|\mathbf{D}^* - \mathcal{P}^{h_n}\mathbf{D}^*\|_{1+\delta,\Omega} + \|\mathcal{P}^{h_n}\mathbf{D}^* - \mathbf{U}_n\|_{1+\delta,\Omega} \\
&\le c_dh_n\|\mathbf{D}^*\|_{2+\delta,\Omega} + \frac{c_i}{{h_n}^\delta}\|\mathcal{P}^{h_n}\mathbf{D}^* - \mathbf{U}_n\|_{1,\Omega} \\
&\le c_dh_n\|\mathbf{D}^*\|_{2+\delta,\Omega} + \frac{c_i}{{h_n}^\delta}\left[\|\mathbf{D}^* - \mathcal{P}^{h_n}\mathbf{D}^*\|_{1,\Omega} + \|\mathbf{D}^* - \mathbf{U}_n\|_{1,\Omega}\right] \\
&\le c_dh_n\|\mathbf{D}^*\|_{2+\delta,\Omega} + \frac{c_i}{{h_n}^\delta}\left[c_d{h_n}^{1+\delta}\|\mathbf{D}^*\|_{2+\delta,\Omega} + 2c_dh_n\right] \\
&\le c_\delta {h_n}^{1-\delta}.
\end{aligned}$$

This lemma confirms max norm $O(h^{1-\delta})$ convergence. □

LEMMA 4.6. *Suppose that* $\mathbf{V}_n \in \mathcal{S}_n$ *for sufficiently small* r. *Let* $\widetilde{\mathbf{U}}_{n+1}$ *denote one exact discrete Newton step with initial guess* $\mathbf{V}_n$, *and let* $\mathbf{V}_{n+1}$ *denote its multigrid approximation. Then*

$$\|\widetilde{\mathbf{U}}_{n+1} - \mathbf{V}_{n+1}\|_{1,\Omega} \le \rho^{\nu_0}\|\widetilde{\mathbf{U}}_{n+1} - \mathbf{V}_n\|_{1,\Omega}. \tag{4.12}$$

Here, $\rho \in [0,1)$ *is a bound on the multigrid convergence factor for any level* n *and any initial guess* $\mathbf{V} \in \mathcal{S}_n$; *it depends only on* $\mathbf{D}^*, \mathbf{E}, r, \Omega, \delta, c_d$, *and* c_i.

Proof. Convergence estimate (4.12) follows from standard multigrid theory (cf. [10]) using the $H^1(\Omega)^q$ equivalence result in (4.1) with $\mathbf{D} = \mathbf{V}_n$. □

We have now established the tools that allow us to prove our central theorems.

4.2. Proofs of Theorems 4.1, 4.2, and 4.3.

Proof of Theorem 4.1. Consider a Taylor expansion for $\mathbf{P}(\mathbf{D}^*)$ about $\mathbf{D}_n$:

$$\mathbf{0} = \mathbf{P}(\mathbf{D}^*) = \mathbf{P}(\mathbf{D}_n) + \mathbf{P}'(\mathbf{D}_n)[\mathbf{D}^* - \mathbf{D}_n] + \frac{1}{2}\mathbf{P}''(\widetilde{\mathbf{D}})[\mathbf{D}^* - \mathbf{D}_n, \mathbf{D}^* - \mathbf{D}_n],$$

where $\widetilde{\mathbf{D}} = \theta\mathbf{D}_n + (1-\theta)\mathbf{D}^*$ for some $\theta \in [0,1]$ is bounded in the $H^{1+\delta}(\Omega)^q$ norm (Lemma 4.4). Then the lower bound in (4.1), combining the above expansion with (3.1), and using (4.2) proves (4.8):

$$\begin{aligned}
\|\mathbf{D}^* - \mathbf{D}_{n+1}\|_{1+\epsilon,\Omega} &\le c_c\|\mathbf{P}'(\mathbf{D}_n)[\mathbf{D}^* - \mathbf{D}_{n+1}]\|_{\epsilon,\Omega} \\
&= \frac{c_c}{2}\|\mathbf{P}''(\widetilde{\mathbf{D}})[\mathbf{D}^* - \mathbf{D}_n, \mathbf{D}^* - \mathbf{D}_n]\|_{\epsilon,\Omega} \\
&\le \frac{c_cc_2}{2}\|\mathbf{D}^* - \mathbf{D}_n\|_{1+\delta,\Omega}\|\mathbf{D}^* - \mathbf{D}_n\|_{1+\epsilon,\Omega}. \qquad (4.13)
\end{aligned}$$

To show that $\mathbf{D}_{n+1} \in B_r$, consider (4.13) with $\epsilon = \delta$. For $\mathbf{D}_n \in B_r$, this reduces to

$$\|\mathbf{D}^* - \mathbf{D}_{n+1}\|_{1+\delta,\Omega} \le \frac{c_cc_2}{2}r^2,$$

and we just require

$$r \le \frac{2}{c_c c_2}. \qquad \square \tag{4.14}$$

Proof of Theorem 4.2. First assume that r is so small that (4.14) is satisfied. Assume also that h_0 is so small that

$$c_\delta h_0^{1-\delta} = 2c_i c_d h_0^{1-\delta} + (1+c_i)c_d\|\mathbf{D}^*\|_{2+\delta,\Omega} h_0 \le r.$$

Hence, by Lemma 4.5 and because we assume that $\mathbf{U}_0 \in \mathcal{S}_0$, we must have $\mathbf{U}_0 \in B_r$. Suppose now that we could show that $\mathbf{U}_n \in \mathcal{S}_n$ implies that $\mathbf{U}_{n+1} \in \mathcal{S}_{n+1}$ for all $n \ge 0$. Then, since $\mathbf{U}_0 \in \mathcal{S}_0$, we would know that $\mathbf{U}_1 \in \mathcal{S}_1$, which in turn would imply that $\mathbf{U}_2 \in \mathcal{S}_2$. Continuing in this way would show that (4.9) holds for all $n \ge 0$.

To this end, assume that (4.9) holds for n replaced by $n-1$, with $\mathbf{U}_n \in \mathcal{S}_n$:

$$\|\mathbf{D}^* - \mathbf{U}_n\|_{1,\Omega} \le 2c_d h_n. \tag{4.15}$$

From Lemma 4.5, we see that

$$\|\mathbf{D}^* - \mathbf{U}_n\|_{1+\delta,\Omega} \le c_\delta h_n^{1-\delta}. \tag{4.16}$$

We bound the left-hand side of (4.9) by first using the triangle inequality:

$$\|\mathbf{D}^* - \mathbf{U}_{n+1}\|_{1,\Omega} \le \|\mathbf{D}^* - \widetilde{\mathbf{D}}_{n+1}\|_{1,\Omega} + \|\widetilde{\mathbf{D}}_{n+1} - \mathbf{U}_{n+1}\|_{1,\Omega}, \tag{4.17}$$

where $\widetilde{\mathbf{D}}_{n+1}$ is the result of infinite-dimensional Newton step (3.2) based on initial guess $\mathbf{U}_n$.

Consider the first term on the right-hand side of (4.17). By Theorem 4.1 with $\epsilon = 0$, (4.11), and (4.15), we have that

$$\begin{aligned}\|\mathbf{D}^* - \widetilde{\mathbf{D}}_{n+1}\|_{1,\Omega} &\le c_q\|\mathbf{D}^* - \mathbf{U}_n\|_{1,\Omega}\|\mathbf{D}^* - \mathbf{U}_n\|_{1+\delta,\Omega}\\ &\le 2c_d c_q c_\delta h_n^{2-\delta}.\end{aligned} \tag{4.18}$$

For the second term on the right-hand side of (4.17), we now show that

$$\|\widetilde{\mathbf{D}}_{n+1} - \mathbf{U}_{n+1}\|_{1,\Omega} \le c_b c_c(4c_d c_q c_\delta h_0^{1-\delta} + c_d h_0^{\delta}\|\mathbf{D}^*\|_{2+\delta,\Omega})h_n. \tag{4.19}$$

To this end, let $\mathbf{f} \equiv \mathbf{P}'(\mathbf{U}_n)[\mathbf{U}_n] - \mathbf{P}(\mathbf{U}_n)$. Then discrete Newton step (3.3) becomes the following: given $\mathbf{U}_n \in H^{h_n}$, find $\mathbf{U}_{n+1} \in H^{h_n}$ such that

$$\big(\mathbf{P}'(\mathbf{U}_n)[\mathbf{K}^{h_n}],\ \mathbf{P}'(\mathbf{U}_n)[\mathbf{U}_{n+1}]\big) = \big(\mathbf{P}'(\mathbf{U}_n)[\mathbf{K}^{h_n}],\ \mathbf{f}\big) \quad \forall\ \mathbf{K}^{h_n} \in H^{h_n}. \tag{4.20}$$

Note that $\widetilde{\mathbf{D}}_{n+1} \in \mathcal{H}_{1+\delta}$, which is generally not in H^{h_n}, is defined by

$$(\mathbf{P}'(\mathbf{U}_n)[\mathbf{K}],\ \mathbf{P}'(\mathbf{U}_n)[\widetilde{\mathbf{D}}_{n+1}]) = (\mathbf{P}'(\mathbf{U}_n)[\mathbf{K}]\ ,\ \mathbf{f}) \quad \forall\ \mathbf{K} \in \mathcal{H}_{1+\delta}. \tag{4.21}$$

Combining (4.20) and (4.21) for $\mathbf{K} = \mathbf{K}^{h_n} \in H^{h_n}$, we have that

$$(\mathbf{P}'(\mathbf{U}_n)[\mathbf{K}^{h_n}],\ \ \mathbf{P}'(\mathbf{U}_n)[\widetilde{\mathbf{D}}_{n+1} - \mathbf{U}_{n+1}]) = 0,$$

from which it follows that

$$\begin{aligned}&(\mathbf{P}'(\mathbf{U}_n)[\mathbf{K}],\ \mathbf{P}'(\mathbf{U}_n)[\widetilde{\mathbf{D}}_{n+1} - \mathbf{U}_{n+1}])\\ &\quad = (\mathbf{P}'(\mathbf{U}_n)[\mathbf{K} - \mathbf{K}^{h_n}],\ \mathbf{P}'(\mathbf{U}_n)[\widetilde{\mathbf{D}}_{n+1} - \mathbf{U}_{n+1}]).\end{aligned}$$

Letting $\mathbf{K} = \widetilde{\mathbf{D}}_{n+1} - \mathbf{U}_{n+1}$ and $\mathbf{K}^{h_n} = \mathcal{Q}^{h_n}(\widetilde{\mathbf{D}}_{n+1} - \mathbf{U}_{n+1}) = \mathcal{Q}^{h_n}\widetilde{\mathbf{D}}_{n+1} - \mathbf{U}_{n+1}$, so that $\mathbf{K} - \mathbf{K}^{h_n} = \widetilde{\mathbf{D}}_{n+1} - \mathcal{Q}^{h_n}\widetilde{\mathbf{D}}_{n+1}$, then yields

$$\begin{aligned} &(\mathbf{P}'(\mathbf{U}_n)[\widetilde{\mathbf{D}}_{n+1} - \mathbf{U}_{n+1}],\ \mathbf{P}'(\mathbf{U}_n)[\widetilde{\mathbf{D}}_{n+1} - \mathbf{U}_{n+1}]) \\ &\quad = (\mathbf{P}'(\mathbf{U}_n)[\widetilde{\mathbf{D}}_{n+1} - \mathcal{Q}^{h_n}\widetilde{\mathbf{D}}_{n+1}],\ \mathbf{P}'(\mathbf{U}_n)[\widetilde{\mathbf{D}}_{n+1} - \mathbf{U}_{n+1}]). \end{aligned} \tag{4.22}$$

Applying the Cauchy–Schwarz inequality to the right-hand side of (4.22), then cancelling the term $\|\mathbf{P}'(\mathbf{U}_n)[\widetilde{\mathbf{D}}_{n+1} - \mathbf{U}_{n+1}]\|_{0,\Omega}$ that results on both sides, yields

$$\|\mathbf{P}'(\mathbf{U}_n)[\widetilde{\mathbf{D}}_{n+1} - \mathbf{U}_{n+1}]\|_{0,\Omega} \le \|\mathbf{P}'(\mathbf{U}_n)[\widetilde{\mathbf{D}}_{n+1} - \mathcal{Q}^{h_n}\widetilde{\mathbf{D}}_{n+1}]\|_{0,\Omega}.$$

However, (4.1) confirms that $\mathbf{P}'(\mathbf{U}_n)[\mathbf{M}]$ is coercive and bounded in the $H^1(\Omega)^q$ norm for $\mathbf{M} \in \mathcal{H}_{1+\delta}$, and so the above bound becomes

$$\|\widetilde{\mathbf{D}}_{n+1} - \mathbf{U}_{n+1}\|_{1,\Omega} \le c_b c_c \|\widetilde{\mathbf{D}}_{n+1} - \mathcal{Q}^{h_n}\widetilde{\mathbf{D}}_{n+1}\|_{1,\Omega}. \tag{4.23}$$

We now bound the right-hand side of (4.23) using the triangle inequality, (4.3), (4.5), (4.18), and relation $h_n \le h_0$:

$$\begin{aligned} \|\widetilde{\mathbf{D}}_{n+1} - \mathcal{Q}^{h_n}\widetilde{\mathbf{D}}_{n+1}\|_{1,\Omega} &\le \|\widetilde{\mathbf{D}}_{n+1} - \mathbf{D}^*\|_{1,\Omega} + \|\mathbf{D}^* - \mathcal{Q}^{h_n}\mathbf{D}^*\|_{1,\Omega} \\ &\qquad + \|\mathcal{Q}^{h_n}(\mathbf{D}^* - \widetilde{\mathbf{D}}_{n+1})\|_{1,\Omega} \\ &\le 2\|\widetilde{\mathbf{D}}_{n+1} - \mathbf{D}^*\|_{1,\Omega} + c_d h_n^{1+\delta}\|\mathbf{D}^*\|_{2+\delta,\Omega} \\ &\le (4c_d c_q c_\delta h_0^{1-\delta} + c_d h_0^\delta \|\mathbf{D}^*\|_{2+\delta,\Omega}) h_n. \end{aligned} \tag{4.24}$$

Bound (4.19) now follows from bounds (4.23) and (4.24).

Combining (4.17), (4.18), (4.19), and relation $h_n \le h_0$ yields

$$\|\mathbf{D}^* - \mathbf{U}_{n+1}\|_{1,\Omega} \le (2(1 + 2c_b c_c)c_q c_\delta h_0^{1-\delta} + c_b c_c \|\mathbf{D}^*\|_{2+\delta,\Omega} h_0^\delta)(c_d h_n).$$

Theorem 4.2 now follows by choosing h_0 perhaps smaller still so that

$$2(1 + 2c_b c_c)c_q c_\delta h_0^{1-\delta} + c_b c_c \|\mathbf{D}^*\|_{2+\delta,\Omega} h_0^\delta \le 1. \qquad \square$$

Proof of Theorem 4.3. As in Theorem 4.2, we need h_0 sufficiently small, but even smaller yet to account for the fact that we do not solve the Newton steps exactly: we need h_0 so small that the error in approximating $\widetilde{\mathbf{U}}_{n+1}$ (the exact discrete Newton step) by $\mathbf{V}_{n+1}$ keeps these iterates in S_{n+1}.

From Theorem 4.2, we have that if $\mathbf{U}_n \in \mathcal{S}_n$, then $\mathbf{U}_{n+1} \in \mathcal{S}_{n+1}$. This result can be tightened by choosing a smaller value for h_0: choosing h_0 such that, say,

$$2(1 + 2c_b c_c)c_q c_\delta h_0^{1-\delta} + c_b c_c \|\mathbf{D}^*\|_{2+\delta,\Omega} h_0^\delta \le \frac{2}{3},$$

means that $\mathbf{U}_n \in \mathcal{S}_n$ implies that

$$\|\mathbf{D}^* - \mathbf{U}_{n+1}\|_{1,\Omega} \le \frac{4}{3}c_d h_{n+1} = \frac{2}{3}c_d h_n. \tag{4.25}$$

As for Theorem 4.2, to prove that (4.10) holds for n, we may assume that it holds for n replaced by $n-1$:

$$\|\mathbf{D}^* - \mathbf{V}_n\|_{1,\Omega} \le 2c_d h_n. \tag{4.26}$$

Letting $\widetilde{\mathbf{U}}_{n+1}$ as before denote one exact discrete Newton step with initial guess $\mathbf{V}_n$, then

$$\|\mathbf{D}^* - \mathbf{V}_{n+1}\|_{1,\Omega} \le \|\mathbf{D}^* - \widetilde{\mathbf{U}}_{n+1}\|_{1,\Omega} + \|\widetilde{\mathbf{U}}_{n+1} - \mathbf{V}_{n+1}\|_{1,\Omega}. \tag{4.27}$$

For sufficiently small h_0, we know that $\mathbf{V}_n$ is in $\mathcal{S}_n \subset B_r$, and, with the reduced value of h_0, the first term on the right-hand side is bounded according to (4.25):

$$\|\mathbf{D}^* - \widetilde{\mathbf{U}}_{n+1}\|_{1,\Omega} \le \frac{2}{3} c_d h_n. \tag{4.28}$$

For the second term, we use (4.12), the triangle inequality, (4.28), and (4.26):

$$\begin{aligned}
\|\mathbf{U}_{n+1} - \mathbf{V}_{n+1}\|_{1,\Omega} &\le \rho^{\nu_0} \|\widetilde{\mathbf{U}}_{n+1} - \mathbf{V}_n\|_{1,\Omega} \\
&\le \rho^{\nu_0} \left[\|\widetilde{\mathbf{U}}_{n+1} - \mathbf{D}^*\|_{1,\Omega} + \|\mathbf{D}^* - \mathbf{V}_n\|_{1,\Omega} \right] \\
&\le \rho^{\nu_0} \left[\frac{2}{3} c_d h_n + 2 c_d h_n \right] \\
&= \frac{8}{3} c_d \rho^{\nu_0} h_n \\
&\le \frac{1}{3} c_d h_n, \qquad (4.29)
\end{aligned}$$

where ν_0 is chosen so large that $\rho^{\nu_0} \le \frac{1}{8}$. Combining bounds (4.27), (4.28), and (4.29) then yields

$$\|\mathbf{D}^* - \mathbf{V}_{n+1}\|_{1,\Omega} \le c_d h_n,$$

which proves Theorem 4.3. □

5. Conclusion.

The general theory developed here applies to virtually any set of quasi-linear partial differential equations that can be reformulated as a first-order system, provided it is amenable to an H^1-elliptic least-squares principle. The approach uses nested iteration based on one Newton step per level, implemented using a fixed number of multigrid V-cycles. The theory shows that, for a sufficiently fine coarsest grid, the method produces a final approximation to the solution of the first-order system that is H^1 accurate to the level of discretization error. Use of this general theory is illustrated in the companion paper [17] by applying it to a first-order system for the elliptic grid generation equations. The companion paper also reports on numerical experiments that support the theory.

REFERENCES

[1] R. A. Adams, *Sobolev Spaces*, Pure Appl. Math. 65, Academic Press, New York, 1975.

[2] E. L. Allgower and K. Böhmer, *Application of the mesh independence principle to mesh refinement strategies*, SIAM J. Numer. Anal., 24 (1987), pp. 1335–1351.

[3] E. L. Allgower, K. Böhmer, F. A. Potra, and W. C. Rheinboldt, *A mesh-independence principle for operator equations and their discretizations*, SIAM J. Numer. Anal., 23 (1986), pp. 160–169.

[4] E. L. Allgower, S. McCormick, and D. Pryor, *A general mesh independence principle for Newton's method applied to second order boundary value problems*, Computing, 23 (1979), pp. 223–246.

[5] R. E. Bank and D. J. Rose, *Global approximate Newton methods*, Numer. Math., 37 (1981), pp. 279–295.

[6] R. E. Bank and D. J. Rose, *Analysis of a multilevel iterative method for nonlinear finite element equations*, Math. Comp., 39 (1982), pp. 435-465.

[7] P. Bochev, Z. Cai, T. A. Manteuffel, and S. F. McCormick, *Analysis of velocity-flux first-order system least-squares principles for the Navier–Stokes equations: Part* I, SIAM J. Numer. Anal., 35 (1998), pp. 990–1009.

[8] P. Bochev, T. A. Manteuffel, and S. F. McCormick, *Analysis of velocity-flux least-squares principles for the Navier–Stokes equations: Part* II, SIAM J. Numer. Anal., 36 (1999), pp. 1125–1144.

[9] D. Braess, *Finite Elements Theory, Fast Solvers, and Applications in Solid Mechanics*, Cambridge University Press, Cambridge, UK, 1997.

[10] J. H. Bramble, *Multigrid Methods*, Pitman Res. Notes in Math. 294, Longman Scientific and Technical, Harlow, Essex, UK, 1993.

[11] S. C. Brenner and L. R. Scott, *The Mathematical Theory of Finite Element Methods*, Texts Appl. Math. 15, Springer-Verlag, New York, 1994.

[12] W. L. Briggs, V. E. Henson, and S. F. McCormick, *A Multigrid Tutorial*, 2nd ed., SIAM, Philadelphia, 2000.

[13] Z. Cai, R. Lazarov, T. A. Manteuffel, and S. F. McCormick, *First-order system least squares for second-order partial differential equations: Part* I, SIAM J. Numer. Anal., 31 (1994), pp. 1785–1799.

[14] Z. Cai, T. A. Manteuffel, and S F McCormick, *First-order system least squares for the Stokes equations, with application to linear elasticity*, Electron. Trans. Numer. Anal., 3 (1995), pp. 150–159.

[15] Z. Cai, T. A. Manteuffel, and S. F. McCormick, *First-order system least squares for second-order partial differential equations: Part* II, SIAM J. Numer. Anal., 34 (1997), pp. 425–454.

[16] A. L. Codd, *Elasticity-Fluid Coupled Systems and Elliptic Grid Generation (EGG) Based on First-Order System Least Squares (FOSLS)*, Ph.D. thesis, Department of Applied Mathematics, University of Colorado at Boulder, Boulder, CO, 2001.

[17] A. L. Codd, T. Manteuffel, S. McCormick, and W. Ruge, *Multilevel first-order system least squares for elliptic grid generation*, SIAM J. Numer. Anal., 41 (2003), pp. 2210–2232.

[18] W. Layton, *A two-level discretization method for the Navier–Stokes equations*, Comput. Math. Appl., 26 (1993), pp. 33–38.

[19] W. Layton and H. W. J. Lenferink, *A multilevel mesh independence principle for the Navier–Stokes equations*, SIAM J. Numer. Anal., 33 (1996), pp. 17–30.

[20] S. McCormick, *A mesh refinement method for $Ax = \lambda Bx$*, Math. Comp., 36 (1981), pp. 485–498.

[21] S. F. McCormick, *A revised mesh refinement strategy for Newton's method applied to nonlinear two-point boundary value problems*, in Numerical Treatment of Differential Equations in Applications, (Proceedings of the Meeting at the Mathematical Research Center, Oberwolfach, 1977), Springer, Berlin, 1978, pp. 15–23.

[22] G. Timmermann, *A cascadic multigrid algorithm for semilinear elliptic problems*, Numer. Math., 86 (2000), pp. 717–731.

SIAM J. Numer. Anal.
Vol. 41, No. 6, pp. 2210–2232

MULTILEVEL FIRST-ORDER SYSTEM LEAST SQUARES FOR ELLIPTIC GRID GENERATION*

A. L. CODD[†], T. A. MANTEUFFEL[‡], S. F. MCCORMICK[‡], and J. W. RUGE[‡]

Abstract. A new fully variational approach is studied for elliptic grid generation (EGG). It is based on a general algorithm developed in a companion paper [A. L. Codd, T. A. Manteuffel, and S. F. McCormick, *SIAM J. Numer. Anal.*, 41 (2003), pp. 2197–2209] that involves using Newton's method to linearize an appropriate equivalent first-order system, first-order system least squares (FOSLS) to formulate and discretize the Newton step, and algebraic multigrid (AMG) to solve the resulting matrix equation. The approach is coupled with nested iteration to provide an accurate initial guess for finer levels using coarse-level computation. The present paper verifies the assumptions of the companion work and confirms the overall efficiency of the scheme with numerical experiments.

Key words. least-squares discretization, multigrid, nonlinear elliptic boundary value problems

AMS subject classifications. 35J65, 65N15, 65N30, 65N50, 65F10

DOI. 10.1137/S0036142902404418

1. Introduction.

A companion paper [10] develops an algorithm using Newton's method, first-order system least squares (FOSLS), and algebraic multigrid (AMG) for efficient solution of general nonlinear elliptic equations. The equations are first converted to an appropriate first-order system, and an approximate solution to the coarsest-grid problem is then computed (by any suitable method such as Newton iteration coupled perhaps with direct solvers, damping, or continuation). The approximation is then interpolated to the next finer level, where it is used as an initial guess for one Newton linearization of the nonlinear problem, with a few AMG cycles applied to the resulting matrix equation. This algorithm repeats itself until the finest grid is processed, again by *one* Newton/AMG step. At each Newton step, FOSLS is applied to the linearized system, and the resulting matrix equation is solved using just a few V-cycles of AMG.

In the present paper, we apply this algorithm to elliptic grid generation (EGG) equations. Grid generation is usually based on a map between a relatively simple *computational* region and a possibly complicated *physical* region. It can be used numerically to create a mesh for a discretization method to solve a given system of equations posed on the physical domain. Alternatively, it can be used to transform equations posed on the physical region into ones posed on the computational region, where the transformed equations are then solved. If the Jacobian of the transformation is positive throughout the computational region, the equation type is unchanged [12]. Actually, the relative minimum value of the Jacobian is important in practice because relatively small values signal small angles between the grid lines and large errors in approximating the equations [20].

*Received by the editors March 21, 2002; accepted for publication (in revised form) April 16, 2003; published electronically December 5, 2003. This work was sponsored by the National Institute of Health under grant 1–R01–EY12291–01, the National Science Foundation under grant DMS–9706866, and the Department of Energy under grant DE–FG03–93ER25165.

http://www.siam.org/journals/sinum/41-6/40441.html

[†] Centre for Mathematics and Its Applications, School of Mathematical Sciences, Australian National University, Canberra, ACT 0200, Australia (andrea.codd@anu.edu.au).

[‡] Department of Applied Mathematics, Campus Box 526, University of Colorado at Boulder, Boulder, CO 80309–0526 (tmanteuf@boulder.colorado.edu, stevem@boulder.colorado.edu, jruge@boulder.colorado.edu).

Our interest is in EGG using the Winslow generator [12], which allows us to specify the boundary maps completely. Moreover, by choosing the two-dimensional computational region to be convex, we can ensure that the Jacobian of the map is positive, which in turn ensures that the map is one-to-one and onto and therefore does not fold [8]. The Winslow generator tends to create smooth grids, with good aspect ratios. The map also tends to control variations in gridline spacing and nonorthogonality of the gridline intersections in the physical space. See Thompson, Warsi, and Mastin [20] and Knupp and Steinberg [12] for background on grid generation in general and EGG in particular. Several discretization methods for the EGG equations together with their associated errors are discussed in [20]. In [12], the EGG equations are derived, and several existing methods are described for solving them.

A brief description of the first-order EGG system is given in section 2. The assumptions needed to apply the theory in [10] are verified in section 3. Section 4 discusses scaling of the functional terms used for the computations as well as numerical results for two representative problems. The last section includes some final remarks.

2. Equations. We use standard notation for the associated spaces. Restricting ourselves to two dimensions, we consider a generic open domain $\Omega \in R^2$, with Lipschitz boundary $\Gamma \in C^{3,1}$. (The superscript 1 indicates Lipschitz continuity of the functions and their derivatives.) Suppose that $m \geq 0$ and $n \geq 1$ are given integers. Let $(\cdot,\cdot)_{0,\Omega}$ denote the inner product on $L^2(\Omega)^n$, $\|\cdot\|_{0,\Omega}$ its induced norm, and $H^m(\Omega)^n$ the standard Sobolev space with norm $\|\cdot\|_{m,\Omega}$ and seminorms $|\cdot|_{i,\Omega}$ $(0 \leq i \leq m)$. (The superscript n is omitted when dependence is clear by context.) For $\delta \in (0,1)$, let $H^{m+\delta}(\Omega)$ (cf. [6]) denote the Sobolev space associated with the norm defined by

$$\|u\|^2_{m+\delta,\Omega} \equiv \|u\|^2_{m,\Omega} + \sum_{|\alpha|=m} \int_\Omega \int_\Omega \frac{|\partial_\alpha u(x) - \partial_\alpha u(y)|^2}{|x-y|^{2(1+\delta)}} dxdy.$$

(This definition allows the use of the "real interpolation" method [1, 6].) Also, let $H^{\frac{1}{2}}(\Gamma)$ denote the trace Sobolev space associated with the norm

$$\|u\|_{\frac{1}{2},\Gamma} \equiv \inf\{\|v\|_{1,\Omega} : v \in H^1(\Omega),\ \text{trace } v = u \text{ on } \Gamma\}.$$

We start by mapping a known convex computational region, $\Omega \in R^2$ with boundary $\Gamma \in C^{3,1}$, to a given physical region, $\Omega_{\mathbf{x}} \in R^2$ with boundary $\Gamma_{\mathbf{x}} \in C^{3,1}$. We define map $\boldsymbol{\xi} : \bar{\Omega}_{\mathbf{x}} \to \bar{\Omega}$ and its inverse $\mathbf{x} : \bar{\Omega} \to \bar{\Omega}_{\mathbf{x}}$. The coordinates in $\Omega_{\mathbf{x}}$ are denoted by the vector of unknowns $\mathbf{x} = (x \quad y)^t$, and those in Ω by $\boldsymbol{\xi} = (\xi \quad \eta)^t$.

For the EGG smoothness or Winslow generator, we choose $\boldsymbol{\xi}$ to be harmonic:

$$\begin{aligned} \Delta\boldsymbol{\xi} &= \mathbf{0} && \text{in} \quad \Omega_{\mathbf{x}}, \\ \boldsymbol{\xi} &= \mathbf{v}(\mathbf{x}) && \text{on} \quad \Gamma_{\mathbf{x}}, \end{aligned} \tag{2.1}$$

where $\mathbf{v} \in H^{\frac{7}{2}}(\Gamma_{\mathbf{x}})$ is a given homeomorphism (continuous and one-to-one) from the boundary of the physical region onto the boundary of the computational region. ($H^{\frac{7}{2}}(\Gamma_{\mathbf{x}})$ is consistent with our boundary smoothness assumption, $\Gamma \in C^{3,1}$.) With $\Omega_{\mathbf{x}}$ bounded, the weak form of Laplace system (2.1) has one and only one solution $\boldsymbol{\xi}^*$ in $H^4(\Omega_{\mathbf{x}})^2$ (see [11]) and, by Weyl's lemma [22], $\boldsymbol{\xi}^* \in C^\infty_{loc}(\Omega_{\mathbf{x}}) \equiv \{\boldsymbol{\xi} \in C^\infty(K)\ \forall\, K \subset \Omega_{\mathbf{x}}\}$.

Map $\boldsymbol{\xi}^*$ is posed on $\Omega_{\mathbf{x}}$, and thus computing an approximation to it would nominally involve specifying a grid on the physical region. But specifying such a grid is the aim of EGG in the first place, and so this formulation is not useful. We therefore choose instead to solve the inverse of problem (2.1), which takes a regular grid in Ω and maps it onto a grid in $\Omega_{\mathbf{x}}$, thus achieving our objective. To this end, we assume $\Omega_{\mathbf{x}}$ and Ω to be simply connected and bounded, and $\bar{\Omega}$ to be convex, so that Γ and $\Gamma_{\mathbf{x}}$ are simple closed curves. Map $\boldsymbol{\xi}^*$ is continuous and harmonic, and $\mathbf{v}$ is a homeomorphism of $\Gamma_{\mathbf{x}}$ onto Γ, so Rado's theorem (cf. [16]) implies the existence of a unique inverse map $\mathbf{x}^*$ from Ω onto $\Omega_{\mathbf{x}}$. An outline of the proof is provided in [14]. It then follows that domain map $\boldsymbol{\xi}^*$ is a diffeomorphism [8, 12], and the associated Jacobian $J^*_{\mathbf{x}} \equiv \xi^*_x \nu^*_y - \xi^*_y \nu^*_x$ is continuous and uniformly positive and bounded on $\Omega_{\mathbf{x}}$. ($J_0 \leq |J^*_{\mathbf{x}}(x,y)| \leq J_1$ for some constants $J_0, J_1 \in R^+$ and all $(x,y) \in \Omega_{\mathbf{x}}$.) The choice of the space for $\mathbf{x}^*$ follows from the assumptions for $\boldsymbol{\xi}^*$, $\Gamma_{\mathbf{x}}$, and $\mathbf{v}$ and is discussed further in section 3.

The inverse map satisfies the following equations (positive Jacobian throughout $\Omega_{\mathbf{x}}$ ensures that the solution of (2.1) is an invertible map):

$$\begin{aligned}
(x_\eta^2 + y_\eta^2)x_{\xi\xi} - (x_\xi x_\eta + y_\xi y_\eta)(x_{\xi\eta} + x_{\eta\xi}) + (x_\xi^2 + y_\xi^2)x_{\eta\eta} &= 0 \quad \text{in} \quad \Omega,\\
(x_\eta^2 + y_\eta^2)y_{\xi\xi} - (x_\xi x_\eta + y_\xi y_\eta)(y_{\xi\eta} + y_{\eta\xi}) + (x_\xi^2 + y_\xi^2)y_{\eta\eta} &= 0 \quad \text{in} \quad \Omega,\\
x &= w_1(\xi,\eta) \quad \text{on} \quad \Gamma,\\
y &= w_2(\xi,\eta) \quad \text{on} \quad \Gamma,
\end{aligned} \tag{2.2}$$

where function $\mathbf{w} = \binom{w_1(\xi,\eta)}{w_2(\xi,\eta)}$ is the inverse of function $\mathbf{v} = \binom{v_1(x,y)}{v_2(x,y)}$ (i.e., $\mathbf{x} = \mathbf{w}(\mathbf{v}(\mathbf{x}))$). See [12] for more detail. The inverse map $\mathbf{x}^*$ exists and solves (2.2). We assume that the Fréchet derivative of the operator in (2.2) at $\mathbf{x}^*$ is one-to-one on $H_0^{2+\delta}(\Omega)^2$ (subscript 0 denoting homogeneous Dirichlet conditions on Γ). This is easily verified when $\mathbf{x}^*$ deviates from a constant map by a sufficiently small amount.

To apply our method, we begin by converting (2.2) to a first-order system. We could write these equations in a simple way using the standard notation of a 2×2 matrix for the Jacobian matrix, but this is not convenient for the linearized equations treated in section 3. Our notation is therefore based primarily on writing the Jacobian matrix as a 4×1 vector:

$$\mathbf{J} = \begin{pmatrix} x_\xi \\ x_\eta \\ y_\xi \\ y_\eta \end{pmatrix} = \begin{pmatrix} J_{11} \\ J_{21} \\ J_{12} \\ J_{22} \end{pmatrix}.$$

On the other hand, at times it is useful to refer to the matrix form of the unknowns. We therefore define the block-structured matrix $\underline{\mathbf{J}}$ and its classical adjoint $\hat{\underline{\mathbf{J}}}$ as follows:

$$\underline{\mathbf{J}} = \begin{pmatrix} J_{11} & J_{21} & 0 & 0 \\ J_{12} & J_{22} & 0 & 0 \\ 0 & 0 & J_{11} & J_{21} \\ 0 & 0 & J_{12} & J_{22} \end{pmatrix} \quad \text{and} \quad \hat{\underline{\mathbf{J}}} = \begin{pmatrix} J_{22} & -J_{21} & 0 & 0 \\ -J_{12} & J_{11} & 0 & 0 \\ 0 & 0 & J_{22} & -J_{21} \\ 0 & 0 & -J_{12} & J_{11} \end{pmatrix}.$$

Note that the Jacobian of the inverse transformation is given by

$$J \equiv x_\xi y_\eta - x_\eta y_\xi = J_{11}J_{22} - J_{21}J_{12} = \sqrt{\det \underline{\mathbf{J}}}.$$

Also, $J = \frac{1}{J_{\mathbf{x}}}$ and $\|J\|_{\infty,\Omega} = \|\frac{1}{J_{\mathbf{x}}}\|_{\infty,\Omega_{\mathbf{x}}} = \frac{1}{\|J_{\mathbf{x}}\|_{\infty,\Omega_{\mathbf{x}}}} > 0$.

In keeping with the vector notation, denote grad, div, and curl, respectively, by

$$\nabla = \begin{pmatrix} \partial_\xi & 0 \\ \partial_\eta & 0 \\ 0 & \partial_\xi \\ 0 & \partial_\eta \end{pmatrix}, \quad \nabla\cdot = \begin{pmatrix} \partial_\xi & \partial_\eta & 0 & 0 \\ 0 & 0 & \partial_\xi & \partial_\eta \end{pmatrix}, \quad \nabla\times = \begin{pmatrix} -\partial_\eta & \partial_\xi & 0 & 0 \\ 0 & 0 & -\partial_\eta & \partial_\xi \end{pmatrix}.$$

The same calculus notation is used in both $\Omega_{\mathbf{x}}$ and Ω (e.g., ∇, $\nabla\cdot$, and $\nabla\times$). Differentiation in $\Omega_{\mathbf{x}}$ is with respect to x and y, and differentiation in Ω is with respect to ξ and η. Let the boundary unit normal vector be denoted by

$$\mathbf{n} = \begin{pmatrix} n_1 & 0 \\ n_2 & 0 \\ 0 & n_1 \\ 0 & n_2 \end{pmatrix}. \tag{2.3}$$

As in previous applications of the FOSLS methodology (cf. [7]), the natural first-order system is often augmented with a curl equation to ensure that the system is elliptic in the H^1 product norm. The augmented system also allows for the possibility of solving for the unknowns in two separate stages: we can solve for $\mathbf{J}$ alone in the first stage, then fix $\mathbf{J}$ and solve for $\mathbf{x}$ alone in the second stage, as the following development shows. The curl-augmented system we consider here is

$$\begin{aligned} \mathbf{J} - \nabla\mathbf{x} &= \mathbf{0} && \text{in} \quad \Omega, \\ (\underline{\hat{\mathbf{J}}}\underline{\hat{\mathbf{J}}}^t\nabla)\cdot\mathbf{J} &= \mathbf{0} && \text{in} \quad \Omega, \\ \nabla\times\mathbf{J} &= \mathbf{0} && \text{in} \quad \Omega, \\ \mathbf{x} &= \mathbf{w} && \text{on} \quad \Gamma, \\ \mathbf{n}\times\mathbf{J} &= \mathbf{n}\times\nabla\mathbf{w} && \text{on} \quad \Gamma. \end{aligned} \tag{2.4}$$

To be very clear about our notation, note that derivatives apply only to terms on their right. Thus, for $(\underline{\hat{\mathbf{J}}}\underline{\hat{\mathbf{J}}}^t\nabla)\cdot$ in the second equation of (2.4), the matrix multiplication is applied first, keeping the order of each entry in the resulting matrix consistent with the multiplication. To perform the dot product, the matrix is transposed without altering the order of the terms in each component. For example, if we write

$$\underline{\hat{\mathbf{J}}}\underline{\hat{\mathbf{J}}}^t = \begin{pmatrix} \alpha & -\beta & 0 & 0 \\ -\beta & \gamma & 0 & 0 \\ 0 & 0 & \alpha & -\beta \\ 0 & 0 & -\beta & \gamma \end{pmatrix},$$

then

$$(\underline{\hat{\mathbf{J}}}\underline{\hat{\mathbf{J}}}^t\nabla)\cdot\mathbf{J} = \begin{pmatrix} \alpha\frac{\partial J_{11}}{\partial\xi} - \beta\frac{\partial J_{11}}{\partial\eta} - \beta\frac{\partial J_{21}}{\partial\xi} + \gamma\frac{\partial J_{21}}{\partial\eta} \\ \alpha\frac{\partial J_{12}}{\partial\xi} - \beta\frac{\partial J_{12}}{\partial\eta} - \beta\frac{\partial J_{22}}{\partial\xi} + \gamma\frac{\partial J_{22}}{\partial\eta} \end{pmatrix}.$$

We consider a two-stage algorithm, but focus only on the following first stage:

$$\begin{aligned} (\underline{\hat{\mathbf{J}}}\underline{\hat{\mathbf{J}}}^t\nabla)\cdot\mathbf{J} &= \mathbf{0} && \text{in} \quad \Omega, \\ \nabla\times\mathbf{J} &= \mathbf{0} && \text{in} \quad \Omega, \\ \mathbf{n}\times\mathbf{J} &= \mathbf{n}\times\nabla\mathbf{w} && \text{on} \quad \Gamma. \end{aligned} \tag{2.5}$$

Note that $\mathbf{x}$ can be recovered from the solution of (2.5) by a second stage that minimizes $\|\nabla\mathbf{x}-\mathbf{J}\|^2_{0,\Omega}+\|\mathbf{x}-\mathbf{w}\|^2_{\frac{1}{2},\Gamma}$ over $\mathbf{x}$ with the computed $\mathbf{J}$ held fixed. The homogeneous part of the first term in this functional is precisely the $H^1(\Omega)^4$ seminorm of $\mathbf{x}$, so minimizing this functional leads to a simple system of decoupled Poisson equations. The remainder of our analysis therefore focuses on (2.5).

To obtain homogeneous boundary conditions, we rewrite the equations in terms of the perturbation $\mathbf{D}$ of a smooth extension of $\mathbf{w}$ into Ω. To this end, suppose that some function $\mathbf{w} \in H^4(\Omega)^2$ is given so that its trace agrees with $\mathbf{w}$ on Γ. Defining $\mathbf{E} \equiv \nabla\mathbf{w} \in H^3(\Omega)^4$, we thus have

$$\mathbf{n}\times\mathbf{E} = \mathbf{n}\times\nabla\mathbf{w} \qquad \text{on } \Gamma. \tag{2.6}$$

(In practice, we do not really need an extension of $\mathbf{w}$, but rather just an extension of its gradient: any $\mathbf{E} \in H^3(\Omega)^4$ that satisfies (2.6) will do. However, if this extension is not necessarily a gradient, then $\mathbf{E}$ must be included in the curl term in (2.7) below.)

In the notation of the companion paper [10], we have

$$\mathbf{P}(\mathbf{D}) \equiv \begin{pmatrix} ((\underline{\hat{\mathbf{E}}}+\underline{\hat{\mathbf{D}}})(\underline{\hat{\mathbf{E}}}+\underline{\hat{\mathbf{D}}})^t\nabla)\cdot(\mathbf{E}+\mathbf{D}) \\ \nabla\times\mathbf{D} \end{pmatrix} = \mathbf{0}, \tag{2.7}$$

with boundary conditions

$$\mathbf{n}\times\mathbf{D} = \mathbf{0}. \tag{2.8}$$

System (2.7)–(2.8) corresponds to the inverse Laplace problem with Dirichlet boundary conditions. Existence of a solution $\mathbf{D}^*$ that yields a positive Jacobian is guaranteed by Rado's theorem. We show in section 3 that $\mathbf{D}^* \in H^3(\Omega)^4$. One implication of this smoothness property is that $(\underline{\hat{\mathbf{E}}}+\underline{\hat{\mathbf{D}}}^*)(\underline{\hat{\mathbf{E}}}+\underline{\hat{\mathbf{D}}}^*)^t$ is a uniformly positive definite and bounded matrix on Ω.

From the companion paper [10], we define

$$\mathcal{H}_{1+\delta} \equiv \{\mathbf{D} \in H^{1+\delta}(\Omega)^4 : \mathbf{n}\times\mathbf{D} = \mathbf{0} \text{ on } \Gamma\}.$$

Restricting $\mathbf{D}$ to $H^{1+\delta}(\Omega)^4$ ensures that $((\underline{\hat{\mathbf{E}}}+\underline{\hat{\mathbf{D}}})(\underline{\hat{\mathbf{E}}}+\underline{\hat{\mathbf{D}}})^t\nabla)\cdot(\mathbf{E}+\mathbf{D}) \in L^2(\Omega)^2$, as the results of the next section show.

The first Fréchet derivative of (2.7) in direction $\mathbf{K}$ is

$$\mathbf{P}'(\mathbf{D})[\mathbf{K}] = \begin{pmatrix} ((\underline{\hat{\mathbf{D}}}+\underline{\hat{\mathbf{E}}})(\underline{\hat{\mathbf{D}}}+\underline{\hat{\mathbf{E}}})^t\nabla)\cdot\mathbf{K}+\mathbf{B}\cdot\mathbf{K} \\ \nabla\times\mathbf{K} \end{pmatrix}, \tag{2.9}$$

where

$$\mathbf{B}\cdot\mathbf{K} \equiv (\underline{\hat{\mathbf{K}}}(\underline{\hat{\mathbf{D}}}+\underline{\hat{\mathbf{E}}})^t\nabla)\cdot(\mathbf{D}+\mathbf{E}) + ((\underline{\hat{\mathbf{D}}}+\underline{\hat{\mathbf{E}}})\underline{\hat{\mathbf{K}}}^t\nabla)\cdot(\mathbf{D}+\mathbf{E}),$$

and the second Fréchet derivative in directions $\mathbf{K}$ and $\mathbf{M}$ is

$$\begin{aligned} \mathbf{P}''(\mathbf{D})[\mathbf{K},\mathbf{M}] = & \begin{pmatrix} (\underline{\hat{\mathbf{M}}}(\underline{\hat{\mathbf{D}}}+\underline{\hat{\mathbf{E}}})^t\nabla)\cdot\mathbf{K} + ((\underline{\hat{\mathbf{D}}}+\underline{\hat{\mathbf{E}}})\underline{\hat{\mathbf{M}}}^t\nabla)\cdot\mathbf{K} \\ \mathbf{0} \end{pmatrix} \\ & + \begin{pmatrix} (\underline{\hat{\mathbf{K}}}(\underline{\hat{\mathbf{D}}}+\underline{\hat{\mathbf{E}}})^t\nabla)\cdot\mathbf{M} + ((\underline{\hat{\mathbf{D}}}+\underline{\hat{\mathbf{E}}})\underline{\hat{\mathbf{K}}}^t\nabla)\cdot\mathbf{M} \\ \mathbf{0} \end{pmatrix} \\ & + \begin{pmatrix} (\underline{\hat{\mathbf{K}}}\underline{\hat{\mathbf{M}}}^t\nabla)\cdot(\mathbf{D}+\mathbf{E}) + (\underline{\hat{\mathbf{M}}}\underline{\hat{\mathbf{K}}}^t\nabla)\cdot(\mathbf{D}+\mathbf{E}) \\ \mathbf{0} \end{pmatrix}. \end{aligned} \tag{2.10}$$

3. The assumptions and their verification. Consider the assumptions made in our companion paper [10]. The first is existence of a solution in $\mathcal{H}_{2+\delta}$. From [16], we know that a unique inverse map exists and that it provides a solution $\mathbf{D}^*$ to (2.7). Recall that $\boldsymbol{\xi}^* \in H^4(\Omega_{\mathbf{x}})^2$. In Lemma 3.4 below, we show that $\mathbf{D}^* \in \mathcal{H}_3$. This establishes our first assumption for the EGG equations for any $\delta \in (0,1)$.

The remaining assumptions we need to establish are, for $\epsilon = 0$ or δ, that $\mathbf{P}[\mathbf{D}] \in H^\epsilon(\Omega)^4$ for every $\mathbf{D} \in B_r$ (Lemma 3.5), that $\|\mathbf{P}'(\mathbf{D})[\,\cdot\,]\|_{\epsilon,\Omega}$ is $H^{1+\epsilon}(\Omega)^4$ equivalent (Lemma 3.6), and that the second Fréchet derivative of $\mathbf{P}(\mathbf{D})$ is bounded for all $\mathbf{D} \in B_r$ (Lemma 3.7). (The discretization assumptions are standard.) But first we need three results that follow directly from a corollary to the Sobolev imbedding theorem [11], which (tailored to our needs) states that the product of a function in $H^{m_1}(\Omega)$ and a function in $H^{m_2}(\Omega)$ is in $H^m(\Omega)$, provided that either $m_1+m_2-m \geq 1$, $m_1 > m$, and $m_2 > m$ or $m_1+m_2-m > 1$, $m_1 \geq m$, and $m_2 \geq m$.

Assume that $r > 0$ is so small that matrix $(\hat{\underline{\mathbf{E}}} + \hat{\underline{\mathbf{D}}})(\hat{\underline{\mathbf{E}}} + \hat{\underline{\mathbf{D}}})^t$ is positive definite and bounded uniformly on Ω and over $\mathbf{D} \in B_r \equiv \{\mathbf{D} \in \mathcal{H}_{1+\delta} : \|\mathbf{D}^* - \mathbf{D}\|_{1+\delta,\Omega} < r\}$. This assumption is possible because it is true at $\mathbf{D} = \mathbf{D}^*$ and because the matrix is continuous as a function defined on B_r. Assume that $a, b, c \in H^{1+\delta}(\Omega)$. For convenience, we let ∂ denote either ∂_x or ∂_y. In the proof of Lemma 3.4, we also use ∂^2 to denote any of the four second partial derivatives, and ∂^3 for any combination of third partial derivatives. Note that $(\partial a)^2$ could mean $a_x a_y$, for example.

Lemma 3.1. *There exists a constant C, depending only on Ω and δ, such that*

$$\|ab\partial c\|_{\epsilon,\Omega} \leq C\|a\|_{1+\delta,\Omega}\|b\|_{1+\delta,\Omega}\|c\|_{1+\epsilon,\Omega}.$$

Proof. Using the corollary to the Sobolev imbedding theorem [11] with $m_1 = 1+\delta,\ m_2 = \epsilon$, and $m = \epsilon$ twice yields

$$\begin{aligned}\|ab\partial c\|_{\epsilon,\Omega} &\leq C\|a\|_{1+\delta,\Omega}\|b\partial c\|_{\epsilon,\Omega} \\ &\leq C\|a\|_{1+\delta,\Omega}\|b\|_{1+\delta,\Omega}\|\partial c\|_{\epsilon,\Omega} \\ &\leq C\|a\|_{1+\delta,\Omega}\|b\|_{1+\delta,\Omega}\|c\|_{1+\epsilon,\Omega}. \qquad \square\end{aligned}$$

Lemma 3.2. *There exists a constant C, depending only on Ω and δ, such that*

$$\|ab\partial c\|_{\epsilon,\Omega} \leq C\|a\|_{1+\delta,\Omega}\|b\|_{1+\epsilon,\Omega}\|c\|_{1+\delta,\Omega}.$$

Proof. Using the corollary to the Sobolev imbedding theorem [11] first with $m_1 = 1+\delta,\ m_2 = \epsilon$, and $m = \epsilon$, then with $m_1 = 1+\epsilon,\ m_2 = \delta$, and $m = \epsilon$ yields

$$\begin{aligned}\|ab\partial c\|_{\epsilon,\Omega} &\leq C\|a\|_{1+\delta,\Omega}\|b\partial c\|_{\epsilon,\Omega} \\ &\leq C\|a\|_{1+\delta,\Omega}\|b\|_{1+\epsilon,\Omega}\|\partial c\|_{\delta,\Omega} \\ &\leq C\|a\|_{1+\delta,\Omega}\|b\|_{1+\epsilon,\Omega}\|c\|_{1+\delta,\Omega}. \qquad \square\end{aligned}$$

Lemma 3.3. *Assume that $a, b \in H^{2+\delta}(\Omega)$ and $k \in H^{1+\delta}(\Omega)$. Then there exists a constant C, depending only on Ω and δ, such that*

$$\|ak\partial b\|_{1,\Omega} \leq C\|a\|_{1+\delta,\Omega}\|b\|_{2+\delta,\Omega}\|k\|_{1,\Omega}.$$

Proof.

$$\begin{aligned}\|ak\partial b\|_{1,\Omega} &\le C\|a\|_{1+\delta,\Omega}\|k\partial b\|_{1,\Omega}\\ &\le C\|a\|_{1+\delta,\Omega}\|\partial b\|_{1+\delta,\Omega}\|k\|_{1,\Omega}\\ &\le C\|a\|_{1+\delta,\Omega}\|b\|_{2+\delta,\Omega}\|k\|_{1,\Omega},\end{aligned}$$

where we have used the corollary to the Sobolev imbedding theorem from [11] with $m_1 = 1+\delta,\ m_2 = 1$, and $m = 1$ twice. □

LEMMA 3.4. *The solution* $\mathbf{D}^*$ *of* (2.7) *is in* $H^3(\Omega)^4$.

Proof. We have

$$\mathbf{J}^* \equiv \mathbf{E} + \mathbf{D}^* = \begin{pmatrix} J^*_{11}\\ J^*_{21}\\ J^*_{12}\\ J^*_{22}\end{pmatrix} = \begin{pmatrix} x^*_\xi\\ x^*_\eta\\ y^*_\xi\\ y^*_\eta\end{pmatrix} = \frac{1}{J^*_{\mathbf{x}}}\begin{pmatrix} \eta^*_y\\ -\xi^*_y\\ -\eta^*_x\\ \xi^*_x\end{pmatrix},$$

where $\mathbf{E} \in H^3(\Omega_{\mathbf{x}})^4$ and $\boldsymbol{\xi}^* \in H^4(\Omega_{\mathbf{x}})^2$. We now show that $\mathbf{J}^* \in H^3(\Omega)^4$, from which follows the result that $\mathbf{D}^* \in H^3(\Omega)^4$.

Since $\boldsymbol{\xi}^* \in H^4(\Omega_{\mathbf{x}})^2$, then $\xi^*_x, \xi^*_y, \eta^*_x, \eta^*_y \in H^3(\Omega_{\mathbf{x}})$. From the corollary to the Sobolev imbedding theorem (with $m_1 = 3,\ m_2 = 3$, and $m = 3$), we must have $J^*_{\mathbf{x}} = \xi^*_x\eta^*_y - \xi^*_y\eta^*_x \in H^3(\Omega_{\mathbf{x}})$. Recall from section 2 that $\mathbf{J}^*$ is continuous and uniformly positive and bounded: $J_0 \le |J^*_{\mathbf{x}}(x,y)| \le J_1$ for some constants $J_0, J_1 \in R^+$ and all $(x,y) \in \Omega_{\mathbf{x}}$.

Dropping the superscript * for convenience, consider J_{11}. (The other entries are treated similarly.) Using the corollary to the Sobolev imbedding theorem [11] with $m_1 = 3,\ m_2 = 3$, and $m = 3$, we get

$$\|J_{11}\|_{3,\Omega} = \left\|\frac{1}{J_{\mathbf{x}}}\eta_y\right\|_{3,\Omega_{\mathbf{x}}} \le C\left\|\frac{1}{J_{\mathbf{x}}}\right\|_{3,\Omega_{\mathbf{x}}}\|\eta_y\|_{3,\Omega_{\mathbf{x}}}.$$

Therefore, we need only show that $\frac{1}{J_{\mathbf{x}}} \in H^3$. But

$$\left\|\frac{1}{J_{\mathbf{x}}}\right\|^2_{3,\Omega_{\mathbf{x}}} = \sum_{i\le 3}\left\|\partial^i\frac{1}{J_{\mathbf{x}}}\right\|^2_{0,\Omega_{\mathbf{x}}}.$$

We consider each order separately. By Theorem 3.2 in [21], for any $a \in C^0(\Omega_{\mathbf{x}})$ and $b \in L^2(\Omega_{\mathbf{x}})$, we have

$$\|ab\|_{0,\Omega_{\mathbf{x}}} \le \|a\|_{\infty,\Omega_{\mathbf{x}}}\|b\|_{0,\Omega_{\mathbf{x}}}. \tag{3.1}$$

For the zeroth-order term, using (3.1) yields

$$\left\|\frac{1}{J_{\mathbf{x}}}\right\|_{0,\Omega_{\mathbf{x}}} \le \left\|\frac{1}{J_{\mathbf{x}}}\right\|_{\infty,\Omega_{\mathbf{x}}}\|1\|_{0,\Omega_{\mathbf{x}}} \le \frac{1}{J_0}\|1\|_{0,\Omega_{\mathbf{x}}}.$$

For the first-order term, we use (3.1) to get

$$\left\|\partial\frac{1}{J_{\mathbf{x}}}\right\|_{0,\Omega_{\mathbf{x}}} = \left\|\frac{-1}{J^2_{\mathbf{x}}}\partial J_{\mathbf{x}}\right\|_{0,\Omega_{\mathbf{x}}} \le \frac{1}{J_0^2}\|\partial J_{\mathbf{x}}\|_{0,\Omega_{\mathbf{x}}} \le \frac{1}{J_0^2}\|J_{\mathbf{x}}\|_{1,\Omega_{\mathbf{x}}}.$$

For the second-order term, we use the triangle inequality, (3.1), and the corollary to the Sobolev imbedding theorem with $m_1 = m_2 = 1$ and $m = 0$ to get

$$\begin{aligned}
\left\|\partial^2 \frac{1}{J_{\mathbf{x}}}\right\|_{0,\Omega_{\mathbf{x}}} &= \left\|\frac{2}{J_{\mathbf{x}}^3}(\partial J_{\mathbf{x}})^2 + \frac{-1}{J_{\mathbf{x}}^2}\partial^2 J_{\mathbf{x}}\right\|_{0,\Omega_{\mathbf{x}}} \\
&\le \left\|\frac{2}{J_{\mathbf{x}}^3}(\partial J_{\mathbf{x}})^2\right\|_{0,\Omega_{\mathbf{x}}} + \left\|\frac{1}{J_{\mathbf{x}}^2}\partial^2 J_{\mathbf{x}}\right\|_{0,\Omega_{\mathbf{x}}} \\
&\le \frac{2}{J_0^3}\left\|(\partial J_{\mathbf{x}})^2\right\|_{0,\Omega_{\mathbf{x}}} + \frac{1}{J_0^2}\left\|\partial^2 J_{\mathbf{x}}\right\|_{0,\Omega_{\mathbf{x}}} \\
&\le C\left(\frac{2}{J_0^3}\left\|\partial J_{\mathbf{x}}\right\|_{1,\Omega_{\mathbf{x}}}^2 + \frac{1}{J_0^2}\left\|J_{\mathbf{x}}\right\|_{2,\Omega_{\mathbf{x}}}\right) \\
&\le C\left(\frac{2}{J_0^3}\left\|J_{\mathbf{x}}\right\|_{2,\Omega_{\mathbf{x}}}^2 + \frac{1}{J_0^2}\left\|J_{\mathbf{x}}\right\|_{2,\Omega_{\mathbf{x}}}\right).
\end{aligned}$$

For the third-order term, we use the triangle inequality, (3.1), and the Corollary to the Sobolev imbedding theorem once with $m_1 = 1$, $m_2 = \frac{3}{4}$, and $m = 0$, once with $m_1 = m_2 = 1$ and $m = 0$, and once with $m_1 = m_2 = 1$ and $m = \frac{3}{4}$ to get

$$\begin{aligned}
\left\|\partial^3 \frac{1}{J_{\mathbf{x}}}\right\|_{0,\Omega_{\mathbf{x}}} &= \left\|\frac{-6}{J_{\mathbf{x}}^4}(\partial J_{\mathbf{x}})^3 + \frac{6}{J_{\mathbf{x}}^3}\partial J_{\mathbf{x}}\partial^2 J_{\mathbf{x}} + \frac{-1}{J_{\mathbf{x}}^2}\partial^3 J_{\mathbf{x}}\right\|_{0,\Omega_{\mathbf{x}}} \\
&\le \left\|\frac{6}{J_{\mathbf{x}}^4}(\partial J_{\mathbf{x}})^3\right\|_{0,\Omega_{\mathbf{x}}} + \left\|\frac{6}{J_{\mathbf{x}}^3}\partial J_{\mathbf{x}}\partial^2 J_{\mathbf{x}}\right\|_{0,\Omega_{\mathbf{x}}} + \left\|\frac{1}{J_{\mathbf{x}}^2}\partial^3 J_{\mathbf{x}}\right\|_{0,\Omega_{\mathbf{x}}} \\
&\le \frac{6}{J_0^4}\left\|(\partial J_{\mathbf{x}})^3\right\|_{0,\Omega_{\mathbf{x}}} + \frac{6}{J_0^3}\left\|\partial J_{\mathbf{x}}\partial^2 J_{\mathbf{x}}\right\|_{0,\Omega_{\mathbf{x}}} + \frac{1}{J_0^2}\left\|\partial^3 J_{\mathbf{x}}\right\|_{0,\Omega_{\mathbf{x}}} \\
&\le C\left(\frac{6}{J_0^4}\left\|\partial J_{\mathbf{x}}\right\|_{1,\Omega_{\mathbf{x}}}^3 + \frac{6}{J_0^3}\left\|\partial J_{\mathbf{x}}\right\|_{1,\Omega_{\mathbf{x}}}\left\|\partial^2 J_{\mathbf{x}}\right\|_{1,\Omega_{\mathbf{x}}} + \frac{1}{J_0^2}\left\|J_{\mathbf{x}}\right\|_{3,\Omega_{\mathbf{x}}}\right) \\
&\le C\left(\frac{6}{J_0^4}\left\|J_{\mathbf{x}}\right\|_{2,\Omega_{\mathbf{x}}}^3 + \frac{6}{J_0^3}\left\|J_{\mathbf{x}}\right\|_{2,\Omega_{\mathbf{x}}}\left\|J_{\mathbf{x}}\right\|_{3,\Omega_{\mathbf{x}}} + \frac{1}{J_0^2}\left\|J_{\mathbf{x}}\right\|_{3,\Omega_{\mathbf{x}}}\right).
\end{aligned}$$

The result follows from these bounds. □

LEMMA 3.5. $\mathbf{P}[\mathbf{D}] \in H^{\epsilon}(\Omega)^p$ *for every* $\mathbf{D} \in B_r$*: there exists a constant* C*, depending only on* $\mathbf{D}^*$*,* $\mathbf{E}$*,* r*,* Ω*, and* δ*, such that*

$$\|\mathbf{P}(\mathbf{D})\|_{\epsilon,\Omega} \le C \quad \forall\ \mathbf{D} \in B_r. \tag{3.2}$$

Proof. The products in (2.7) are of the form treated in Lemma 3.1. In fact, there exists a constant C, depending only on Ω and δ, such that

$$\|\mathbf{P}(\mathbf{D})\|_{\epsilon,\Omega} \le C(\|\mathbf{D}+\mathbf{E}\|_{1+\delta,\Omega}^2\|\mathbf{D}+\mathbf{E}\|_{1+\epsilon,\Omega} + \|\mathbf{D}\|_{1+\epsilon,\Omega}), \tag{3.3}$$

and so (3.2) follows because $\mathbf{D} \in B_r$. □

Next we establish uniform coercivity and continuity of $\mathbf{P}'$ in a neighborhood of $\mathbf{D}^*$. This result needs the assumption that $\mathbf{P}'(\mathbf{D}^*)[\,\cdot\,]$ is one-to-one on $\mathcal{H}_{1+\delta}$, which is a consequence of an analogous assumption on the original EGG equations.

LEMMA 3.6 (ellipticity property). $\|\mathbf{P}'(\mathbf{D})[\,\cdot\,]\|_{\epsilon,\Omega}$ *is* $H^{1+\epsilon}(\Omega)^4$ *equivalent: there exist constants* c_c *and* c_b*, depending only on* $\mathbf{D}^*, \mathbf{E}, r, \Omega$*, and* δ*, such that*

$$\frac{1}{c_c}\|\mathbf{K}\|_{1+\epsilon,\Omega} \le \|\mathbf{P}'(\mathbf{D})[\mathbf{K}]\|_{\epsilon,\Omega} \le c_b\|\mathbf{K}\|_{1+\epsilon,\Omega} \quad \forall\ \mathbf{K} \in \mathcal{H}_{1+\epsilon}. \tag{3.4}$$

Proof. The products in (2.9) are of the form treated in Lemmas 3.1 and 3.2. In fact, there exists a constant C, depending only on Ω and δ, such that

$$\|\mathbf{P}'(\mathbf{D})[\mathbf{K}]\|_{\epsilon,\Omega} \leq C(\|\mathbf{D}+\mathbf{E}\|^2_{1+\delta,\Omega}\|\mathbf{K}\|_{1+\epsilon,\Omega} + \|\mathbf{K}\|_{1+\epsilon,\Omega}). \tag{3.5}$$

Proof of the lower bound follows from Theorem 10.5 of [2], as we now show. We first need to prove H^{m+1} boundedness and coercivity for $\mathbf{P}'(\mathbf{D}^*)[\mathbf{K}]$: there exist constants c_1 and c_3, depending only on $\mathbf{D}^*, \mathbf{E}, m$, and Ω, such that

$$\frac{1}{c_1}\|\mathbf{K}\|_{m+1,\Omega} \leq \|\mathbf{P}'(\mathbf{D}^*)[\mathbf{K}]\|_{m,\Omega} \leq c_3\|\mathbf{K}\|_{m+1,\Omega} \quad \forall\ \mathbf{K} \in \mathcal{H}_{m+1}, \tag{3.6}$$

for any $m \in [0,1]$. The upper bound is simply an application of the corollary to the Sobolev imbedding theorem similar to Lemmas 3.1 and 3.2. Consider the lower bound. It would be a simple matter to just assume $\mathbf{D}^* \in \mathcal{H}_{3+\delta}$ and then, because the coefficients would be sufficiently smooth, apply the theory of [2] (herafter referred to as ADN2 theory) for both $m = 0$ and $m = 1$. Instead, we just have $\mathbf{D}^* \in \mathcal{H}_{2+\delta}$, so while the higher-order coefficients are in C^1, the lower-order coefficients are only in C^0. This means that we need more care.

First consider $m = 0$. What follows for this case is a straightforward application of ADN2 theory to the entire system because all of the coefficients are sufficiently smooth. Recall that Ω is a bounded open subset of R^2 with $C^{3,1}$ boundary Γ. We write the system as

$$\begin{array}{rcl} \mathcal{L}\mathbf{K} = \mathbf{f} & \text{in} & \Omega, \\ \mathcal{B}\mathbf{K} = \mathbf{g} & \text{on} & \Gamma, \end{array} \tag{3.7}$$

where $\mathcal{L} \equiv \mathbf{P}'(\mathbf{D}^*)$ and $\mathcal{B} = \mathbf{n}\times$. (Recall that $\mathbf{n}$ is the outward unit normal on Γ (2.3).) For convenience, we write the coefficients using $\mathbf{J}^* = \mathbf{D}^* + \mathbf{E}$ and drop the * from the components. Note that $\mathcal{L} = \mathcal{L}_1 + \mathcal{L}_2$ and $l_{ij} = l'_{ij} + l''_{ij}$, where

$$\mathcal{L}_1 = (l'_{ij}(\boldsymbol{\xi}, \partial)) = \begin{pmatrix} \alpha\partial_\xi - \beta\partial_\eta & -\beta\partial_\xi + \gamma\partial_\eta & 0 & 0 \\ 0 & 0 & \alpha\partial_\xi - \beta\partial_\eta & -\beta\partial_\xi + \gamma\partial_\eta \\ -\partial_\eta & \partial_\xi & 0 & 0 \\ 0 & 0 & -\partial_\eta & \partial_\xi \end{pmatrix},$$

$$\begin{aligned} \mathcal{L}_2 &= (l''_{ij}(\boldsymbol{\xi}, \partial)) \\ &= \begin{pmatrix} 2J_{11}J_{21,\eta} - J_{21}(J_{11,\eta} + J_{21,\xi}) & 2J_{11}J_{22,\eta} - J_{21}(J_{12,\eta} + J_{22,\xi}) & 0 & 0 \\ 2J_{21}J_{11,\xi} - J_{11}(J_{11,\eta} + J_{21,\xi}) & 2J_{21}J_{12,\xi} - J_{11}(J_{12,\eta} + J_{22,\xi}) & 0 & 0 \\ 2J_{12}J_{21,\eta} - J_{22}(J_{11,\eta} + J_{21,\xi}) & 2J_{12}J_{22,\eta} - J_{22}(J_{12,\eta} + J_{22,\xi}) & 0 & 0 \\ 2J_{22}J_{11,\xi} - J_{12}(J_{11,\eta} + J_{21,\xi}) & 2J_{22}J_{12,\xi} - J_{12}(J_{12,\eta} + J_{22,\xi}) & 0 & 0 \end{pmatrix}^t, \end{aligned}$$

$$\alpha = J_{21}^2 + J_{22}^2, \qquad \beta = J_{11}J_{21} + J_{12}J_{22}, \qquad \gamma = J_{11}^2 + J_{12}^2.$$

Note that

$$\mathcal{B} = (b_{ij}(\boldsymbol{\xi}, \partial)) = \begin{pmatrix} -n_2 & n_1 & 0 & 0 \\ 0 & 0 & -n_2 & n_1 \end{pmatrix}. \tag{3.8}$$

In ADN2 theory, three types of integer weights are used to determine the leading order terms for boundary value problem (3.7). Weight $s_i \leq 0$ refers to the ith equation, weight $t_j \geq 0$ to the jth dependent variable, and weight r_k to the kth boundary condition. These weights are chosen as small as possible but so that

$$\deg l_{ij}(\boldsymbol{\xi}, \partial) \leq s_i + t_j, \quad \deg b_{kj}(\boldsymbol{\xi}, \partial) \leq r_k + t_j, \quad i, j = 1, 2, 3, 4, \quad k = 1, 2,$$

where deg refers to the order of the derivatives. Our weights are

$$s_i = 0, \qquad t_j = 1, \qquad r_k = -1, \qquad i, j = 1, 2, 3, 4, \quad k = 1, 2.$$

The leading order part of $\mathcal{L}$ consists of the elements l_{ij} for which $\deg l_{ij}(\boldsymbol{\xi}, \partial) = s_i + t_j = 1$. Therefore, $\mathcal{L}_1$ is the leading order (in this case, first-order) part. The leading order part of $\mathcal{B}$ consists of elements b_{kj} for which $\deg b_{kj}(\boldsymbol{\xi}, \partial) = r_k + t_j = 0$. Therefore, the leading order (in this case, zeroth-order) part of $\mathcal{B}$ is $\mathcal{B}$ itself.

We must show that $\mathcal{L}_1$ satisfies two ADN2 conditions: the *supplementary condition on its determinant* and *uniform ellipticity*. ($\mathcal{L}_1$ will then automatically be elliptic.) ADN2 also requires that the system of equations and boundary conditions be well posed. This means that $\mathcal{L}_1$ and $\mathcal{B}$, when combined, must satisfy the *complementing boundary condition*. Let L denote the determinant of $\mathcal{L}_1$:

$$L(\boldsymbol{\xi}, \partial) = \det(l'_{ij}) = -(\alpha\partial_\xi^2 - 2\beta\partial_\xi\partial_\eta + \gamma\partial_\eta^2)^2.$$

Since $J > 0$ (see section 2), then

$$\begin{aligned} \beta^2 - \alpha\gamma &= (x_\xi x_\eta + y_\xi y_\eta)^2 - (x_\eta^2 + y_\eta^2)(x_\xi^2 + y_\xi^2) \\ &= -(x_\xi y_\eta - y_\xi x_\eta)^2 = -J^2 < 0. \end{aligned} \tag{3.9}$$

Let $\mathbf{d} = (d \quad e)^t$ and $\mathbf{p} = (p \quad q)^t$ be any two linearly independent vectors. To aid clarity of the following discussion, we first define some quantities:

$$\begin{aligned} \mathcal{A} &= \alpha dp - \beta(pe + dq) + \gamma eq, \\ \mathcal{B} &= \sqrt{\mathcal{DC} - \mathcal{A}^2} = J|pe - dq|, \\ \mathcal{C} &= \alpha d^2 - 2\beta de + \gamma e^2, \\ \mathcal{D} &= \alpha p^2 - 2\beta pq + \gamma q^2. \end{aligned}$$

Note that $\mathcal{B} > 0$ since $J > 0$ and linear independence of $\mathbf{p}$ and $\mathbf{d}$ implies $|pe - dq| > 0$.

The *supplementary condition on L* requires the equation

$$L(\boldsymbol{\xi}, \mathbf{d} + \tau\mathbf{p}) = -\{\mathcal{C} + 2\tau\mathcal{A} + \tau^2\mathcal{D}\}^2 = 0 \tag{3.10}$$

to have exactly two roots in τ with positive imaginary part. Polynomial (3.10) has two double roots ($\iota = \sqrt{-1}$),

$$\tau = \frac{-\mathcal{A} \pm \iota\mathcal{B}}{\mathcal{D}},$$

which form two complex conjugate pairs. Thus, (3.10) does indeed have exactly two roots with positive imaginary part (one such double root).

To satisfy *uniform ellipticity*, we need to show that

$$\frac{1}{C}\|\mathbf{d}\|^4 \leq |L(\boldsymbol{\xi}, \mathbf{d})| \leq C\|\mathbf{d}\|^4, \tag{3.11}$$

with $\|\mathbf{d}\| = \sqrt{d^2 + e^2}$, for all vectors $\mathbf{d} \neq \mathbf{0}$ and points $\boldsymbol{\xi}$ in Ω.

To prove the left bound in (3.11), let $\rho = \frac{|\beta|}{\sqrt{\alpha\gamma}} < 1$ (see (3.9)). Then

$$\begin{aligned}
|L\,(\boldsymbol{\xi}, \mathbf{d})| &= (\alpha d^2 - 2\beta de + \gamma e^2)^2 \\
&\geq (\alpha d^2 - 2|\beta||d||e| + \gamma e^2)^2 \\
&= (\alpha d^2 - 2\rho\sqrt{\alpha\gamma}|d||e| + \gamma e^2)^2 \\
&= ((1-\rho)(\alpha d^2 + \gamma e^2) + \rho(\sqrt{\alpha} d - \sqrt{\gamma} e)^2)^2 \\
&\geq ((1-\rho)(\alpha d^2 + \gamma e^2))^2 \\
&\geq \min((1-\rho)^2\alpha^2, (1-\rho)^2\gamma^2)(d^2 + e^2)^2 \\
&= \min((1-\rho)^2\alpha^2, (1-\rho)^2\gamma^2)\|\mathbf{d}\|^4.
\end{aligned}$$

To prove the right bound in (3.11), note that Hölder's inequality implies that

$$\begin{aligned}
|L\,(\boldsymbol{\xi}, \mathbf{d})| &\leq (\alpha d^2 + 2|\beta||d||e| + \gamma e^2)^2 \\
&\leq (\alpha d^2 + |\beta|(d^2 + e^2) + \gamma e^2)^2 \\
&\leq \max((\alpha + |\beta|)^2, (\gamma + |\beta|)^2)(d^2 + e^2)^2 \\
&= \max((\alpha + |\beta|)^2, (\gamma + |\beta|)^2)\|\mathbf{d}\|^4.
\end{aligned}$$

We then establish (3.11) by choosing

$$C = \max\left\{\frac{1}{(1-\rho)^2\alpha^2}, \frac{1}{(1-\rho)^2\gamma^2}, (\alpha + |\beta|)^2, (\gamma + |\beta|)^2\right\}.$$

This shows that operator L satisfies the two conditions of ADN2. We now prove that the problem is well posed by showing that $\mathcal{L}_1$ and $\mathcal{B}$ satisfy the *complementing boundary condition*. This condition involves comparing two polynomials. We consider a point on the boundary with normal $\mathbf{d} = (d \ \ e)^t$ and tangent $\mathbf{p} = (p \ \ q)^t$ vectors. The first polynomial is formed from the roots of (3.10) with positive imaginary parts:

$$M^+(\boldsymbol{\xi}, \mathbf{d}, \tau) = \left[\tau + \frac{\mathcal{A} - \iota\mathcal{B}}{\mathcal{D}}\right]^2. \tag{3.12}$$

The second polynomial is formed from the leading order elements of $\mathcal{L}$ and $\mathcal{B}$:

$$\sum_{k=1}^{2} a_k (BL)_{km}, \tag{3.13}$$

where

$$(BL)_{km} = \sum_{j=1}^{4} b_{kj}(\boldsymbol{\xi}, \mathbf{d} + \tau\mathbf{p}) l^{jm}(\boldsymbol{\xi}, \mathbf{d} + \tau\mathbf{p}),$$

and $l^{jm}(\boldsymbol{\xi}, \mathbf{d} + \tau\mathbf{p})$ are the elements of the (classical) adjoint $(l'_{ij} l^{jm} = \delta_i^m L,\ i, j, m = 1, 2, 3, 4)$ of $l'_{ij}(\boldsymbol{\xi}, \mathbf{d} + \tau\mathbf{p})$ and b_{kj} is defined in (3.8).

The polynomials for (3.13) are

$$\left(\sum_{k=1}^{2} a_k (BL)_{km}\right) = \mathcal{D}\left[\tau + \frac{\mathcal{A} + \iota\mathcal{B}}{\mathcal{D}}\right]\left[\tau + \frac{\mathcal{A} - \iota\mathcal{B}}{\mathcal{D}}\right]\begin{pmatrix} a_1(pe - qd) \\ a_2(pe - qd) \\ a_1(\mathcal{A} + \tau\mathcal{D}) \\ a_2(\mathcal{A} + \tau\mathcal{D}) \end{pmatrix}^t. \tag{3.14}$$

Comparing polynomials (3.12) and (3.14) and noting that $\mathcal{B} > 0$, we have that (3.12) is not a factor of (3.14). Thus, the *complementing boundary condition* is satisfied.

Theorem 10.5 from [2] implies that, for $l_{ij} \in C^m(\overline{\Omega}), b_{kj} \in C^{m+1}(\Gamma)$, there exists a constant c_1 that depends only on $\mathbf{D}^*, \mathbf{E}, r, \Omega$, and δ such that, if $K_j \in H^1(\Omega)$, $1 \leq j \leq 4$, solves (3.7) and is unique, then $K_j \in H^{m+1}(\Omega)$ and

$$\|K_j\|_{m+1,\Omega} \leq \frac{c_1}{4} \left[\sum_{i=1}^{4} \|f_i\|_{m,\Omega} + \sum_{i=1}^{2} \|g_i\|_{m+\frac{1}{2},\Gamma} \right],$$

where f_i and g_i are the components of $\mathbf{f}$ and $\mathbf{g}$, respectively, in (3.7). The coefficients of $\mathcal{L}$ are at least in $H^{1+\delta}(\Omega)$ and $C^0(\Omega)$. For the boundary conditions, we have $\mathcal{B} = \mathbf{n}\times$ and $\Gamma \in C^{3,1}$, and thus we get $b_{kj} \in C^2(\Gamma)$. The boundary conditions are homogeneous, and so we can drop the boundary term in the inequality. We therefore have

$$\|\mathbf{K}\|_{1,\Omega} \leq c_1 \|\mathcal{L}(\mathbf{D}^*)[\mathbf{K}]\|_{0,\Omega} \quad \forall\ \mathbf{K} \in \mathcal{H}_1(\Omega)^4. \tag{3.15}$$

Now consider $m = 1$. We cannot simply apply ADN2 to the whole system because the coefficients are not sufficiently smooth. Instead, we split the operator according to $\mathcal{L} = \mathcal{L}_1 + \mathcal{L}_2$ and restrict our ADN2 result to reduced system

$$\begin{aligned} \mathcal{L}_1\mathbf{K} = \mathbf{f} \quad &\text{in} \quad \Omega, \\ \mathcal{B}\mathbf{K} = \mathbf{g} \quad &\text{on} \quad \Gamma. \end{aligned}$$

Operator $\mathcal{L}_1$ satisfies the ADN2 conditions (as illustrated for case $m = 0$), and thus

$$\|\mathbf{K}\|_{2,\Omega} \leq c_1 \|\mathcal{L}_1(\mathbf{D}^*)[\mathbf{K}]\|_{1,\Omega} \quad \forall\ \mathbf{K} \in \mathcal{H}_2. \tag{3.16}$$

The coefficients of $\mathcal{L}_1$ are in $H^{2+\delta}(\Omega)$ and $C^1(\Omega)$. For the boundary conditions, we have $\mathcal{B} = \mathbf{n}\times$ and $\Gamma \in C^{3,1}$, and thus we get $b_{kj} \in C^2(\Gamma)$. The boundary conditions are homogeneous, and so we can drop the boundary term in the inequality.

We use Lemma 3.3 to obtain

$$\|\mathcal{L}_2(\mathbf{D}^*)[\mathbf{K}]\|_{1,\Omega} \leq c_2 \|\mathbf{K}\|_{1,\Omega} \quad \forall\ \mathbf{K} \in \mathcal{H}_2. \tag{3.17}$$

Note that c_2 depends continuously on $\sup_{\mathbf{D} \in B_r} \|\mathbf{D}^* + \mathbf{E}\|_{2+\delta,\Omega}$, and thus it depends on $\mathbf{D}^*, \mathbf{E}, r$, Ω, and δ.

Combining (3.16) and (3.17) yields

$$\|\mathbf{K}\|_{2,\Omega} \leq C(\|\mathbf{P}'(\mathbf{D}^*)[\mathbf{K}]\|_{1,\Omega} + \|\mathbf{K}\|_{1,\Omega}) \quad \forall\ \mathbf{K} \in \mathcal{H}_2. \tag{3.18}$$

This is a Gårdings inequality (cf. [13, 19]), which allows us now to prove that

$$\frac{1}{c_c} \|\mathbf{K}\|_{2,\Omega} \leq \|\mathbf{P}'(\mathbf{D}^*)[\mathbf{K}]\|_{1,\Omega} \quad \forall\ \mathbf{K} \in \mathcal{H}_2. \tag{3.19}$$

To this end, assume that (3.19) is not true. Then there exists a sequence $\mathbf{K}_j \in \mathcal{H}_2$ such that

$$\|\mathbf{K}_j\|_{2,\Omega} = 1 \tag{3.20}$$

and

$$\|\mathbf{P}'(\mathbf{D}^*)[\mathbf{K}_j]\|_{1,\Omega} = \frac{1}{j}, \quad j = 1, 2 \ldots. \tag{3.21}$$

Now, because $H^2(\Omega)$ is compactly imbedded in $H^1(\Omega)$ (cf. the Rellich selection theorem [5]), then (3.20) implies that there exists a limit $\hat{\mathbf{K}} \in \mathcal{H}_2$ of a subsequence $\mathbf{K}_{j_k} \to \hat{\mathbf{K}}$ in the $H^1(\Omega)$ norm. Combining this with (3.18) and (3.21), we know that $\mathbf{K}_{j_k}$ must also be a Cauchy sequence in the $H^2(\Omega)$ norm with some limit $\bar{\mathbf{K}}$. However, from the upper bound in (3.4), we have

$$\begin{aligned}\lim_{j_k\to\infty} \|\mathbf{P}'(\mathbf{D}^*)[\mathbf{K}_{j_k}] - \mathbf{P}'(\mathbf{D}^*)[\bar{\mathbf{K}}]\|_{1,\Omega} &= \lim_{j_k\to\infty} \|\mathbf{P}'(\mathbf{D}^*)[\mathbf{K}_{j_k} - \bar{\mathbf{K}}]\|_{1,\Omega} \\ &\le \lim_{j_k\to\infty} \|\mathbf{K}_{j_k} - \bar{\mathbf{K}}\|_{2,\Omega} = 0.\end{aligned}$$

From (3.21), we thus obtain

$$\|\mathbf{P}'(\mathbf{D}^*)[\bar{\mathbf{K}}]\|_{1,\Omega} \le \lim_{j_k\to\infty} \{\|\mathbf{P}'(\mathbf{D}^*)[\mathbf{K}_{j_k}] - \mathbf{P}'(\mathbf{D}^*)[\bar{\mathbf{K}}]\|_{1,\Omega} + \|\mathbf{P}'(\mathbf{D}^*)[\mathbf{K}_{j_k}]\|_{1,\Omega}\} = 0.$$

However, $\mathbf{P}'(\mathbf{D}^*)[\,\cdot\,]$ is one-to-one. Hence, $\mathbf{P}'(\mathbf{D}^*)[\bar{\mathbf{K}}] = \mathbf{0}$ implies that $\bar{\mathbf{K}} = \mathbf{0}$, which in turn implies that $\|\bar{\mathbf{K}}\|_{1,\Omega} = 0$, contradicting (3.20). Thus, (3.4) and the lemma are established for $m = 1$. We have thus established (3.6) for both $m = 0$ and $m = 1$.

For the general case of $m \in [0, 1]$, bound (3.6) follows from the results in [17, 3, 18] and the following proof of elliptic regularity of the formal adjoint problem.

Consider boundary value problem (3.7), for $\mathbf{K} \in \mathcal{H}_{m+1}$ and $\mathcal{L}\mathbf{K} \in \mathcal{V}_m = \{\mathbf{D} \in H^m(\Omega)^4\}$. From [2], we know that $\mathcal{L}\mathbf{K} = \mathbf{f}$ is onto. This system has normal boundary conditions, and hence, the formal adjoint problem has normal boundary conditions of the same type [17]. We thus consider

$$\begin{aligned}\mathcal{L}^*\mathbf{M} = \mathbf{f}_1 \quad &\text{in} \quad \Omega,\\ \mathcal{B}^*\mathbf{M} = \mathbf{0} \quad &\text{on} \quad \Gamma,\end{aligned}$$

for $\mathbf{M} \in (\mathcal{V}_m)^* = \{\mathbf{M} \in H^{-m}(\Omega)^4 : \mathcal{B}^*\mathbf{M} = \mathbf{0} \text{ on } \Gamma\}$ and $\mathcal{L}^*\mathbf{M} \in (\mathcal{H}_{m+1})^*$. The system is both Petrovskii elliptic, because $s_1 = s_2 = s_3 = s_4 = 0$, and homogeneous elliptic, because $t_1 = t_2 = t_3 = t_4$; cf. [17]. Thus, the adjoint system is elliptic [17] and has a similar ellipticity result in the dual space: for all $\mathbf{M} \in (\mathcal{V}_m)^*$ we have

$$\begin{aligned}\|\mathbf{M}\|_{-m,\Omega} &= \sup_{\mathbf{V}\neq 0\in\mathcal{V}_m} \frac{(\mathbf{M},\mathbf{V})}{\|\mathbf{V}\|_{m,\Omega}}\\ &= \sup_{\mathbf{K}\neq 0\in\mathcal{H}_{m+1}} \frac{(\mathbf{M},\mathcal{L}\mathbf{K})}{\|\mathcal{L}\mathbf{K}\|_{m,\Omega}}\\ &\le c_1 \sup_{\mathbf{K}\neq 0\in\mathcal{H}_{m+1}} \frac{(\mathcal{L}^*\mathbf{M},\mathbf{K})}{\|\mathbf{K}\|_{m+1,\Omega}}\\ &= c_1\|\mathcal{L}^*\mathbf{M}\|_{-(m+1),\Omega}\end{aligned}$$

and

$$\begin{aligned}\|\mathbf{M}\|_{-m,\Omega} &= \sup_{\mathbf{V}\neq 0\in\mathcal{V}_m} \frac{(\mathbf{M},\mathbf{V})}{\|\mathbf{V}\|_{m,\Omega}}\\ &= \sup_{\mathbf{K}\neq 0\in\mathcal{H}_{m+1}} \frac{(\mathbf{M},\mathcal{L}\mathbf{K})}{\|\mathcal{L}\mathbf{K}\|_{m,\Omega}}\\ &\ge \frac{1}{c_3} \sup_{\mathbf{K}\neq 0\in\mathcal{H}_{m+1}} \frac{(\mathcal{L}^*\mathbf{M},\mathbf{K})}{\|\mathbf{K}\|_{m+1,\Omega}}\\ &= \frac{1}{c_3}\|\mathcal{L}^*\mathbf{M}\|_{-(m+1),\Omega}.\end{aligned}$$

The result for all $m \in [0,1]$ now follows from interpolation [15], use of local maps, and a partition of unity. (If we had assumed $\mathbf{D}^* \in C^\infty(\Omega)^4$ and $\Gamma \in C^\infty$, then the ellipticity result would hold for all real m; we only need this result for $m \in [0,1]$, and so we are able to reduce the continuity requirements of [15], as we have.)

We now generalize the result for $\mathbf{D} \in B_r$. Using a Taylor expansion, the triangle inequality, Lemmas 3.5 and 3.7, and (3.6), we have (for $\epsilon = 0$ or δ)

$$\begin{aligned}
\|\mathbf{P}'(\mathbf{D})[\mathbf{K}]\|_{\epsilon,\Omega} &= \|\mathbf{P}'(\mathbf{D}^*)[\mathbf{K}] + \mathbf{P}''(\widetilde{\mathbf{D}})[\mathbf{K}, \mathbf{D}^* - \mathbf{D}]\|_{\epsilon,\Omega} \\
&\geq \|\mathbf{P}'(\mathbf{D}^*)[\mathbf{K}]\|_{\epsilon,\Omega} - \|\mathbf{P}''(\widetilde{\mathbf{D}})[\mathbf{K}, \mathbf{D}^* - \mathbf{D}]\|_{\epsilon,\Omega} \\
&\geq c_1\|\mathbf{K}\|_{1+\epsilon,\Omega} - c_4\|\widetilde{\mathbf{D}} + \mathbf{E}\|_{1+\delta,\Omega}\|\mathbf{K}\|_{1+\epsilon,\Omega}\|\mathbf{D}^* - \mathbf{D}\|_{1+\delta,\Omega} \\
&\geq \|\mathbf{K}\|_{1+\epsilon,\Omega}(c_1 - c_4 r(\|\mathbf{D}^*\|_{1+\delta,\Omega} + \|\mathbf{E}\|_{1+\delta,\Omega} + r)) \\
&\geq \tfrac{1}{c_c}\|\mathbf{K}\|_{1+\epsilon,\Omega}
\end{aligned} \tag{3.22}$$

for sufficiently small r. The lemma now follows. □

LEMMA 3.7. *The second Fréchet derivative of* $\mathbf{P}(\mathbf{D})$ *is bounded for all* $\mathbf{D} \in B_r$*: for every* $\mathbf{D} \in B_r$ *there exists a constant* c_2*, depending only on* $\mathbf{D}^*$, $\mathbf{E}$, r, Ω, *and* δ, *such that*

$$\|\mathbf{P}''(\mathbf{D})[\mathbf{K}, \mathbf{K}]\|_{\epsilon,\Omega} \leq c_2\|\mathbf{K}\|_{1+\delta,\Omega}\|\mathbf{K}\|_{1+\epsilon,\Omega} \quad \forall\, \mathbf{K} \in \mathcal{H}_{1+\epsilon}(\Omega). \tag{3.23}$$

Here, $\mathbf{P}''(\mathbf{D})[\mathbf{K}, \mathbf{K}]$ *denotes the second Fréchet derivative of* $\mathbf{P}(\mathbf{D}_n)$ *with respect to* $\mathbf{D}_n$ *in directions* $\mathbf{K}$ *and* $\mathbf{K}$.

Proof. The products in (2.10) are of the form treated in Lemmas 3.1 and 3.2. In fact, there exists a constant C, depending only on Ω and δ, such that

$$\|\mathbf{P}''(\mathbf{D})[\mathbf{K}, \mathbf{K}]\|_{\epsilon,\Omega} \leq C\|\mathbf{D} + \mathbf{E}\|_{1+\delta,\Omega}\|\mathbf{K}\|_{1+\delta,\Omega}\|\mathbf{K}\|_{1+\epsilon,\Omega} \qquad \forall\, \mathbf{K} \in H^{1+\delta}(\Omega).$$

The lemma now follows. □

4. Numerical results. Here we validate our algorithm with numerical tests. Define H^h as the space of continuous piecewise bilinear functions corresponding to a uniform grid. Note that $H^h \subset H^{1+\delta}(\Omega)$ for any $\delta < \frac{1}{2}$. The functional to be minimized is

$$\begin{aligned}
&\mathbf{G}(\mathbf{x}_{n+1}, \mathbf{J}_{n+1}; \mathbf{x}_n, \mathbf{J}_n, \mathbf{w}) \\
&\quad = \varepsilon\|\mathbf{J}_{n+1} - \nabla\mathbf{x}_{n+1}\|^2_{0,\Omega} \\
&\quad + (1-\varepsilon)\left\|\tfrac{1}{J_n}\left[(\underline{\hat{\mathbf{J}}_n\hat{\mathbf{J}}^t_n}\nabla)\cdot\mathbf{J}_{n+1} + (\underline{\hat{\mathbf{J}}_{n+1}\hat{\mathbf{J}}^t_n}\nabla)\cdot\mathbf{J}_n + (\underline{\hat{\mathbf{J}}_n\hat{\mathbf{J}}^t_{n+1}}\nabla)\cdot\mathbf{J}_n - 2(\underline{\hat{\mathbf{J}}_n\hat{\mathbf{J}}^t_n}\nabla)\cdot\mathbf{J}_n\right]\right\|^2_{0,\Omega} \\
&\quad + (1-\varepsilon)\|\nabla \times \mathbf{J}_{n+1}\|^2_{0,\Omega} + \varepsilon\|\mathbf{x}_{n+1} - \mathbf{w}\|^2_{\frac{1}{2},\Gamma} + (1-\varepsilon)\|\mathbf{n} \times \mathbf{J}_{n+1} - \mathbf{n} \times \nabla\mathbf{w}\|^2_{\frac{1}{2},\Gamma}.
\end{aligned} \tag{4.1}$$

There are three aspects of (4.1) worth noting. The first is that we are solving for $\mathbf{x}_{n+1}, \mathbf{J}_{n+1}$ and not $\mathbf{D}_{n+1}$ as we did for the theory. While it was more convenient in the theory to incorporate the boundary conditions into the equations, here we enforce them, so that the last two terms in (4.1) vanish. A second aspect is the interstage scale factor ε. In [10], we discussed the two-stage algorithm, where, in the first stage, we set $\varepsilon = 0$ and solve for $\mathbf{J}_{n+1}$ and, in the second, we set $\varepsilon = 1$ and solve for $\mathbf{x}_{n+1}$. The second stage amounts to a simple system of decoupled Poisson equations. For $\varepsilon \in (0,1)$, minimizing (4.1) amounts to a single-stage algorithm. In section 4.1, we compare performance of the first stage of the two-stage algorithm with the single-stage algorithm for the pinched square (Figure 4.1, below). For the single stage, we set $\varepsilon = \frac{1}{2}$ and multiply the entire functional by 2 for fair comparison. Results for both algorithms are similar, and thus the remaining tests are for the single-stage algorithm.

The third aspect of (4.1) to notice is the presence of the equation scale factor $\frac{1}{J_n}$ in the second functional term. The EGG equations are derived from the well-understood Laplace equations, so we exploit this correspondence now to guide the choice of scales. First note that the augmented first stage of the first-order system [7] associated with the Laplace equations (2.1) that define $\boldsymbol{\xi}$ is (ignoring boundary conditions and with $\boldsymbol{\Psi} = \nabla\boldsymbol{\xi}$)

$$\begin{pmatrix} \nabla \cdot \boldsymbol{\Psi} \\ \nabla \times \boldsymbol{\Psi} \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}.$$

Transforming this system to that for the gradient $\mathbf{J}$ of the inverse map (without cancelling terms) yields

$$\text{(4.2)} \qquad \frac{1}{J^2}\underline{\hat{\mathbf{J}}} \begin{pmatrix} \frac{1}{J}[(\underline{\hat{\mathbf{J}}\hat{\mathbf{J}}}^t \nabla) \cdot \mathbf{J}] \\ \nabla \times \mathbf{J} \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}.$$

The key in scaling this new system is to understand the relative balance between its two equations. Thus, $\frac{1}{J^2}\underline{\hat{\mathbf{J}}}$ can be dropped in deference to the relative scale reflected in the $\frac{1}{J}$ term in the first equation of (4.2). To mimic the Laplace scaling for the EGG system, we thus choose to scale the second term in (4.1) by $\frac{1}{J}$. This expresses the scaling we use in the numerical experiments. To improve performance in practice, we use $\frac{1}{J}$ just to scale the norm; it is *not* involved in the linearization process. (Presence of the scale factor $\frac{1}{J}$ in this way does not affect the theoretical results, so it was omitted in the analysis to simplify the calculations.) The scaling effect is demonstrated in section 4.3, where we study the convergence factors for increasingly distorted maps for the pinched square using both unscaled and scaled functionals. In both sections 4.1 and 4.2, we measure actual errors as well as functional values and validate the equivalence of the square root of the functional and the H^1 errors as proved theoretically in section 3.

In section 4.2, we first test the performance of AMG on the one-sided pinched square with grid size $h = \frac{1}{64}$. We compare the performance of V(q,s)-cycles with $q + s \leq 3$, where q is the number of relaxation steps before coarse grid correction and s is the number after. We use V(1,1)-cycles for the rest of our tests because these initial results suggest that it is one of the most efficient of these choices. We then test dependence of the linear solver on grid size. We study how the convergence factor for linear solves suffers with increasingly large perturbations from the identity map for several different grid sizes.

The method we use to obtain an approximation to $\mathbf{D}^*$ (or $\mathbf{J}^*$) is discussed in some detail in [10]. Here we give a brief overview. We use a nested sequence of $m+1$ rectangular grids with continuous piecewise bilinear function subspaces of $\mathcal{H}_{1+\delta}$ denoted by $H^{h_0} \subset H^{h_1} \subset \cdots \subset H^{h_m} \subset \mathcal{H}_{1+\delta}$, where $h_n = 2^{-n}h_0$, $0 \leq n \leq m$. Let $\mathbf{V}_0$ denote the initial guess in H^{h_0} obtained by solving the problem on the coarsest subspace, H^{h_0}. In practice, we simply iterate with a discrete Newton iteration until the error in the approximation is below discretization error. The result, $\mathbf{V}_1$, becomes the initial guess for level h_1, where the process continues. In general, the initial guess for AMG on level h_n comes from the final AMG approximation on level h_{n-1}: $\mathbf{V}_n$.

In sections 4.2 and 4.4, we study performance of the NI algorithm. Here we use transfinite interpolation (TFI), which is analogous to linear interpolation, to form the initial guess. The basic principle is to add the linear interpolant between the north and south boundary maps to the linear interpolant between the east and west boundary

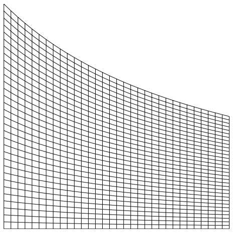


FIG. 4.1. *Pinched square with $a = 1$.*

maps, then subtract the interpolant between the four corners. The computational domain is the unit square. For boundary conditions $\mathbf{x} = \mathbf{w}(\boldsymbol{\xi})$, we get

$$\begin{aligned}\mathbf{x} = \eta\mathbf{w}_n(\xi) &+ (1-\eta)\mathbf{w}_s(\xi) + \xi\mathbf{w}_e(\eta) + (1-\xi)\mathbf{w}_w(\eta)\\ &- \left[\xi\eta\mathbf{w}_a + \xi(1-\eta)\mathbf{w}_b + (1-\xi)(1-\eta)\mathbf{w}_c + (1-\xi)\eta\mathbf{w}_d\right],\end{aligned}$$

where we define $\mathbf{w}_n$, $\mathbf{w}_s$, $\mathbf{w}_e$, and $\mathbf{w}_w$ as the boundary maps on the north, south, east, and west boundaries, respectively, and $\mathbf{w}_a$, $\mathbf{w}_b$, $\mathbf{w}_c$, and $\mathbf{w}_d$ as the values on the northeast, southeast, southwest, and northwest corners, respectively. The initial condition we use for $\mathbf{J}$ is the Jacobian of this map.

On the north and south boundaries, boundary conditions are needed for x, y, J_{11}, and J_{12}. On the east and west boundaries, boundary conditions are needed for x, y, J_{21}, and J_{22}. Boundary conditions are imposed on the finite element space.

We first establish the similarity between the first stage of the two-stage algorithm ($\varepsilon = 0$) and the single-stage algorithm ($\varepsilon = \frac{1}{2}$ and the functional in (4.1) multiplied by 2). Second, we test the effect of different numbers of relaxation sweeps for multigrid V-cycles to suggest a good choice for the remainder of the tests. Third, we study performance of the AMG solver for increasingly distorted grids for the pinched square. Finally, we study the algorithm on the arch. Further results can be found in [9].

4.1. First-stage and single-stage algorithms. The one-sided pinched square map has the following exact solution:

$$\begin{aligned}&x = \xi, && y = \frac{\eta}{a\xi+1},\\ &J_{11} = 1, && J_{21} = 0,\\ &J_{12} = \frac{a\eta}{(a\xi+1)^2}, && J_{22} = \frac{1}{a\xi+1},\end{aligned} \tag{4.3}$$

where $a \in [0, 1]$. The physical domain is a square for $a = 0$, with the pinch increasing as a increases. See Figure 4.1.

To test performance of the first-stage and single-stage algorithms for standard Newton iterations and NI on the pinched square with $a = 1.0$, we add a varying amount of small error at each grid point (except for those on the boundary) to TFI (the exact solution in this case) to form the initial guess:

$$\begin{aligned}&x = \xi + \mathbf{g}\sin(b\xi + c\eta), && y = \frac{\eta}{\xi+1} + \mathbf{g}\sin(d\xi + e\eta),\\ &J_{11} = 1 + \mathbf{g}\sin(b\xi + c\eta), && J_{21} = \mathbf{g}\sin(b\xi + c\eta),\\ &J_{12} = \frac{\eta}{(\xi+1)^2} + \mathbf{g}\sin(d\xi + e\eta), && J_{22} = \frac{1}{\xi+1} + \mathbf{g}\sin(d\xi + e\eta),\end{aligned}$$

TABLE 4.1
Asymptotic convergence factors for $V(1,1)$-cycles, with varying grid size and Newton iterations.

Newton	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{64}$	$\frac{1}{128}$
1	0.96	0.93	0.90	0.95
2	0.49	0.85	0.56	0.50
3	0.24	0.47	0.41	0.40
4	0.25	0.33	0.40	0.43
5	0.25	0.32	0.40	0.41
6	0.25	0.33	0.40	0.44

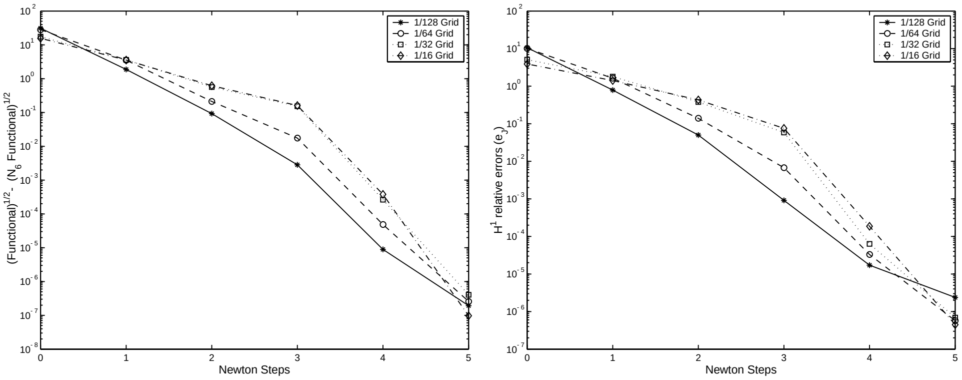


FIG. 4.2. *First-stage functional. Newton convergence, using standard Newton iterations with imposed boundary conditions, in both functional (left) and H^1 error (right) measures. Differences between the values at the current and sixth Newton steps are plotted.*

where

$$\begin{aligned}&\mathbf{g} = f\xi\eta(1-\xi)(1-\eta),\\ &b = 12967493.946193764, \quad c = 491843027.481264509,\\ &d = 184625498.4710938, \quad e = 174365204.5761938,\end{aligned}$$

with $f = 2$ for grid $h = \frac{1}{128}$, $f = 4$ for grids $h = \frac{1}{64}$ and $h = \frac{1}{32}$, $f = 7$ for grid $h = \frac{1}{16}$, and $f = 24$ for NI (with coarsest grid $h = \frac{1}{4}$). Note that the exact solutions for x, J_{11}, and J_{21} are in the finite-dimensional subspaces.

Consider the first-stage one-sided pinched square. (Recall that there are no x or y terms.) Table 4.1 depicts asymptotic convergence factors for the AMG solver. Note the poor performance shown in the early Newton steps. This degradation probably occurs because the functional is suffering from loss of elliptic character due to the crude initial guess inheriting poor values for the Jacobian map. Nested iteration tends to ameliorate this potential difficulty, so we may focus on later Newton iterations, where these results suggest that two $V(1,1)$-cycles yield overall convergence factors of about 0.2. We use two $V(1,1)$-cycles in the tests that follow.

Figure 4.2 depicts Newton convergence results for grids $h = \{\frac{1}{16}, \frac{1}{32}, \frac{1}{64}, \frac{1}{128}\}$. We study performance in terms of both the functional error measure (i.e., square root of the functional) and the relative H^1 errors in $\mathbf{J}$,

$$e_{\mathbf{J}} \equiv \frac{\|\mathbf{J}^* - \mathbf{J}_n\|_{1,\Omega}}{\sqrt{\|\mathbf{J}^*\|_{1,\Omega}\|\mathbf{J}_n\|_{1,\Omega}}}.$$

The graphs show the differences between the values at the current and sixth Newton steps. We are interested in the functional measure because it is equivalent to the H^1

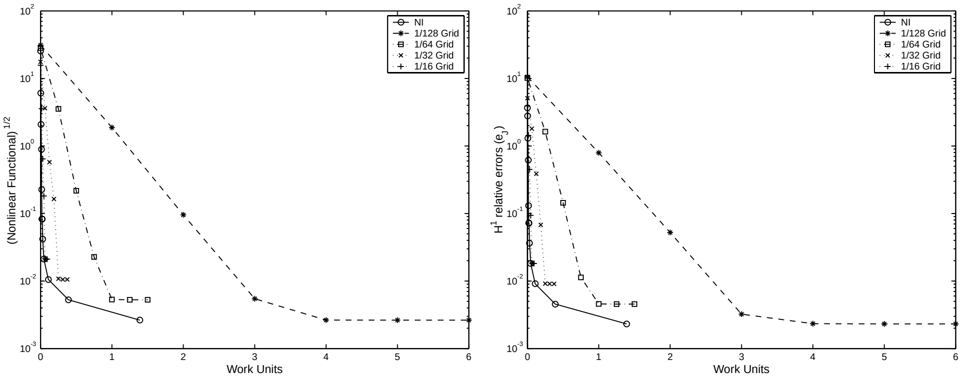


FIG. 4.3. *First-stage functional. Functional and H^1 error measures for standard Newton iterations and NI with imposed boundary conditions. One work unit is the equivalent of one step on the $h = \frac{1}{128}$ grid using two $V(1,1)$-cycles.*

norm of the errors, as we established theoretically in [10] and as these graphs suggest. The left-hand graph contains the functional values, and the right-hand graph contains the errors. Convergence appears to be approximately linear, which is consistent with the theoretical result. The factors also appear to be bounded independent of grid size.

Figure 4.3 compares standard Newton and NI results. We again report on the functional and relative H^1 error measures in $\mathbf{J}$. For proper comparison of cost, we now base the data on a *work unit*, defined to be the equivalent of one Newton step on the $h = \frac{1}{128}$ grid. (One Newton step has two V(1,1)-cycles.) We thus count one Newton step on the $h = \frac{1}{64}$ grid as $\frac{1}{4}$ of a work unit, $\frac{1}{16}$ on the next coarser grid, and so on. After about the sixth standard Newton step for each of the grid sizes, the change in the functional value (and the H^1 error) at each iteration is very small relative to the functional value itself. The exact solution is only approximated by the finite-dimensional subspace. Thus, while the functional value for the exact solution is zero, the minimum on the finite-dimensional subspace is not. With more Newton steps, we can thus get as close as we choose to the finite-dimensional approximation of the exact solution, but the decrease in the functional and, hence, in the error, stalls because discretization error is reached. The ratios of the functional and the relative H^1 error measures in $\mathbf{J}$ are about 1.16 near the solution for grids $h = \{\frac{1}{16}, \frac{1}{32}, \frac{1}{64}\}$ and 1.14 for grid $h = \frac{1}{128}$. After the third Newton step, this ratio is a constant for all grid sizes, which affirms H^1 equivalence.

Next we study performance of the single-stage algorithm, with the same map and initial guess. Again, we report on functional and relative H^1 error measures in $\mathbf{J}$. Figure 4.4 contains graphs of differences between these values at the current and sixth Newton steps. Consistent with the theory, convergence using this measure appears to be approximately linear, with factors bounded independent of grid size.

Figure 4.5 compares standard Newton iterations with NI based on work units as defined above. Behavior of the errors for the single stage is essentially the same as for the first stage. The ratios of functional measures of the single stage to the first stage varies between 0.87 and 1.12, fixing at 1.09 after Newton step 4 for each grid. For NI, the ratio is 1.09 for all the finer grids.

The relative error in the computed solution does not appear to vary with grid size because the variation is small compared to the error. Again, on any finite-dimensional subspace, we cannot expect to reduce the error to zero because of discretization error.

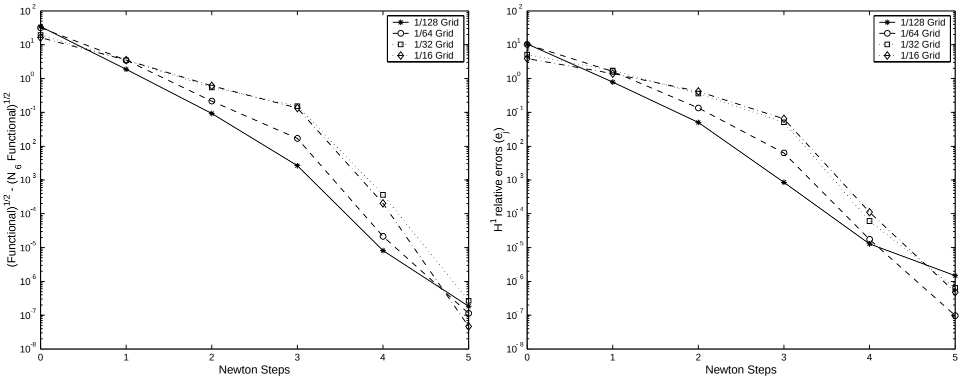


FIG. 4.4. *Newton convergence, using standard Newton iterations with imposed boundary conditions, in both functional (left) and H^1 error (right) measures. Differences between the value at the current and sixth Newton steps are plotted.*

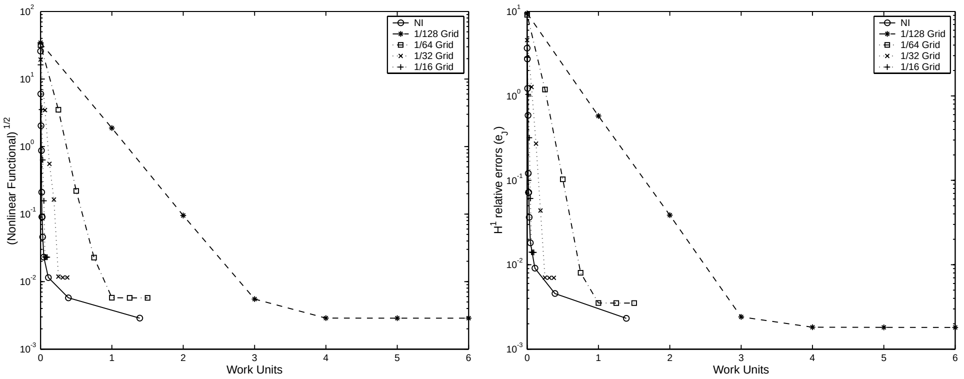


FIG. 4.5. *Newton versus NI results, with imposed boundary conditions, in both functional (left) and H^1 error (right) measures.*

The ratios of the functional and H^1 error measures in $\mathbf{J}$ are about 1.3 near the solution for grids $h = \{\frac{1}{16}, \frac{1}{32}, \frac{1}{64}\}$ and 1.2 for grid $h = \frac{1}{128}$. After the third Newton step, this ratio is a constant for all grid sizes, which affirms H^1 equivalence. We need at least 4 standard Newton steps to reduce the functional to about the same level for which NI needed an equivalent of only about 1.5 steps. We expect this difference to widen for larger problems, where the required steps for standard Newton would tend to grow, while NI would probably remain below an equivalent of two.

4.2. V-cycle tests. To determine which V(q,s)-cycle is most efficient, we study asymptotic convergence factors with $q + s \leq 3$ and $h = \frac{1}{64}$. Here we linearize the equations about the exact solution, set the right-hand side to zero, start with a random initial guess, and then observe residual reduction factors after many V-cycles. Table 4.2 shows the observed V-cycle convergence factors for different values of a. The cycles with more relaxation sweeps naturally have better convergence factors but involve more computation. We thus consider a measure of the time required to reduce the initial residual by a factor of 10. Since we are interested only in comparisons, we choose the *relative* measure $t \equiv (q+s+c)\ln(0.1)/\ln(r)$, where r is the observed asymptotic convergence factor for the V(q,s)-cycle and c estimates the fixed cost of a cycle. We choose $c = 2$ because of residual calculations and intergrid transfers. Observed

Table 4.2
Asymptotic convergence factors for different V-cycles and values of a; grid size $h = \frac{1}{64}$.

a	V(0,1)	V(1,0)	V(0,2)	V(1,1)	V(2,0)	V(0,3)	V(1,2)	V(2,1)	V(3,0)
0.0	0.33	0.23	0.20	0.13	0.13	0.16	0.11	0.11	0.11
0.1	0.33	0.25	0.21	0.14	0.14	0.16	0.12	0.12	0.12
0.4	0.38	0.30	0.26	0.18	0.19	0.21	0.15	0.15	0.16
0.7	0.47	0.39	0.34	0.26	0.25	0.28	0.22	0.22	0.22
1.0	0.60	0.53	0.45	0.40	0.39	0.37	0.32	0.32	0.28

Table 4.3
Relative time to reduce residual by a factor of 10 *with various a for different V-cycles and $h = \frac{1}{64}$.*

a	V(0,1)	V(1,0)	V(0,2)	V(1,1)	V(2,0)	V(0,3)	V(1,2)	V(2,1)	V(3,0)
0.0	6.2	4.7	5.7	4.5	4.5	6.3	5.2	5.2	5.2
0.1	6.2	5.0	5.9	4.7	4.7	6.3	5.4	5.4	5.4
0.4	7.1	5.7	6.8	5.4	5.5	7.4	6.1	6.1	6.3
0.7	9.1	7.3	8.5	6.8	6.6	9.0	7.6	7.6	7.6
1.0	13.5	10.9	11.5	10.1	9.8	11.6	10.1	10.1	9.0

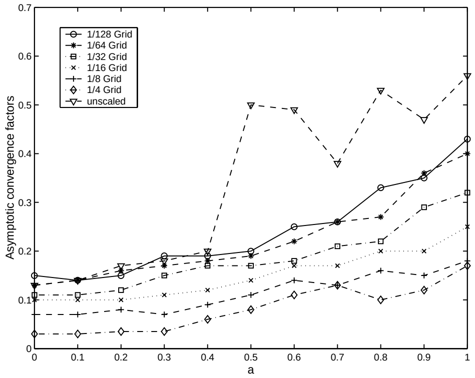


Fig. 4.6. *Asymptotic convergence factors for various grid sizes and values of a. The scaled functional was used in all but the test for $h = \frac{1}{64}$.*

values for t for the $h = \frac{1}{64}$ grid and different values of a are given in Table 4.3. While performance of the V(1,1)- or V(2,0)-cycles were similar, we chose the V(1,1)-cycle for the remainder of our tests.

4.3. AMG tests. We next test the performance of the linear solver with varying h. Again, the equations are linearized about the solution and the right-hand side is set to zero. We study the deterioration in asymptotic convergence factors as a increases from zero to one; the results are plotted in Figure 4.6. In all but one test, we used the scaled functional in (4.1) with the boundary conditions enforced so that the boundary terms vanish. For the test marked “unscaled” and for which $h = \frac{1}{64}$, factor $\frac{1}{J_n}$ in the second term of the functional in (4.1) was omitted. For the scaled functional, asymptotic convergence factors increase as a increases, as expected. At $a = 0$, which corresponds to the identity map, convergence factors are similar to those for the Laplace problem. There is some variation with respect to the grid size, although for smaller h $\left(\frac{1}{64}\text{ and }\frac{1}{128}\right)$ the factors are similar. The results for the unscaled and scaled convergence factors for the $h = \frac{1}{64}$ grid confirm that scaling the second term of the functional in (4.1) by $\frac{1}{J_n}$ significantly improves convergence factors: the unscaled factors are significantly larger than the scaled factors for $a \geq 0.5$.

TABLE 4.4
Asymptotic convergence factors for the $V(1,1)$-cycle, with varying grid size and Newton iterations, for the arch.

Newton	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{64}$	$\frac{1}{128}$
1	0.59	0.50	0.43	0.59
2	0.63	0.65	0.67	0.61
3	0.66	0.67	0.68	0.67
4	0.63	0.67	0.66	0.67
5	0.66	0.68	0.66	0.68

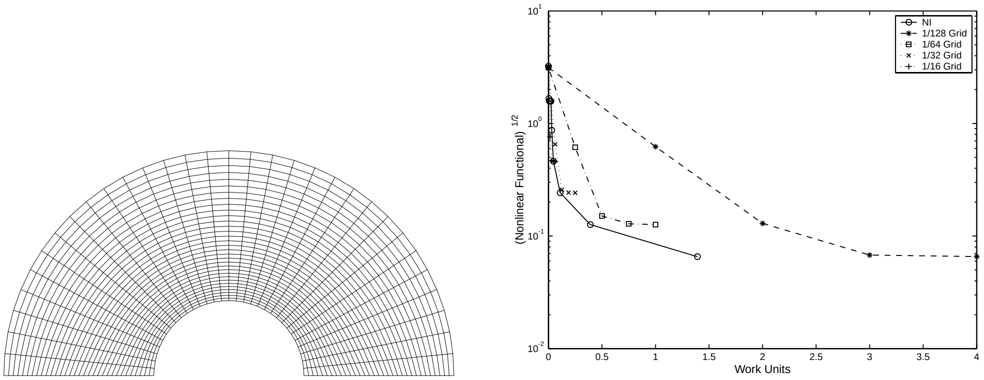


FIG. 4.7. *NI and standard Newton methods for the arch.*

4.4. Nested iteration for the arch. We compare standard Newton iterations for $h = \frac{1}{128}, \frac{1}{64}, \frac{1}{32}$, and $\frac{1}{16}$ to NI with $h = \frac{1}{4}$ for the coarsest grid and $h = \frac{1}{128}$ for the finest grid. The initial guess for the arch is

$$\begin{aligned} x &= 1.5 + (1.5 - \xi)\cos(\pi(1-\eta)), & y &= (1.5 - \xi)\sin(\pi(1-\eta)), \\ J_{11} &= -\cos(\pi(1-\eta)), & J_{21} &= \pi(1.5 - \xi)\sin(\pi(1-\eta)), \\ J_{12} &= -\sin(\pi(1-\eta)), & J_{22} &= -\pi(1.5 - \xi)\cos(\pi(1-\eta)). \end{aligned}$$

Choices of the numbers of V(1,1)-cycles per iteration and Newton steps on each grid are currently made by observation. In the theoretical section of [10], we suggested $\rho^{\nu_0} \leq \frac{1}{8}$ as a criterion, where ρ is the convergence factor and ν_0 is the number of V-cycles. A significantly larger value would allow the iterates to wander too far from the true solution as the grid was refined. We could choose a smaller value for ρ^{ν_0} so that the multigrid solutions would shadow the exact finite-dimensional solutions more closely. But too small a value would likely be less efficient than simply proceeding to finer meshes. NI required significantly less work to obtain the same discretization error than did the standard Newton method. Standard Newton needed just a few steps to reach discretization error for our tests anyway, but the savings afforded by NI for smaller h should be much larger still. More results can be found in [9].

Table 4.4 depicts asymptotic convergence factors for standard Newton iterations. These factors are not small enough to allow just one V-cycle per Newton step. We thus used three V(1,1)-cycles to solve each Newton step. Thus one work unit is three V(1,1)-cycles on the $\frac{1}{128}$ grid. Here we performed two coarsest-grid Newton iterations, with only one on all finer grids.

Three standard Newton steps were required to reach discretization error, while NI required less than one-and-a-half equivalents (see Figure 4.7). The final functional value decreases by about a factor of four as the grid size is halved, which confirms the $\mathcal{O}(h)$ approximation in the $H^1(\Omega)$ norm.

5. Conclusion. We showed theoretically that the nested iteration process involving only one discrete Newton step on each level produces a result on the finest level that is within discretization error of the exact solution. We also showed this result numerically using an $H^{1+\delta}(\Omega)$ discrete space for each of the unknowns. Future directions involve automating the numerical tests to include the following choices: number of relaxations before and after coarsening, number of V-cycles, number of Newton steps on each grid, size and choice of solvers for the coarsest grid, parameterization of the boundary maps, and adaptive mesh refinement.

The first three choices dictate the overall efficiency of the algorithm and should be considered carefully for maximum effectiveness. Automation would require heuristics to sense performance of smoothing and coarse-grid correction, as well as linearization trade-offs. We used one Newton step on all but the coarsest grid in our examples and theory, but severely distorted regions may dictate more such steps to improve effectiveness, and possibly other continuation methods to address the Newton method's local convergence characteristics. In any case, the special ability of the FOSLS functional to signal errors could be exploited to make these choices in an effective and automatic way. The fourth coarsest-grid choice rests heavily on the geometry of the particular map. Complex regions may require a fairly small coarsest grid and a significant amount of effort to solve the nonlinear problem there. Damped Newton methods and various forms of continuation techniques may come into play. Of course, complicated regions generally require very fine meshes to supply meaningful simulations, so the relative cost of such coarsest-grid effort may again be fairly minimal. Moreover, the special properties of the FOSLS functional may also be exploited for these choices. The fifth choice would be to use a parameterization of the boundary in the associated terms of the functional that would allow concentration of grid points near special boundary features. The final choice of adaptive mesh refinement can be served by noting that the functional value on each element is a sharp measure of the error on that element, which makes it suitable as a measure for determining which elements need to be further subdivided (cf. [4]).

REFERENCES

[1] R.A. Adams, *Sobolev Spaces*, Pure Appl. Math. 65, Academic Press, New York, 1975.

[2] S. Agmon, A. Douglis, and L. Nirenberg, *Estimates near the boundary for solutions of elliptic partial differential equations satisfying general boundary conditions* II, Comm. Pure Appl. Math., 17 (1963), pp. 35–92.

[3] A.K. Aziz, *The Mathematical Foundations of the Finite Element Method with Application to Partial Differential Equations*, Academic Press, New York, 1972.

[4] M. Berndt, T.A. Manteuffel, and S.F. McCormick, *Local error estimates and adaptive refinement for first-order system least-squares (FOSLS)*, Electron. Trans. Numer. Anal., 6 (1997), pp. 35–43.

[5] D. Braess, *Finite Elements Theory, Fast Solvers, and Applications in Solid Mechanics*, Cambridge University Press, Cambridge, UK, 1997.

[6] S.C. Brenner and L.R. Scott, *The Mathematical Theory of Finite Element Methods*, Texts Appl. Math. 15, Springer-Verlag, New York, 1994.

[7] Z. Cai, T.A. Manteuffel, and S.F. McCormick, *First-order system least squares for second-order partial differential equations: Part* II, SIAM J. Numer. Anal., 34 (1997), pp. 425–454.

[8] J.E. Castillo, ed., *Mathematical Aspects of Numerical Grid Generation*, Frontiers Appl. Math. 8, SIAM, Philadelphia, 1991.

[9] A.L. Codd, *Elasticity-Fluid Coupled Systems and Elliptic Grid Generation (EGG) based on First-Order System Least Squares (FOSLS)*, Ph.D. thesis, Department of Applied Mathematics, University of Colorado at Boulder, Boulder, CO, 2001.

[10] A.L. Codd, T.A. Manteuffel, and S.F. McCormick, *Multilevel first-order system least squares for nonlinear elliptic partial differential equations*, SIAM J. Numer. Anal., 41 (2003), pp. 2197–2209.

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

[12] P. Knupp and S. Steinberg, *Fundamentals of Grid Generation*, CRC Press, Boca Raton, FL, 1993.

[13] B. Lee, T.A. Manteuffel, S.F. McCormick, and J. Ruge, *First-order system least-squares for the Helmholtz equation*, SIAM J. Sci. Comput., 21 (2000), pp. 1927–1949.

[14] G. Liao, *On harmonic maps*, in Mathematical Aspects of Numerical Grid Generation, Frontiers Appl. Math. 8, J.E. Castillo, ed., SIAM, Philadelphia, 1991, pp. 123–130.

[15] J.L. Lions and E. Magenes, *Non-Homogeneous Boundary Value Problems and Applications*, Vol. 1, Die Grundlehren der mathematischen Wissenschaften in Einzeldarstellungen 181, Springer-Verlag, Berlin, English edition, 1972.

[16] T. Rado, *Aufgabe* 41, Jahresber. Deutsch. Math.-Verein., 35 (1926), p. 49.

[17] Ja. A. Roitberg and Z.G. Seftel, *A theorem on homeomorphisms for elliptic systems and its applications*, Math. USSR-Sbornik, 7 (1969), pp. 439–465.

[18] Ya. A. Roitberg, *A theorem about the complete set of isomorphisms for systems elliptic in the sense of Douglis and Nirenberg*, Ukrainian Math. J., 25 (1973), pp. 396–405.

[19] M. Schechter, *Solution of Dirichlet problem for systems not necessarily strongly elliptic*, Comm. Pure Appl. Math., 12 (1959), pp. 241–247.

[20] J.F. Thompson, Z.U.Z. Warsi, and C.W. Mastin, *Numerical Grid Generation*, North–Holland, New York, 1985.

[21] J. Wloka, *Partial Differential Equations*, Cambridge University Press, Cambridge, UK, 1987.

[22] K. Yosida, *Functional Analysis*, 6th ed., Springer-Verlag, Berlin, 1965.

SIAM J. Numer. Anal.
Vol. 41, No. 6, pp. 2233–2261

NUMERICAL DISCRETIZATION OF BOUNDARY CONDITIONS FOR FIRST ORDER HAMILTON–JACOBI EQUATIONS[*]

RÉMI ABGRALL[†]

Abstract. We provide two simple ways of discretizing a large class of boundary conditions for first order Hamilton–Jacobi equations. We show the convergence of the numerical scheme under mild assumptions. However, many types of such boundary conditions can be written in this way. Some provide "good" numerical results (i.e., without boundary layers), whereas others do not. To select a good one, we first give some general results for monotone schemes which mimic the maximum principle of the continuous case, and then we show in particular cases that no boundary layer can exist. Some numerical applications illustrate the method. An extension to a geophysical problem is also considered.

Key words. Hamilton–Jacobi equations, approximation of boundary conditions

AMS subject classifications. 65M60, 65N99, 35L60, 65Z05, 35Q58

DOI. 10.1137/S0036142998345980

1. Introduction.

The problem of discretizing first order Hamilton–Jacobi equations in $\mathbb{R}^N$ has been considered by several authors (see, e.g., [8, 9, 3]) on various types of meshes (see the previous references and [1]). However, in our knowledge, the discretization of boundary conditions has not yet been considered in a systematic way. The aim of this paper is to provide a simple and systematic way of discretizing a wide variety of boundary conditions. This is done in the framework of discontinuous viscosity solutions [4]. More precisely, we consider the following problem:

$$\begin{cases} H(x,u,Du)=0, & x\in\Omega,\\ F(x,u,Du)=0, & x\in\partial\Omega, \end{cases} \tag{1.1}$$

where the Hamiltonian is continuous on $\overline{\Omega}\times\mathbb{R}\times\mathbb{R}^N$ and the boundary condition F is continuous on $\partial\Omega\times\mathbb{R}\times\mathbb{R}^N$.

For any function z, we consider the upper semicontinuous (u.s.c) and lower semicontinuous (l.s.c) envelopes of z with respect to all variables. These are defined by

$$z^*(x)=\limsup_{x\to y} z(y) \quad\text{and}\quad z_*(x)=\liminf_{x\to y} z(y).$$

Following [4], we introduce the function G:

$$G(x,u,p)=\begin{cases} H(x,u,p), & x\in\Omega,\\ F(x,u,p), & x\in\partial\Omega. \end{cases}$$

A locally bounded u.s.c function u defined on $\overline{\Omega}$ is a viscosity subsolution of (1.1) if and only if, for any $\phi\in C^1(\overline{\Omega})$, if $x_0\in\overline{\Omega}$ is a local maximum of $u-\phi$, then

$$G_*(x_0,u(x_0),D\phi(x_0))\le 0. \tag{1.2}$$

[*] Received by the editors October 12, 1998; accepted for publication (in revised form) March 31, 2003; published electronically December 5, 2003. This research was supported in part by D.O.E grant DE–FG02–88ER25053 when the author was at the Courant Institute, on leave from INRIA Sophia Antipolis, then during a visit at RIACS, NASA Ames Research Center, Moffett Field, CA.

http://www.siam.org/journals/sinum/41-6/34598.html

[†] Mathématiques Appliquées de Bordeaux, Université Bordeaux I, 351 Cours de la Libération, 33 405 Talence cedex, France, and Institut Universitaire de France (abgrall@math.u-bordeaux.fr).

Similarly, u, a locally bounded l.s.c. function defined on $\overline{\Omega}$, is a viscosity supersolution of (1.1) if and only if, for any $\phi \in C^1(\overline{\Omega})$, if $x_0 \in \overline{\Omega}$ is a local minimum of $u - \phi$, then

$$G^*(x_0, u(x_0), D\phi(x_0)) \geq 0. \tag{1.3}$$

The computation of G_* and G^* is easy, and we have

$$\begin{cases} G_*(x,u,p) = G^*(x,u,p) = H(x,u,p) & \text{if } x \in \Omega, \\ G_*(x,u,p) = \min(H(x,u,p), F(x,u,p)) & \text{if } x \in \partial\Omega, \\ G^*(x,u,p) = \max(H(x,u,p), F(x,u,p)) & \text{if } x \in \partial\Omega. \end{cases} \tag{1.4}$$

More specifically, we consider the cases of the Dirichlet and Neumann boundary conditions, but the results of this paper may extend to more general boundary conditions, provided they are of the form (1.1) and if some regularity on F is assumed. In the case of Dirichlet boundary conditions, namely $u = \varphi$, we have

$$F(x,u,p) = u(x) - \varphi(x), \tag{1.5}$$

and for Neumann boundary conditions we have

$$F(x,u,p) = \frac{\partial u}{\partial n} - g(x), \tag{1.6}$$

where g is defined on $\partial\Omega$ and continuous.

This paper is organized as follows. We first recall a convergence result by Barles and Souganidis [5]. Then, starting from the dynamical programming principle, we indicate a way of discretizing general boundary conditions, and show the convergence of this scheme. In a second part, we describe several particular cases for convex and nonconvex Hamiltonians. A particular emphasis is set on the Dirichlet boundary conditions because it is more difficult to provide effective boundary conditions in that case, at least more difficult than for Neumann conditions. This problem is explained and has many similarities with the technical difficulties encountered in the study of these conditions in the continuous case. We provide numerical illustrations that show the effectiveness of the schemes. It is known that first order Hamilton–Jacobi equations have many similarities with a particular class of hyperbolic systems. Because of that, one might think that boundary conditions built on the structure of inflow and outflow characteristics would be efficient enough. This is true if the structure of the solution is known a priori. This is rarely the case in practice, and we provide an example where the structure of the solution at the boundary is not known a priori, so more sophisticated approximations are required. Another example has its origin in seismology problems.

Throughout the paper, we consider an open and bounded domain Ω. To simplify the presentation, we assume $\Omega \subset \mathbb{R}^2$, but our results are also valid for $\mathbb{R}^N$, $N \geq 2$. The open set Ω is discretized by a triangulation $\mathcal{T}_\rho$. The nodes of the mesh are denoted by $x_i, i = 1, \dots, n_s$; the triangles are denoted by $T_k, k = 1, \dots, n_T$. The vertices of T are denoted by $x_{i_k}, k = 1, \dots, 3$. The parameter ρ above is, for example, the largest radius of the circumscribed circles of T_k, $k = 1, \dots, n_T$.

2. A convergence result.

2.1. Preliminaries. All our results rely on the following one by Barles and Souganidis [5]. The symbol $B(\overline{\Omega})$ denotes the set of bounded functions over $\overline{\Omega}$.

They consider approximations schemes of the form

$$S(\rho, x, u^\rho(x), u^\rho) = 0 \qquad \text{in } \overline{\Omega}, \tag{2.1}$$

where S maps $\mathbb{R}^+ \times \overline{\Omega} \times \mathbb{R} \times B(\overline{\Omega})$ onto $\mathbb{R}$, is locally bounded, and has the following properties:

1. monotonicity: if $u \geq v$, for all $\rho \geq 0$, $x \in \overline{\Omega}$, $t \in \mathbb{R}$, and $u, v \in L^\infty(\overline{\Omega})$ we have

$$S(\rho, x, t, u) \leq S(\rho, x, t, v); \tag{2.2}$$

2. stability: for all $\rho > 0$ there exists a solution $u^\rho \in L^\infty(\overline{\Omega})$ to (2.1) with a bound independent of ρ;
3. consistency: for all $x \in \overline{\Omega}$ and $\phi \in C_b^\infty(\overline{\Omega})$ (the set of C^∞ bounded functions),

$$\limsup_{\rho \to 0, y \to x, \xi \to 0} S(\rho, y, \phi(y) + \xi, \phi + \xi) \leq G^*(x, \phi(x), D\phi(x)) \tag{2.3}$$

and

$$\liminf_{\rho \to 0, y \to x, \xi \to 0} S(\rho, y, \phi(y) + \xi, \phi + \xi) \geq G_*(x, \phi(x), D\phi(x)); \tag{2.4}$$

4. strong uniqueness principle: if $u \in L^\infty(\overline{\Omega})$ is an u.s.c subsolution of (1.1) and $v \in L^\infty(\overline{\Omega})$ is an l.s.c supersolution of (1.1), then $u \leq v$ on $\overline{\Omega}$.

THEOREM 2.1 (from Barles and Souganidis). *Assuming the monotonicity, consistency, and stability of the scheme* (2.1) *and the strong uniqueness property of the problem* (1.1), *then the solution* u^ρ *of* (2.1) *converges locally uniformly to the unique continuous viscosity solution of* (1.1).

The stability, (2.2), (2.3), and (2.4) imply that the functions

$$\bar{u} = \limsup_{\rho \to 0, y \to x} u^\rho(y) \qquad \text{and} \qquad \underline{u} = \liminf_{\rho \to 0, y \to x} u^\rho(y)$$

are defined on $\overline{\Omega}$; they are, respectively, u.s.c. subsolutions and l.s.c. supersolutions of (1.1). By definition, we have $\underline{u} \leq \overline{u}$. The opposite inequality follows from the uniqueness property. Note that if we have only this uniqueness property on Ω, as is the case for Dirichlet boundary conditions, the same argument shows that $\underline{u} = \overline{u}$ on Ω.

2.2. Two numerical schemes. We consider a bounded open domain Ω that is discretized by means of a triangulation $\mathcal{T}_\rho$. The parameter ρ is the maximum, on the elements T of $\mathcal{T}_\rho$, of the radius of the smallest disk containing T.

We consider a scheme for $H(x, u, Du) = 0$ that is defined for any point of the mesh except perhaps for the boundary nodes. It is written as

$$S_H(\rho, x, u^\rho(x), u^\rho) = 0. \tag{2.5}$$

We also consider an approximation of the boundary conditions that is defined for any node of the triangulation on the boundary of Ω,

$$S_F(\rho, x, u^\rho(x), u^\rho) = 0. \tag{2.6}$$

Let $(x, t, p) \mapsto H_b(x, t, p)$ be a Hamiltonian defined at least in a neighborhood of $\partial\Omega \times \mathbb{R} \times \mathbb{R}^N$. It fulfills the same assumptions as H. We also have a numerical scheme S_{H_b} for the Hamiltonian H_b. We define the following scheme for (1.1):

$$0 = S(\rho, x, u^\rho(x), u^\rho) = \begin{cases} S_H(\rho, x, u^\rho(x), u^\rho) & \text{if } x \in \Omega, \\ \max(S_{H_b}(\rho, x, u^\rho(x), u^\rho), S_F(\rho, x, u^\rho(x), u^\rho)) & \text{if } x \in \partial\Omega. \end{cases} \tag{2.7}$$

Another a priori reasonable scheme could also be

$$0 = S(\rho, x, u^\rho(x), u^\rho) = \begin{cases} S_H(\rho, x, u^\rho(x), u^\rho) & \text{if } x \in \Omega, \\ \min(S_{H_b}(\rho, x, u^\rho(x), u^\rho), S_F(\rho, x, u^\rho(x), u^\rho)) & \text{if } x \in \partial\Omega. \end{cases} \tag{2.8}$$

The questions are the following: On which conditions can the scheme (2.7) or (2.8) be considered as a good numerical approximation of (1.1)? Can we identify criteria for preferring scheme (2.7) to (2.8)?

Before giving conditions that ensure the convergence of the schemes (2.7) and (2.8), we motivate the "max" condition of (2.7) in the case of a convex Hamiltonian. The justification comes from the dynamical programming principle and is therefore valid for convex Hamiltonians. We could make the same type of justification for the "min" condition of (2.8) for concave Hamiltonians.

Dynamical programming principle. We assume that the Hamiltonian is given by

$$H(x, u, p) = \sup_{v \in V} \left\{ -b(x, v) \,.\, p + \lambda u - f(x, v) \right\},$$

where the space of controls V is compact, and we have standard assumptions on b and f. We also assume $\lambda > 0$. For the Dirichlet condition (1.5), the solution of (1.1) is given by, for any $T > 0$,

$$0 = u(x) - \inf_{v(.)} \left[\int_0^{\min(T,\tau)} f(y_x(t), v(t)) e^{-\lambda t} dt + 1_{\{T<\tau\}} u(y_x(T)) e^{-\lambda T} + 1_{\{T \geq \tau\}} \varphi(y_x(\tau)) e^{-\lambda \tau} \right]. \tag{2.9}$$

As usual, the trajectory $y_x(.)$ satisfies $y_x(0) = x \in \Omega$ and

$$\frac{d}{dt} y_x(t) = b(y_x(t), v(t)) \quad \text{for } t > 0.$$

The exit time τ is

$$\tau = \inf\{t \geq 0, y_x(t) \notin \Omega\}.$$

Now, the set of controls can be split into two parts: the set V_1 for which $T < \tau$, and V_2 for which $T \geq \tau$. Hence,

$$u(x) = \min \left(\inf_{v \in V_1} [\cdots], \inf_{v \in V_2} [\cdots] \right).$$

Let $\vec{n}$ be the interior normal to Ω at $x \in \overline{\Omega}$. Since T is arbitrary, it can be chosen as small as possible. In the limit $T \to 0$, the set V_1 would be the set of controls for which

$b(x,v)\,.\,\vec{n} > 0$, i.e., the control for which the trajectory goes into Ω. The dynamical programming principle $\inf_{v\in V_1}[\cdots] - u(x) = 0$ corresponds to the Hamiltonian

$$H_b(x,t,p) = \sup_{v\in V_1}\{b(x,v)\,.\,p + \lambda t - f(x,v)\}.$$

We also have the relation $H_b \leq H$.

The "inf" on V_2 can be approximated, if T is small, by $\varphi(y_x(\tau))$. Since $T \leq \tau$ and if we can choose controls for which $T \simeq \tau$, we get

$$\varphi(y_x(\tau)) \simeq \varphi(x)$$

because φ is continuous. Thus, by setting $S_F = u(x) - \varphi(x)$, we see that (2.9) can be approximated by

$$0 = \max(S_{H_b}, S_F),$$

which is want we wanted. We have the following result.

THEOREM 2.2. *Assume that*

1. $H_b \leq H$*;*
2. S_H*,* S_{H_b}*, and* S_F *are monotone and stable;*
3. *for all* $\phi \in C_b^\infty(\overline{\Omega})$*, we have*

$$\begin{aligned}
&\textit{for any } x \in \overline{\Omega},\\
&\lim_{\rho\to 0, y\to x, \xi\to 0} S_H(\rho, y, \varphi(y)+\xi, \varphi+\xi) = H(x, \varphi(x), D\varphi(x)),\\
&\textit{for any } x \textit{ in a neighborhood of } \partial\Omega,\\
&\lim_{\rho\to 0, y\to x, \xi\to 0} S_{H_b}(\rho, y, \varphi(y)+\xi, \varphi+\xi) = H_b(x, \varphi(x), D\varphi(x)),\\
&\textit{for any } x \in \partial\Omega,\\
&\lim_{\rho\to 0, y\to x, \xi\to 0} S_F(\rho, y, \varphi(y)+\xi, \varphi+\xi) = F(x, \varphi(x), D\varphi(x));
\end{aligned}$$

4. *the equation* (1.1) *has a uniqueness principle.*

Then the family u^ρ *defined by* (2.7) *converges locally uniformly to the solution of* (1.1) *in* Ω*. We have the same result for* (2.8)*, provided that the condition* 1 *is replaced by* $H \leq H_b$*.*

Proof. We make the proof for the scheme (2.7). The proof for (2.8) is similar.

We first note that, on the boundary,

$$\begin{aligned}
&\limsup_{\rho\to 0, y\to x, \xi\to 0} S(\rho, y, \varphi(y)+\xi, \varphi+\xi)\\
&\qquad = \max\big(H(x,\varphi(x),D\varphi(x)), \max(H_b(x,\varphi(x),D\varphi(x)), F(x,\varphi(x),D\varphi(x)))\big),
\end{aligned}$$

$$\begin{aligned}
&\liminf_{\rho\to 0, y\to x, \xi\to 0} S(\rho, y, \varphi(y)+\xi, \varphi+\xi)\\
&\qquad = \min\big(H(x,\varphi(x),D\varphi(x)), \max(H_b(x,\varphi(x),D\varphi(x)), F(x,\varphi(x),D\varphi(x)))\big),
\end{aligned}$$

while in the interior points,

$$\lim_{\rho\to 0, y\to x\xi\to 0} S(\rho, y, \varphi(y)+\xi, \varphi+\xi) = H(x,\varphi(x),D\varphi(x)). \tag{2.10}$$

Then we proceed as in [4]. We define

$$\overline{u}(x) = \limsup_{y\to x, \rho\to 0} u^\rho(y) \quad \text{and} \quad \underline{u}(x) = \liminf_{y\to x, \rho\to 0} u^\rho(y).$$

They are defined on $\overline{\Omega}$ because u^ρ has bounds independent of ρ. We will show now that the functions $\overline{u}$ and $\underline{u}$ are, respectively, sub- and supersolutions of (1.1). In fact, we show first that if $x_0 \in \partial\Omega$ is a local minimum of $\underline{u} - \phi$, then

$$\max\big(H(x_0, \underline{u}(x_0), D\varphi(x_0)), \max(H_b(x_0, \underline{u}(x_0), D\varphi(x_0)), F(x_0, \underline{u}(x_0), D\varphi(x_0)))\big) \geq 0, \tag{2.11}$$

while if $x_0 \in \partial\Omega$ is a local maximum of $\overline{u} - \phi$ for some $\phi \in C^b_\infty(\overline{\Omega})$, then

$$\min\big(H(x_0, \overline{u}(x_0), D\varphi(x_0)), \max(H_b(x_0, \underline{u}(x_0), D\varphi(x_0)), F(x_0, \overline{u}(x_0), D\varphi(x_0)))\big) \leq 0. \tag{2.12}$$

To show (2.11), we repeat Barles and Souganidis's arguments. Equation (2.12) is obtained in the same way. We may assume that x_0 is a strict minimum, $\underline{u}(x_0) = \phi(x_0)$, and $\phi \leq 2\inf_\rho ||u^\rho||_\infty$ outside of $B(x_0, r)$, where r is such that

$$\underline{u}(x) - \phi(x) \geq \underline{u}(x_0) - \phi(x_0) = 0 \quad \text{in } B(x_0, r).$$

There exist sequences ρ_n and $y_n \in \overline{\Omega}$ such that $n \to +\infty$, $\rho_n \to 0$, $y_n \to x_0$, $u^{\rho_n}(y_n) \to \underline{u}(x_0)$, and y_n is a global minimum of $u^{\rho_n} - \phi$. We denote by ξ_n the quantity $u^{\rho_n}(y_n) - \phi(y_n)$. We have $\xi_n \to 0$ and $u^{\rho_n}(y) \geq \phi(y) + \xi_n$ in $B(x_0, r)$. Since S is monotone, we get

$$0 \leq \limsup_n S(\rho_n, y_n, \phi(y_n) + \xi_n, \phi + \xi_n) \leq \limsup_{\rho\to 0, y\to x_0, \xi\to 0} S(\rho, y, \varphi(y) + \xi, \varphi + \xi)$$

$$= \max(H(x_0, \varphi(x_0), D\varphi(x_0)), \max(H_b(x_0, \varphi(x_0), D\varphi(x_0)), F(x, \varphi(x_0), D\varphi(x_0)))).$$

If $x_0 \in \Omega$ is a local maximum (resp., minimum) of $\overline{u} - \phi$ (resp., $\underline{u} - \phi$), we use (2.10) and the same arguments as above to get

$$H(x_0, \overline{u}(x_0), D\phi(x_0)) \leq 0 \qquad (\text{resp., } H(x_0, \underline{u}(x_0), D\phi(x_0)) \geq 0). \tag{2.13}$$

Now we have to check that the condition (2.12) (resp., (2.11)) implies the supersolution (resp., subsolution) condition.

- *Inequality* (2.12). If $F(x_0, \underline{u}(x_0), D\phi(x_0)) \leq 0$, there is nothing to prove. We assume $F(x_0, \underline{u}(x_0), D\phi(x_0)) > 0$. We have either

  $$H(x_0, \underline{u}(x_0), D\varphi(x_0)) \leq 0 \tag{2.14}$$

  or

  $$\max(H_b(x_0, \underline{u}(x_0), D\phi(x_0)), F(x_0, \underline{u}(x_0), D\phi(x_0))) \leq 0.$$

  In the second case, we necessarily have (2.14), and in both cases the inequality holds.

- *Inequality* (2.11). If $F(x_0, \overline{u}(x_0), D\phi(x_0)) \geq 0$, there is nothing to prove. If we assume $F(x_0, \overline{u}(x_0), D\phi(x_0)) < 0$, then we must have either $H(x_0, \overline{u}(x_0), D\phi(x_0)) \geq 0$ or

$$\max(H_b(x_0, \overline{u}(x_0), D\phi(x_0)), F(x_0, \overline{u}(x_0), D\phi(x_0))) \geq 0.$$

Since $F < 0$, this inequality implies $H_b \geq 0$, so that

$$H(x_0, \overline{u}(x_0), D\phi(x_0)) \geq H_b(x_0, \overline{u}(x_0), D\phi(x_0)) \geq 0.$$

Thus, in both cases, we get $H(x_0, \overline{u}(x_0), D\phi(x_0)) \geq 0$, which is what we wanted.

This shows that $\underline{u}$ is a supersolution and $\overline{u}$ is a subsolution of (1.1). The strong uniqueness principle enables us to conclude. □

In the following section, we explain the role of H_b and give some examples for (2.7). These examples can easily be extended to (2.8). In section 4.1, we provide some simple criteria on H for choosing between (2.7) and (2.8).

3. Some examples. In [9, 1], two classes of numerical Hamiltonian were considered, Godunov and Lax–Friedrichs Hamiltonians. Here, we recall the main results of [1] because they can be applied to a more general setting than those of [9] from which they are inspired. In both cases, only the case of the domain $\mathbb{R}^N$ has been studied, i.e., in the present setting, the case of *interior* nodes.

In order to discretize the problem

$$\begin{cases} u_t + H(Du) = 0, \\ u(x,0) = u_0(x) \end{cases}$$

for $x \in \mathbb{R}^N$ and $t > 0$, where u_0 is Lipschitz continuous, we have considered the scheme

$$\begin{aligned} &u_i^0 = u_0(x_i), \\ &u_i^{n+1} = u_i^n - \Delta t \mathcal{H}_\rho(D_{T_1}u^n, \dots, D_{T_{k_i}}u^n). \end{aligned} \tag{3.1}$$

In (3.1), u^n represents the piecewise linear interpolant of (u_j^n), the set $\{T_1, \dots, T_{k_i}\}$ is the set of triangles that contain x_i, and $D_T u^n$ represents the (constant) gradient of u^n in the triangle T. The parameter ρ describes the local geometry of the mesh. In the examples to come, we specify this parameter; see Remark 1. For any $R > 0$, let us introduce the set $\mathcal{C}_R$ of continuous piecewise linear functions defined by

$$\mathcal{C}_R = \{u \text{ continuous piecewise linear s.t. } ||D_T u|| \leq R \text{ for any triangle } T\}. \tag{3.2}$$

We have to define the numerical Hamiltonians $(p_1, \dots, p_k) \mapsto \mathcal{H}_\rho(p_1, \dots, p_{k_i})$. They have been designed to have the following properties:

1. consistency: $\mathcal{H}_\rho(p, \dots, p) = H(p)$,
2. monotonicity: there exists Δt_R such that for all $\Delta t \leq \Delta t_R$, if $u^n, v^n \in \mathcal{C}_R$ and if $u_i^n \leq v_i^n$ for all i, then $u_i^{n+1} \leq v_i^{n+1}$,
3. intrinsicness: the definition of $\mathcal{H}_\rho$ does not depend on the geometrical description of u^n. For any vertex x_i, for any triangle T such that x_i is a vertex of T, if T is split into two triangles T_1 and T_2 for which x_i is still a vertex, then the value of the numerical Hamiltonian is not modified; see Figure 1.

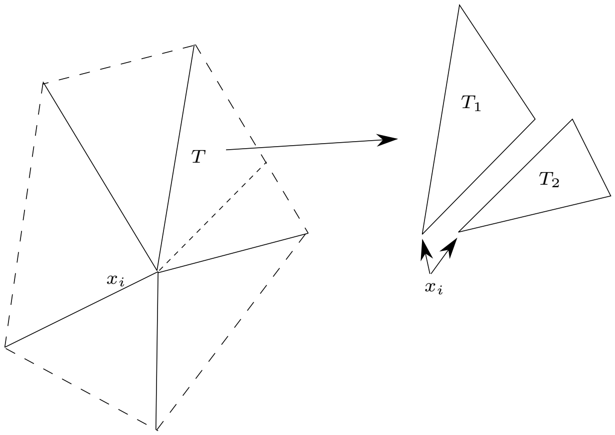


Fig. 1. *Geometrical elements for the intrinsic property: the numerical Hamiltonian is not modified if the triangle T is split into T_1 and T_2 and the value of u at the new vertex is evaluated by linear interpolation.*

We also assume that $\mathcal{H}_\rho$ is *uniformly* continuous in the p and ρ variables. Of course, the number of arguments changes from one mesh point to the other, but if the mesh is regular, the number of neighbors is bounded above, and, thanks to the "intrinsicness" property, we can think of $\mathcal{H}_\rho$ as the same function everywhere. Assuming these properties, it is possible to show the convergence of the scheme (3.1) and to give an error estimate [8, 1].

Notation. In what follows, when we consider numerical schemes that can be put in the form (3.1), sometimes the $D_{T_l}u^n$'s are rewritten in terms of u_i^n and the values of u^n for the neighboring nodes of x_i. For the sake of convenience, we denote the set of the neighbors of x_i (excluding x_i) by $\mathcal{N}_i$, and we rewrite the scheme as

$$u_i^{n+1} = G_\rho(u_i^n, \{u_j^n, j \in \mathcal{N}_i\}; \Delta t). \tag{3.3}$$

More generally, when the Hamiltonian is of the form $H(x, u(x), Du(x))$, the scheme is sometimes rewritten as

$$u_i^{n+1} = G_\rho(x_i, u_i^n, \{u_j^n, j \in \mathcal{N}_i\}; \Delta t) = u_i^n - \Delta t \mathcal{H}_\rho \left(x_i, u_i^n, \{u_j^n, j \in \mathcal{N}_i\}\right). \tag{3.4}$$

In the G-function, $G_\rho(x, t, \{t_l, l \in \mathcal{N}\}; \Delta t)$, t is similar to $u^\rho(x)$, and the variables t, $\{t_l, l \in \mathcal{N}\}$, provide a description of u^ρ in a neighborhood of x_i.

When we are interested in steady problems, the scheme, in the most general case considered in the paper, is

$$\mathcal{H}_\rho(x_i, u_i^n, \{u_j^n, j \in \mathcal{N}_i\}) = 0. \tag{3.5}$$

Similar definitions are also considered for implicit schemes.

In the case of schemes (3.3) and (3.4), the monotonicity condition is equivalent to the following property of G_ρ: $(x, t, \{t_l, l \in \mathcal{N}_i\}) \mapsto G_\rho(x, t, \{t_l, l \in \mathcal{N}_i\}; \Delta t)$. For any fixed grid point $x = x_i$, G should be increasing in t and $\{t_l, l \in \mathcal{N}_i\}$. In practice, the numerical Hamiltonian $\mathcal{H}_\rho$ is an increasing function of $\{t_l, l \in \mathcal{N}_i\}$ and decreasing in t, so that the monotonicity condition for the explicit scheme is true, provided that a CFL-type condition on the time step holds. In the case of schemes (3.5), the monotonicity condition stated in Theorem 2.1 is less restrictive than for unsteady problems: $\mathcal{H}_\rho$ is decreasing with respect to t and increasing with respect to t_l.

Two types of Hamiltonians have been constructed so far, and for the sake of simplicity we describe them in the simplest case. The general case can be treated by "freezing" the x and $u(x)$ variables. They all satisfy the following "translation invariance" property, which mimics the facts that the t-arguments are used to approximate a gradient:

$$\forall x, t, t_l, C \in \mathbb{R}, \ \ G_\rho(x, t+C, \{t_l + C, l \in \mathcal{N}_i\}; \Delta t) = G_\rho(x, t, \{t_l, l \in \mathcal{N}_i\}; \Delta t). \tag{3.6}$$

Godunov Hamiltonians. If $H = H_1 + H_2$, where H_1 (resp., H_2) is convex (resp., concave),[1] then we set

$$\mathcal{H}^G_\rho(p_1, \dots, p_{k_i}) = \inf_{q \in \mathbb{R}^2} \max_{0 \le l \le k_i} \sup_{y \in -\Omega_l + q} \left[(p_i \mid y - q) - H_1^*(y) - H_2^*(q)\right], \tag{3.7}$$

where Ω_l, $l = 1, \dots, k_1$, are the angular sectors defined by the triangles $T_1, \dots, T_{k_i}$ at node x_i; H_1^*, for any l, $-\Omega_l$ is the symmetric of Ω_l with respect to x_i; and H_2^* are the Legendre transforms of H_1 and H_2. We have denoted by $(x \mid y)$ the dot product of x and y.

If h is the smallest radius of the circles of center x_i contained in $\cup_{i=1}^{k_i} T_i$, and if L_1 and L_2 are Lipschitz constants for H_1 and H_2, then the scheme is monotone, provided that the time step satisfies

$$\frac{\Delta t}{h}(L_1 + L_2) \le \frac{1}{2}.$$

The numerical Hamiltonian (3.7) is obtained by saying that $H_1 + H_2$ is bounded below by the convex functions $H_q(p) = H_1(p) - (p \mid q) + H_2^*(q)$. Another monotone Hamiltonian can also be obtained, as in [1], by saying that $H_1 + H_2$ is bounded above by the concave functions $H_q(p) = H_2(q) + (p \mid q) - H_1^*(p)$.

Lax–Friedrichs Hamiltonians. Here we set

$$\mathcal{H}^{LF}_\rho(p_1, \dots, p_{k_i}) = H(\bar{U}) - \frac{\epsilon}{h} \oint_{C_h} [u(x) - u(x_i)] dl,$$

where C_h (resp., D_h) is a circle (resp., disk) of center x_i and radius h,

$$\widehat{U} = \frac{\int_{D_h} Du \, dxdy}{\pi h^2},$$

and ϵ is larger than any Lipschitz constant of H divided by 2π.

Remark 1. For the Godunov and Lax–Friedrichs Hamiltonians, and at a mesh node x, the ρ parameter is the set of unit vectors defining the edges of the triangles at this node and the angles (at node x) of the triangles; see Figure 2.

3.1. Godunov boundary Hamiltonians for convex Hamiltonians. We look for a Hamiltonian $\mathcal{H}^b_\rho$ of the form

$$\mathcal{H}^b_\rho(p_1, \dots, p_k) = \max_{0 \le l \le k_i} \sup_{z \in -\Omega_l} \{(p_l \mid z) - H_b^*(z)\},$$

where the p_i's are the local gradients of a piecewise linear continuous function defined on $\overline{\Omega}$ and the Ω_l's are the angular sectors as before. We need that

$$\mathcal{H}^b_\rho(p, \dots, p) = H_b(p) \le H(p).$$

[1] In the case of a Cartesian mesh, this assumption can be relaxed as shown in [9].

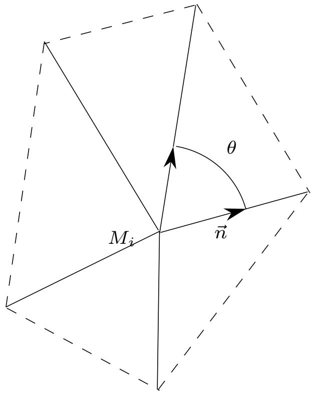


FIG. 2. *A description of ρ.*

If we assume that H_b is convex, this inequality implies $H_b^* \geq H^*$; the most natural choice is to take

$$\begin{cases} H_b^*(q) = H^*(q) & \text{if } q \in \cup_{j=1}^{k_i} \Omega_j, \\ H_b^*(q) = +\infty & \text{otherwise,} \end{cases} \tag{3.8}$$

but any convex Hamiltonian K such that $K^* \geq H_b^*$, which domain is included in $\cup_{j=1}^{k_i}\Omega_j$, would also be a solution. The monotonicity condition is automatically satisfied, thanks to the Hopf formula. In the case of an unsteady problem, the same CFL condition is valid; i.e., if L is a Lipschitz constant of H, the time step satisfies

$$\frac{\Delta t}{h} L \leq \frac{1}{2},$$

because the Lipschitz constant of H_b is at most that of H. If L is a Lipschitz constant of H, $H^*(p)$ may be finite only if p belongs to the ball $B(0, L)$ of center 0 and radius L. Since H_b^* is finite when H is finite, a Lipschitz constant of H is a Lipschitz constant for H_b.

Note that the Hamiltonian (3.8) is the largest possible choice and is the one suggested by the analysis from the dynamical programming principle. It can be interpreted by saying that we take into account all the outgoing rays.

3.2. Godunov boundary Hamiltonians for concave Hamiltonians. The analysis via the dynamical programming principle suggests choosing the boundary condition (2.8). We define H_b by (3.8), where the $+\infty$ condition is replaced by $-\infty$.

3.3. Lax–Friedrichs boundary Hamiltonians for convex Hamiltonians. Here, we are looking for Hamiltonians of the type

$$\mathcal{H}_\rho^b(p_1, \dots, p_k) = K(\bar{U}) - \frac{\epsilon_b}{h} \oint_{C_h} [u(x) - u(x_i)] dl,$$

where K is unknown, as well as the numerical dissipation ϵ_b. The average state is once more defined by

$$\bar{U} = \frac{\int_{D_h \cap \Omega} Du \, dxdy}{|D_h \cap \Omega|}.$$

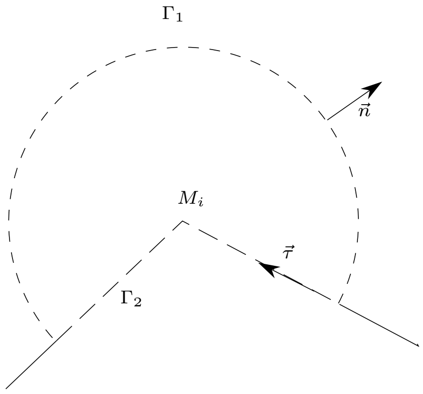


FIG. 3. *Definition of* Γ_1 *and* Γ_2.

The monotonicity condition is satisfied, provided that, L_b being a Lipschitz constant of K,

$$\epsilon_b \geq \frac{L_b\, h}{|D_h \cap \Omega)|}.$$

The area $\partial(D_h \cap \Omega)$ is $\theta\, h$, where θ is the angle of $D_h \cap \Omega$ at x_i. This can be seen by using the same arguments as in [1].

We denote by Γ_1 the part of $\partial(D_h \cap \Omega)$ which is inside Ω, and by Γ_2 the boundary part; see Figure 3. To determine K, we consider the consistency condition. It is easy to see that

$$H_b(p) \equiv \mathcal{H}^b_\rho(p, \dots, p) = K(p) - \frac{\epsilon_b}{h}\left(p\Big|\left\{\int_{\Gamma_1} \vec{n} dl - \int_{\Gamma_2} \vec{\tau} dl\right\}\right),$$

where $\vec{n}$ is the outward unit normal to Γ_1 and $\vec{\tau}$ is the unit tangent vector to Γ_2. The vector

$$\vec{N} = -\frac{1}{h}\left(\int_{\Gamma_1} \vec{n} dl - \int_{\Gamma_2} \vec{\tau} dl\right)$$

enters into Ω if $\partial\Omega$ is regular enough.

The convergence property of Theorem 2.2 is satisfied if $H_b \leq H$. When H_b is assumed to be convex, this condition is equivalent to asking for K to be convex and

$$\left(K + \epsilon_b \left(\vec{N}\mid\ .\ \right)\right)^*(q) \geq H^*(q) \quad \forall q \in \mathbb{R}^2.$$

The Legendre transform of $x \mapsto K(x) + \epsilon_b(\vec{N}|x)$ is

$$q \mapsto K^*(q + \epsilon_b \vec{N}),$$

and consequently K^* is defined by the relation

$$K^*(q) \geq H^*(q - \epsilon_b \vec{N}). \tag{3.9}$$

Let us call $\mathrm{Dom}(H^*)$ (resp., $\mathrm{Dom}(H_b)$) the subset of $\mathbb{R}^2$ for which $H^*(q)$ (resp., $(H^b)^*$) is finite. If L is a Lipschitz constant of H, $\mathrm{Dom}(H^*) \subset B(L)$. A similar result holds

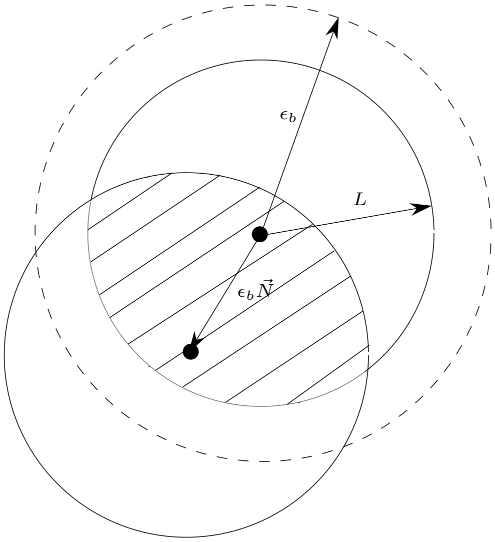


FIG. 4. *Geometrical representation of the conditions.*

for K. These sets are convex. There is a solution to the problem (different from $K = -\infty$) if and only if we can find ϵ_B such that

$$(\text{Dom}(H) - \epsilon_b\vec{N}) \cap \text{Dom}(H) \neq \emptyset, \qquad (3.10)$$
$$(\text{Dom}(H) - \epsilon_b\vec{N}) \cap B(\epsilon_b) \neq \emptyset.$$

See Figure 4 for a representation of these conditions.

In some cases, there is *no* solution at all. The simplest counterexample is given by

$$H(x) = (\vec{a} \mid x)$$

with $\vec{a} \neq 0$. In this case, $\text{Dom}(H) = \{\vec{a}\}$. The first condition implies $\epsilon_b = 0$ (thus $||\vec{a}|| = 0$); the second one gives $\vec{a} = 0$.

In some cases, there are solutions. An example is given by any Hamiltonian for which $\min H > -\infty$: since $0 \in \text{Dom}(H)$, we can set $\epsilon_b = 0$ and $K \equiv \min H$. Another example is provided by $H(x) = ||x||$. Here, we can choose any $\epsilon_b \in [0, \frac{1}{2||\vec{N}||}]$.

3.4. Other choices. In sections 3.1 and 3.3, a very obvious choice would be $H_b \equiv -\infty$. This choice enables us to satisfy our convergence conditions. In fact we have

$$\min(H, \max(-\infty, F)) = \min(H, F),$$
$$\max(H, \max(-\infty, F)) = \max(H, F),$$

so that the viscosity inequalities are obviously satisfied. This reduces to strongly imposing the boundary conditions. However, this is not be the best choice, since a numerical boundary layer may be generated, especially in the case of Dirichlet boundary conditions. The scheme converges, but very slowly, as can be seen by numerical experiments; see section 5. Moreover, the results of this paper have been formally extended to more general cases, particularly the case of discontinuous Dirichlet conditions. In this particular case, the choice $H_b = -\infty$ may prevent convergence, whereas the Godunov or Lax Friedrichs boundary Hamiltonian seems to ensure convergence.

It now becomes clear that some selection procedures must be established. We will provide some, in special cases.

4. Some selection criteria. In this section, we discuss the problem of finding suitable boundary Hamiltonians and the question of selecting between the min and max conditions. These two questions are related, but the most difficult one is to find a "good" boundary Hamiltonian. If no care is taken, the numerical solution may develop a boundary layer structure, especially in the case of Dirichlet conditions; i.e., the gradient of the numerical solution may become unbounded when the mesh size tends to zero. If Theorem 2.2 provides some necessary conditions for convergence, this is not acceptable in general because practical calculations are done with finite but nonvanishing mesh sizes.

This situation is very similar to the technical difficulties encountered in the analysis of the boundary problem in the continuous case; see [4]. To explain this point, let us consider a simple one dimensional example.

If, for example, we think of a numerical scheme for

$$\begin{cases} |u'(x)| = 1, & x \in]0,1[, \\ u(0) = 0, & u(1) = 2, \end{cases}$$

where the boundary conditions are strongly imposed as being modeled by

$$\begin{cases} |u'(x)| - \epsilon u_{xx} = 1, & x \in]0,1[, \\ u(0) = 0, & u(1) = 2, \end{cases} \tag{4.1}$$

the solution should look like

$$u_\epsilon(x) = x + \frac{\exp\left(\frac{x-1}{\epsilon}\right) - \exp\left(-\frac{1}{\epsilon}\right)}{1 - \exp\left(-\frac{1}{\epsilon}\right)},$$

and hence a boundary layer exists: the derivative of u is not bounded at $x = 1$ when $\epsilon \to 0$. Its thickness tends to 0 as $\epsilon \to 0$.

This simple example is quite generic from the numerical point of view. Assume that the numerical solution u^ρ converges in the neighborhood of the boundary to a regular solution u. Then, up to second order truncation errors, the numerical Hamiltonian behaves like

$$H(p_1, \dots, p_{k_i}) \simeq H(Du) - \epsilon(\rho) D^2 u,$$

where D^2u represents some elliptic operator and $\epsilon(\rho) \to 0$ as the mesh size tends to zero. In [8], some numerical schemes are constructed by directly using this idea. Because of that, if one sets the boundary condition strongly on $\partial\Omega$, as in the example (4.1), a boundary layer must exist in the vicinity of $\partial\Omega$. Its thickness tends to 0 as $\epsilon \to 0$. Its thickness also tends to 0 as $\epsilon(\rho) \to 0$.

4.1. Choosing between the "min" and "max" conditions for Dirichlet boundary conditions. In some situations, the choice can be motivated by some a priori knowledge of the behavior of the exact solution. For example, if one makes the following assumption—there exists $R \in]0, +\infty[$ such that

$$\lim_{\lambda \to +\infty} H(x, u, p - \lambda \vec{n}) = +\infty$$

uniformly on x in a neighborhood of $\partial\Omega$, $-R \le u \le R$ and p bounded[2]—then one can prove that if φ is continuous and u is the solution of

$$\begin{aligned} &H(x,u(x),Du(x)) = 0, && x \in \Omega, \\ &u(x) = \varphi(x), && x \in \partial\Omega, \end{aligned}$$

then $u(x) \le \varphi(x)$ on $\partial\Omega$ (see [4]). The boundary condition (2.7) implies that $u \le \phi$ at the discrete level, while (2.8) implies the opposite inequality. Hence, when the above assumption is true, the boundary condition (2.7) is the natural one to consider. This situation is encountered for nonbounded convex Hamiltonians.

When we have

$$\lim_{\lambda\to+\infty} H(x,u,p-\lambda\vec{n}) = -\infty,$$

the boundary condition (2.8) is the natural one to consider. This situation is encountered for nonbounded concave Hamiltonians.

4.2. The case of coercive Hamiltonians and boundary conditions (2.7). We are not able to provide an error bound between the numerical and the exact solutions. However, when the Hamiltonians H and H_b are coercive, we can show that no numerical boundary layer can appear; i.e., the gradient of the numerical solution is bounded when the mesh size tends to zero.

We say that H is coercive if

$$H(x,u,p) \to +\infty \quad \text{when } ||p|| \to +\infty$$

uniformly for $x \in \Omega$, $u \in [-R,R]$, $R \in]0,+\infty[$. We say that the boundary Hamiltonian is coercive if

$$H_b(x,u,p) \to +\infty \quad \text{when } ||p|| \to +\infty \text{ and } (p \mid \vec{n}) \ge 0$$

uniformly for $x \in \partial\Omega$, $u \in [-R,R]$ for all $R \in [0,+\infty[$. Here $\vec{n}$ is the inward unit vector at point $x \in \partial\Omega$. We have implicitly assumed that $\partial\Omega$ is C^1. In what follows, we consider the Dirichlet problem

$$\text{(4.2)} \quad \begin{cases} \mathcal{H}_\rho(x_i,u_i,D_{T_1}u,\dots,D_{T_{i_k}}u) = 0, & x_i \text{ interior node and } i_l \in \mathcal{N}_i, \\ \max(\mathcal{H}_\rho(x_i,u_i,D_{T_1}u,\dots,D_{T_{i_k}}u), u_i - \varphi(x_i)) = 0, & x_i \text{ boundary node and } i_l \in \mathcal{N}_i \end{cases}$$

but our results clearly extend to the more general case considered in this paper.

PROPOSITION 4.1. *Let $\mathcal{H}_\rho$ and $\mathcal{H}_\rho^b$ be monotone Hamiltonians consistent with H and H_b. Assume that $\mathcal{H}_\rho$ and $\mathcal{H}_\rho^b$ also satisfy* (3.6), *and that H and H_b are continuous, convex, and coercive. Assume also that the mesh is regular. Then the scheme* (2.7) *is convergent and the maximum over the triangles T of the norm of the numerical solution, when $h \to 0$ remains bounded.*

The boundary of $\Omega_h = \cup_{T\in\mathcal{T}_h} T$ is denoted by Γ_h.

Proof. The convergence is a consequence of Theorem 2.2. The uniform boundedness of the gradients is a consequence of the following lemma. □

LEMMA 4.2. *If H, $\mathcal{H}_\rho$, $\mathcal{H}_\rho^b$, and the mesh satisfy the assumptions of Proposition* 4.1, *and if u is a subsolution of* (4.2), *then there exists C independent of h such that for any two mesh points M_i, M_j we have*

$$|u_i - u_j| \le CM_iM_j.$$

[2] $\vec{n}$ is the inward unit normal to $\partial\Omega$.

Proof. For the sake of simplicity, we assume that H and H_b depend only on the p variable.

Let $K > 0$ and x_i be a mesh point. For now we let K be free. Since Ω_h has a finite number of points, there exists M', a mesh point such that $u_l - KM_lM_i$ is maximum at M':

$$u_l - KM_lM_i \leq u_{M'} - KM'M_i.$$

This indicates that $v_l = (u(x') - K\, ||x' - x_i||) + K\, ||x_l - x_i||$ is greater that u, with an equality at node x'. Hence, using the same techniques as in Appendix A, if x' is an interior node,

$$0 \geq \mathcal{H}_\rho(u) \geq \mathcal{H}_\rho(K\, ||x - x_i||), \tag{4.3}$$

where we have written $\mathcal{H}_\rho(u)$ instead of $\mathcal{H}_\rho(D_{T_{i_1}}u, \dots, D_{T_{i_k}}u)$, for short. Similarly, since $\max(\mathcal{H}^b_\rho(u), u - \varphi) \leq 0$, we have $\mathcal{H}^b_\rho(u)$, and by the monotonicity of $\mathcal{H}^b_\rho$, we get

$$0 \geq \mathcal{H}_\rho(K\, ||x - x_i||) \tag{4.4}$$

at x' if it is on the boundary. Assume that $x' \neq x_i$. We show that if K is large enough, we have a contradiction. We can assume that $x_i = 0$ so that we have to deal with the piecewise interpolant $\pi_h||x||$ of the convex function $x \mapsto ||x||$. Note that $0 = x_i$ does not lie in the interior of any triangle. A simple consequence of the Taylor formula [7] shows that there exists $C' > 0$ such that if the mesh is regular,

$$\left|\left| D_T\pi_h|x| - \frac{x_G}{||x_G||} \right|\right| \leq C_1 h,$$

where x_G is the gravity center of T.

Since H is regular, there exist $C_2 > 0$ such that

$$\left| \mathcal{H}_\rho(\pi_h||x||) - \mathcal{H}_\rho\left(K\frac{x_{G_1}}{||x_{G_1}||}, \dots, K\frac{x_{G_k}}{||x_{G_k}||} \right) \right| \leq C_2 h.$$

The same inequality is also true for $\mathcal{H}^b_\rho$. Since the scheme is consistent with uniformly continuous numerical Hamiltonians, and because $O = x_i$ does not belong to the interior of the molecule associated with x', we can replace $\mathcal{H}_\rho(K\frac{x_{G_1}}{||x_{G_1}||}, \dots, K\frac{x_{G_k}}{||x_{G_k}||})$ by $H(K\, \frac{x'}{||x'||})$ up to an $O(h)$ term. The same is true for $\mathcal{H}^b_\rho$. Hence we get that $H(K\, \frac{x'}{||x'||}) \leq O(h)$, which is impossible if K is large enough. This shows that if $K = C + 1$, where C is chosen so that

$$H(C\, p) < 0 \text{ with } ||p|| = 1 \quad \text{and} \quad H_b(C\, p) < 0 \text{ with } ||p|| = 1 \text{ and } (p \mid \vec{n}) < 0,$$

we have that $x' = x_i$, and then

$$u_l - u_i \leq K||x_l - x_i||$$

when h is small enough. The conclusion holds by symmetry. □

An example. We consider the example of a convex Hamiltonian. The boundary Hamiltonian consistent with the Godunov boundary Hamiltonian is

$$H_b(p) = \max_{(y \mid \vec{n}) \leq 0} \left((y \mid p) - H^*(y) \right),$$

where $\vec{n}$ is the inward unit vector. If we assume that H is smooth enough, the optimal ray is $p^* = DH(p)$. A sufficient (and crude) condition to ensure that H_b is coercive if H is coercive is to state that $(p^* \mid \vec{n}) \leq 0$, because in this case $H_b(p) = H(p)$. To obtain this sufficient condition, it is enough to say that

$$\lambda \in \mathbb{R}^+ \mapsto H(p + \lambda \vec{n})$$

is monotone increasing. This has to be connected to the conditions of section 4.1. An example where $(p^* \mid \vec{n}) \leq 0$ is given by the eikonal Hamiltonian because $p^* = \frac{p}{||p||}$.

4.3. Extension to nonconvex Hamiltonians. Let us consider a problem where $H = H_1 + H_2$, with H_1 convex and H_2 concave. Following (3.7), a natural boundary Hamiltonian is also

$$\mathcal{H}^b_\rho(p_1, \dots, p_{k_i}) = \inf_{q \in \mathbb{R}^2} \max_{0 \leq l \leq k_i} \sup_{y \in -\Omega_l + q} \left[(p_i \mid y - q) - H_1^*(y) - H_2^*(q) \right].$$

Since $\cup_l \Omega_l$ is a strict subset of $\mathbb{R}^2$, we have

$$H^b(p) \equiv \inf_{q \in \mathbb{R}^2} \sup_{(y \mid \vec{n}) \geq 0} \left[(p \mid y - q) - H_1^*(y) - H_2^*(q) \right] \leq H(p),$$

and the conditions of Theorem 2.2 are satisfied.

If the family $\{H_1(p) - (p \mid q)\}_{q \in \mathbb{R}^2}$ is uniformly coercive, then the numerical solution develops no numerical layer; this is a simple consequence of Proposition 4.1.

5. Applications.

5.1. Some numerical tests. We have not been able to get error estimates for the schemes presented above. Even in the case of the crudest approximations of the boundary condition, i.e., by taking $H_b = -\infty$, we can show the convergence of the numerical solution. However, we have shown in a special case that no numerical boundary layer exists even for Dirichlet conditions when the Hamiltonians are coercive.

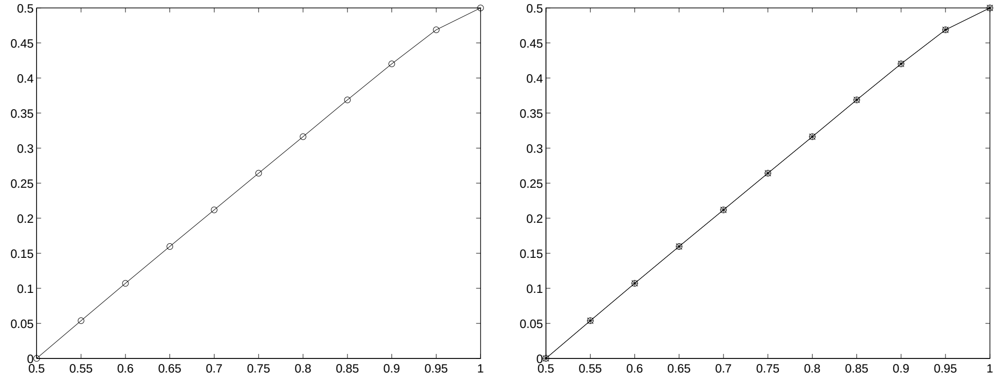
The purpose of this paragraph is to illustrate the various phenomena that we have encountered, for Dirichlet and Neuman conditions. In each case, the strong boundary conditions are obtained with $H_b = -\infty$, and the weak ones with H_b being the Godunov Hamiltonian. In sections 5.1.1 and 5.1.2, the interior Hamiltonian is the Lax–Friedrichs one. In section 5.2, it is the Godunov Hamiltonian. Other experiments with the Lax–Friedrichs condition have been done, but they are not reported here. They provide the same results. We also show the behavior of the schemes on a problem with nonconvex Hamiltonians and Dirichlet conditions; this is the subject of section 5.1.3.

5.1.1. Dirichlet conditions. The domain Ω is limited by two "concentric" circles of radius 0.5 and 1. It is discretized by a finite element–type mesh, but this is not essential. The problems are

$$\begin{cases} ||Du|| = 1 & \text{in } \Omega, \\ u = 0 & \text{for } ||x|| = 0.5, \\ u = C & \text{for } ||x|| = 1. \end{cases} \tag{5.1}$$

Table 5.1
Boundary condition for the Dirichlet boundary conditions.

Case	1	2	3	4
C	0.5	10	0.25	-11



(a) (b)

Fig. 5. *Comparison of different implementations of Dirichlet conditions for problem* (5.1) *and Case* 1: (a) *weak conditions,* (b) *comparison of weak* ($\circ$) *and strong* ($*$).

The constant C takes the values displayed in Table 5.1.

The viscosity solution is given as follows:

- Case 1 and 2: $u(x) = ||x|| - 0.5$;
- Case 3: $u(x) = ||x|| - \frac{1}{2}$ if $||x|| \in [\frac{1}{2}, \frac{7}{8}]$ and $u(x) = -||x|| + \frac{5}{4}$ if $||x|| \in [\frac{7}{8}, 1]$;
- Case 4: $u(x) = -||x|| + \frac{1}{2}$.

The difference between these test cases is that for Cases 1 and 3, the boundary conditions on $||x|| = 1$ are enforced strongly, whereas for 2 and 4, they are enforced in the viscosity sense only.

We plot the cross section only in the y-direction and positive abscissa. Two kinds of tests have been done. In the first, we have strongly imposed the boundary conditions; i.e., we have taken $H_b = -\infty$. In the second test, the conditions have been imposed weakly, with the Godunov boundary Hamiltonian.

Comparison of Figures 5, 6, 7, 8 clearly shows that when the boundary condition is *strongly* enforced by the viscosity solution, no special treatment is needed. On the contrary, when it is only weakly enforced, then a special treatment is mandatory, otherwise a boundary layer–type phenomenon is observed.

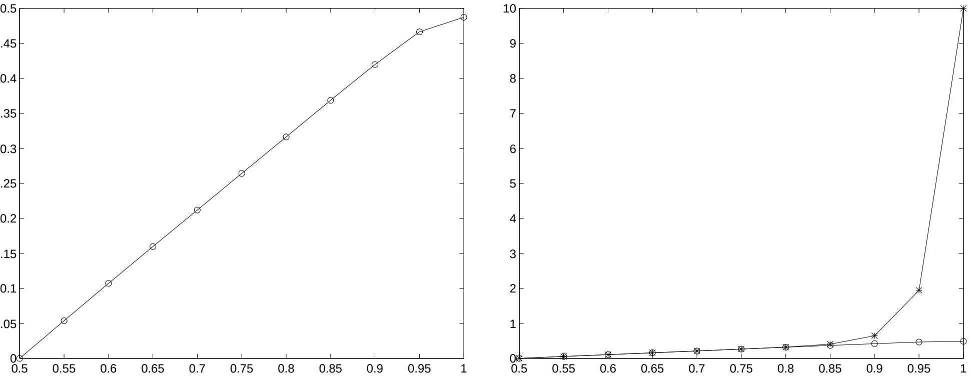
5.1.2. Neumann conditions. Here, we test the problem

$$\left\{ \begin{array}{ll} ||Du|| = 1 & \text{in } \Omega, \\ u = 0 & \text{for } ||x|| = 1, \\ \frac{\partial u}{\partial n} = 0 & \text{for } ||x|| = 0.5. \end{array} \right. \tag{5.2}$$

Its solution is $u(x) = -||x|| + \frac{1}{2}$.

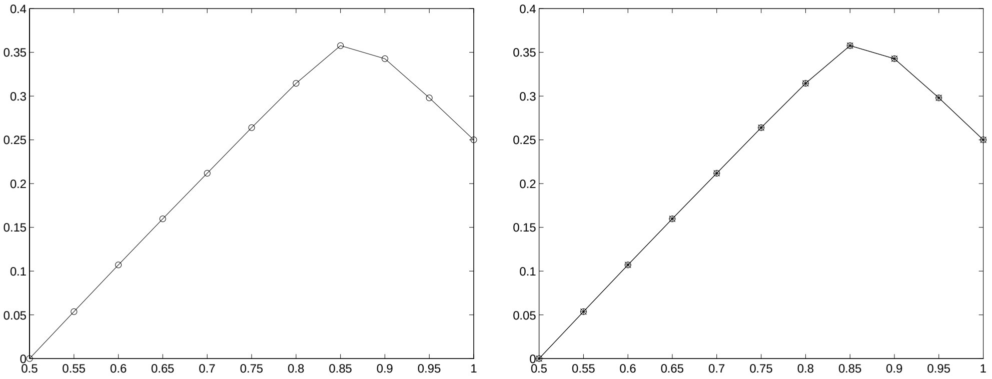
The viscosity solution is given by the solution for Case 1. Once more, the numerical solution is obtained by imposing the boundary conditions either strongly or weakly.

The problem $\frac{\partial u}{\partial n} = g$ is approximated at node A in the following way (see Figure 9). The outward unit normal is approximated as $\vec{n}_A = \overrightarrow{AB}^{\perp} + \overrightarrow{AC}^{\perp}$, which is



(a) (b)

Fig. 6. *Comparison of different implementations of Dirichlet conditions for problem* (5.1) *and Case* 2: (a) *weak conditions,* (b) *comparison of weak* ($\circ$) *and strong* ($*$).



(a) (b)

Fig. 7. *Comparison of different implementation of Dirichlet conditions for problem* (5.1) *and Case* 3: (a) *weak conditions,* (b) *comparison of weak* ($\circ$) *and strong* ($*$).

normalized. Here, $\vec{x}^{\perp}$ is the orthogonal vector to $\vec{x}$ such that $(\vec{x}, \vec{x}^{\perp})$ is positive. Then we consider a node D which is on the side of the triangle opposite to D, which is cut by $-\vec{n}_A$. We then set

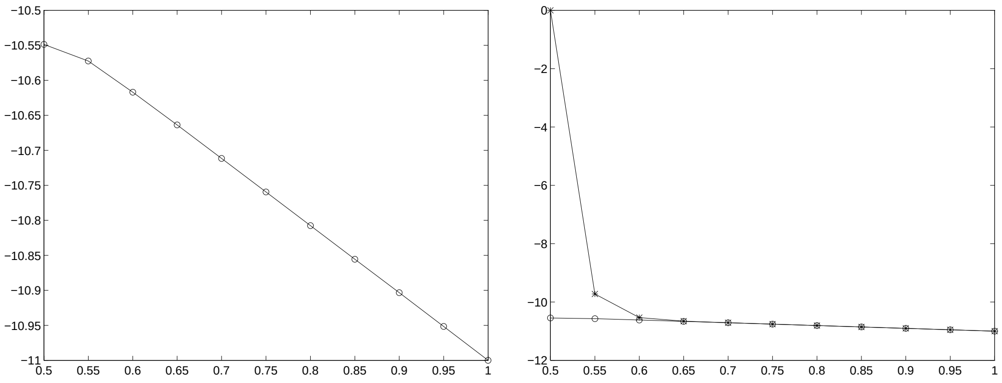
$$u(A) = ||\overrightarrow{AD}||g(A) + u(C). \tag{5.3}$$

Here, u is the piecewise linear interpolation of the data.

In the strong formulation, we use (5.3) directly. In the weak formulation, we set

$$\max\left(\frac{u_A^{n+1} - u_A^n}{\Delta t} - \mathcal{H}_\rho^b(u^n), \frac{u_A^{n+1} - u_C^{n+1}}{AC} - g(u_A^n)\right) = 0.$$

From Figure 10, it is clear that the weak formulation gives much better results. However, the difference between the two formulations is not as important as for the Dirichlet problem, as expected.



(a) (b)

FIG. 8. *Comparison of different implementation of Dirichlet conditions for problem* (5.1) *and Case* 4: (a) *weak conditions,* (b) *comparison of weak* (○) *and strong* (∗).

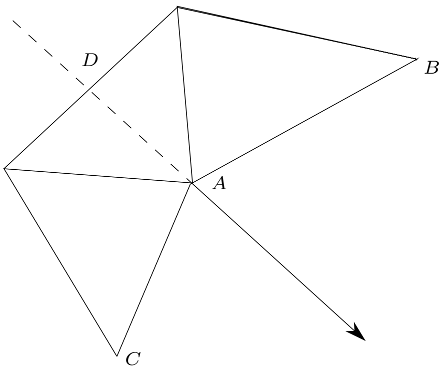


FIG. 9. *Schema for the approximation of the Neuman boundary conditions.*

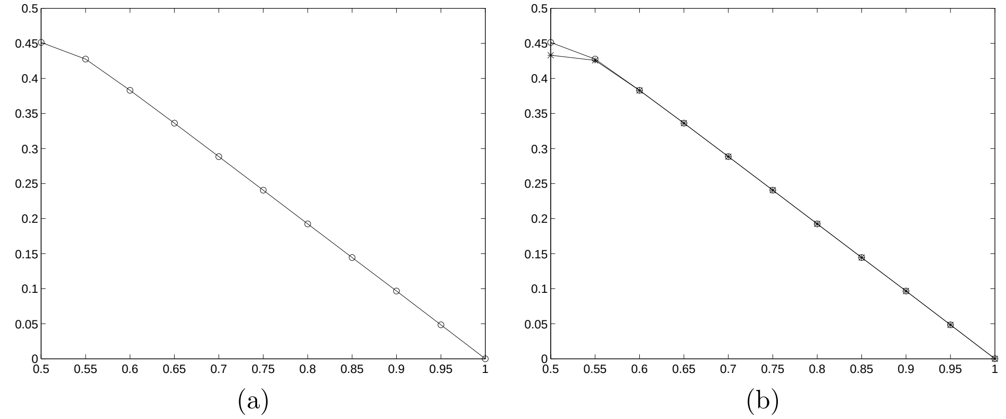
5.1.3. The case of a nonconvex Hamiltonian and Dirichlet conditions. In general, it is difficult to compute analytically the solution of a first order Hamilton–Jacobi equation, and the situation is even worse when the Hamiltonian is not convex (nor concave), because the analogy with hyperbolic systems becomes looser in general. Hence, it becomes more difficult to judge the quality of numerical results. To overcome this difficulty in a special case, we consider $H(p) = (||p|| - 1)^3$ and the problem

$$\begin{array}{ll} H(Du) = 0 & \text{on } \Omega, \\ u = 0 & \text{on } \Gamma_1, \\ u = 10 & \text{on } \Gamma_2, \end{array} \tag{5.4}$$

where Ω is depicted in Figure 11. Since $t \mapsto t^3$ is monotone increasing, u is a solution of (5.4) if and only if it is a solution of

$$\begin{array}{ll} ||Dv|| - 1 = 0 & \text{on } \Omega, \\ v = 0 & \text{on } \Gamma_1, \\ v = 10 & \text{on } \Gamma_2. \end{array} \tag{5.5}$$

The solution of (5.4) and (5.5) is the distance to Γ_1.



(a)

(b)

FIG. 10. *Comparison of different implementations of homogeneous Neumann conditions for problem* (5.2): (a) *weak conditions,* (b) *comparison of weak* (○) *and strong* (∗).

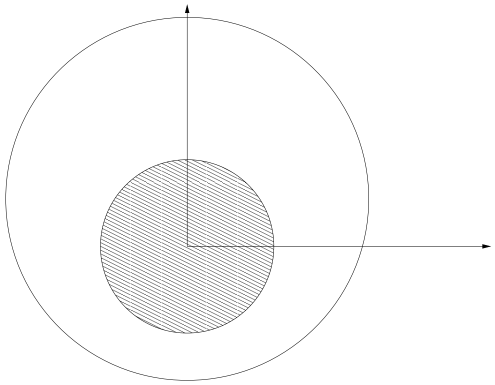


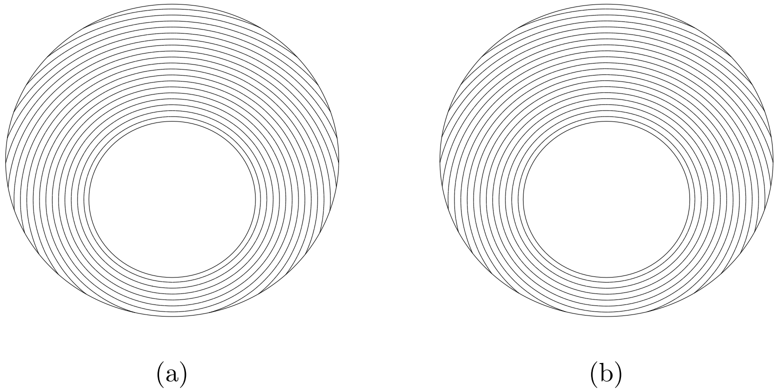
FIG. 11. *Computational domain for problem* (5.4). Γ_1 *is the inner circle of center* $(0,0)$ *and radius* $r = 1$, Γ_2 *is the outer circle of center* $(0,0.5)$ *and radius* $r = 3$.

In order to discretize (5.4), we write $H = H_1 + H_2$, with $H_1(p) = \max(|p|| - 1, 0)^3$ and $H_2(p) = \min(||p|| - 1, 0)^3$. These functions are respectively convex and concave. The numerical Hamiltonian and the boundary Hamiltonian are the same as in sections 3 and 4.3. The numerical solution is displayed in Figure 12(a). The solution of (5.5) with the Godunov Hamiltonian is provided in Figure 12(b). A close comparison shows that they are (almost) identical.

Another application of the boundary conditions developed in this paper is given by the approximation of the following problem, on the same geometry:

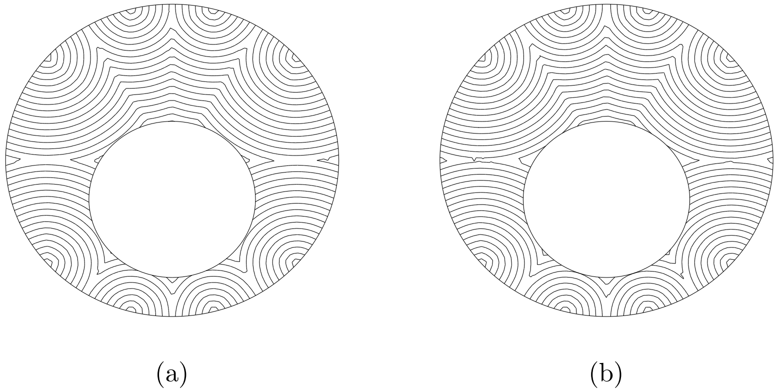
$$\begin{aligned} &H(Du) = 0 && \text{on } \Omega,\\ &u(x,y) = 0, && (x,y) \in \Gamma_1,\\ &u(x,y) = 3\cos(2\pi x), && (x,y) \in \Gamma_2. \end{aligned} \tag{5.6}$$

Since H is nonconvex, it is difficult to know a priori what the value of the solution on the boundary would be. The computed solution is given in Figure 13(a). It can be seen that the solution satisfies the boundary condition strongly on Γ_2 and only weakly on Γ_1 (in contrast to the previous example). Note, however, that the boundary conditions have been numericaly *weakly* imposed on Γ_1 and Γ_2. The solution is also



(a) (b)

FIG. 12. (a) *Solution of problem* (5.4)*,* min = 0*,* max = 1.48*.* (b) *Solution of problem* (5.10)*,* min = 0*,* max = 1.504*.*



(a) (b)

FIG. 13. (a) *Solution of problem* (5.6)*,* min = −3*,* max = −1.53*.* (b) *Solution of problem* (5.7)*,* min = −3*,* max = −1*,* 47*.*

in very good agreement with the one obtained from the discretization of

$$\begin{array}{ll} ||Dv|| - 1 = 0 & \text{on } \Omega, \\ v(x,y) = 0, & (x,y) \in \Gamma_1, \\ v(x,y) = 3\cos(2\pi x), & (x,y) \in \Gamma_2, \end{array} \tag{5.7}$$

which is displayed in Figure 13(b).

5.2. Application to a problem in geophysics. In [6] is developed a technique to compute the multivalued solutions τ of the Eikonal equation with an initial condition

$$\tau(x_S) = 0$$

at the source term x_S. This corresponds to the problem of computing the very high frequency approximation of the wave equation in a possibly inhomogeneous media, when the source term is located at a single point with a Dirac source term. In this case, the solution consists of a wave front that might have a very complex structure.

The solution of this problem is important in geophysics applications; it is the core of an inverse method for reconstructing the index of the media knowing only the arrival times of the wave fronts at the ground.

In Benamou's method [6], we need to be able to solve, in several arbitrary domains Ω containing x_S, the following problem:

$$\left\{ \begin{array}{ll} ||D\tau|| - n(x) = 0 & \text{in } \Omega, \\ \tau(x_S) = 0, & \\ \tau(x) = +\infty & \text{if } x \in \partial\Omega - \{x_S\}. \end{array} \right. \tag{5.8}$$

The boundary conditions have to be understood in the viscosity sense. In particular, the second boundary solution corresponds to the Soner boundary condition. The solution to this problem is known,

$$\tau(x) = \inf_{y_x} \left[\int_0^{\min(T,\zeta)} n(y_x(s)) \left| \frac{dy_x}{ds} \right| ds \right], \tag{5.9}$$

where the trajectory y_x starts at x_S for $s = 0$, and ζ is its first exit time, i.e.,

$$\zeta = \inf\{s \geq 0; y_x(s) \notin \overline{\Omega}\}.$$

In other words, we do not take into account the rays that start at x_S and come into Ω.

The idea is to characterize the solution of (5.8) as the steady solution of

$$\left\{ \begin{array}{ll} u_t + ||Du|| - n(x), & t > 0 \text{ and } x \in \Omega, \\ u(x, t = 0) = 0, & x \in \Omega, \\ u(x_S, t) = 0 & \text{at } x_S \\ u(x, t) = +\infty, & x \in \partial\Omega - \{x_S\}. \end{array} \right. \tag{5.10}$$

It is clear that neither (5.9) nor (5.10) falls into the framework that we have considered here. The idea is to introduce an approximation of τ_ρ^e, the solution of

$$\left\{ \begin{array}{ll} ||D\tau|| - 1 = 0 & \text{in } \Omega, \\ \tau(x_S) = 0, & \\ \tau(x) = 0, & x \in \partial\Omega - \{x_S\}, \end{array} \right. \tag{5.11}$$

given by (5.9) for $n \equiv 1$. We then show that the scheme (5.12) can be rewritten as

$$\left\{ \begin{array}{ll} \frac{u_i^{n+1} - u_i^n}{\Delta t} + \mathcal{H}_\rho(D_{T_1}u^n, \dots, D_{T_1}u^n) - n(x_i), & n > 0 \text{ and } x_i \text{ interior point}, \\ u_i^0 = 0 & \text{for all } i, \\ u_{x_S}^n = 0 & \text{for } n \geq 1, \\ u_i^n = \min(u_i^n - \Delta t \mathcal{H}_\rho^b(D_{T_1}u^n, \dots, D_{T_1}u^n) - n(x_i), K\tau_\rho^e), & n > 0 \text{ and } x_i \neq x_S, \end{array} \right. \tag{5.12}$$

for K large enough, uniformly in ρ. Once this is shown, we can apply the arguments of (2.2) to conclude. In what follows, we restrict ourselves to the case of the Godunov Hamiltonian.

5.2.1. Finding an elementary supersolution of (5.11) when $n \equiv 1$.

Our aim is to find an elementary supersolution of

$$\left\{ \begin{array}{ll} \mathcal{H}_\rho(u, x_i) = 1 & \text{for any node different from } x_S, \\ & \\ u(x_S) = 0. & \end{array} \right. \tag{5.13}$$

Here, the numerical Hamiltonian is the Godunov Hamiltonian for $H(x,p) = ||p|| - 1$.

For any mesh points x_i and x_j, we consider a path $P(x_i \to x_j) = x_i \dots P_k \dots x_j$ connecting x_i and x_j. The points P_k of $P(x_i \to x_j)$ are nodes of the mesh. We define $\#(P(x_i \to x_j))$ as the number of nodes that define the path $P(x_i \to x_j)$. We consider u defined by

$$u^e(x_i) = \min_{P(x_i \to x_S))} \left(\sum_{l=1}^{\#(P(x_i \to x_S))} P_l P_{l+1} \right), \tag{5.14}$$

with the convention $P_1 = x_i$ and $P_{\#(P(x_i \to x_S))} = x_S$.

LEMMA 5.1. *Let $k \geq 0$. We have, for any mesh point x_ℓ,*

$$\tau_\rho^e(x_i) = \min_{P(x_i \to x_S)} \left(\sum_{l=0}^{\#(P(x_i \to x_S))} P_j P_{j+1} + \tau_\rho^e(x_\ell) \right).$$

Proof. Let P be an optimal path that connects x_ℓ to x_S, and P' be a path that connects x_i to x_ℓ. The path $P \cup P'$ connects x_i to x_S, and we have

$$\tau_\rho^e(x_i) \leq \sum_{j=1}^{\#(P(x_i \to x_S))} P_j P_{j+1} + \tau_\rho^e(x_\ell).$$

By taking the infinum, we have the first inequality. Let P now be an optimal path for

$$v(x_i) = \min_{P(x_i \to x_\ell)} \left(\sum_{l=0}^{\#(P(x_i \to x_S))} P_j P_{j+1} + \tau_\rho^e(x_\ell) \right).$$

If P' is an optimal path for $u(x_\ell)$, by connecting the two paths, we get the opposite inequality. □

We show that τ_ρ^e defined in (5.14) is a supersolution of the problem. First it is clear that $u(x_S) = 0$. For any node x_i, we denote by $\mathcal{N}(x_i)$ the neighboring nodes of x_i in the mesh. We show now that

$$\max_{x \in \mathcal{N}(x_i)} \left(\frac{\tau_\rho^e(x_i) - \tau_\rho^e(x)}{||x_i - x||} \right) \geq 1. \tag{5.15}$$

This is a direct consequence of Lemma 5.1. Since $\mathcal{H}_\rho \geq \max_{x \in \mathcal{N}(x_i)} \left(\frac{\tau_\rho^e(x_i) - \tau_\rho^e(x)}{||x_i - x||} \right)$,[3] the function τ_ρ^e is a supersolution of (5.12) when $n \equiv 1$.

5.2.2. Study of the scheme (5.12).

For any $K > \max_\Omega n(x)$, we consider the scheme (5.12). By using the maximum principle, it is easy to get the next result.

PROPOSITION 5.2. *Under the CFL restriction $\frac{\Delta t}{h} \leq 1/2$, the numerical values $(u_i^n)_{x_i, n \geq 0}$ satisfy the following:*

- *at the source point x_S, $u_{x_S}^n = 0$ for any $n \geq 0$.*

[3] This is true because on each angular sector the maximum is reached by one of the terms $\frac{\tau_\rho^e(x_i) - \tau_\rho^e(x)}{||x_i - x||}$ for $x \in \mathcal{N}_i$ or by the gradient of τ_ρ^e in this angular sector. If we take $\vec{e} = \frac{x - x_i}{||x - x_i||}$, we have $\frac{\tau_\rho^e(x_i) - \tau_\rho^e(x)}{||x_i - x||} \leq |\,(Du \mid \vec{e})\,| \leq ||D\tau_\rho^e||$, and the inequality follows.

- $0 \leq u_i^n \leq K\tau_\rho^e(x_i)$ *for any* x_i, $t_n = n\Delta t$.
- *For any* i, *the sequence* (u_i^n) *has a limit when* $n \to +\infty$, *which is the solution of*

$$\begin{cases} \mathcal{H}_\rho(D_{T_1}u, \dots, D_{T_1}u) - n(x_i) = 0 & \textit{for any } x_i, \\ u_{x_S} = 0 & \textit{at } x_S, \\ \max\left(\mathcal{H}_\rho^b(D_{T_1}u, \dots, D_{T_1}u) - n(x_i), K\tau_\rho^e(x_i)\right) & \textit{on the boundary.} \end{cases}$$

In particular, u_i^n is *independent* of K and ρ when $K \geq \max_\Omega n(x)$.

Proof.

- $0 \leq u_i^n$. This is obvious since the scheme is monotone and $u_i^0 = 0$.
- $u_i^n \leq K\tau_\rho^e(x_i)$. The previous results show that $K\tau_\rho^e$ is a supersolution of (5.12) when $K \geq \max_\Omega n(x)$. The uniqueness principle shows that $u_i^n \leq K\tau_\rho^e(x_i)$.
- At the source point, $u_{x_S}^n = 0$. This is true for $n = 0$. Assume that $u_{x_S}^n = 0$. Since $0 \leq u_i^n$, and thanks to the monotonicity property of $\mathcal{H}_\rho$, we have $\mathcal{H}_\rho \leq 0$ at x_S. Since $\tau_\rho^e(x_S) = 0$, we have

$$u_{x_S}^{n+1} = \min(-\Delta t \mathcal{H}_\rho + \Delta t n(x_S), 0) = 0.$$

The last statement is obvious by continuity. □

By applying the result of Theorem 2.2, we conclude the following.

PROPOSITION 5.3. *The solution of the scheme*

$$\begin{cases} \mathcal{H}_\rho(D_{T_1}u, \dots, D_{T_1}u) - n(x_i) = 0 & \textit{for any } x_i, \\ u_{x_S} = 0 & \textit{at } x_S, \\ \max\left(\mathcal{H}_\rho^b(D_{T_1}u, \dots, D_{T_1}u) - n(x_i), K\tau_\rho^e(x_i)\right) & \textit{on the boundary} \end{cases}$$

converges, as $\rho \to 0$, *to the function* (5.9) *when* Ω *is smooth enough.*

5.2.3. Numerical application. We have considered in numerical applications [2] the index n given by a realistic model of the underground of the Gabon gulf, the Marmousi model developed by the French Petroleum Institute (IFP). Since there is no exact solution in closed form for this problem, it is probably more enlightening to consider a more academical problem where an exact solution is known. The computational domain is represented in Figure 14, and the solution at any point M is the distance between the point S and M. A mesh is displayed in Figure 15. The numerical solution is shown in Figure 16. The boundary conditions are very well taken into account: there is no boundary layer, and the isolines of the solution are orthogonal to the circle, as they should be. In Figure 17, we display the isolines of the logarithm of the error between the exact and the computed solutions.

6. Conclusion and summary.

In this paper, we have described two ways of discretizing boundary conditions for first order Hamilton–Jacobi equations for which convergence can be proved. This is done through a boundary numerical Hamiltonian. In the case of convex or concave Hamiltonians, we have given explicit formulas. In the case of a coercive Hamiltonian, we have shown that the natural boundary conditions prevent the appearance of numerical boundary layers. Then we have illustrated the schemes by simple numerical examples. An extension to geophysics is also provided.

Appendix A. Some properties of monotone numerical schemes.

The aim of this section is to provide some properties of the maximum principle type for monotone numerical schemes:

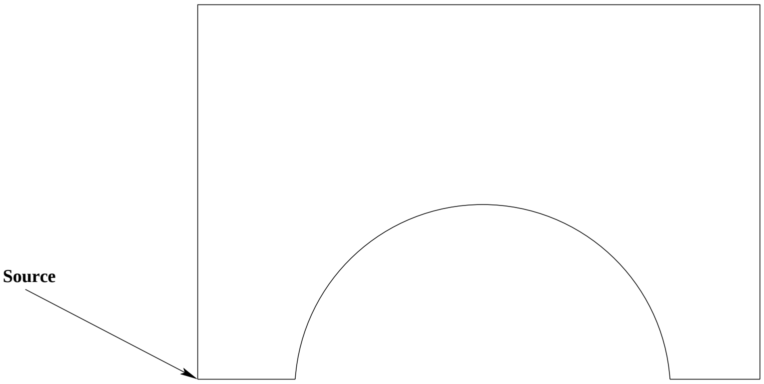


FIG. 14. *Test case for the Soner/source boundary conditions. The Soner condition is imposed everywhere except at the source.*

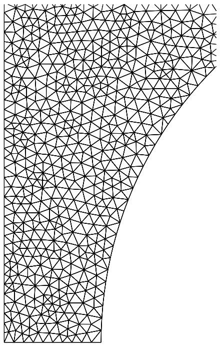


FIG. 15. *Zoom of the mesh around the source. Number of vertices:* 5906; *number of triangles:* 11430.

- For steady problems,

$$\mathcal{H}_\rho(x_i, u_i^n, \{u_j^n, j \in \mathcal{N}_i\}) = 0. \tag{A.1}$$

- For unsteady problems,

$$u_i^{n+1} = G_\rho(x_i, u_i^n, \{u_j^n, j \in \mathcal{N}_i\}; \Delta t) = u_i^n - \Delta t \mathcal{H}_\rho\left(x_i, u_i^n, \{u_j^n, j \in \mathcal{N}_i\}\right). \tag{A.2}$$

These results are useful in section 4. The notation is the same as in section 3.

As in the continuous case, we say that a piecewise linear function u is a discrete subsolution of (A.1) if we have

$$\text{for any } x_i, \quad \mathcal{H}_\rho(x_i, u(x_i), D_{T_1}u, \dots, D_{T_k}u) \le 0.$$

It is a supersolution of (A.1) if

$$\text{for any } x_i, \quad \mathcal{H}_\rho(x_i, u(x_i), D_{T_1}u, \dots, D_{T_k}u) \ge 0.$$

Similarly, let $R > 0$ and $\Delta t \le \Delta t_R$ to ensure the monotonicity of the operator G in (A.2). We say that $u \in \mathcal{C}_R$ is a subsolution of the explicit scheme (A.2) if for all

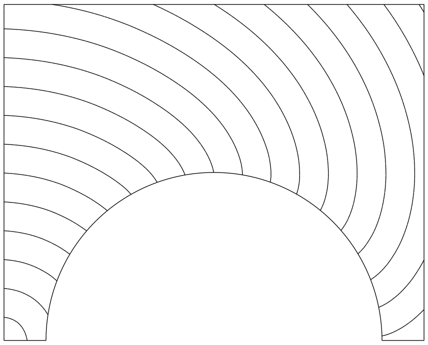


FIG. 16. *Numerical solution.*

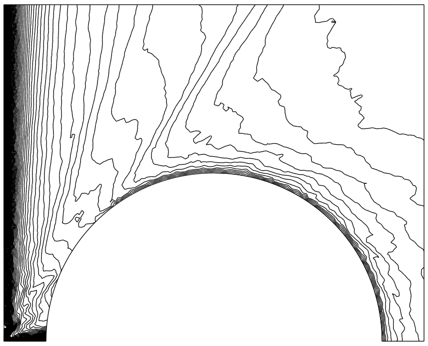


FIG. 17. *Isolines of* $\log_{10}(\tau^{exact} - \tau^{num})$, *max* $= -1.82$.

$n \geq 0$,

$$\text{for any } x_i, \quad \frac{u_i^{n+1} - u_i^n}{\Delta t} + \mathcal{H}_\rho(x_i, u^n(x_i), D_{T_1} u^n, \dots, D_{T_k} u^n) \leq 0.$$

$v \in \mathcal{C}_R$ is a supersolution when the opposite inequality holds. The case of implicit schemes is dealt with the same way.

A solution is obviously a sub- and supersolution, and by maximum principle we understand the following: if u (resp., v) is a sub- (resp., super-)solution of (A.1) such that for any x_i on the boundary of Ω_h we have

$$u(x_i) \leq v(x_i),$$

then the same inequality is true for any node of Ω_h. We say that there is a maximum principle on (A.2) and (A.1) if when u and v are sub- and supersolutions of (A.2) and (A.1) with $u_i^n \leq v_i^n$ on the boundary nodes of Ω_h and for any node of Ω_h at $t = 0$ or t_N, then $u_i^n \leq v_i^n$ everywhere.

We want to show that under the assumptions (H0)–(H1) or (H0)–(H2) below, we have a maximum principle for some classes of schemes of the type

$$G_\rho(x_i, \xi; \{\zeta_l\}_{l \in \mathcal{N}_i}) = 0. \tag{A.3}$$

(H0) The numerical Hamiltonian $\mathcal{H}_\rho$ is monotone increasing in ζ_i and monotone decreasing in $\zeta_l, l \neq i$, for any i. It also satisfies the following: for any i, $\xi \in \mathbb{R}$, and any $(\zeta_1, \dots, \zeta_k)$, $\mathcal{H}_\rho$ is invariant by translation on the ζ_ks.

(H1) For any $R > 0$, for any u, v such that $-R \leq v \leq u \leq R$, for any $p_1, \dots, p_k$ vectors, and for any mesh point x_i, we have

$$\gamma_R(u - v) \leq \mathcal{H}_\rho(x_i, u, p_1, \dots, p_k) - \mathcal{H}_\rho(x_i, v, p_1, \dots, p_k).$$

Here, k is the number of triangles having x_i as vertex.

(H2) The Hamiltonian $\mathcal{H}_\rho$ is convex in the variables $p_1, \dots, p_k$, and there exists a subsolution $\Phi_i, i = 1, \dots, n_s$, and $\alpha < 0$ such that $\mathcal{H}_\rho(x_i, D_{T_1}\Phi, \dots, D_{T_k}\Phi) \leq \alpha$ for any i.

The assumptions (H1) and (H2) are only the discrete analogue of classical assumptions on the Hamiltonian H.

Here, we provide a maximum principle for a monotone Hamiltonian and a *fixed* mesh only. The arguments are too crude to pass to the limit.

A.1. Maximum principle in the steady case.

THEOREM A.1. *We assume that the scheme satisfies* (H0) *and* (H1). *If* $(u_i)_{x_i}$ *and* $(v_i)_{x_i}$ *are sub- (super-)solutions of* (A.3) *in the interior nodes of* Ω_h *and satisfy* $u_i \leq v_i$ *on its boundary nodes, then* $u_i \leq v_i$ *everywhere.*

Proof. Let $R = \max_{x_i, i=1,\dots,n_s}(|u_i|, |v_i|)$ and x_{i_0} be the mesh point where $\{u_i - v_i\}_i$ reaches its maximum. Let us call this maximum M. If x_{i_0} belongs to the boundary, then $M \leq 0$ and we are done. If x_{i_0} is an interior point, then we have

$$\phi = v_{i_0} - u_{i_0} + u = -M + u \leq v$$

on Ω_h, by assumption. Moreover, $\phi(x_{i_0}) = v_{i_0}$.

Since $\mathcal{H}_\rho$ is monotone, we have

$$\begin{aligned} 0 &\leq G_\rho(x_{i_0}, v_{i_0}, v_{i_1}, \dots, vi_k) \\ &\leq G_\rho(x_{i_0}, v_{i_0}, \phi(x_{i_1}), \dots, \phi(x_{i_k})) \\ &= G_\rho(x_{i_0}, \phi(x_{i_0}), \phi(x_{i_1}), \dots, \phi(x_{i_k})) \\ &= \mathcal{H}_\rho(x_{i_0}, D_{T_1}u, \dots, D_{T_k}u). \end{aligned}$$

Then, since u is a subsolution, assuming $M > 0$, we have

$$\gamma_R(u_{i_0} - v_{i_0}) \leq \mathcal{H}_\rho(x_{i_0}, u_{i_0}; D_{T_1}u, \dots, D_{T_k}u) - \mathcal{H}_\rho(x_{i_0}, v_{i_0}; D_{T_1}u, \dots, D_{T_k}u) \leq 0,$$

which is absurd. □

THEOREM A.2. *Under* (H0)–(H2), *there is a maximum principle.*

Proof. We consider u and v a sub- and supersolution of $\mathcal{H}_\rho = 0$ with $u \leq v$ on the boundary of Ω_h. We can assume that $\phi \leq v$ on the boundary of Ω_h, thanks to (H0).

Let $\lambda \in [0, 1]$ and $u_i^\lambda = \lambda u_i + (1 - \lambda)\phi_i$. It is clear that u^λ is a subsolution of $\mathcal{H}_\rho = (1 - \lambda)\alpha$ and $u^\lambda \leq v$ on the boundary of Ω_h. Let us assume that $u^\lambda - v$ reaches its maximum C_λ at an interior point x_{i_0}.

By assumption, we have $\mathcal{H}_\rho(x_{i_0}, v_{i_0}, v_{i_1}, \dots, v_k) \geq 0$ and $\mathcal{H}_\rho(x_{i_0}, u_{i_0}^\lambda, u_{i_1}^\lambda, \dots, u_k^\lambda) \leq (1 - \lambda)\alpha < 0$. The same arguments as in the proof of Theorem A.1 show that

$$(1 - \lambda)\alpha \geq \mathcal{H}_\rho(x_{i_0}, u_{i_0}^\lambda, u_{i_1}^\lambda, \dots) \geq \mathcal{H}_\rho(x_{i_0}, u_{i_0}^\lambda, v_{i_1} + C_\lambda, \dots) = \mathcal{H}_\rho(x_{i_0}, v_{i_0}, v_{i_1}, \dots),$$

and we have a contradiction. Thus, $u^\lambda - v$ is maximum on the boundary, and then

$$u^\lambda \leq v$$

in Ω. By taking the limit when $\lambda \to 1$, we conclude that $u \leq v$ in Ω. □

A.2. Maximum principle in the unsteady case.

Theorem A.3. *If the scheme* (A.2) *is monotone under $\Delta t \leq \Delta t_R$, then we have a maximum principle. If $u^n \in \mathcal{C}_R$ (resp., $v^n \in \mathcal{C}_R$) is a subsolution (resp., supersolution) of* (A.2) *such that $u_i^0 \leq v_i^0$ for all i and $u_i^n \leq v_i^n$ for each $n \geq 0$ and boundary node, then $u_i^n \leq v_i^n$ for all $n \geq 0$ and i.*

The same result holds for an implicit scheme.

Proof. We give the proof for the scheme (A.2). The proof for an implicit scheme is the same. We proceed by induction on n. For $n = 0$, the result is true by assumption. Since $u_i^0 \leq v_i^0$ for the interior points and the boundary points, we have $u_i^1 \leq v_i^1$ for the *interior* nodes. By assumption, $u_i^1 \leq v_i^1$ for the boundary points, and the result follows by induction. □

A.3. A uniqueness principle. We consider the scheme

$$\text{(A.4)} \qquad \begin{cases} \mathcal{H}_\rho^b(x_i, u_i; u_i, \{u_l, l \in \mathcal{N}_i\}) = 0 & \text{if } x_i \text{ interior node} \\ \\ \max(\mathcal{H}_\rho(x_i, u_i; u_i, \{u_l, l \in \mathcal{N}_i\}), u_i - \varphi(x_i)) = 0 & \text{otherwise,} \end{cases}$$

which discretizes the Dirichlet problem

$$\begin{array}{ll} H(x, u(x), Du(x)) = 0 & \text{if } x \in \Omega, \\ u = \phi & \text{otherwise.} \end{array}$$

We have the following result.

Theorem A.4. *Under the assumptions* (H0)–(H1) *or* (H0)–(H2) *for $\mathcal{H}_\rho$ and $\mathcal{H}_\rho^b$, if u (resp., v) is a subsolution (resp., supersolution) of* (A.4), *then $u_i \leq v_i$ for any mesh point x_i.*

A direct consequence of this result is that if (A.4) has a solution, it is unique.

Proof. We consider $M = \max_{x_i}(u_i - v_i)$, and we assume $M > 0$. Since the set x_i is finite, the maximum is reached at x_{i_0}. If x_{i_0} is not on the boundary, then we can repeat the arguments for the discrete maximum principle. Thus we can assume that x_{i_0} is on the boundary, and we have

$$\begin{array}{l} \max(\mathcal{H}_\rho^b(x_{i_0}, u_{i_0}; u_{i_0}, \{u_l, l \in \mathcal{N}_{i_0}\}), u_{i_0} - \varphi(x_{i_0})) \leq 0, \\ \max(\mathcal{H}_\rho^b(x_{i_0}, v_{i_0}; v_{i_0}, \{v_l, l \in \mathcal{N}_{i_0}\}), u_{i_0} - \varphi(x_{i_0})) \geq 0. \end{array}$$

The conditions on u give

$$\mathcal{H}_\rho^b(x_{i_0}, u_{i_0}; u_{i_0}, \{u_l, l \in \mathcal{N}_{i_0}\}) \leq 0 \quad \text{and} \quad u_{i_0} \leq \varphi(x_{i_0}).$$

Those on v give

$$\mathcal{H}_\rho^b(x_{i_0}, u_{i_0}; u_{i_0}, \{u_l, l \in \mathcal{N}_{i_0}\}) \geq 0 \quad \text{or} \quad v_{i_0} \geq \varphi(x_{i_0}).$$

First case. $u_{i_0} \leq \varphi(x_{i_0})$ and $v_{i_0} \geq \varphi(x_{i_0})$. There is nothing to prove; $M \leq 0$.

Second case. $\mathcal{H}_\rho^b(x_{i_0}, u_{i_0}; u_{i_0}, \{u_l, l \in \mathcal{N}_{i_0}\}) \leq 0$ and $\mathcal{H}_\rho^b(x_{i_0}, u_{i_0}; u_{i_0}, \{u_l, l \in \mathcal{N}_{i_0}\}) \geq 0$. By using the same arguments as in the maximum principle, we get an absurdity when $M > 0$.

Thus we have proved that $M \leq 0$, i.e., $u_i \leq v_i$ for any x_i. □

Acknowledgements. A.P. Blanc and G. Barles (University of Tours, France) are acknowledged for many interesting and useful discussions.

REFERENCES

[1] R. Abgrall, *Numerical discretization of first order Hamilton–Jacobi equations on triangular meshes*, Comm. Pure Appl. Math., 49 (1996), pp. 1339–1373

[2] R. Abgrall and J.D. Benamou, *Big ray tracing and eikonal solver on unstructured grids: Application to the computation of a multi-valued travel-time field*, Geophysics, 64 (1999), pp. 230–239.

[3] M. Bardi and S. Osher, *The nonconvex multi-dimensional Riemann problem for Hamilton–Jacobi equations*, SIAM J. Math. Anal., 22 (1991), pp. 344–351.

[4] G. Barles, *Solutions de viscosité des équations de Hamilton–Jacobi*, Math. Appl., Springer Verlag, Paris, 1994.

[5] G. Barles and P.E. Souganidis, *Convergence of approximation schemes for fully nonlinear second order equations*, Asymptot. Anal., 4 (1991), pp. 271–283.

[6] J.D. Benamou, *Big ray tracing: Multivalued travel time field computations using viscosity solutions of the eikonal equation*, J. Comput. Phys., 128 (1996), pp. 463–474.

[7] P.G. Ciarlet and P.A. Raviart, *General Lagrange and Hermite interpolation in $\mathbb{R}^n$ with application to finite element methods*, Arch. Ration. Mech. Anal., 42 (1972), pp. 177–199.

[8] M.G. Crandall and P.-L. Lions, *Two approximations of solutions of Hamilton–Jacobi equations*, Math. Comp., 43 (1984), pp. 1–19.

[9] S. Osher and C.W. Shu, *High–order essentially nonoscillatory schemes for Hamilton–Jacobi equations*, SIAM J. Numer. Anal., 28 (1991), pp. 907–922.

SIAM J. Numer. Anal.
Vol. 41, No. 6, pp. 2262–2293

ENTROPY FORMULATION FOR PARABOLIC DEGENERATE EQUATIONS WITH GENERAL DIRICHLET BOUNDARY CONDITIONS AND APPLICATION TO THE CONVERGENCE OF FV METHODS*

ANTHONY MICHEL† AND JULIEN VOVELLE‡

Abstract. This paper is devoted to the analysis and the approximation of parabolic hyperbolic degenerate problems defined on bounded domains with *nonhomogeneous* boundary conditions. It consists of two parts. The first part is devoted to the definition of an original notion of entropy solutions to the continuous problem, which can be adapted to define a notion of measure-valued solutions, or entropy process solutions. The uniqueness of such solutions is established. In the second part, the convergence of the finite volume method is proved. This result relies on (weak) estimates and on the theorem of uniqueness of the first part. It also entails the existence of a solution to the continuous problem.

Key words. parabolic degenerate equations, boundary conditions, finite volume methods

AMS subject classifications. 35K65, 35F30, 35K35, 65M12

DOI. 10.1137/S0036142902406612

1. Introduction.

Let Ω be an open bounded polyhedral subset of $\mathbb{R}^d$ and $T \in \mathbb{R}_+^*$. Let us denote by Q the set $(0,T)\times\Omega$, and by Σ the set $(0,T)\times\partial\Omega$.

We consider the following parabolic-hyperbolic problem:

$$(1) \qquad \begin{cases} u_t + \operatorname{div}(F(t,x,u)) - \Delta\varphi(u) = 0, & (t,x)\in Q,\\ u(0,x) = u_0(x), & x\in\Omega,\\ u(t,x) = \bar{u}(t,x), & (t,x)\in\Sigma\,. \end{cases}$$

Such an equation of quasilinear advection with degenerate diffusion governs the evolution of the saturation of the wetting fluid in the study of diphasic flow in porous media [GMT96], [Mic01], [EHM01]. In that case, the function φ can be expressed using the capillary pressure and the relative mobilities. The function φ is only supposed to be a nondecreasing Lipschitz continuous function. In particular, the study of problem (1) includes the study of nonlinear hyperbolic problems (cases where $\varphi' = 0$).

The analysis of the approximation of nonlinear hyperbolic problems via the finite volume (FV) method began in the mid 1980s, involving several authors including, for example, Cockburn, Coquel, and LeFloch [CCL95], Szepessy [Sze91], Vila [Vil94], Kröner, Rokyta, and Wierse [KRW96], and Eymard, Gallouët, and Herbin [EGH00]. Results on the convergence of FV schemes for degenerate problems in general came to light in more recent years [EGHM02], [Ohl01]. See also [BGN00], [EK00] for other methods of approximation.

When the function φ is strictly increasing, problem (1) is of parabolic type. In that case, the existence of a unique weak solution is well known. In the case where $\varphi' = 0$, problem (1) is a nonlinear hyperbolic problem, the uniqueness of a weak

[*] Received by the editors April 29, 2002; accepted for publication (in revised form) April 1, 2003; published electronically December 5, 2003.

http://www.siam.org/journals/sinum/41-6/40661.html

[†] Departement de Mathematiques, Université de Montpellier II, CC 051–Place Eugène Bataillon, F-34095 Montpellier cedex 5, France (amichel@math.univ-montp2.fr).

[‡] Université de Provence, Centre de Mathématiques et d'Informatique, F-13453 Marseille, France (vovelle@cmi.univ-mrs.fr).

solution is not ensured, and one has to define a notion of entropy solutions to recover uniqueness [Kru70]. Therefore, it is quite difficult to define a notion of solution in the case where φ is merely a nonincreasing function. In fact, as far as the Cauchy problem in the whole space is concerned, such a definition has been done for a long time, since Volpert and Hudjaev [VH69], but uniqueness with nonlinear parabolic terms has only been proved recently by Carrillo [Car99] (see also [KO01], [KR00]).

Another difficulty in the study of degenerate parabolic problems is analysis of the boundary conditions (see [LBS93], [RG99]). It is not always easy to give a correct formulation of the boundary conditions, or of the way they have to be taken into account. In the case where the function φ is strictly increasing, the classical framework of variational solutions of parabolic equations is enough to satisfy this wish. In the case where $\varphi' = 0$, things are completely different. Even if the (entropy) solution u of problem (1) admits a trace (say, γu) on Σ, the equality $\gamma u = \overline{u}$ on Σ does not necessarily hold. Actually, a condition on Σ can be given, which is known as the BLN condition [BLN79]: this is the right way to formulate boundary conditions in the study of scalar hyperbolic problems. However, the notion of entropy solution to nonlinear Cauchy–Dirichlet hyperbolic problems given by Bardos, LeRoux, and Nédélec is not really suitable to the study of FV schemes since it requires that the solution u be in a space BV (because the trace of u is involved in the formulation of the BLN condition), and it is known that it is difficult to get BV estimates on the numerical approximations given by the FV method on non-Cartesian grids. Actually, Otto gave an integral formulation of entropy solutions to scalar hyperbolic problems with boundary conditions [Ott96], and this indeed allows us to prove the convergence of the FV method [Vov02].

To our knowledge, the problem that we deal with (convergence of the FV method for degenerate parabolic equations with nonhomogeneous boundary conditions) has never been considered before. Nevertheless, in [MPT02], the authors give a definition of entropy solution for which uniqueness and consistency with the parabolic approximation are proved. This definition is not completely in integral form and therefore not suitable for proving the convergence of the FV method, since only poor compactness results are available on the numerical approximation. That is why we give an original definition of the problem (see Definition 3.1). This complete integral formulation includes the definition of Otto but not exactly the one of Carrillo (see the comments that follow Definition 3.1). It is well suited to the study of the convergence of several approximations of problem (1) and is used, for example, in [GMT02] to prove the convergence of a discrete Bhatnagar–Gross–Krook (BGK) model (see also [MPT02] for the parabolic approximation).

Notice that some particular cases have been fully treated: in [EGHM02], the authors prove the convergence of the FV method in the case where $F(x,t,s) = \mathbf{q}(x,t)f(s)$, $\operatorname{div}(\mathbf{q}) = 0$, with $\mathbf{q} \cdot \mathbf{n} = 0$ on Σ. In that case, the boundary condition does not act on the hyperbolic part of the equation. From a technical point of view, this means that the influence of the boundary condition appears in the terms related to the parabolic degenerate part of the equation. These parabolic degenerate terms are estimated by following the methods of Carrillo in [Car99], who deals with homogeneous boundary conditions. On the other hand, in [Vov02], the author proves the convergence of an FV method in the case where $\varphi' = 0$, adapting the ideas of Otto [Ott96]. In that case, the effects of the boundary condition in the hyperbolic equation are the center of the work. In this paper we mix these two precedent approaches to deal with the parabolic degenerate problem with general boundary conditions.

We will make the following assumptions on the data:

(H1) $F : (t, x, s) \mapsto F(t, x, s) \in C^1(\mathbb{R}^+ \times \mathbb{R}^d \times \mathbb{R})\,, \quad \text{div}_x F = 0,$
$\dfrac{\partial F}{\partial s}$ is locally Lipschitz continuous uniformly with respect to (t, x);

(H2) $\varphi : s \mapsto \varphi(s)$ is a nondecreasing Lipschitz continuous function;

(H3) $u_0 : x \mapsto u_0(x) \in L^\infty(\Omega)$; and

(H4) the function $\bar{u} : (x, t) \mapsto \bar{u}(x, t) \in L^\infty(\Sigma)$ and is the trace of a function $\bar{u} \in L^\infty(Q)$with $\varphi(\bar{u}) \in L^2(0, T; H^1(\Omega))\,.$

To prove the convergence of the FV method, we will also assume that the boundary datum satisfies

(H5) the function $\bar{u} : (x, t) \mapsto \bar{u}(x, t) \in L^\infty(\Sigma)$ and is the trace of a function $\bar{u} \in L^\infty(Q)$ with $\varphi(\bar{u}) \in L^2(0, T; H^1(\Omega))\,,\ \nabla\bar{u} \in L^2(Q)\,,\ \bar{u}_t \in L^1(Q)\,.$

In the course of the proof of uniqueness of the entropy process solution (Theorem 4.1), additional hypotheses on the boundary datum are required. Using the notation defined in subsection 4.1, they read

$$\text{(H6)}\ \overline{u}_\Sigma \in W^{1,1}((0, T) \times B \cap Q) \quad \text{and} \quad \Delta\varphi(\overline{u}_\Sigma) \in L^1((0, T) \times B \cap Q)\,.$$

Remark 1.1. As suggested by Porretta [MPT02], hypothesis (H6) may be relaxed as

$$\text{(H6Bis)}\ \overline{u}_\Sigma \in W^{1,1}((0, T) \times B \cap Q) \quad \text{and}$$
$$\Delta\varphi(\overline{u}_\Sigma) \text{ is a bounded Radon measure on } (0, T) \times \Pi\,.$$

We do not give a justification of this assertion now. Indeed, hypothesis (H6) is involved in the proof of Lemma 4.2, and we have waited until Remark 4.1, just after this proof, to specify to what extent hypothesis (H6Bis) is admissible.

Under assumptions (H3)–(H4), there exists $(A, B) \in \mathbb{R}^2$ such that

$$A \le \min\left(\operatorname*{ess\,inf}_\Omega(u_0), \operatorname*{ess\,inf}_Q(\bar{u})\right) \le \max\left(\operatorname*{ess\,sup}_\Omega(u_0), \operatorname*{ess\,sup}_Q(\bar{u})\right) \le B, \tag{2}$$

and we set

$$M = \max\left\{\left|\frac{\partial F}{\partial s}(t, x, s)\right|\,,\ (t, x, s) \in Q \times [A, B]\right\}\,.$$

We introduce the function ζ defined by $\zeta' = \sqrt{\varphi'}$. (This makes sense in view of (H2).) We will derive $L^2(0, T; H^1)$ estimates on nonlinear quantities such as $\zeta(u)$. A simple explanation for this fact is the following. Consider the equation $u_t - \Delta\varphi(u) = 0$ on $(0, T) \times \Omega$. Multiply it by u, and sum the result with respect to $x \in \Omega$. The formal identity $\int_\Omega \nabla\varphi(u) \cdot \nabla u = \int_\Omega |\nabla\zeta(u)|^2$ then leads to $\frac{1}{2}\frac{d}{dt}\int_\Omega u^2 dx + \int_\Omega |\nabla\zeta(u)|^2 \le 0$, from which can be derived an energy estimate.

Notice that the hypothesis $\text{div}_x F = 0$ can be relaxed, and source terms can be considered in the right-hand side of (1).

The assumption that $\bar{u}$ is the trace of an L^∞ function $\bar{u}$ such that $\varphi(\bar{u}) \in L^2(0, T; H^1(\Omega))$ is a necessary condition for the existence of solutions to problem (1);

the additional hypotheses introduced in (H5) are involved in the proofs of different estimates on the approximate solution, defined thanks to the FV method.

As implied at the beginning of this introduction, one of the main points in the study of problem (1) is the definition of a notion of solution suitable for the classical techniques of convergence of FV schemes. This point is specified in section 2. In section 3, we introduce and define a notion of entropy process solutions (a concept similar to the concept of measure-valued solutions), and in section 4 we prove the uniqueness of such solutions (see Theorem 4.1). Section 5 is devoted to the FV scheme used to approximate problem (1); a priori estimates are derived and the convergence is proved.

2. Entropy weak solution. Here, as in the study of purely hyperbolic problems, the concept of weak solutions is not sufficient since the uniqueness of such solutions may fail. Thus, we turn to the notion of weak entropy solutions. The entropy-flux pairs considered in the definition of this solution are the so-called Kruzhkov semi entropy-flux pairs $(\eta_\kappa^\pm, \Phi_\kappa^\pm)$ (see [Car99], [Ser96], [Vov02]). They are defined by the formula

$$\begin{cases}\eta_\kappa^+(s) = (s-\kappa)^+ = s\top\kappa - \kappa,\\ \eta_\kappa^-(s) = (s-\kappa)^- = \kappa - s\bot\kappa,\end{cases} \qquad \begin{cases}\Phi_\kappa^+(t,x,s) = (s-\kappa)^+ = F(t,x,s\top\kappa) - F(t,x,\kappa),\\ \Phi_\kappa^-(t,x,s) = (s-\kappa)^- = F(t,x,\kappa) - F(t,x,s\bot\kappa),\end{cases}$$

with $a\top b = \max(a,b)$ and $a\bot b = \min(a,b)$. Notice that, in the case where κ is considered as a variable, for example when the doubling variable technique of Kruzhkov is used, the entropy-fluxes will be written

$$\Phi^+(x,t,s,\kappa) = \Phi_\kappa^+(t,x,s) \quad \text{and} \quad \Phi^-(x,t,s,\kappa) = \Phi_\kappa^-(t,x,s).$$

Definition 2.1 (entropy weak solution). *A function u of $L^\infty(Q)$ is said to be an entropy weak solution to problem* (1) *if it is a weak solution of problem* (1)*, that is, if $\varphi(u) - \varphi(\overline{u}) \in L^2(0,T;H_0^1(\Omega))$ and*

$$\forall\theta \in \mathcal{C}_c^\infty([\,0,T)\times\Omega)\,, \quad \int_Q u\,\theta_t + (F(t,x,u) - \nabla\varphi(u))\cdot\nabla\theta \;\,dx\,dt \,+ \int_\Omega u_0\,\theta(0,x)\,dx \,= 0, \tag{3}$$

and if it satisfies the following entropy inequalities for all $\kappa \in [A,B]$, for all $\psi \in \mathcal{C}_c^\infty([\,0,T)\times\mathbb{R}^d)$ such that $\psi \geq 0$ and $\operatorname{sgn}^\pm(\varphi(\overline{u}) - \varphi(\kappa))\psi = 0$ a.e. on Σ:

$$\int_Q \eta_\kappa^\pm(u)\,\psi_t + (\,\Phi_\kappa^\pm(t,x,u) - \nabla\,(\varphi(u)-\varphi(\kappa))^\pm\,)\cdot\nabla\psi\;dx\,dt \,+ \int_\Omega \eta_\kappa^\pm(u_0)\,\varphi(0,x)\,dx \\ + M\int_\Sigma \eta_\kappa^\pm(\bar{u})\psi\,d\gamma(x)\,dt \geq 0. \tag{4}$$

Notice that the weak equation (3) is superfluous, for it is a consequence of (4). However, if the function φ were (strictly) increasing, (3) would be enough to define a notion of the solution of problem (1) for which existence and uniqueness hold: in that case, problem (1) would merely be a nonlinear parabolic problem. For general φ, the uniqueness of the solution will be a consequence of the entropy inequalities (4); indeed, the class of Kruzhkov semi entropy-flux pairs is wide enough to ensure the uniqueness of the weak entropy solution, while—and we stress this fact—the class of classical Kruzhkov entropy-flux pairs $s \mapsto |s-\kappa|$ is not.

Also notice that, first, in the homogeneous case $\bar{u} = 0$, the previous definition is slightly different from the original definition given by Carrillo [Car99] and that, second, if $\varphi' = 0$ (problem (1) becomes hyperbolic), then the previous definition of the entropy solution coincides with the definition of a solution suitable for hyperbolic problems; see Otto [Ott96] and [Vov02]. A notion of an entropy solution for degenerate parabolic problems with nonhomogeneous boundary conditions has also been defined by Mascia, Porretta, and Terracina in [MPT02]. It is interesting to notice that, in their definition, they directly require that the entropy condition satisfy the entropy condition on the boundary (14) as stated in Proposition 4.1. We prove that this property (14) is, in fact, a consequence of the entropy inequalities (4) and then follow the main lines of the uniqueness theorem proved in [MPT02].

3. Entropy process solution. The proof of the existence of a weak entropy solution to problem (1) lies in the study of the numerical solution u_D defined by an FV method for problem (1) (see section 5.2). Theorem 5.1 states that the numerical solution satisfies approximate entropy inequalities (see (50)), but the bounds on u_D (a bound in $L^\infty(Q)$ and a bound on the discrete H^1-norm of $\varphi(u_D)$) do not give strong compactness, only weak compactness. Therefore, in order to be able to take the limit of the nonlinear terms of u_D (as $\Phi_\kappa^\pm(u_D)$, in particular), we have to turn to the notion of measure-valued solutions (see DiPerna [DiP85], Szepessy [Sze91]) or, equivalently, to the notion of entropy process solution defined by Eymard, Gallouët, and Herbin [EGH00]. In light of the following theorem, it appears that the notion of entropy process solution is indeed well suited to compensate for the weakness of the compactness estimates on the approximate solution u_D and to deal with nonlinear expressions of u_D.

Theorem 3.1 (nonlinear convergence for the weak-$\star$ topology). *Let $\mathcal{O}$ be a Borel subset of $\mathbb{R}^m$, R be positive, and (u^n) be a sequence of $L^\infty(\mathcal{O})$ such that, for all $n \in \mathbb{N}$, $||u^n||_{L^\infty} \le R$. Then there exists a subsequence, still denoted by (u^n) and $\mu \in L^\infty(\mathcal{O} \times (0,1))$, such that*

$$\forall g \in \mathcal{C}(\mathbb{R}), \quad g(u^n) \longrightarrow \int_0^1 g(\mu(.,\alpha))\, d\alpha \quad \text{in } L^\infty(\mathcal{O}) \text{ weak-}\star\,.$$

Now the notion of an entropy process solution can be defined.

Definition 3.1 (weak entropy process solution). *Let u be in $L^\infty(Q \times (0,1))$. The function u is said to be an entropy process solution to problem* (1) *if*

$$\varphi(u) - \varphi(\overline{u}) \in L^2(0,T; H_0^1(\Omega)) \tag{5}$$

and if u satisfies the following entropy inequalities for all $\kappa \in [A,B]$, for all $\psi \in \mathcal{C}_c^\infty([\,0,T) \times \mathbb{R}^d)$ such that $\psi \ge 0$ and $\operatorname{sgn}^\pm(\varphi(\overline{u}) - \varphi(\kappa))\psi = 0$ a.e. on Σ:

$$\begin{aligned}\int_Q \int_0^1 & \eta_\kappa^\pm(u(t,x,\alpha))\, \psi_t(t,x) \\ & + \left(\Phi_\kappa^\pm(t,x,u(t,x,\alpha)) - \nabla\left(\varphi(u)(t,x) - \varphi(\kappa)\right)^\pm\right) \cdot \nabla\psi(t,x) d\alpha dx dt \\ & + \int_\Omega \eta_\kappa^\pm(u_0)\, \psi(0,x)\, dx + M \int_\Sigma \eta_\kappa^\pm(\bar{u})\psi\, d\gamma(x) dt \ge 0\,.\end{aligned} \tag{6}$$

Notice that if the function u is an entropy process solution of problem (1), then it satisfies condition (5), which means in particular that $\varphi(u)$ does not depend on the last variable α and is denoted by $\varphi(u)(t,x)$.

Notation. We set $\mathcal{Q} = Q \times (0,1)$.

We will now show that any entropy *process* solution actually reduces to an entropy *weak* solution.

4. Uniqueness of the entropy process solution.

THEOREM 4.1 (uniqueness of the entropy process solution). *Let u, $v \in L^\infty(Q \times (0,1))$ be two entropy process solutions of problem* (1) *in accordance with Definition* 3.1. *Suppose that Ω is either a polyhedral open subset of $\mathbb{R}^d$ or a strong $\mathcal{C}^{1,1}$ open subset of $\mathbb{R}^d$, and assume hypotheses* (H1), (H2), (H3), (H4), *and* (H6) *(or* (H6Bis)*). Then there exists a function $w \in L^\infty(Q)$ such that*

$$u(t,x,\alpha) = w(t,x) = v(t,x,\beta) \text{ for almost every } (t,x,\alpha,\beta) \in Q \times (0,1)^2 .$$

COROLLARY 4.1 (uniqueness of the weak entropy solution). *If Ω is either a polyhedral open subset of $\mathbb{R}^d$ or a strong $\mathcal{C}^{1,1}$ open subset of $\mathbb{R}^d$, and under hypotheses* (H1), (H2), (H3), (H4), *and* (H6) *(or* (H6Bis)*), problem* (1) *admits at most one weak entropy solution.*

In the case where Ω is a polyhedral open subset of $\mathbb{R}^d$, the proof of Theorem 4.1 is slightly more complicated than the proof in the case where Ω is a strong $\mathcal{C}^{1,1}$ open subset of $\mathbb{R}^d$. Besides, although the study of the FV method applied to (1) relies on Theorem 4.1 only in the case of Ω polyhedral, we wish to specify the validity of Theorem 4.1 when Ω is $\mathcal{C}^{1,1}$. Indeed, problem (1) may of course be posed on such an open set, and, in that case, Theorem 4.1 would be one of the major steps in the proof of the convergence of such an approximation, as for the vanishing viscosity approximation, for example.

We therefore explain the proof of Theorem 4.1 in the case where Ω is $\mathcal{C}^{1,1}$ and then indicate how to adapt it to the case where Ω is a polyhedral open subset of $\mathbb{R}^d$ (see subsection 4.6).

4.1. Proof of Theorem 4.1: Definitions and notation.

4.1.1. Localization near the boundary. We suppose that Ω is a strong $\mathcal{C}^{1,1}$ open subset of $\mathbb{R}^d$. In that case, there exists a finite open cover $(B_\nu)_{0,\dots,N}$ of $\overline{\Omega}$ and a partition of unity $(\lambda_\nu)_{0,\dots,N}$ on $\overline{\Omega}$ subordinate to $(B_\nu)_{0,\dots,N}$ such that, for $\nu \geq 1$, up to a change of coordinates represented by an orthogonal matrix A_ν, the set $\Omega \cap B_\nu$ is the epigraph of a $\mathcal{C}^{1,1}$-function $f_\nu : \mathbb{R}^{d-1} \to \mathbb{R}$; that is,

$$\begin{aligned} \Omega \cap B_\nu &= \{x \in B_\nu \ ; \ (A_\nu x)_d > f_\nu(\overline{A_\nu x})\} \quad \text{and} \\ \partial\Omega \cap B_\nu &= \{x \in B_\nu \ ; \ (A_\nu x)_d = f_\nu(\overline{A_\nu x})\} , \end{aligned}$$

where $\overline{y}$ stands for $(y_i)_{1,d-1}$ if $y \in \mathbb{R}^d$.

Until the end of the proof of Theorem 4.1, the problem will be localized with the help of a function λ_ν. We drop the index ν and, for the sake of clarity, suppose that the change of coordinates is trivial: $A = I_d$. We denote by $\Pi = \{\bar{x}, x \in \Omega \cap B\} \subset \mathbb{R}^{d-1}$ the projection of $B \cap \Omega$ onto the $(d-1)$ first components, and $\Pi_\lambda = \{\bar{x}, x \in \text{supp}\,(\lambda) \cap \Omega\}$ (see Figure 1). If a function ψ is defined on Σ, we denote by ψ_Σ the function defined on $[0,T) \times B \cap Q$ by $\psi_\Sigma(t,x) = \psi(t,\bar{x},f(\overline{x}))$. Notice that the function ψ_Σ does not depend on x_d and that, by abusing the notation, we shall also denote by ψ_Σ the restriction of ψ_Σ to $[0,T) \times \Pi$. In the same way, if L_i is defined on $[0,T] \times \Pi$, we also denote by L_i the function defined on $[0,T) \times B \cap Q$ by $L_i(t,x) = L_i(t,\overline{x})$.

4.1.2. Weak notion of trace. An important step in the proof of the uniqueness of entropy process solutions is the derivation of the condition satisfied by any entropy

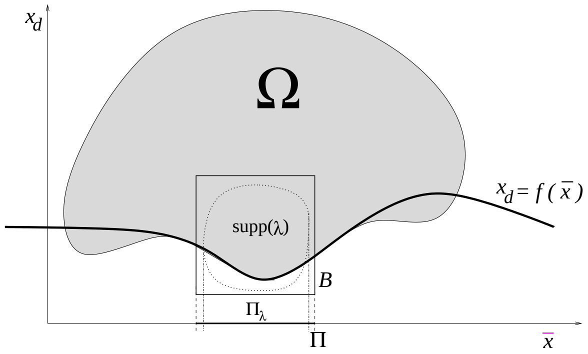


FIG. 1. *Localization by λ in the ball B.*

process solution on the boundary of the domain. This condition is the matter of Proposition 4.1. (It can be viewed as a kind of BLN condition [BLN79], balanced by second order terms issued from the degenerate parabolic part of the equation of (1).) In the course of the proof of Proposition 4.1, we need to define the normal trace of certain fluxes $(\Phi_\kappa^+(t,x,u) - \nabla(\varphi(u) - \varphi(\kappa))^+$, among others, for example) and, more precisely, to ensure the consistency of this definition of the normal trace with different approximations. For that purpose, we turn to the work of Chen and Frid [CF02]. Adapted to our context, the main theorem of [CF02] is the following.

THEOREM 4.2 (see Chen and Frid [CF02]). *Recall that $Q = (0,T)\times\Omega$, and denote by ν the outward unit normal to Q. Let $\mathcal{F} \in (L^2(Q))^{d+1}$ be such that* div$\mathcal{F}$ *is a bounded Radon measure on Q. Then there exists a linear functional $\mathcal{T}_\nu$ on $W^{1/2,2}(\partial Q)\cap\mathcal{C}(\partial Q)$ which represents the normal traces $\mathcal{F}\cdot\nu$ on ∂Q in the sense that, first, the following Gauss–Green formula holds: for all $\psi \in \mathcal{C}_c^\infty(\overline{Q})$,*

$$\langle \mathcal{T}_\nu, \psi\rangle = \int_Q \psi \,\mathrm{div}\mathcal{F} + \int_Q \nabla\psi\cdot\mathcal{F}. \tag{7}$$

Second, $\langle \mathcal{T}_\nu, \psi\rangle$ depends only on $\psi_{|\partial Q}$, while, third, if (B,λ,f) is as above (subsection localization near the boundary), then for all $\psi \in \mathcal{C}_c^\infty([0,T)\times\overline{\Omega})$,

$$\begin{aligned}\langle \mathcal{T}_\nu, \psi\lambda\rangle = -\lim_{s\to 0}\frac{1}{s}\Bigg(&\int_s^T\int_\Pi\int_{f(\overline{x})}^{f(\overline{x})+s}\mathcal{F}\cdot\begin{pmatrix}-\nabla f(\overline{x})\\ 1\\ 0\end{pmatrix}\psi\lambda\,dx_d\,d\overline{x}\,dt\\ &+\int_0^s\int_\Omega\mathcal{F}\cdot\begin{pmatrix}0\\0\\1\end{pmatrix}\psi\lambda\,dx\,dt\Bigg).\end{aligned} \tag{8}$$

Let u be an entropy weak solution of problem (1). The entropy inequality (4) shows that the divergence of the field

$$\mathcal{F}_\kappa^+(t,x) = \begin{pmatrix}(u-\kappa)^+\\ \Phi_\kappa^+(t,x,u) - \nabla(\varphi(u)-\varphi(\kappa))^+\end{pmatrix}$$

is a bounded Radon measure on Q. This field belongs to $(L^2(Q))^{d+1}$, and according to the previous theorem, there exists a linear functional $\mathcal{T}_{\nu,\kappa}^+$ on $W^{1/2,2}(\partial Q)\cap\mathcal{C}(\partial Q)$ which represents $\mathcal{F}_\kappa^+(t,x)\cdot\nu$. Then, to define a notion of the normal trace of the flux

$\Phi^+_\kappa(t,x,u) - \nabla(\varphi(u)-\varphi(\kappa))^+$, we set

$$\langle \mathcal{T}^+_{n,\kappa}, \psi\rangle = \langle \mathcal{T}^+_{\nu,\kappa}, \psi\rangle + \int_\Omega (u_0-\kappa)^+\psi(0,x)\,dx \qquad \forall \psi \in \mathcal{C}^\infty_c([0,T)\times\overline{\Omega})\,. \tag{9}$$

This definition makes sense because the entropy weak solution assumes the values of the initial data u_0:

$$\lim_{s\to 0}\frac{1}{s}\int_0^s\int_\Omega (u-\kappa)^+\psi\,dx\,dt = \int_\Omega (u_0-\kappa)^+\psi(0,x)\,dx\,,$$

as can be seen by choosing $\frac{s-t}{s}\chi_{(0,s)}(t)\psi$ as a test-function in (4). In particular, $\langle \mathcal{T}^+_{n,\kappa}, \psi\lambda\rangle$ depends only on $\psi_{|\Sigma}$, and from (8) we can derive the formula

$$\begin{aligned}&\langle \mathcal{T}^+_{n,\kappa}, \psi\lambda\rangle\\ &= -\lim_{s\to 0}\frac{1}{s}\int_s^T\int_\Pi\int_{f(\overline{x})}^{f(\overline{x})+s}(\Phi^+_\kappa(t,x,u)-\nabla(\varphi(u)-\varphi(\kappa))^+)\cdot\begin{pmatrix}-\nabla f(\overline{x})\\ 1\end{pmatrix}\psi\lambda\,dx_d\,d\overline{x}\,dt\end{aligned}$$

for all $\psi\in\mathcal{C}^\infty_c([0,T)\times\overline{\Omega})$.

4.1.3. Mollifiers ρ_n and the cut-off function ω_ε. Technically, the heart of the proof of uniqueness is the doubling of variables. This technique involves mollifiers, which are defined as $\rho_n(t) = n\rho(nt)$, where ρ is a nonnegative function of $\mathcal{C}^\infty_c(-1,0)$ such that $\int_{-1}^0 \rho(t)\,dt = 1$. (Notice that the the support of the function ρ is located to the left of zero.) For ε a positive number, ρ_ε naturally denotes the map $t\mapsto \frac{1}{\varepsilon}\rho(\frac{t}{\varepsilon})$, and we define $R_n(t) = \int_{-\infty}^{-t}\rho_n(s)\,ds$. Since the technique of doubling of variables interferes with a certain evaluation of the boundary behavior of the entropy process solution (described by (14)), we need to define a cut-off function ω_ε built upon the sequence of mollifiers. We set

$$\omega_\varepsilon(x) = \int_{f(\overline{x})-x_d}^0 \rho_\varepsilon(z)\,dz = \int_{\frac{f(\overline{x})-x_d}{\varepsilon}}^0 \rho(z)\,dz. \tag{10}$$

On $\Omega\cap B$, the function ω_ε vanishes in a neighborhood of $\partial\Omega$ and equals 1 if $\mathrm{dist}(x,\partial\Omega) > \varepsilon$; in particular, $\omega_\varepsilon\to 1$ in $L^1(\Omega\cap B)$ and, if $\psi\in H^1(\Omega)$, then

$$\int_\Omega \lambda\,\psi\cdot\nabla\omega_\varepsilon = -\int_\Omega \mathrm{div}(\lambda\,\psi)\,\omega_\varepsilon \overset{\varepsilon\to 0}{\to} -\int_\Omega \mathrm{div}(\lambda\,\psi) = -\int_{\partial\Omega}\lambda\,\psi\cdot\mathbf{n}\,.$$

Roughly speaking, if $F:\Omega\to\mathbb{R}^d$, then $-F\cdot\nabla\omega_\varepsilon$ approaches the normal trace $F\cdot\mathbf{n}$. To make this idea more precise, for the field $F = \Phi^+_\kappa(t,x,u) - \nabla(\varphi(u)-\varphi(\kappa))^+$ we call upon the notion of normal trace defined above (subsection 4.1.2). Let $\psi\in\mathcal{C}^\infty_c([0,T)\times\overline{\Omega})$. Since $\psi = \psi(1-\omega_\varepsilon)$ on Σ, $\langle \mathcal{T}^+_{n,\kappa}, \psi\lambda\rangle = \langle \mathcal{T}^+_{n,\kappa}, \psi\lambda(1-\omega_\varepsilon)\rangle$. The definition of $\mathcal{T}^+_{n,\kappa}$ (see (9)) and the Gauss–Green formula (7) yield

$$\langle \mathcal{T}^+_{n,\kappa}, \psi\lambda\rangle = \int_Q \psi(1-\omega_\varepsilon)\lambda\,\mathrm{div}\mathcal{F}^+_\kappa + \int_Q \nabla(\psi(1-\omega_\varepsilon)\lambda)\cdot\mathcal{F}^+_\kappa + \int_\Omega (u_0-\kappa)^+\psi(1-\omega_\varepsilon)\lambda\,dx.$$

Since $0\le 1-\omega_\varepsilon\le 1$ and $\omega_\varepsilon(x)\to 1$ for all $x\in\Omega\cap B$, the dominated convergence theorem ensures that $\lim_{\varepsilon\to 0}\int_Q \psi(1-\omega_\varepsilon)\lambda\,\mathrm{div}\mathcal{F}^+_\kappa = 0$ and

$$\langle \mathcal{T}^+_{n,\kappa}, \psi\lambda\rangle = -\lim_{\varepsilon\to 0}\int_Q [\Phi^+_\kappa(t,x,u) - \nabla(\varphi(u)-\varphi(\kappa))^+]\cdot\nabla\omega_\varepsilon\,\psi\lambda\,dx\,dt\,.$$

4.1.4. Otto entropy-fluxes. Let $u \in L^\infty(Q \times (0,1))$ be an entropy process solution of problem (1) and $\kappa \in [A,B]$. Set $\Phi = \Phi^+ + \Phi^-$. We denote by $\mathcal{G}_x(t,x,u,\kappa)$ the quantity

$$\mathcal{G}_x(t,x,u,\kappa) = \Phi(t,x,u(t,x,\alpha),\kappa) - \nabla_x|\varphi(u)(t,x) - \varphi(\kappa)|\,. \tag{11}$$

For $w \in \mathbb{R}$, the function $\mathcal{F}_\varphi$ is defined by the formula

$$\mathcal{F}_\varphi(t,x,u,\kappa,w) = \mathcal{G}_x(t,x,u,\kappa) + \mathcal{G}_x(t,x,u,w) - \mathcal{G}_x(t,x,\kappa,w)\,. \tag{12}$$

4.2. A result of approximation.

LEMMA 4.1. *Let U be a bounded open subset of $\mathbb{R}^q$, $q \geq 1$. If $f \in L^\infty \cap BV(U)$, then, given $\varepsilon > 0$, there exists $g \in \mathcal{C}(\overline{U})$ such that*

$$g \geq f \quad \text{a.e. on } U \qquad \text{and} \qquad \int_U (g(x) - f(x))\,dx < \varepsilon\,.$$

This result may be false if $f \notin BV(U)$ (consider $f = \mathbb{1}_{\mathbb{Q}\cap(0,1)}$ on $U = (0,1)$), but this is not a necessary condition, because, on $U = (0,1)$, the function $f = \mathbb{1}_K$, where K is the triadic Cantor, can be approximated in $L^1(0,1)$ by continuous functions g such that $g \geq f$ a.e. Indeed, we claim that, if E is a measurable subset of U, then $f = \mathbb{1}_E$ satisfies the conclusion of Lemma 4.1 if and only if

$$m(E) = \inf\{m(K)\,;\ E \subset K\,,\ K \text{ compact}\}\,. \tag{13}$$

(Here, m denotes the Lebesgue measure on $\mathbb{R}^q$.)

Before proving Lemma 4.1, let us justify this assertion. If (13) holds, then, given $\varepsilon > 0$, there exists a compact K of U such that $E \subset K$ and $m(K \setminus E) < \varepsilon$. Since the Lebesgue measure is regular, there exists an open subset V of U such that $K \subset V \subset \overline{V} \subset U$ and $m(V \setminus K) < \varepsilon$. Then the function $g : x \mapsto d(x, \mathbb{R}^q \setminus V)/(d(x,K) + d(x, \mathbb{R}^q \setminus V))$ is continuous on $\mathbb{R}^q$, $g \geq \mathbb{1}_E$, and $\int_U (g - \mathbb{1}_E) < 2\varepsilon$.

Conversely, suppose that, given $\varepsilon > 0$, there exists $g \in \mathcal{C}(\overline{U})$ such that $g \geq \mathbb{1}_E$ and $\int_U (g - \mathbb{1}_E) < \varepsilon$. Then $K = \{x \in \overline{U}\,;\ g(x) \geq 1\}$ is compact, $E \subset K$, and $m(K \setminus E) < \varepsilon$.

Proof of Lemma 4.1. Notice that, if E is a measurable subset of U such that $m(\partial E) = 0$, then (13) holds (consider the compact $\overline{E}$). If E is a level set of a BV function, then E has almost surely a finite perimeter and, consequently, $m(\partial E) = 0$, which ensures that $\mathbb{1}_E$ satisfies the conclusion of Lemma 4.1. This result may be seen as the heart of the proof. Indeed, first suppose that $0 \leq f(x) \leq 1$ for every $x \in U$. For $t \in [0,1]$, set $E_t = \{x \in U\,;\ f(x) < t\}$. Then, for almost every t, E_t is a set with finite perimeter since $f \in BV(U)$. Let (t_n) be a sequence of reals dense in $[0,1]$ and such that $t_1 = 1$; E_{t_n} is a set with finite perimeter for every n. We will define a sequence of simple functions $\theta_n = \sum_{i=1}^n \alpha_i^n \mathbb{1}_{A_i^n}$ which approximate f from above and such that each set A_i^n is built upon the level sets E_{t_i}. To that purpose, first define $\theta_1(x) = 1$ for all $x \in U$. If $n > 1$, let $\{k_1, \dots, k_n\}$ be an enumeration of $\{1, \dots, n\}$ such that $t_{k_1} > \cdots > t_{k_n}$. Set

$$\begin{aligned} &A_i^n = E_{t_{k_i}} \setminus E_{t_{k_{i+1}}} \quad \text{if} \quad 1 \leq i < n\,,\\ &A_n^n = E_{t_{k_n}} \end{aligned}$$

and $\theta_n = \sum_{i=1}^n t_{k_i} \mathbb{1}_{A_i^n}$. Notice that $(A_i^n)_{1\leq i\leq n}$ is a partition of U and that $A_i^n \subset E_{t_{k_i}}$; therefore, if $x \in U$, say $x \in A_i^n$, then $\theta_n(x) = t_{k_i} > f(x)$ and $\theta_n \geq f$. Besides, the sequence $(E_{t_{k_i}})_{1\leq i\leq n}$ is decreasing, and this, together with the definition of A_i^n, ensures that $\theta_n(x) \leq t_i$ if $x \in E_{t_i}$ for $1 \leq i \leq n$. Now, given $x \in U$ and $\varepsilon > 0$, there exists n_0 such that $f(x) + \varepsilon > t_{n_0} > f(x)$. Then, for every $n \geq n_0$, $x \in E_{t_{n_0}}$ and, consequently, $\theta_n(x) \leq t_{n_0} < f(x) + \varepsilon$. Thus, (θ_n) converges to f everywhere on U (in fact, the convergence is monotone, but we do not prove this fact), and, since $0 \leq \theta_n \leq 1$, the dominated convergence theorem shows that

$$\lim_{n\to+\infty} \int_U \theta_n - f = 0\,.$$

However, for each fixed n, the function θ_n satisfies the conclusion of the lemma. Indeed, let $\varepsilon > 0$ be fixed. Since $E_{t_{k_{i+1}}} \subset E_{t_{k_i}}$, we have $\mathbb{1}_{A_i^n} = \mathbb{1}_{E_{t_{k_i}}} - \mathbb{1}_{E_{t_{k_{i+1}}}}$. The functions $\mathbb{1}_{E_{t_{k_i}}}$ and $\mathbb{1}_{E_{t_{k_{i+1}}}}$ are in $BV(U)$, by the definition of a set with finite perimeter. Thus $\mathbb{1}_{A_i^n}$ is BV too, and A_i^n is a set with finite perimeter. As noticed in the beginning of the proof, A_i^n satisfies (13), and there exists $g_i \in \mathcal{C}(\overline{U})$ such that $g_i \geq t_{k_i}\,\mathbb{1}_{A_i^n}$ and $\int_U (g_i - t_{k_i}\,\mathbb{1}_{A_i^n}) < \varepsilon/n$. Moreover, we can suppose that $g_i \leq t_{k_i}$ for every i. Set $g = \max_{1\leq i\leq n} g_i$. The function g is continuous on $\overline{U}$, and $g \geq \theta_n$ on U by construction. It remains to compute $||g - \theta_n||_{L^1(U)}$. If $x \in A_i^n$, then $g_i(x) = t_{k_i}$, and the condition $g_j \leq t_{k_j}$ enforces the maximum of the $g_j(x)$ to be reached for $j \in \{i, \dots, n\}$. We then have

$$(g - \theta_n)(x) = g_j(x) - t_{k_i} \leq g_j(x) - t_{k_j}\,\mathbb{1}_{A_j^n}(x).$$

Indeed, if $j = i$, this is obvious, and if $j > i$, we have $\mathbb{1}_{A_j^n}(x) = 0$, while $t_{k_i} \geq 0$. Consequently, $(g - \theta_n)(x) \leq \sum_{i=1}^n (g_j - t_{k_j}\,\mathbb{1}_{A_j^n})(x)$ and $\int_U (g - \theta_n) < n \times \varepsilon/n = \varepsilon$. If n has been chosen such that $\int_U (\theta_n - f) < \varepsilon$, then g is relevant to the conclusion of the lemma.

We suppose that $0 \leq f(x) \leq 1$ for every $x \in U$. For a general function $f \in L^\infty \cap BV(U)$, we can suppose, after an adequate modification of the function on a set of negligible measure, that $-M \leq f(x) \leq M$ for every $x \in U$, where $M = ||f||_{L^\infty(U)}$. Then we consider the function $f_1 = (f + M)/(2M)$. Given $\varepsilon > 0$, there exists $g_1 \in \mathcal{C}(\overline{U})$ such that $g_1(x) \geq f_1(x)$ and $||g_1 - f_1||_{L^1(U)} < \varepsilon/(2M)$ and $g = 2Mg_1 - M$ is convenient. $\square$

4.3. Proof of Theorem 4.1 (preliminary): Boundary condition.

PROPOSITION 4.1 (boundary condition). *Let $u \in L^\infty(Q \times (0,1))$ be an entropy process solution of problem* (1)*, and let $\mathcal{F}_\varphi$ be defined by* (12)*. Assume hypotheses* (H1)*,* (H2)*,* (H3)*,* (H4)*, and* (H6) *(or* (H6Bis)*). Then, for all $\kappa \in [A, B]$, for all nonnegative $\psi \in \mathcal{C}_c^\infty([0,T) \times \mathbb{R}^d)$,*

$$\lim_{\varepsilon\to 0} \int_Q \mathcal{F}_\varphi(t, x, u(t, x, \alpha), \kappa, \overline{u}_\Sigma(t, x)) \cdot \nabla\omega_\varepsilon(x)\,\psi(t, x)\,\lambda(x)\,d\alpha\,dx\,dt \;\leq 0\,. \tag{14}$$

In the case of a purely hyperbolic problem ($\varphi' = 0$), inequality (14) is the boundary condition written by Otto [Ott96], equivalent to the BLN condition [BLN79] for BV solutions. If the problem is strictly parabolic (that is, $\varphi'(u) \geq \Phi_{min} > 0$), then inequality (14) is trivially satisfied by any weak solution of the problem (1). In [MPT02], the condition (14) is listed among the conditions that an entropy solution should satisfy *by definition.* We refer to [MPT02] for a complete discussion of (14).

Proof of Proposition 4.1. We first aim to prove the following result: for every $\tilde\kappa \in [A,B]$, for every nonnegative $\psi \in \mathcal{C}_c^\infty([0,T)\times\mathbb{R}^d)$,

(15)

$$\lim_{\varepsilon\to 0}\int_Q [\Phi^+(t,x,u,\tilde\kappa\top\overline{u}_\Sigma) - \nabla(\varphi(u)-\varphi(\tilde\kappa\top\overline{u}_\Sigma))^+]\cdot\nabla\omega_\varepsilon(x)\,\psi(t,x)\,\lambda(x)\,d\alpha\,dx\,dt \ \le 0.$$

Fix $\tilde\kappa\in[A,B]$. In subsections 4.1.2 and 4.1.3, we defined a notion of normal trace for the flux $\Phi^+_\kappa(t,x,u)-\nabla(\varphi(u)-\varphi(\kappa))^+$ when u is an entropy weak solution of problem (1). Of course, the same can be done when u is an entropy process solution of problem (1); this time just consider the field $\mathcal{F}^+_\kappa$ defined by

$$\mathcal{F}^+_\kappa = \begin{pmatrix} \int_0^1 (u-\kappa)^+\,d\alpha \\ \int_0^1(\Phi^+_\kappa(t,x,u)-\nabla(\varphi(u)-\varphi(\kappa))^+)d\alpha \end{pmatrix}.$$

Moreover, if $\mathcal{T}^+_{n,\kappa}$ still denotes the normal trace of the spatial part of $\mathcal{F}^+_\kappa$, for all $\psi\in\mathcal{C}_c^\infty([\,0,T)\times\mathbb{R}^d)$,

(16)

$$\langle\mathcal{T}^+_{n,\kappa},\psi\lambda\rangle$$
$$=-\lim_{s\to 0}\frac{1}{s}\int_s^T\int_0^1\int_\Pi\int_{f(\overline{x})}^{f(\overline{x})+s}(\Phi^+_\kappa(t,x,u)-\nabla(\varphi(u)-\varphi(\kappa))^+)\cdot\begin{pmatrix}-\nabla f(\overline{x})\\ 1\end{pmatrix}\psi\lambda\,dx_d\,d\overline{x}\,dt\,d\alpha$$

and

$$\text{(17)}\quad \langle\mathcal{T}^+_{n,\kappa},\psi\lambda\rangle = -\lim_{\varepsilon\to 0}\int_Q[\Phi^+_\kappa(t,x,u)-\nabla(\varphi(u)-\varphi(\kappa))^+]\cdot\nabla\omega_\varepsilon\,\psi\lambda\,dx\,dt\,d\alpha\,.$$

Therefore, if ψ is a nonnegative function of $\mathcal{C}_c^\infty([\,0,T)\times\mathbb{R}^d)$ such that $\mathrm{sgn}^+(\varphi(\overline{u})-\varphi(\kappa))\overline{\psi}=0$ a.e. on $(0,T)\times\partial\Omega$, then, choosing $\psi(1-\omega_\varepsilon)$ as a test-function in (6), we get

$$\text{(18)}\qquad -\langle\mathcal{T}^+_{n,\kappa},\psi\lambda\rangle \le M\int_\Sigma(\overline{u}-\kappa)^+\,\psi\,\lambda\,d\gamma(x)\,dt\,.$$

Now, we intend to define a notion of normal trace for the flux $\Phi^+(t,x,u,\overline{u}_\Sigma\top\tilde\kappa)-\nabla(\varphi(u)-\varphi(\overline{u}_\Sigma\top\tilde\kappa))^+$. To that purpose, we set

$$\text{(19)}\qquad \overline{\mathcal{F}}^+ = \begin{pmatrix}\int_0^1(u-\overline{u}_\Sigma\top\tilde\kappa)^+\,d\alpha \\ \int_0^1(\Phi^+(t,x,u,\overline{u}_\Sigma\top\tilde\kappa)-\nabla(\varphi(u)-\varphi(\overline{u}_\Sigma\top\tilde\kappa))^+)d\alpha\end{pmatrix},$$

and we prove the following lemma.

Lemma 4.2. *Let $u\in L^\infty(Q\times(\,0,1))$ be an entropy process solution of problem* (1)*, and let the field $\overline{\mathcal{F}}^+\in(L^2((0,T)\times B\cap Q))^{d+1}$ be defined by* (19)*. Assume hypotheses* (H1)*,* (H2)*,* (H3)*,* (H4)*, and* (H6) *(or* (H6Bis)*). Then, for every open subset D of B such that $\overline{D}\subset B$, the divergence of $\overline{\mathcal{F}}^+$ is a bounded Radon measure on $(0,T)\times D\cap Q$.*

Proof of Lemma 4.2. Set $g=\partial_t\,\overline{u}_\Sigma+\mathrm{div}_x\,F(t,x,\overline{u}_\Sigma)-\Delta\,\varphi(\overline{u}_\Sigma)$. From hypothesis (H6) we have $g\in L^1((0,T)\times B\cap Q)$, and the function $\overline{u}_\Sigma$ (which, we recall, belongs to $W^{1,1}((0,T)\times B\cap Q)))$ can be seen as an entropy solution of the equation $\partial_t\,w+$

$\operatorname{div}_x F(t,x,w) - \Delta\,\varphi(w) = g$ with unknown w. The identity $(\overline{u}_\Sigma \top \tilde{\kappa} - \kappa)^- = (\overline{u}_\Sigma - \kappa)^- - (\overline{u}_\Sigma - \tilde{\kappa} \bot \kappa)^-$ ensures that the function $\overline{u}_\Sigma \top \tilde{\kappa}$ satisfies the entropy inequality

$$\int_Q \left[(\overline{u}_\Sigma \top \tilde{\kappa} - \kappa)^-\, \theta_t + [\Phi^-(t,x,\overline{u}_\Sigma \top \tilde{\kappa}, \kappa) - \nabla(\varphi(\overline{u}_\Sigma \top \tilde{\kappa}) - \varphi(\kappa))^-] \cdot \nabla\theta \right] d\alpha\, dx\, dt$$
$$+ \int_\Omega (\overline{u}_\Sigma \top \tilde{\kappa}(0,x) - \kappa)^-\, \theta(0)\, dx + \int_Q \operatorname{sgn}^-(\overline{u}_\Sigma \top \tilde{\kappa} - \kappa)\, g\, \theta\, dx\, dt\, d\alpha \ \geq 0$$

for every $\kappa \in [A,B]$ and nonnegative function $\theta \in \mathcal{C}_c^\infty([0,T) \times B \cap Q)$. Now we use a result of comparison and assert that, for any nonnegative function $\theta \in \mathcal{C}_c^\infty([0,T) \times B \cap Q)$, we have

$$\int_Q \left[(u - \overline{u}_\Sigma \top \tilde{\kappa})^+\, \theta_t + [\Phi^+(t,x,u,\overline{u}_\Sigma \top \tilde{\kappa}) - \nabla(\varphi(u) - \varphi(\overline{u}_\Sigma \top \tilde{\kappa}))^+] \cdot \nabla\theta \right] d\alpha\, dx\, dt$$
$$+ \int_\Omega (u_0 - \overline{u}_\Sigma \top \tilde{\kappa}(0,x))^+\, \theta(0)\, dx + \int_Q \operatorname{sgn}^+(u - \overline{u}_\Sigma \top \tilde{\kappa})\, g\, \theta\, dx\, dt\, d\alpha \ \geq 0. \tag{20}$$

This result of comparison, proved in [Car99] for entropy weak solution, remains true when applied to entropy process solutions. Notice that we state a result of comparison *inside* $[0,T) \times \Omega$ (the previous function θ vanishes on $[0,T) \times \partial\Omega$); this point is crucial. A result of comparison on the whole domain Q is the object of Theorem 4.1, which we are actually proving. As a matter of fact, we would like to rule out the hypothesis that θ vanishes on $[0,T) \times \partial\Omega$. Toward that end, first notice that (20) is still true if $\theta \in \mathcal{C}_c^1([0,T) \times (B \cap \overline{\Omega}))$ and $\theta = 0$ on $[0,T) \times (B \cap \partial\Omega)$. Let $\tilde{\theta} \in \mathcal{C}_c^\infty([0,T) \times (B \cap \overline{\Omega}))$, define, for $s > 0$, $h_s(x) = \min([x_d - f(\overline{x})]/s, 1)$, and choose $\theta = \tilde{\theta}\, h_s$ in (20) to get

$$\int_Q \left[(u - \overline{u}_\Sigma \top \tilde{\kappa})^+\, \tilde{\theta}_t + [\Phi^+(t,x,u,\overline{u}_\Sigma \top \tilde{\kappa}) - \nabla(\varphi(u) - \varphi(\overline{u}_\Sigma \top \tilde{\kappa}))^+] \cdot \nabla\tilde{\theta} \right] h_s\, d\alpha\, dx\, dt$$
$$+ \int_\Omega (u_0 - \overline{u}_\Sigma \top \tilde{\kappa}(0,x))^+\, \tilde{\theta}(0)\, h_s\, dx + \int_Q \operatorname{sgn}^+(u - \overline{u}_\Sigma \top \tilde{\kappa})\, g\, \tilde{\theta}\, h_s\, dx\, dt\, d\alpha \ \geq A_s + B_s, \tag{21}$$

where

$$A_s = -\frac{1}{s} \int_0^T \int_0^1 \int_\Pi \int_{f(\overline{x})}^{f(\overline{x})+s} \Phi^+(t,x,u,\overline{u}_\Sigma \top \tilde{\kappa}) \cdot \nabla_x(x_d - f(\overline{x}))\, \tilde{\theta}\, dx_d\, d\overline{x}\, dt\, d\alpha,$$
$$B_s = \frac{1}{s} \int_0^T \int_\Pi \int_{f(\overline{x})}^{f(\overline{x})+s} \nabla(\varphi(u) - \varphi(\overline{u}_\Sigma \top \tilde{\kappa}))^+ \cdot \nabla_x(x_d - f(\overline{x}))\, \tilde{\theta}\, dx_d\, d\overline{x}\, dt.$$

Let C be a bound of $\Phi^+(t,x,z,w) \cdot \nabla_x(x_d - f(\overline{x}))$ in $L^\infty(Q \times [A,B]^2)$. Such a bound exists and, for every s,

$$A_s \geq -C\, T\, |\Pi|\, ||\tilde{\theta}||_{L^\infty([0,T) \times (B \cap \overline{\Omega}))}\,. \tag{22}$$

On the other hand, the term B_s can be decomposed as $B_s = \overline{B_s} + B_s^d$, where

$$\overline{B_s} = -\frac{1}{s} \int_0^T \int_\Pi \int_{f(\overline{x})}^{f(\overline{x})+s} \nabla_{\overline{x}}(\varphi(u) - \varphi(\overline{u}_\Sigma \top \tilde{\kappa}))^+ \cdot \nabla_{\overline{x}} f(\overline{x})\, \tilde{\theta}\, dx_d\, d\overline{x}\, dt,$$
$$B_s^d = \frac{1}{s} \int_0^T \int_\Pi \int_{f(\overline{x})}^{f(\overline{x})+s} \partial_{x_d}(\varphi(u) - \varphi(\overline{u}_\Sigma \top \tilde{\kappa}))^+\, \tilde{\theta}\, dx_d\, d\overline{x}\, dt.$$

Integration by parts with respect to $\overline{x}$ in $\overline{B_s}$ and integration by parts with respect to x_d in B_s^d (we use the fact that $\varphi(u)(t,\overline{x},f(\overline{x}))=\varphi(\overline{u}_\Sigma)(t,\overline{x})$) yields the following: for almost every positive s (small enough),

$$\begin{aligned}\overline{B_s} &= \frac{1}{s}\int_0^T\!\!\int_\Pi\int_{f(\overline{x})}^{f(\overline{x})+s}(\varphi(u)-\varphi(\overline{u}_\Sigma\top\tilde{\kappa}))^+\mathrm{div}_{\overline{x}}\nabla_{\overline{x}}f(\overline{x})\,\tilde{\theta}\,dx_d\,d\overline{x}\,dt,\\ B_s^d &= -\frac{1}{s}\int_0^T\!\!\int_\Pi\int_{f(\overline{x})}^{f(\overline{x})+s}(\varphi(u)-\varphi(\overline{u}_\Sigma\top\tilde{\kappa}))^+\,\partial_{x_d}\tilde{\theta}\,dx_d\,d\overline{x}\,dt\\ &\quad+\frac{1}{s}\int_0^T\!\!\int_\Pi(\varphi(u)-\varphi(\overline{u}_\Sigma\top\tilde{\kappa}))^+(t,\overline{x},f(\overline{x})+s)\,\tilde{\theta}(\overline{x},f(\overline{x})+s)d\overline{x}\,dt.\end{aligned}$$

Notice that, first, the second term on the right-hand side of the previous equality in nonnegative; that, second, $\lim_{s\to0}\overline{B_s}=0$ and $\lim_{s\to0}\frac{1}{s}\int_0^T\int_\Pi\int_{f(\overline{x})}^{f(\overline{x})+s}(\varphi(u)-\varphi(\overline{u}_\Sigma\top\tilde{\kappa}))^+\,\partial_{x_d}\tilde{\theta}\,dx_d\,d\overline{x}\,dt\;=0$ (because the trace of $\varphi(u)$ is $\varphi(\bar{u})$); and that, third, h_s converge to 1 in $L^1(B\cap\Omega)$. Consequently, letting s go to zero on both sides of inequality (21) yields

$$\begin{aligned}&\int_Q\Big[(u-\overline{u}_\Sigma\top\tilde{\kappa})^+\,\tilde{\theta}_t+[\Phi^+(t,x,u,\overline{u}_\Sigma\top\tilde{\kappa})-\nabla(\varphi(u)-\varphi(\overline{u}_\Sigma\top\tilde{\kappa}))^+]\cdot\nabla\tilde{\theta}\Big]\,d\alpha\,dx\,dt\\ &+\int_\Omega(u_0-\overline{u}_\Sigma\top\tilde{\kappa}(0,x))^+\,\tilde{\theta}(0)\,dx+\int_Q\mathrm{sgn}^+(u-\overline{u}_\Sigma\top\tilde{\kappa})\,g\,\tilde{\theta}\,dx\,dt\,d\alpha\geq\liminf_{s\to0}A_s.\end{aligned}$$

Let D be an open subset of B whose closure is a subset of B too. From (22), it appears that $\liminf_{s\to0}A_s$ can be viewed as the action of a certain distribution A_∞ on $\tilde{\theta}$ and that A_∞ is a bounded Radon measure on $[0,T)\times D\cap Q$. Since $\int_0^1\mathrm{sgn}^+(u-\overline{u}_\Sigma)\,g\,d\alpha$ and $(u_0-\overline{u}_\Sigma\top\tilde{\kappa}(0,x))^+\delta_{t=0}$ are bounded Radon measures on $[0,T)\times D\cap Q$, the previous inequality shows that the divergence of the field $\overline{\mathcal{F}}^+$ is a bounded Radon measure on $[0,T)\times D\cap Q$. This ends the proof of Lemma 4.2. □

Remark 4.1. If $\overline{u}_\Sigma$ satisfies (H6Bis) instead of (H6), then $\overline{u}_\Sigma$ can be seen as the entropy solution of the equation $\partial_t\,w+\mathrm{div}_x\,F(t,x,w)-\Delta\,\varphi(w)=g$, with a source term g which is a bounded Radon measure on $(0,T)\times B\cap Q$. In the proof of the previous lemma we used a theorem of comparison of Carrillo (Theorem 8 in [Car99]) between two entropy solutions u_i ($i\in\{1,2\}$) of the equation

$$\partial_tu_i+\mathrm{div}F(t,x,u_i)-\Delta\varphi(u_i)=f_i$$

(where $f_i\in L^1$) to derive the inequality (20). A careful study of the proof of the result of comparison given by Carrillo shows that it still holds if $f_1=0$ and f_2 is a bounded Radon measure. Consequently, inequality (20) remains true under hypothesis (H6Bis) and Lemma 4.2 also.

As a consequence of this lemma, we can define a functional $\overline{\mathcal{T}}^+_{n,\tilde{\kappa}}$, which represents the normal trace of the flux $\Phi^+(t,x,u,\overline{u}_\Sigma\top\tilde{\kappa})-\nabla(\varphi(u)-\varphi(\overline{u}_\Sigma\top\tilde{\kappa}))^+$ on $(0,T)\times(\partial\Omega\cap D)$ and satisfies the analogue of the relations (16) and (17), where κ has been replaced by $\overline{u}_\Sigma\top\tilde{\kappa}$ in these latter. (We use the fact that there exists an open set D such that $\mathrm{supp}(\lambda)\subset D\subset\overline{D}\subset B$ to ensure that these limits make sense.)

Now, denote by $\overline{\mathcal{S}}$ the set of all the functions $v:(0,T)\times\Pi\to\mathbb{R}$ satisfying

$$v(t,x)=\sum_{i=1}^{N_v}w_i\,L_i(t,x), \tag{23}$$

where

$$\forall i, \quad w_i \in \mathbb{R},\, L_i \in \mathcal{C}^\infty([0,T]\times\Pi),\; L_i \geq 0, \text{ and } \sum_{i=1}^{N_v} L_i = 1 \quad \text{on } [0,T]\times\Pi_\lambda. \tag{24}$$

We say that $v \in \overline{\mathcal{S}}^+$ if $v \in \overline{\mathcal{S}}$ and admits a decomposition as (23) such that $w_i \geq \overline{u}_\Sigma$ a.e. on $\operatorname{supp}(L_i)$ for all i. If $v \in \overline{\mathcal{S}}$ and satisfies (23), we set

$$\langle \mathcal{T}^+_{n,v\top\tilde\kappa}, \psi\lambda\rangle \;=\; \sum_{i=1}^{N_v} \langle \mathcal{T}^+_{n,w_i\top\tilde\kappa}, L_i\psi\lambda\rangle.$$

Notice that this is a *notation* and not a *definition*, because the decomposition (23) with w_i, L_i satisfying (24) is not unique. An immediate consequence of (18) is the following: if $v \in \overline{\mathcal{S}}^+$, then

$$-\langle \mathcal{T}^+_{n,v\top\tilde\kappa}, \psi\lambda\rangle \leq 0 \quad \forall\psi \in \mathcal{C}^\infty_c([0,T)\times\mathbb{R}^d),\; \psi \geq 0. \tag{25}$$

Furthermore, we claim that, if $v \in \overline{\mathcal{S}}^+$, then

$$\langle \mathcal{T}^+_{n,v\top\tilde\kappa} - \overline{\mathcal{T}}^+_{n,\tilde\kappa}, \psi\lambda\rangle \leq M\sqrt{1+||\nabla_{\overline{x}} f||^2_\infty}\; \sum_{i=1}^{N_v} \int_0^T\!\!\int_\Pi |w_i - \overline{u}_\Sigma|\, \psi\lambda\, L_i\, d\overline{x}\, dt. \tag{26}$$

Let us prove this result: from (16) we have $\langle \mathcal{T}^+_{n,v\top\tilde\kappa} - \overline{\mathcal{T}}^+_{n,\tilde\kappa}, \psi\lambda\rangle = -\lim\limits_{s\to 0} \sum_{i=1}^{N_v} (H_i(s) + P_i(s))$, where

$$\begin{aligned} H_i(s) = \frac{1}{s}\int_s^T\!\!\int_0^1\!\!\int_\Pi \int_{f(\overline{x})}^{f(\overline{x})+s} &(\, \Phi\,^+(t,x,u,w_i\top\tilde\kappa) \\ &- \Phi^+(t,x,u,\overline{u}_\Sigma\top\tilde\kappa)) \cdot \begin{pmatrix} -\nabla f(\overline{x}) \\ 1 \end{pmatrix} L_i\psi\lambda\, dx_d\, d\overline{x}\, dt\, d\alpha, \end{aligned}$$

$$\begin{aligned} P_i(s) = \frac{1}{s}\int_s^T\!\!\int_\Pi \int_{f(\overline{x})}^{f(\overline{x})+s} &\nabla((\, \varphi\,(u) - \varphi(\overline{u}_\Sigma\top\tilde\kappa))^+ \\ &- (\varphi(u) - \varphi(w_i\top\tilde\kappa))^+) \cdot \begin{pmatrix} -\nabla f(\overline{x}) \\ 1 \end{pmatrix} L_i\psi\lambda\, dx_d\, d\overline{x}\, dt. \end{aligned}$$

Since the function $\Phi^+(t,x,u,v)$ is M-Lipschitz continuous with respect to v, uniformly with respect to $(t,x,u) \in Q\times[A,B]$, we have

$$\begin{aligned} H_i(s) &\geq -\frac{1}{s} M\sqrt{1+||\nabla_{\overline{x}} f||^2_\infty} \int_0^T\!\!\int_\Pi \int_{f(\overline{x})}^{f(\overline{x})+s} |w_i\top\tilde\kappa - \overline{u}_\Sigma\top\tilde\kappa|\, L_i\psi\lambda\, dx_d\, d\overline{x}\, dt \\ &\geq -\frac{1}{s} M\sqrt{1+||\nabla_{\overline{x}} f||^2_\infty} \int_0^T\!\!\int_\Pi \int_{f(\overline{x})}^{f(\overline{x})+s} |w_i - \overline{u}_\Sigma|\, L_i\psi\lambda\, dx_d\, d\overline{x}\, dt\,. \end{aligned}$$

Consequently,

$$\sum_{i=1}^{N_v} H_i(s) \geq -\frac{1}{s} M\sqrt{1+||\nabla_{\overline{x}} f||^2_\infty} \sum_{i=1}^{N_v} \int_0^T\!\!\int_\Pi \int_{f(\overline{x})}^{f(\overline{x})+s} |w_i - \overline{u}_\Sigma|\, \psi\lambda\, L_i\, dx_d\, d\overline{x} dt,$$

and the limit of the right-hand side of this latter inequality can be explicitly computed since the function $\psi\lambda\, L_i$ is smooth:

$$\lim_{s\to 0}\sum_{i=1}^{N_v} H_i(s) \geq -M\sqrt{1+||\nabla_{\overline{x}} f||_\infty^2}\sum_{i=1}^{N_v}\int_0^T\!\!\int_\Pi |w_i-\overline{u}_\Sigma|\,\psi\lambda\, L_i\, d\overline{x}\, dt\,. \tag{27}$$

On the other hand, we have $\limsup_{s\to 0} P_i(s) \geq 0$. We will not detail the proof of this result, for it is identical to the justification of the fact that $\limsup_{s\to 0} B_s \geq 0$ in the proof of Lemma 4.2. Together with (27), the result $\limsup_{s\to 0} P_i(s) \geq 0$ yields (26). Furthermore, (26) combined with (25) shows that, if $v \in \overline{\mathcal{S}}^+$($v$ satisfies (23), with $w_i \geq \overline{u}_\Sigma$ a.e. on $\mathrm{supp}(L_i)$), then

$$-\langle \overline{\mathcal{T}}^+_{n,\tilde{\kappa}}, \psi\lambda\rangle \leq M\sqrt{1+||\nabla_{\overline{x}} f||_\infty^2}\sum_{i=1}^{N_v}\int_0^T\!\!\int_\Pi (w_i-\overline{u}_\Sigma)\,\psi\lambda\, L_i\, d\overline{x}\, dt. \tag{28}$$

Since

$$\langle \overline{\mathcal{T}}^+_{n,\tilde{\kappa}}, \psi\lambda\rangle = -\lim_{\varepsilon\to 0}\int_Q [\Phi^+(t,x,u,\overline{u}_\Sigma\top\tilde{\kappa}) - \nabla(\varphi(u)-\varphi(\overline{u}_\Sigma\top\tilde{\kappa}))^+]\cdot\nabla\omega_\varepsilon\,\psi\lambda\, dx\, dt\, d\alpha\,,$$

our first aim, which is the proof of (15), will be reached if the right-hand side of (28) can be made as small as desired. Let us prove this fact: $\varepsilon > 0$. Since $\overline{u}_\Sigma \in L^\infty\cap W^{1,1}((0,T)\times\Pi)$ (hypothesis (H6)), we have $\overline{u}_\Sigma \in L^\infty \cap BV((0,T)\times\Pi)$, and Lemma 4.1 shows that there exists $g\in\mathcal{C}([0,T]\times\overline{\Pi})$ such that $g\geq\overline{u}_\Sigma$ a.e. on $(0,T)\times\Pi$ and $\int_{(0,T)\times\Pi} g-\overline{u}_\Sigma < \varepsilon$. Let η be a modulus of uniform continuity of g on $[0,T]\times\overline{\Pi}$. The set $(0,T)\times\Pi$ (with compact closure) can be covered by a finite number of balls with radius η centered in $(0,T)\times\Pi$, say $V_1,\dots,V_Q$. Let $(L_i)_{1,Q}$ be a regular partition of unity subordinate to the open coverage (V_i) of $[0,T]\times\overline{\Pi}$. For a certain $(t_i,\overline{x}_i)\in V_i$, set $w_i = g(t_i,\overline{x}_i)+\varepsilon$ and define $v=\sum_{i=1}^Q w_i\, L_i$. Then $v\in\overline{\mathcal{S}}^+$ and

$$\begin{aligned}\sum_{i=1}^{Q}\int_0^T\!\!\int_\Pi (w_i-\overline{u}_\Sigma)\,\psi\lambda\, L_i\, d\overline{x}\, dt &= \int_0^T\!\!\int_\Pi (v-\overline{u}_\Sigma)\,\psi\lambda\, d\overline{x}\, dt\\ &= \int_0^T\!\!\int_\Pi (v-g)\,\psi\,\lambda\, d\overline{x}\, dt + \int_0^T\!\!\int_\Pi (g-\overline{u}_\sigma)\,\psi\,\lambda\, d\overline{x}\, dt\\ &\leq 2||\psi\,\lambda||_\infty\, T\,|\Pi|\,\varepsilon.\end{aligned}$$

This completes the proof of (15). Similarly, we can prove

$$\lim_{\varepsilon\to 0}\int_Q [\Phi^-(t,x,u,\tilde{\kappa}\bot\overline{u}_\Sigma) - \nabla(\varphi(u)-\varphi(\tilde{\kappa}\bot\overline{u}_\Sigma))^-]\cdot\nabla\omega_\varepsilon(x)\,\psi(t,x)\,\lambda(x)\, d\alpha\, dx\, dt \leq 0 \tag{29}$$

for every $\tilde{\kappa}\in[A,B]$ and for every nonnegative $\psi\in\mathcal{C}_c^\infty([0,T)\times\mathbb{R}^d)$. Then Proposition 4.1 follows from the formula

$$\begin{aligned}\mathcal{F}_\varphi(t,x,u,\kappa,w) = &\left[\,\Phi^{\,+}(t,x,u,\kappa\top w) - \nabla(\varphi(u)-\varphi(\kappa\top w))^+\right]\\ &+\left[\Phi^-(t,x,u,\kappa\bot w) - \nabla(\varphi(u)-\varphi(\kappa\bot w))^-\right].\qquad\square\end{aligned}$$

4.4. Proof of Theorem 4.1 (step 1): Inner comparison. Let u and $v \in L^\infty(Q\times(0,1))$ be two entropy process solutions of problem (1). The following result of comparison between u and v involving test-functions which *vanish* on the boundary of Ω can be proved (see [Car99] or [EGHM02]).

PROPOSITION 4.2 (inner comparison). *Let u and $v \in L^\infty(Q\times(0,1))$ be two entropy process solutions of problem* (1). *Assume hypotheses* (H1), (H2), (H3), *and* (H4). *Let ζ be a nonnegative function of $\mathcal{C}^\infty([0,T)\times\mathbb{R}^d\times[0,T)\times\mathbb{R}^d)$ such that*

$$\begin{cases} \forall (s,y)\in Q\,,\ (t,x)\longmapsto \zeta(t,x,s,y)\in\mathcal{C}_c^\infty([0,T)\times\Omega), \\ \forall (t,x)\in Q\,,\ (s,y)\longmapsto \zeta(t,x,s,y)\in\mathcal{C}_c^\infty([0,T)\times\Omega). \end{cases}$$

Then we have

$$\begin{aligned} &\int_Q\!\!\int_Q \left[\begin{array}{c} |u(t,x,\alpha)-v(s,y,\beta)|(\zeta_t+\zeta_s) \\ +\,\mathcal{G}_x(t,x,u(t,x,\alpha),v(s,y,\beta))\cdot\nabla_x\zeta \\ +\,\mathcal{G}_y(s,y,v(s,y,\beta),u(t,x,\alpha))\cdot\nabla_y\zeta \\ -\,\nabla_x|\varphi(u)(t,x)-\varphi(v)(s,y)|\cdot\nabla_y\zeta \\ -\,\nabla_y|\varphi(u)(t,x)-\varphi(v)(s,y)|\cdot\nabla_x\zeta \end{array}\right] d\alpha dx dt d\beta dy ds \\ &+\int_Q\int_\Omega |u_0(x)-v(s,y,\beta)|\,\zeta(0,x,s,y)\,dx\,d\beta\,dy\,ds \\ &+\int_Q\int_\Omega |u_0(y)-u(t,x,\alpha)|\,\zeta(t,x,0,y)\,dy\,d\alpha\,dx\,dt\ \geq 0. \end{aligned} \tag{30}$$

4.5. Proof of Theorem 4.1 (step 2): General test-function. We now follow the lines of the proof of uniqueness given by Mascia, Porretta, and Terracina in [MPT02].

First, we would like to consider test-functions which do not necessarily vanish on $\partial\Omega$ and are localized into the ball B. For $\overline{x}\in\mathbb{R}^{d-1}$, set $\rho_m(\overline{x})=\rho_m(x_1)\cdots\rho_m(x_{d-1})$ and define the function ξ by

$$\xi(t,s,x,y)=\psi(t,x)\,\rho_l(t-s)\,\rho_m(\overline{x}-\overline{y})\,\rho_n(x_d-y_d)\,. \tag{31}$$

We took care to choose ρ satisfying $\operatorname{supp}(\rho)\subset[-1,0)$ to ensure

$$\begin{aligned} &\forall (t,x)\in Q\,,\ (s,y)\longmapsto \xi(t,s,x,y)\in\mathcal{C}_c^\infty(Q)\,, \\ &\forall (t,s,x)\in[0,T)\times[0,T)\times\operatorname{supp}(\lambda),\ \operatorname{supp}_y\xi(t,s,x,\cdot)\subset B. \end{aligned} \tag{32}$$

For $\varepsilon>0$ define ζ to be the function

$$\zeta\,:\,(t,s,x,y)\longmapsto \omega_\varepsilon(x)\,\xi(t,s,x,y)\,\lambda(x).$$

Then, for m large enough compared with n, the assumptions of Proposition 4.2 are satisfied, and, with this particular choice of function ζ, inequality (30) turns into the inequality

$$\int_{\mathcal{Q}}\int_{\mathcal{Q}}\left[\begin{array}{l} \qquad\qquad |u-\widehat{v}|\,\omega_\varepsilon(x)\,((\xi\lambda)_t+(\xi\lambda)_s) \\ +\Big(\mathcal{G}_x(t,x,u,\widehat{v})\cdot\nabla_x(\xi\,\lambda) \\ \qquad\qquad +\mathcal{G}_y(t,y,\widehat{v},u)\cdot\nabla_y(\xi\,\lambda)\Big)\,\omega_\varepsilon(x) \\ -\Big(\nabla_x|\varphi(u)-\varphi(\widehat{v})|\cdot\nabla_y(\xi\,\lambda) \\ \qquad\qquad +\nabla_y|\varphi(u)-\varphi(\widehat{v})|\cdot\nabla_x(\xi\,\lambda)\Big)\,\omega_\varepsilon(x) \end{array}\right] dxdtd\alpha dydsd\beta$$
$$+\int_{\mathcal{Q}}\int_{\mathcal{Q}} \mathcal{G}_x(t,x,u,\widehat{v})\cdot\nabla\omega_\varepsilon(x)\;\xi\,\lambda\,d\alpha\,dx\,dt\;\,d\beta\,dy\,ds$$
$$-\int_{\mathcal{Q}}\int_{\mathcal{Q}} \nabla_y|\varphi(u)-\varphi(\widehat{v})|\cdot\nabla\omega_\varepsilon(x)\,\xi\,\lambda\,dx\,dt\,d\alpha\,dy\,ds\,d\beta$$
$$+\int_{\Omega}\int_{\mathcal{Q}} |u_0(x)-\widehat{v}|\,(\xi\,\lambda)(0,x,y)\,\omega_\varepsilon(x)\,dx\delta\beta\,dy\,ds\geq 0,$$

where

$$u=u(t,x,\alpha)\quad\text{and}\quad \widehat{v}=v(s,y,\beta).$$

Using formula (12), this inequality can be rewritten as

$$\int_{\mathcal{Q}}\int_{\mathcal{Q}}\left[\begin{array}{l} \qquad\qquad |u-\widehat{v}|\,\omega_\varepsilon(x)\,((\xi\,\lambda)_t+(\xi\lambda)_s) \\ +(\mathcal{G}_x(t,x,u,\widehat{v})\cdot\nabla_x(\xi\,\lambda)+\mathcal{G}_y(t,y,\widehat{v},u)\cdot\nabla_y(\xi\,\lambda))\,\omega_\varepsilon(x) \\ -(\nabla_x|\varphi(u)-\varphi(\widehat{v})|\cdot\nabla_y(\xi\,\lambda) \\ \quad +\nabla_y|\varphi(u)-\varphi(\widehat{v})|\cdot\nabla_x(\xi\,\lambda))\,\omega_\varepsilon(x) \end{array}\right] d\alpha dxdtd\beta dyds$$

(33)

$$+\int_{\mathcal{Q}}\int_{\mathcal{Q}} \mathcal{F}_\varphi(t,x,u,\widehat{v},\overline{u}_\Sigma)\cdot\nabla\omega_\varepsilon(x)\,\xi\,\lambda\,d\alpha\,dx\,dt\,d\beta\,dy\,ds$$
$$+\int_{\Omega}\int_{\mathcal{Q}} |u_0(x)-\widehat{v}|\,(\xi\,\lambda)(0,x,y)\,\omega_\varepsilon(x)\,dx\,dy\,ds\,d\beta\;\geq\;A+B+C,$$

where

$$A=\int_Q\int_Q \nabla_y|\varphi(u)-\varphi(\widehat{v})|\cdot\nabla\omega_\varepsilon(x)\,\xi\,\lambda\,dx\,dt\,dy\,ds,$$
$$B=-\int_{\mathcal{Q}}\int_{\mathcal{Q}} \mathcal{G}_x(t,x,\widehat{v},\overline{u}_\Sigma)\cdot\nabla\omega_\varepsilon(x)\,\xi\,\lambda\,d\alpha\,dx\,dt\,d\beta\,dy\,ds,$$
$$C=\int_{\mathcal{Q}}\int_{\mathcal{Q}} \mathcal{G}_x(t,x,u,\overline{u}_\Sigma)\cdot\nabla\omega_\varepsilon(x)\,\xi\,\lambda\,d\alpha\,dx\,dt\,d\beta\,dy\,ds.$$

Using Proposition 4.1 and taking the limit of both sides of the previous inequality with respect to ε then yields

$$\int_{\mathcal{Q}}\int_{\mathcal{Q}}\left[\begin{array}{c} |u-\widehat{v}|\,((\xi\,\lambda)_t+(\xi\,\lambda)_s) \\ +\mathcal{G}_x(t,x,u,\widehat{v})\cdot\nabla_x(\xi\,\lambda)+\mathcal{G}_y(t,y,\widehat{v},u)\cdot\nabla_y(\xi\,\lambda) \\ -\nabla_x|\varphi(u)-\varphi(\widehat{v})|\cdot\nabla_y(\xi\,\lambda)+\nabla_y|\varphi(u)-\varphi(\widehat{v})|\cdot\nabla_x(\xi\,\lambda) \end{array}\right] d\alpha\,dx\,dt\,d\beta\,dy\,ds$$
$$+\int_{\Omega}\int_{\mathcal{Q}} |u_0(x)-\widehat{v}|\,(\xi\,\lambda)(0,x,y)\,d\beta\,dy\,ds\,dx\geq\lim_{\varepsilon\to 0}\,(A+B+C\,),$$

or (using formula (11))

$$\begin{aligned}&\int_{\mathcal{Q}}\int_{\mathcal{Q}}\left[\begin{array}{c}|u-\widehat{v}|\,((\xi\,\lambda)_t+(\xi\,\lambda)_s)\\ +\,\Phi(t,x,u,\widehat{v})\cdot\nabla_x(\xi\,\lambda)+\Phi(t,y,\widehat{v},u)\cdot\nabla_y(\xi\,\lambda)\\ -\,(\nabla_x|\varphi(u)-\varphi(\widehat{v})|+\nabla_y|\varphi(u)-\varphi(\widehat{v})|)\cdot(\nabla_y+\nabla_x)(\xi\,\lambda)\end{array}\right]d\alpha\,dx\,dt\,d\beta\,dy\,ds\\ &+\int_\Omega\int_{\mathcal{Q}}|u_0(x)-\widehat{v}|\,(\xi\,\lambda)(0,x,y)\,d\beta\,dy\,ds\,dx\geq\lim_{\varepsilon\to0}\,(A+B+C\,).\end{aligned}\tag{34}$$

Now, we intend to pass to the limit on l, m, and n in the previous inequality. We will do so (on l and m and, eventually, on n), but notice that the study of the behavior of A, B, and C as $[\varepsilon\to0]$ and the doubling variable technique itself interfere with each other.

Using the definition of ξ from (31), it appears that C does not depend on l, m, and n:

$$C=\int_{\mathcal{Q}}\mathcal{G}_x(t,x,u,\overline{u}_\Sigma)\cdot\nabla\omega_\varepsilon(x)\,\psi\,\lambda\,d\alpha\,dx\,dt.$$

Moreover, inequality (34) can be rewritten as

$$\begin{aligned}&\int_{\mathcal{Q}}\int_{\mathcal{Q}}\left[\begin{array}{c}|u-\widehat{v}|\,\rho_l\,\rho_m\,\rho_n(\psi\,\lambda)_t\\ +\,\Phi(t,x,u,\widehat{v})\cdot\nabla_x(\psi\,\lambda)\rho_l\,\rho_m\,\rho_n\\ -\,(\nabla_x|\varphi(u)-\varphi(\widehat{v})|+\nabla_y|\varphi(u)-\varphi(\widehat{v})|)\cdot\nabla_x(\psi\,\lambda)\,\rho_l\,\rho_m\,\rho_n\end{array}\right]d\alpha\,dx\,dt\,d\beta\,dy\,ds\\ &\quad+\int_\Omega\int_\Omega|u_0(x)-u_0(y)|\,(\psi\,\lambda)(0,x)\,\rho_m\,\rho_n\,dx\,dy\;\geq\;\lim_{\varepsilon\to0}\,(A+B+C\,)+D+E,\end{aligned}\tag{35}$$

where

$$D=-\int_{\mathcal{Q}}\int_{\mathcal{Q}}[\Phi(t,x,u,\widehat{v})-\Phi(t,y,u,\widehat{v})]\cdot\nabla_x(\rho_l\,\rho_m\,\rho_n)\,\psi\,\lambda\,d\alpha\,dx\,dt\,d\beta\,dy\,ds,$$

$$E=\int_\Omega\int_{\mathcal{Q}}|u_0(y)-\widehat{v}|\,(\psi\,\lambda)(0,x)\,\rho_l(-s)\,\rho_m\,\rho_n\;d\beta\,dy\,ds\,dx.$$

The term E can be estimated by using the fact that the solution v completely satisfies the initial condition, which means, for example, that $\operatorname{ess\,lim}_{s\to0^+}\int_\Omega\int_0^1|v(s,y,\alpha)-u_0(y)|\,d\beta\,dy=0$. On the other hand, if the flux function F does not depend on the (t,x)-variables, then $D=0$, and more generally, one can prove (see [CH99]) $D+E\geq H$, where

$$\begin{aligned}H=-C(F,\psi)\sup\Bigg\{&\int_{\mathcal{Q}}|v(s,\overline{y},y_d,\beta)-v(s+\sigma,\overline{y}+\overline{h},y_d+k,\beta)|ds\,d\overline{y}\,dy_d\,d\beta\,;\\ &|\sigma|\leq\frac{1}{l},|\overline{h}|\leq\frac{1}{m},|k|\leq\frac{1}{n}\Bigg\}.\end{aligned}\tag{36}$$

Notice that, by continuity of the translations in L^1, we have $\lim_{l,m,n\to+\infty}H=0$.

4.5.1. Study of $A+B$. Going back to the study of A, B, we write $A+B = I + J^y + J^x$, where

$$\begin{aligned}
I &= -\int_{\mathcal{Q}}\int_{\mathcal{Q}} (\Phi(t,x,\widehat{v},\overline{u}_\Sigma(t,\overline{x}))\cdot\nabla\omega_\varepsilon(x)\,\xi\,\lambda\,d\alpha\,dx\,dt\,d\beta\,dy\,ds,\\
J^y &= \int_Q\int_Q \nabla_y|\varphi(u)(t,x)-\varphi(v)(s,y)|\cdot\nabla\omega_\varepsilon(x)\,\xi\,\lambda\,dx\,dt\,dy\,ds,\\
J^x &= \int_Q\int_Q \nabla_x|\varphi(\widehat{v})-\varphi(\overline{u}_\Sigma(t,\overline{x}))|\cdot\nabla\omega_\varepsilon(x)\,\xi\,\lambda\,dx\,dt\,dy\,ds.
\end{aligned}$$

Recall that

$$\nabla\omega_\varepsilon(x) = \rho_\varepsilon(f(\overline{x})-x_d)\begin{pmatrix}-\nabla f(\overline{x})\\ 1\end{pmatrix},$$

so that

$$\begin{aligned}
\widetilde{I} &= \lim_{\varepsilon\to 0} I\\
&= -\int_{\mathcal{Q}}\int_{[0,T)\times\Pi\times(0,1)} (\Phi(t,\overline{x},f(\overline{x}),\widehat{v},\overline{u}_\Sigma(t,\overline{x}))\cdot\begin{pmatrix}-\nabla f(\overline{x})\\ 1\end{pmatrix}(\xi\,\lambda)_{\Sigma_x}d\alpha d\overline{x}dtd\beta dyds,
\end{aligned}$$

where the index Σ_x indicates that the transformation concerns only the x variable. Here, for example, $(\xi\,\lambda)_{\Sigma_x}(t,x,y) = \xi(t,\overline{x},f(\overline{x}),y)\lambda(\overline{x},f(\overline{x}))$. To study the term J^x, we notice that the function $\overline{u}_\Sigma$ does not depend on x_d, and thus

$$\widetilde{J^x} = \lim_{\varepsilon\to 0} J^x = -\int_{[0,T)\times\Pi}\int_Q \nabla_{\overline{x}}\,|\varphi(\widehat{v})-\varphi(\overline{u}_\Sigma(t,\overline{x}))|\cdot\nabla f(\overline{x})\,(\xi\,\lambda)_{\Sigma_x}\,d\overline{x}\,dt\,dy\,ds.$$

Integration by parts with respect to $\overline{x}$ in $\widetilde{J^x}$ yields $\widetilde{J^x} = \widetilde{J^x_f} + \widetilde{J^x_\psi} + \widetilde{J^x_{\rho_m}} + \widetilde{J^x_{\rho_n}}$, where

$$\begin{aligned}
\widetilde{J^x_f} &= \int_{[0,T)\times\Pi}\int_Q |\varphi(\widehat{v})-\varphi(\overline{u}_\Sigma)|\Delta f(\overline{x})\,(\psi\,\lambda)_{\Sigma_x}\,\rho_l(t-s)\\
&\qquad\times\rho_m(\overline{x}-\overline{y})\,\rho_n(f(\overline{x})-y_d)\,d\overline{x}\,dt\,dy\,ds,\\
\widetilde{J^x_\psi} &= \int_{[0,T)\times\Pi}\int_Q |\varphi(\widehat{v})-\varphi(\overline{u}_\Sigma)|\nabla f(\overline{x})\\
&\qquad\cdot\nabla_{\overline{x}}\,((\psi\,\lambda)_{\Sigma_x})\rho_l(t-s)\,\rho_m(\overline{x}-\overline{y})\rho_n(f(\overline{x})-y_d)d\overline{x}dtdyds,\\
\widetilde{J^x_{\rho_m}} &= \int_{[0,T)\times\Pi}\int_Q |\varphi(\widehat{v})-\varphi(\overline{u}_\Sigma)|\nabla f(\overline{x})\\
&\qquad\cdot\nabla_{\overline{x}}\,\rho_m(\overline{x}-\overline{y})\,\rho_n(f(\overline{x})-y_d)\,\rho_l(t-s)\,\psi\,\lambda\,d\overline{x}\,dt\,dy\,ds,\\
\widetilde{J^x_{\rho_n}} &= \int_{[0,T)\times\Pi}\int_Q |\varphi(\widehat{v})-\varphi(\overline{u}_\Sigma)||\nabla f(\overline{x})|^2\rho_l(t-s)\\
&\qquad\times\rho_m(\overline{x}-\overline{y})\,\rho'_n(f(\overline{x})-y_d)\,(\psi\,\lambda)_{\Sigma_x}d\overline{x}dtdyds.
\end{aligned}$$

On the other hand, via integration by parts in J^y with respect to y, and recalling that the boundary condition $\varphi(u)=\varphi(\overline{u})$ on Σ is strongly satisfied according to Definition 3.1, we get

$$\begin{aligned}
\widetilde{J^y} &= \lim_{\varepsilon\to 0} J^y\\
&= -\int_{[0,T)\times\Pi}\int_Q |\varphi(\overline{u}_\Sigma(t,\overline{x}))-\varphi(\widehat{v})|\begin{pmatrix}-\nabla f(\overline{x})\\ 1\end{pmatrix}\cdot\nabla_y(\xi\,\lambda)(t,s,\overline{x},f(\overline{x}),y)\,d\overline{x}\,dt\,dy\,ds,
\end{aligned}$$

and, developing the scalar product,

$$\begin{aligned}\widetilde{J^y} = &\int_{[0,T)\times\Pi}\int_Q |\varphi(\overline{u}_\Sigma(t,\overline{x})) - \varphi(\widehat{v})|\nabla f(\overline{x})\cdot\nabla_{\overline{y}}(\xi\,\lambda)(t,s,\overline{x},f(\overline{x}),y)dy\,d\overline{x}\,dt\,ds\\ &-\int_{[0,T)\times\Pi}\int_Q |\varphi(\overline{u}_\Sigma(t,\overline{x})) - \varphi(\widehat{v})|\partial_{y_d}(\xi\,\lambda)(t,s,\overline{x},f(\overline{x}),y)\,dy\,d\overline{x}\,dt\,ds\\ = &-\widetilde{J^x_{\rho_m}} + \int_{[0,T)\times\Pi}\int_Q |\varphi(\overline{u}_\Sigma) - \varphi(\widehat{v})|\rho_l(t-s)\\ &\qquad\times\ \rho_m(\overline{x}-\overline{y})\,\rho_n'(f(\overline{x})-y_d)\,(\psi\,\lambda)_{\Sigma_x}dy\,d\overline{x}\,dt\,ds,\end{aligned}$$

so that

$$\begin{aligned}\widetilde{J^x}+\widetilde{J^y} = \widetilde{J^x_f}+\widetilde{J^x_\psi} + &\int_{[0,T)\times\Pi}\int_\Omega |\varphi(\overline{u}_\Sigma) - \varphi(\widehat{v})|\,(1+|\nabla f(\overline{x})|^2)\\ &\times\rho_l(t-s)\,\rho_m(\overline{x}-\overline{y})\rho_n'(f(\overline{x})-y_d)\,(\psi\,\lambda)_{\Sigma_x}\,d\overline{x}\,dt\,dy\,ds.\end{aligned}$$

In particular, no derivatives of the functions ρ_m or ρ_l appear in J^x+J^y. Hence, summing up by $\widetilde{v}$ the quantity $v(t,\overline{x},y_d,\beta)$ and passing to the limit $[l,m\to+\infty]$ in $\lim_{\varepsilon\to 0}(A+B) = \widetilde{I}+\widetilde{J^x}+\widetilde{J^y}$, we get

$$\lim_{l,m\to+\infty}\lim_{\varepsilon\to 0}(A+B) = \overline{I}+\overline{J_f}+\overline{J_\psi}+\overline{J_{\rho_n}}\,,$$

with

$$\begin{aligned}\overline{I} = -\int_{[0,T)\times\Pi\times(0,1)}&\int_0^\infty\int_0^1 \Phi(t,\overline{x},f(\overline{x}),\widetilde{v},\overline{u}_\Sigma)\\ &\cdot\begin{pmatrix}-\nabla f(\overline{x})\\ 1\end{pmatrix}\rho_n(f(\overline{x})-y_d)(\psi\,\lambda)_{\Sigma_x}d\overline{x}dtd\alpha dy_d d\beta,\end{aligned}$$

$$\overline{J_f} = \int_{[0,T)\times\Pi}\int_0^\infty |\varphi(\widetilde{v})-\varphi(\overline{u}_\Sigma)|\Delta f(\overline{x})\,(\psi\,\lambda)_{\Sigma_x}\,\rho_n(f(\overline{x})-y_d)\,d\overline{x}\,dt\,dy_d,$$

$$\overline{J_\psi} = \int_{[0,T)\times\Pi}\int_0^\infty |\varphi(\widetilde{v})-\varphi(\overline{u}_\Sigma)|\nabla f(\overline{x})\cdot\nabla_{\overline{x}}\,((\psi\,\lambda)_{\Sigma_x})\,\rho_n(f(\overline{x})-y_d)\,d\overline{x}\,dt\,dy_d,$$

$$\overline{J_{\rho_n}} = \int_{[0,T)\times\Pi}\int_0^\infty |\varphi(\widetilde{v})-\varphi(\overline{u}_\Sigma)|\,(1+|\nabla f(\overline{x})|^2)\,\rho_n'(f(\overline{x})-y_d)\,(\psi\,\lambda)_{\Sigma_x}\,d\overline{x}\,dt\,dy_d.$$

To compute the limit as n tends to $+\infty$ of the four preceding terms, first recall that $\text{trace}((\varphi(v))-\varphi(\overline{u}_\Sigma)) = 0$, and that, consequently,

$$\lim_{n\to+\infty}\overline{J_f} = 0 \quad\text{and}\quad \lim_{n\to+\infty}\overline{J_\psi} = 0.$$

Besides, we note that

$$\Delta\,\omega_{1/n}(x) = -\rho_n'(f(\overline{x})-x_d)\,(1+|\nabla f(\overline{x})|^2) + \rho_n(f(\overline{x})-x_d)\,\Delta f(\overline{x}),$$

so that, replacing y_d by x_d in $\overline{J_{\rho_n}}$, we have

$$\begin{aligned}\overline{J_{\rho_n}} &= -\int_Q |\varphi(v)-\varphi(\overline{u}_\Sigma(t,\overline{x}))|\,\Delta\,\omega_{1/n}(x)\,(\psi\,\lambda)(t,\overline{x},f(\overline{x}))\,dx\,dt+\overline{J_f}\\ &= \int_Q \nabla|\varphi(v)-\varphi(\overline{u}_\Sigma(t,\overline{x}))|\,\nabla\,\omega_{1/n}(x)\,(\psi\,\lambda)(t,\overline{x},f(\overline{x}))\,dx\,dt+\overline{\varepsilon^1_n}\,.\end{aligned}$$

Here, the quantity $\overline{\varepsilon_n^1} = \overline{J_f} + \int_Q |\varphi(v) - \varphi(\overline{u}_\Sigma(t,\overline{x}))|\, \nabla\omega_{1/n}(x)\cdot\nabla(\psi\,\lambda)_{\Sigma_x}\,dx\,dt$ tends to zero when $n\to+\infty$. Moreover,

$$\overline{I} = -\int_{\mathcal{Q}} \Phi(t,x,v,\overline{u}_\Sigma)\cdot\nabla\omega_{1/n}(x)\,(\psi\,\lambda)_{\Sigma_x}\,d\beta\,dx\,dt\ +\overline{\varepsilon_n^2}\,,$$

where $\overline{\varepsilon_n^2} = \int_{\mathcal{Q}} (\Phi(t,x,v,\overline{u}_\Sigma) - \Phi(t,\overline{x},f(\overline{x}),v,\overline{u}_\Sigma))\cdot\nabla\omega_{1/n}(x)\,(\psi\,\lambda)_{\Sigma_x} d\beta\,dx\,dt$ tends to zero when $n\to+\infty$.

Using formula (11), we get

$$\begin{aligned}&\liminf_{n\to+\infty}\ \lim_{l,m\to+\infty}\ \lim_{\varepsilon\to 0}(A+B)\\ &\qquad= -\limsup_{n\to+\infty}\int_{\mathcal{Q}} \mathcal{G}_x(t,x,v(t,x,\beta),\overline{u}_\Sigma)\cdot\nabla\omega_{1/n}(x)\,(\psi\,\lambda)_\Sigma\,dx\,dt\,d\beta\,.\end{aligned}$$

Starting from inequality (35) and taking the limit with respect to l, m, then the limit with respect to n of both sides yields

$$\int_Q\int_0^1\int_0^1 \left[|u-v|\,(\psi\lambda)_t + \mathcal{G}_x(t,x,u,v)\cdot\nabla(\psi\lambda)\right]\,d\beta\,d\alpha\,dx\,dt \tag{37}$$

$$\geq \left[\begin{array}{l} -\displaystyle\lim_{n\to+\infty}\int_Q\int_0^1 \mathcal{G}_x(t,x,v(t,x,\beta),\overline{u}_\Sigma(t,\overline{x}))\cdot\nabla\omega_{1/n}\,(\psi\,\lambda)(t,\overline{x},f(\overline{x}))\,d\beta\,dx\,dt\\ +\displaystyle\lim_{\varepsilon\to 0}\int_Q\int_0^1 \mathcal{G}_x(t,x,u,\overline{u}_\Sigma(t,\overline{x}))\cdot\nabla\omega_\varepsilon(x)\,(\psi\,\lambda)(t,\overline{x},f(\overline{x}))\,d\alpha\,dx\,dt\\ +\displaystyle\lim_{n\to+\infty}\ \lim_{l,m\to+\infty} H\end{array}\right].$$

Since $\lim_{n\to+\infty}\lim_{l,m\to+\infty} H = 0$ (see (36)), the right-hand side of (37) is an antisymmetric function in (u,v), while the left-hand side of (37) is a symmetric function of (u,v). We therefore have

$$\int_{\mathcal{Q}}\int_0^1 \left[|u-v|\,(\psi\lambda)_t + \mathcal{G}_x(t,x,u,v)\cdot\nabla(\psi\lambda)\right]\,d\beta\,d\alpha\,dx\,dt \geq 0. \tag{38}$$

Now, recall that $\lambda=\lambda_\alpha$ is an element of the partition of unity $(\lambda_\alpha)_{0\leq\alpha\leq N}$; summing the previous inequality over $\alpha\in 0,\dots,N$ yields

$$\int_{\mathcal{Q}}\int_0^1 \left[|u-v|\,\psi_t + \mathcal{G}_x(t,x,u,v)\cdot\nabla\psi\right]\,d\beta\,d\alpha\,dx\,dt \geq 0. \tag{39}$$

We define the nonnegative function ψ_0 by $\psi_0(t,x) = \psi_0(t) = (T-t)\chi_{(0,T)}(t)$, and apply (39) with ψ_0 as a test-function to get

$$\int_0^T\int_\Omega\int_0^1\int_0^1 |u(t,x,\alpha)-v(t,x,\beta)|\,d\beta\,d\alpha\ dx\,dt \leq 0.$$

Consequently, we have $u(t,x,\alpha) = v(t,x,\beta)$ for a.e. $(t,x,\alpha,\beta)\in Q\times(\,0,1)\times(\,0,1)$. Defining the function w by the formula

$$w(t,x) = \int_0^1 u(t,x,\alpha)\,d\alpha$$

and accounting for the product structure of the measurable space $Q\times(\,0,1)\times(\,0,1)$, we conclude

$$u(t,x,\alpha)\ =\ w(t,x)\ =\ v(t,x,\beta)\ \text{for a.e.}\ \ (t,x,\alpha,\beta)\in Q\times(\,0,1)^2. \qquad\square$$

4.6. Proof of Theorem 4.1 for Ω a bounded polyhedral subset. Let d be the Euclidean distance on $\mathbb{R}^d$. Denote by $(\partial\Omega_i)_{i=1,\dots,N}$ the faces of Ω, and by $\mathbf{n}_i$ the outward unit normal to Ω along $\partial\Omega_i$. For $\varepsilon > 0$ small, let B_i^ε be the subset of all $x \in \Omega$ such that $d(x, \partial\Omega_i) < \varepsilon$ and $d(x, \partial\Omega_i) < d(x, \partial\Omega_j)$ if $i \neq j$; define G_i^ε to be the largest cylinder generated by $\mathbf{n}_i$ included in B_i^ε, and set $\Delta_i^\varepsilon = B_i^\varepsilon \setminus G_i^\varepsilon$, $\Omega_\varepsilon = \Omega \setminus (\cup_{1,N} \Delta_i^\varepsilon)$, and $b^\varepsilon = \mathbb{1}_{\Omega_{\varepsilon/2}} \star \rho_{\varepsilon/4}$. We have $\text{meas}(\Omega \setminus \Omega_\varepsilon) \leq C\varepsilon^2$. If $\lambda_i \in \mathcal{C}_c^\infty(\mathbb{R}^d)$ is such that $\text{supp}(\lambda_i) \cap \partial\Omega \subset \partial\Omega_i$ and such that the orthogonal projection of $\text{supp}(\lambda_i)$ on the affine hyperplane determined by $\partial\Omega_i$ is included in $\partial\Omega_i$, then of course the whole previous proof explained in the case where Ω is $\mathcal{C}^{1,1}$ applies here (we look at a half-space), to give a result of comparison on $\text{supp}(\lambda)$. Otherwise, for such a choice of function λ_i, (38) is true. Equation (38) is also still true if $\lambda = \lambda_0$, where $\lambda_0 \in \mathcal{C}_c^\infty(\mathbb{R}^d)$ and $\text{supp}(\lambda_0) \subset \Omega$ (use Proposition 4.2). Since the function b^ε can be written as $b^\varepsilon = \sum_{i=0,N} \lambda_i$ for functions λ_i as above, we have

$$\int_{\mathcal{Q}} \int_0^1 \left[|u - v| \, (\psi b^\varepsilon)_t + \mathcal{G}_x(t, x, u, v) \cdot \nabla(\psi b^\varepsilon) \right] d\beta \, d\alpha \, dx \, dt \geq 0. \tag{40}$$

Equation (40) can be rewritten as

$$\int_{\mathcal{Q}} \int_0^1 \left[|u - v| \, \psi_t + \mathcal{G}_x(t, x, u, v) \cdot \nabla\psi \right] d\beta \, d\alpha \, dx \, dt \geq \alpha_\varepsilon,$$

where $\alpha_\varepsilon = \int_{\mathcal{Q}} \int_0^1 \mathcal{G}_x(t, x, u, v) \cdot \nabla b^\varepsilon \psi \, d\beta \, d\alpha \, dx \, dt$ tends to zero when $\varepsilon \to 0$. Indeed, we have $\nabla b^\varepsilon = 0$ on Ω_ε, so that, setting $\mathcal{R}_\varepsilon = (0, T) \times (\Omega \setminus \Omega_\varepsilon) \times (0, 1)^2$, we have

$$\begin{aligned} \alpha_\varepsilon &\leq ||\psi||_{L^\infty} ||\mathcal{G}_x(t, x, u, v)||_{L^1(\mathcal{R}_\varepsilon)} ||\nabla b^\varepsilon||_{L^\infty(\mathcal{R}_\varepsilon)} \\ &\leq ||\psi||_{L^\infty} \text{meas}(\mathcal{R}_\varepsilon)^{1/2} ||\mathcal{G}_x(t, x, u, v)||_{L^2(\mathcal{R}_\varepsilon)} || \mathbb{1}_{\Omega_{\varepsilon/2}} ||_{L^\infty(\mathcal{R}_\varepsilon)} ||\nabla \rho_{\varepsilon/4}||_{L^1(\mathcal{R}_\varepsilon)} \\ &\leq C(T, \psi) \;\; \varepsilon \cdot ||\mathcal{G}_x(t, x, u, v)||_{L^2(\mathcal{R}_\varepsilon)} \cdot \tfrac{1}{\varepsilon}, \end{aligned}$$

and we conclude by using $||\mathcal{G}_x(t, x, u, v)||_{L^2(\mathcal{R}_\varepsilon)} \to 0$ when $\varepsilon \to 0$. We thus obtain (39), from which Theorem 4.1 follows. □

5. The FV scheme.

The mesh used to discretize problem (1) has to be regular enough to ensure the consistency of the numerical fluxes, mainly because a second order problem is considered (at least when the function φ is not constant). This is specified in the following section.

5.1. Assumptions and notation.

We set d to be the Euclidean distance on $\mathbb{R}^d$ and denote by γ the $(d-1)$-Hausdorff measure on $\partial\Omega$.

Definition 5.1 (admissible mesh of Ω). *An admissible mesh of Ω consists of a set $\mathcal{T}$ of open bounded polyhedral convex subsets of Ω called control volumes, a family $\mathcal{E}$ of subsets of $\bar{\Omega}$ contained in hyperplanes of $\mathbb{R}^d$ with positive measure, and a family of points (the "centers" of control volumes) satisfying the following properties:*

(i) *The closure of the union of all control volumes is $\bar{\Omega}$.*

(ii) *For any $K \in \mathcal{T}$, there exists a subset $\mathcal{E}_K$ of $\mathcal{E}$ such that $\partial K = \bar{K} \backslash K = \cup_{\sigma \in \mathcal{E}_K} \bar{\sigma}$. Furthermore, $\mathcal{E} = \cup_{K \in \mathcal{T}} \mathcal{E}_K$.*

(iii) *For any $(K, L) \in \mathcal{T}^2$ with $K \neq L$, either the "length" (i.e., the $(d-1)$-dimensional Lebesgue measure) of $\bar{K} \cap \bar{L}$ is 0 or $\bar{K} \cap \bar{L} = \bar{\sigma}$ for some $\sigma \in \mathcal{E}$. In the latter case, we shall write $\sigma = K|L$ and $\mathcal{E}_{int} = \{\sigma \in \mathcal{E}, \exists (K, L) \in \mathcal{T}^2, \sigma = K|L\}$. For any $K \in \mathcal{T}$, we shall denote by $\mathcal{N}_K$ the set of neighbor control volumes of K, i.e., $\mathcal{N}_K = \{L \in \mathcal{T}, K|L \in \mathcal{E}_K\}$.*

(iv) *The family of points $(x_K)_{K\in\mathcal{T}}$ is such that $x_K \in K$ (for all $K \in \mathcal{T}$), and, if $\sigma = K|L$, it is assumed that the straight line (x_K, x_L) is orthogonal to σ.*

Given a control volume $K \in \mathcal{T}$, we will denote by $m(K)$ its measure and by $\mathcal{E}_{ext,K}$ the subset of the edges of K included in the boundary $\partial\Omega$. If $L \in \mathcal{N}_K$, $m(K|L)$ will denote the measure of the edge between K and L, and $T_{K|L}$ the "transmissibility" through $K|L$, defined by $T_{K|L} = \frac{m(K|L)}{d(x_K,x_L)}$. Similarly, if $\sigma \in \mathcal{E}_{ext,K}$, we will denote by $m(\sigma)$ its measure and by τ_σ the "transmissibility" through σ, defined by $\tau_\sigma = \frac{m(\sigma)}{d(x_K,\sigma)}$. One also denotes by $\mathcal{E}_{ext}$ the union of the edges included in the boundary of Ω: $\cup_{K\in\mathcal{T}}\mathcal{E}_{ext,K}$. The size of the mesh $\mathcal{T}$ is defined by

$$\text{size}(\mathcal{T}) = \max_{K\in\mathcal{T}} \text{diam}(K),$$

and we introduce the following geometrical factor, linked with the regularity of the mesh, defined by

$$\text{reg}(\mathcal{T}) = \min_{K\in\mathcal{T},\sigma\in\mathcal{E}_K} \frac{d(x_K,\sigma)}{\text{diam}(K)}.$$

Remark 5.1. Some examples of meshes satisfying these assumptions are the triangular meshes, which verify the acute angle condition (in fact this condition may be weakened to the Delaunay condition), the rectangular meshes, or the Voronoï meshes; see [EGH99] or [EGH00] for more details.

Definition 5.2 (time discretization of $(0,T)$). *A time discretization of $(0,T)$ is given by an integer value N and by an increasing sequence of real values $(t^n)_{n\in[\![0,N+1]\!]}$ with $t^0 = 0$ and $t^{N+1} = T$. The time steps are then defined by $\delta t^n = t^{n+1} - t^n$, for $n \in [\![0,N]\!]$.*

Definition 5.3 (space-time discretization of Q). *A finite volume discretization D of Q is a family $D = (\mathcal{T}, \mathcal{E}, (x_K)_{K\in\mathcal{T}}, N, (t^n)_{n\in[\![0,N]\!]})$, where $\mathcal{T}$, $\mathcal{E}$, $(x_K)_{K\in\mathcal{T}}$ is an admissible mesh of Ω according to Definition 5.1 and N, $(t^n)_{n\in[\![0,N+1]\!]}$ is a time discretization of $(0,T)$ according to Definition 5.2. For a given FV discretization D, one defines*

$$\text{size}(D) = \max(\text{size}(\mathcal{T}), (\delta t^n)_{n\in[\![0,N]\!]}) \quad \textit{and} \quad \text{reg}(D) = \text{reg}(\mathcal{T}).$$

5.2. The FV scheme. We may now define the FV discretization of (1). Let D be a FV discretization of Q according to Definition 5.3. First, the initial and boundary data are discretized by setting

$$U_K^0 = \frac{1}{m(K)} \int_K u_0(x)dx \quad \forall K \in \mathcal{T} \tag{41}$$

and

$$\bar{U}_\sigma^{n+1} = \frac{1}{\delta t^n\ m(\sigma)} \int_{t^n}^{t^{n+1}} \int_\sigma \bar{u}(t,x)d\gamma(x)dt \quad \forall\sigma \in \mathcal{E}_{ext}, \forall n \in [\![0,N]\!]. \tag{42}$$

An *implicit FV scheme* for the discretization of problem (1) is given by the following set of nonlinear equations with unknowns $U_D = (U_K^{n+1})_{K\in\mathcal{T},n\in[\![0,N]\!]}$: $\forall K \in \mathcal{T}, \forall n \in [\![0,N]\!]$,

(43)
$$\frac{U_K^{n+1} - U_K^n}{\delta t^n}\ m(K) + \sum_{\sigma\in\mathcal{E}_K} m(\sigma) F_{K,\sigma}^{n+1}(U_K^{n+1}, U_{K_\sigma}^{n+1}) - \sum_{\sigma\in\mathcal{E}_K} \tau_\sigma(\varphi(U_{K_\sigma}^{n+1}) - \varphi(U_K^{n+1})) = 0,$$

where

$$U_{K_\sigma}^{n+1} = \begin{cases} U_L^{n+1} \ \textit{if } \sigma = K|L, \\ \bar{U}_\sigma^{n+1} \ \textit{if } \sigma \in \mathcal{E}_{ext}, \end{cases} \tag{44}$$

and where the function $F_{K,\sigma}^{n+1}$ is a monotonous flux consistent with the function F, which means that

- for all $v \in \mathbb{R}$, $u \mapsto F_{K,\sigma}^{n+1}(u,v)$ is a nondecreasing function and for all $u \in \mathbb{R}$, $v \to F_{K,\sigma}^{n+1}(u,v)$ is a nonincreasing function,
- $F_{K,\sigma}^{n+1}(u,v) = -F_{K,\sigma}^{n+1}(v,u)$ for all $(u,v) \in \mathbb{R}^2$,
- $F_{K,\sigma}^{n+1}$ is M-Lipschitz continuous with respect to each variable,
- $F_{K,\sigma}^{n+1}(s,s) = \frac{1}{\delta t_n}\frac{1}{m(\sigma)}\int_{t^n}^{t^{n+1}}\int_\sigma F(x,t,s)\cdot \mathbf{n}_{K,\sigma} d\gamma(x)\,dt$.

The Godunov scheme and the splitting flux scheme of Osher may be the most common examples of schemes with monotone fluxes.

We call an *approximate solution* the piecewise constant function u_D defined a.e. on Q by

$$u_D(t,x) = U_K^{n+1}\,, \quad t \in (t_n, t_{n+1}),\ x \in K. \tag{45}$$

5.3. Monotony of the scheme and direct consequences. As already said in the introduction, it is a necessity to select a physically admissible solution by means of the entropy inequalities. The schemes with monotonous fluxes are well known to add numerical viscosity to the equations. They are L^∞ stable, and they are monotonous so that they respect discrete entropy inequalities. In other words, continuous entropy inequalities have their discrete analogue, and they are respected by any solution of (41)–(44). This is summarized in the following proposition.

PROPOSITION 5.1 (monotony). *Assume hypotheses* (H1), (H2)*,* (H3)*, and* (H4)*. Then there exists a unique solution to the scheme. Moreover, this solution satisfies the following maximum principle and discrete entropy inequalities:* $\forall K \in \mathcal{T}, \forall n \in [\![0,N]\!]$*,*

$$A \le U_K^{n+1} \le B, \tag{46}$$

$$\begin{aligned} \frac{\eta_\kappa^\pm(U_K^{n+1}) - \eta_\kappa^\pm(U_K^n)}{\delta t^n} m(K) + \sum_{\sigma\in\mathcal{E}_K} m(\sigma)\Phi_{K,\sigma,\kappa}^{\pm,n+1}(U_K^{n+1}, U_{K_\sigma}^{n+1}) \\ - \sum_{\sigma\in\mathcal{E}_K} \tau_\sigma\big(\eta_\kappa^\pm(\varphi(U_{K_\sigma}^{n+1})) - \eta_\kappa^\pm(\varphi(U_K^{n+1}))\big) \le 0, \end{aligned} \tag{47}$$

where $\Phi_{K,\sigma,\kappa}^{+,n+1}$ *and* $\Phi_{K,\sigma,\kappa}^{-,n+1}$ *are the numerical entropy-fluxes defined by*

$$\begin{aligned} \Phi_{K,\sigma,\kappa}^{+,n+1}(u,v) &= F_{K,\sigma}^{n+1}(u\top\kappa, v\top\kappa) - F_{K,\sigma}^{n+1}(\kappa,\kappa) \quad \textit{and} \\ \Phi_{K,\sigma,\kappa}^{-,n+1}(u,v) &= F_{K,\sigma}^{n+1}(\kappa,\kappa) - F_{K,\sigma}^{n+1}(u\bot\kappa, v\bot\kappa). \end{aligned} \tag{48}$$

Proof. We give only some elements of the proof of this proposition because it consists of rewriting the proofs of three lemmas that can be found in [EGHM02] (Lemmas 3.1, 3.3, and 3.4 there) in the case where the convective flux $\mathbf{q}(x,t)f(u)$ is replaced by a more general flux $F(x,t,u)$ and the Kruzhkov entropies are replaced by the semi-Kruzhkov entropies as in the work of Vovelle [Vov02].

We follow the classical framework of implicit FV schemes for conservation laws (see [EGH00]). The function U_D is defined in an implicit way, so we first show, using the monotony of the scheme, that if a function U_D is a solution to the scheme, then it satisfies the discrete inequalities (47). Then we derive the maximum principle (46) that provides a result of existence by use of the Leray–Schauder theorem. Uniqueness of U_D is proved by using a method analogous to the one used to prove the discrete entropy inequalities. □

5.4. A priori estimates. The inequalities derived from the properties of monotony and local conservation are L^∞ and L^1 estimates. We will prove now L^2 estimates. We introduce a discretization $\bar{U}_D = (\bar{U}_K^{n+1})_{\{K\in\mathcal{T}, n\in[\![0,N]\!]\}}$ of $\bar{u}$ defined by

$$\bar{U}_K^{n+1} = \frac{1}{\delta t_n}\frac{1}{m(K)}\int_{t^n}^{t^{n+1}}\int_K \bar{u}\,dx\,dt \quad \forall K\in\mathcal{T},\ \forall n\in[\![0,N]\!].$$

Proposition 5.2 ($L^2(0,T,H^1(\Omega))$ and weak BV estimate). *Assume hypotheses* (H1), (H2), (H3), (H4), *and* (H5). *Let u_D be the approximate solution defined by* (41)–(44), *and assume that $reg(D)\geq\xi$, where $\xi>0$. Then there exists a constant C depending only on ξ, T, Ω, $Lip(\varphi)$, M, $\bar{u}$, A, B such that*

$$\big(\mathcal{N}_D(\zeta(u_D))\big)^2 = \sum_{n=0}^{N}\delta t_n\sum_{K\in\mathcal{T}}\left(\frac{1}{2}\sum_{\sigma\in\mathcal{E}_{int,K}}\tau_\sigma(\zeta(U_K^{n+1})-\zeta(U_{K_\sigma}^{n+1}))^2 + \sum_{\sigma\in\mathcal{E}_{ext,K}}\tau_\sigma(\zeta(U_K^{n+1})-\zeta(U_{K_\sigma}^{n+1}))^2\right)\leq C$$

and

$$\sum_{n=0}^{N}\delta t_n\sum_{K\in\mathcal{T}}\frac{1}{2}\sum_{\sigma\in\mathcal{E}_{int,K}}m(\sigma)\max_{U_K^{n+1}\leq c\leq d\leq U_{K_\sigma}^{n+1}}\big((F_{K,\sigma}^{n+1}(d,c)-F_{K,\sigma}^{n+1}(d,d))^2 + (F_{K,\sigma}^{n+1}(d,c)-F_{K,\sigma}^{n+1}(c,c))^2\big)\leq C. \tag{49}$$

Remark 5.2. The inequality (49) is called the "weak BV inequality." See [EGH00], [CGH93], or [CH99].

Proof. As for Proposition 5.1, the proof has already been done in a simpler case in [EGHM02] (Proposition 3.1). The details of the proof differ only by some arguments that can be found in [Vov02].

These estimates are discrete energy estimates. They are obtained by multiplying (41)–(44) by $\delta t_n(U_K^{n+1}-\bar{U}_K^{n+1})$ and summing over $K\in\mathcal{T}$ and $n\in[\![0,N]\!]$. In the proof, we separate terms that contain only U_D from terms containing U_D and $\bar{U}_D$. Then we use the Cauchy–Schwarz inequality and regularity hypotheses (H5) on $\bar{u}$ to control the second type of terms. To get a bound on $\mathcal{N}_D(\bar{U}_D)$, which is a discrete $L^2(0,T,H^1)$-norm for $\bar{U}_D$, we use the following inequality proved in [EGH99]:

$$\mathcal{N}_D(\bar{u})\leq C(reg(D))\|\nabla\bar{u}\|_{L^2(Q)}.$$

This is a consequence of the local conservativity of the scheme combined with the consistency of the numerical fluxes.

The last ingredient is the assumption $\mathrm{div}_x(F(x,t,u))=0$, which ensures that the boundary terms in the discrete integrations-by-parts concerning the hyperbolic terms can be controlled. The constant C depends on ξ, $m(\Omega)$, T, B, A, $Lip(F_{K,\sigma}^{n+1})$, $\|\bar{u}_t\|_{L^1(Q)}$, and on $\|\nabla\bar{u}\|_{L^2(Q)}$. □

5.5. Continuous entropy inequalities. From the discrete entropy inequalities we deduce continuous approximate entropy inequalities. The following theorem is central in the proof of the convergence of the scheme.

THEOREM 5.1 (continuous approximate entropy inequalities). *Assume hypotheses* (H1), (H2), (H3), (H4), *and* (H5). *Let D be an admissible discretization of Q, and let u_D be the corresponding approximate solution defined above. Then u_D satisfies the following approximate entropy inequalities: for all $\kappa \in \mathbb{R}$, for all $\psi \in \mathcal{C}^\infty(\mathbb{R}_+ \times \mathbb{R}^d)$ such that $\psi \geq 0$ and $(\varphi(\bar{u}) - \varphi(\kappa))^{\pm}\,\psi = 0$ a.e. on Σ,*

$$\int_Q \eta_\kappa^\pm(u_D)\psi_t + \Phi_\kappa^\pm(t,x,u_D)\cdot\nabla\psi + \eta_{\varphi(\kappa)}^\pm(\varphi(u_D))\Delta\psi\, dxdt - \int_\Sigma \eta_{\varphi(\kappa)}^\pm(\varphi(\bar{u}))\nabla\psi\cdot\mathbf{n}d\gamma(x)dt$$

$$+\int_\Omega \eta_\kappa^\pm(u_0)\psi(0)dx + M\int_\Sigma \eta_\kappa^\pm(\bar{u})\psi\, d\gamma(x)dt \geq -\mathcal{E}_D^\pm(\psi). \tag{50}$$

Also assume that a uniform CFL condition $\delta t_n \leq C\mathrm{size}(\mathcal{T})$ for all n holds true (with a CFL number C that can be as large as desired). Then, for a given ψ, $\mathcal{E}_D^\pm(\psi)$ tends to zero when the size of the discretization tends to zero.

Proof. The proof of Theorem 5.1 is quite similar to the proof of Theorem 5.1 in [EGHM02], except for the boundary terms, which require extra care. We will therefore stress the analysis of these terms and make reference to [EGHM02] when needed. Of course, we can also limit ourselves to giving the proof of (50) when the nonnegative Kruzhkov entropy pairs are under consideration.

Let $\kappa \in \mathbb{R}$, and let $\psi \in \mathcal{C}^\infty(\mathbb{R}_+ \times \mathbb{R}^d)$ be a nonnegative function satisfying $(\varphi(\bar{u}) - \varphi(\kappa))^+\psi = 0$ a.e. on Σ. We define discrete values of ψ with respect to the mesh as

$$\Psi_K^0 = \psi(0,x_K) \quad \forall K \in \mathcal{T},$$
$$\Psi_K^{n+1} = \frac{1}{\delta t_n}\int_{t^n}^{t^{n+1}} \psi(t,x_K)dt \quad \forall K \in \mathcal{T}, \forall n \in [\![0,N]\!],$$
$$\psi_\sigma^{n+1} = \frac{1}{\delta t_n}\int_{t^n}^{t^{n+1}} \psi(t,x_\sigma)dt \quad \forall \sigma \in \mathcal{E}_{ext}, \forall n \in [\![0,N]\!]$$

and set $\Psi_{K,\sigma}^{n+1} = \Psi_L^{n+1}$ if $\sigma = K|L$ and $\Psi_{K,\sigma}^{n+1} = \psi_\sigma^{n+1}$ if $\sigma \in \mathcal{E}_{ext,K}$.

The definition of the numerical flux $\Phi_{K,\sigma,\kappa}^{+,n+1}$ (see (48)) ensures that it is a conservative flux, consistent with the function Φ_κ^+. Therefore, we have

$$\sum_{\sigma\in\mathcal{E}_K} m(\sigma)\Phi_{K,\sigma,\kappa}^{+,n+1}(U_K^{n+1}, U_K^{n+1}) = 0 \quad \forall K \in \mathcal{T}, n \in [\![0,N]\!],$$

and the discrete entropy inequality (47) can then be rewritten as

$$\frac{\eta_\kappa^+(U_K^{n+1}) - \eta_\kappa^+(U_K^n)}{\delta t^n}m(K) + \sum_{\sigma\in\mathcal{E}_K} m(\sigma)(\Phi_{K,\sigma,\kappa}^{+,n+1}(U_K^{n+1}, U_{K_\sigma}^{n+1}) - \Phi_{K,\sigma,\kappa}^{+,n+1}(U_K^{n+1}, U_K^{n+1}))$$
$$- \sum_{\sigma\in\mathcal{E}_K} \tau_\sigma\big(\eta_{\varphi(\kappa)}^+(\varphi(U_{K_\sigma}^{n+1})) - \eta_{\varphi(\kappa)}^+(\varphi(U_K^{n+1}))\big) \leq 0. \tag{51}$$

Multiplying (51) by $\delta t_n \Psi_K^{n+1}$ and summing over $K \in \mathcal{T}$ and $n \in [\![0,N]\!]$ yields

$$A1 + A2 + A3 \leq 0,$$

where

$$A1 = \sum_{n=0}^{N} \sum_{K\in\mathcal{T}} m(K)(\eta_\kappa^+(U_K^{n+1}) - \eta_\kappa^+(U_K^n))\Psi_K^{n+1}\,,$$

and, summing over the edges, $A2 = A2\text{int} + A2\text{ext}$, with

$$A2\text{int} = \sum_{n=0}^{N} \delta t_n \sum_{K\in\mathcal{T}} \frac{1}{2} \sum_{\sigma\in\mathcal{E}_{int,K}} m(\sigma)\big(\Psi_K^{n+1}(\Phi_{K,\sigma,\kappa}^{+,n+1}(U_K^{n+1}, U_{K_\sigma}^{n+1}) - \Phi_{K,\sigma,\kappa}^{+,n+1}(U_K^{n+1}, U_K^{n+1})) \\ - \Psi_{K,\sigma}^{n+1}(\Phi_{K,\sigma,\kappa}^{+,n+1}(U_K^{n+1}, U_{K_\sigma}^{n+1}) - \Phi_{K,\sigma,\kappa}^{+,n+1}(U_{K_\sigma}^{n+1}, U_{K_\sigma}^{n+1}))\big)$$

and

$$A2\text{ext} = \sum_{n=0}^{N} \delta t_n \sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}_{ext,K}} m(\sigma)\Psi_K^{n+1}\big(\Phi_{K,\sigma,\kappa}^{+,n+1}(U_K^{n+1}, U_{K_\sigma}^{n+1}) - \Phi_{K,\sigma,\kappa}^{+,n+1}(U_K^{n+1}, U_K^{n+1})\big).$$

Similarly, $A3$ admits the decomposition $A3 = A3\text{int} + A3\text{ext}$, with

$$A3\text{int} = \sum_{n=0}^{N} \delta t_n \sum_{K\in\mathcal{T}} \frac{1}{2} \sum_{\sigma\in\mathcal{E}_{int,K}} \tau_\sigma\big(\eta_{\varphi(\kappa)}^+(\varphi(U_K^{n+1})) - \eta_{\varphi(\kappa)}^+(\varphi(U_{K_\sigma}^{n+1}))\big)(\Psi_K^{n+1} - \Psi_{K,\sigma}^{n+1})$$

and

$$A3\text{ext} = \sum_{n=0}^{N} \delta t_n \sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}_{ext,K}} \tau_\sigma\big(\eta_{\varphi(\kappa)}^+(\varphi(U_K^{n+1})) - \eta_{\varphi(\kappa)}^+(\varphi(U_{K_\sigma}^{n+1}))\big)\Psi_K^{n+1}\,.$$

Now, set

$$\begin{aligned} I1 &= -\int_Q \eta_\kappa^+(u_D)\psi_t\,dx\,dt - \int_\Omega \eta_\kappa^+(u_0)\psi(0,x)\,dx, \\ I2 &= -\int_Q \Phi_\kappa^+(t,x,u_D)\cdot\nabla\psi\,dx\,dt - M\int_\Sigma \eta_\kappa^+(\bar{u})\psi\,d\gamma(x)dt, \\ I3 &= -\int_Q \eta_{\varphi(\kappa)}^+(\varphi(u_D))\Delta\psi\,dx\,dt + \int_\Sigma \eta_{\varphi(\kappa)}^+(\varphi(\bar{u}))\nabla\psi\cdot\mathbf{n}\,d\gamma(x)\,dt\,. \end{aligned}$$

We aim at proving the estimate $I1 + I2 + I3 \leq \mathcal{E}_D^+(\psi)$ and, to that purpose, compare $I1$ to $A1$, $I2$ to $A2$, and $I3$ to $A3$, respectively.

A discrete integration by parts leads to $|I1 - A1| \leq \mathcal{E}_{1,D}(\psi)$, with $\mathcal{E}_{1,D}(\psi) \to 0$ as $\text{size}(D) \to 0$ (see [EGHM02]).

Using integration by parts in $I2$ and the fact that u_D is piecewise constant, we obtain

$$I2 = I2\text{int} + I2\text{ext},$$

where $I2$ext is the boundary term and $I2$int gathers the sums on the internal edges. Precisely, we have

$$I2\text{int} = -\sum_{n=0}^{N} \sum_{K\in\mathcal{T}} \frac{1}{2} \sum_{\sigma\in\mathcal{E}_{int,K}} \left(\int_{t^n}^{t^{n+1}}\int_\sigma \Phi_\kappa^+(t,x,U_K^{n+1})\cdot\mathbf{n}_{K,\sigma}\psi\,d\gamma(x)\,dt \right. \\ \left. -\int_{t^n}^{t^{n+1}}\int_\sigma \Phi_\kappa^+(t,x,U_{K_\sigma}^{n+1})\cdot\mathbf{n}_{K,\sigma}\psi\,d\gamma(x)\,dt\right)$$

and

$$I2\text{ext} = -\sum_{n=0}^{N}\sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{E}_{ext,K}}\int_{t^n}^{t^{n+1}}\int_{\sigma}\Phi_{\kappa}^{+}(t,x,U_K^{n+1})\cdot\mathbf{n}_{K,\sigma}\psi d\gamma(x)dt - M\int_{\Sigma}\eta_{\kappa}^{+}(\bar{u})\psi d\gamma(x)dt.$$

As in [EGHM02], we prove $|I2\text{int} - A2\text{int}| \le \mathcal{E}_{2,D}^{\text{int}}(\psi)$, with $\mathcal{E}_{2,D}^{\text{int}}(\psi) \to 0$ as $\text{size}(D) \to 0$.

The comparison of $I2$ext with $A2$ext involves a term corresponding to the consistency error, and three terms related to the approximation of the boundary data:

$$I2\text{ext} - A2\text{ext} \le \mathcal{E}_{2,D}^{c1,\text{ext}}(\psi) + \mathcal{E}_{2,D}^{b1,\text{ext}}(\psi) + \mathcal{E}_{2,D}^{b2,\text{ext}}(\psi) - T_{2,D}^{b2,\text{ext}}(\psi),$$

where

$$\mathcal{E}_{2,D}^{c1,\text{ext}}(\psi) = \sum_{n=0}^{N}\sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{E}\text{ext}\kappa}\left|\int_{t^n}^{t^{n+1}}\int_{\sigma}(\Psi_K^{n+1}-\psi)\Phi_{\kappa}^{+}(\cdot,\cdot,U_K^{n+1})\cdot\mathbf{n}_{K,\sigma}\,d\gamma(x)\,dt\right|,$$

$$\mathcal{E}_{2,D}^{b1,\text{ext}}(\psi) = \sum_{n=0}^{N}\delta t_n\sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{E}\text{ext}\kappa} m(\sigma)|(\Psi_K^{n+1}-\psi_{\sigma}^{n+1})\Phi_{K,\sigma,\kappa}^{+,n+1}(U_K^{n+1},U_{K_\sigma}^{n+1})|,$$

and

$$\mathcal{E}_{2,D}^{b2,\text{ext}}(\psi) = M\sum_{n=0}^{N}\sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{E}\text{ext}\kappa}\Bigg|\int_{t^n}^{t^{n+1}}\int_{\sigma}(\bar{u}-\kappa)^{+}\psi\,d\gamma(x)\,dt$$
$$-\,\delta t_n\,m(\sigma)(U_{K_\sigma}^{n+1}-\kappa)^{+}\psi_{\sigma}^{n+1}\Bigg|$$

are three terms converging to zero when size$(D) \to 0$ and

$$T_{2,D}^{b2,\text{ext}}(\psi) = \sum_{n=0}^{N}\delta t_n\sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{E}\text{ext}\kappa} m(\sigma)\psi_{\sigma}^{n+1}\left(\Phi_{K,\sigma,\kappa}^{+,n+1}(U_K^{n+1},U_{K_\sigma}^{n+1}) + M(U_{K_\sigma}^{n+1}-\kappa)^{+}\right).$$

From the definition of $\Phi_{K,\sigma,\kappa}^{+,n+1}$ (see (48)) and from the monotony of the scheme,

$$\Phi_{K,\sigma,\kappa}^{+,n+1}(a,b) = F_{K,\sigma}^{n+1}(a\top\kappa, b\top\kappa) - F_{K,\sigma}^{n+1}(\kappa,\kappa) \ge -Lip(F_{K,\sigma}^{n+1})(b-\kappa)^{+}$$

follows, and this entails $T_{2,D}^{b2,\text{ext}}(\psi) \ge 0$.

Now, to compare $I3$ to $A3$ we make the distinction between the different contributions of the terms (inside and on the boundary of Ω). Indeed, since the approximate solution u_D is piecewise constant, the term $I3$ reads as $I3 = I3\text{int} + I3\text{ext}$, where

$$I3\text{int} = \sum_{n=0}^{N}\sum_{K\in\mathcal{T}}\frac{1}{2}\sum_{\sigma\in\mathcal{E}\text{int}\kappa}(\eta_{\varphi(\kappa)}^{+}(\varphi(U_{K_\sigma}^{n+1})) - \eta_{\varphi(\kappa)}^{+}(\varphi(U_K^{n+1})))\int_{t^n}^{t^{n+1}}\int_{\sigma}\nabla\psi\cdot\mathbf{n}_{K,\sigma}\,d\gamma(x)\,dt$$

and

$$I3\text{ext} = \sum_{n=0}^{N}\sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{E}\text{ext}\kappa}\int_{t^n}^{t^{n+1}}\int_{\sigma}(\eta_{\varphi(\kappa)}^{+}(\varphi(\bar{u})) - \eta_{\varphi(\kappa)}^{+}(\varphi(U_K^{n+1})))\nabla\psi\cdot\mathbf{n}_{K,\sigma}\,d\gamma(x)\,dt.$$

A consistency error term controls the proximity of $A3\text{int}$ to $I3\text{int}$:

$$|I3\text{int} - A3\text{int}| \leq \mathcal{E}_{3,D}^{c,\text{int}}(\psi),$$

with $\mathcal{E}_{3,D}^{c,\text{int}}(\psi) \to 0$ when size$(D) \to 0$ [EGHM02].

In order to compare $I3\text{ext}$ and $A3\text{ext}$, rearrange the term $I3\text{ext}$, up to consistency or approximation errors, to get

$$I3\text{ext} \leq \sum_{n=0}^{N} \delta t_n \sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}\text{ext}\kappa} \tau_\sigma \left(\eta^+_{\varphi(\kappa)}(\varphi(U^{n+1}_{K_\sigma})) - \eta^+_{\varphi(\kappa)}(\varphi(U^{n+1}_K))\right)(\Psi^{n+1}_{K,\sigma} - \Psi^{n+1}_K) \\ + \mathcal{E}_{3,D}^{c,\text{ext}}(\psi) + \mathcal{E}_{3,D}^{b1,\text{ext}}(\psi),$$

where

$$\mathcal{E}_{3,D}^{c,\text{ext}}(\psi) \\ = \sum_{n=0}^{N} \sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}\text{ext}\kappa} 2 \max_{u\in[A,B]} \eta^+_{\varphi(\kappa)}(\varphi(u)) \left| \int_{t^n}^{t^{n+1}} \int_\sigma \left(\nabla\psi\cdot\mathbf{n} - \frac{\psi^{n+1}_\sigma - \Psi^{n+1}_K}{d_K,\sigma}\right) d\gamma(x)\, dt \right|,$$

$$\mathcal{E}_{3,D}^{b1,\text{ext}}(\psi) = \sum_{n=0}^{N} \sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}\text{ext}\kappa} \int_{t^n}^{t^{n+1}} \int_\sigma \left|\varphi(\bar{u}) - \varphi(\bar{U}^{n+1}_\sigma)\right| \left|\nabla\psi\cdot\mathbf{n}\right| d\gamma(x)\, dt.$$

Then we have

$$I3\text{ext} - A3\text{ext} = \sum_{n=0}^{N} \delta t_n \sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}\text{ext}\kappa} \tau_\sigma(\eta^+_{\varphi(\kappa)}(\varphi(U^{n+1}_{K_\sigma})) - \eta^+_{\varphi(\kappa)}(\varphi(U^{n+1}_K)))\Psi^{n+1}_{K,\sigma} \\ + \mathcal{E}_{3,D}^{c,\text{ext}}(\psi) + \mathcal{E}_{3,D}^{b1,\text{ext}}(\psi)\,.$$

Now, either $\eta^+_{\varphi(\kappa)}(\varphi(U^{n+1}_{K_\sigma})) = 0$, and in that case

$$(\eta^+_{\varphi(\kappa)}(\varphi(U^{n+1}_{K_\sigma})) - \eta^+_{\varphi(\kappa)}(\varphi(U^{n+1}_K))\Psi^{n+1}_{K,\sigma}) \leq 0,$$

or $\eta^+_{\varphi(\kappa)}(\varphi(U^{n+1}_{K_\sigma})) > 0$. In the latter case, the condition $(\varphi(\bar{u}) - \varphi(\kappa))^+\,\psi = 0$ a.e. on Σ ensures that there exists $(t,x) \in [t^n, t^{n+1}] \times \sigma$ such that $\psi(t,x) = 0$. Consequently, we have

$$\Psi^{n+1}_{K,\sigma} \leq Lip(\psi)(\delta t_n + diam(\sigma))\,.$$

This estimate, combined with the inequality

$$\eta^+_{\varphi(\kappa)}(\varphi(U^{n+1}_{K_\sigma})) - \eta^+_{\varphi(\kappa)}(\varphi(U^{n+1}_K)) \leq (\eta^+_{\varphi(\kappa)})'(\varphi(\bar{U}^{n+1}_\sigma))(\varphi(\bar{U}^{n+1}_\sigma) - \varphi(U^{n+1}_K)),$$

which is consequence of the convexity of the function $\eta^+_{\varphi(\kappa)}$, leads to

$$I3\text{ext} - A3\text{ext} \leq \mathcal{E}_{3,D}^{b2,\text{ext}}(\psi) + \mathcal{E}_{3,D}^{c,\text{ext}}(\psi) + \mathcal{E}_{3,D}^{b1,\text{ext}}(\psi),$$

where

$$\mathcal{E}_{3,D}^{b2,\text{ext}}(\psi) = \sum_{n=0}^{N} \delta t_n \sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}\text{ext}\kappa} \tau_\sigma Lip(\psi)(\delta t_n + diam(\sigma))|\varphi(\bar{U}^{n+1}_\sigma) - \varphi(U^{n+1}_K)|\,.$$

Using the Cauchy–Schwarz inequality, together with the $L^2(0,T;H_0^1(\Omega))$ estimate of Proposition 5.2 and the inequality $\varphi(a)-\varphi(b) \le \sqrt{Lip(\varphi)}(\zeta(a)-\zeta(b))$, yields

$$\mathcal{E}_{3,D}^{b2,\mathrm{ext}}(\psi) \le C\sum_{n=0}^{N}\delta t_n \sum_{K\in\mathcal{T}}\ \sum_{\sigma\in\mathcal{E}\mathrm{ext}\kappa} \tau_\sigma(\delta t_n + diam(\sigma))^2 \,.$$

Therefore, a simple way to ascertain that $\mathcal{E}_{3,D}^{b2,\mathrm{ext}}(\psi)$ converges to zero is to suppose a uniform CFL condition such as $\delta t_n \le C\mathrm{size}(\mathcal{T})$ for all n (where the CFL number C can be as large as desired). Then we conclude the proof of Theorem 5.1 by defining $\mathcal{E}_D^+(\psi)$ as the sum of the errors $\mathcal{E}_{1,D}(\psi)$, $\mathcal{E}_{2,D}^{\mathrm{int}}(\psi)$, $\mathcal{E}_{2,D}^{c1,\mathrm{ext}}(\psi)$, $\mathcal{E}_{2,D}^{b1,\mathrm{ext}}(\psi)$, $\mathcal{E}_{2,D}^{b2,\mathrm{ext}}(\psi)$, $\mathcal{E}_{3,D}^{c,\mathrm{int}}(\psi)$, $\mathcal{E}_{3,D}^{c,\mathrm{ext}}(\psi)$, $\mathcal{E}_{3,D}^{b1,\mathrm{ext}}(\psi)$, and $\mathcal{E}_{3,D}^{b2,\mathrm{ext}}(\psi)$. □

5.6. Convergence of the scheme. Let D_n be a sequence of discretizations, such that size(D_n) tends to zero. We wish to prove the convergence of u_{D_n} to an entropy solution of problem (1). For that purpose, in view of the uniqueness Theorem 4.1, it suffices to show that, up to a subsequence, u_{D_n} tends in the nonlinear weak-$\star$ sense to an entropy process solution of (1). We obtain compactness properties using estimates on u_{D_n} derived from discrete estimates on U_{D_n}, then pass to the limit in inequality (50).

5.6.1. Nonlinear weak-$\star$ compactness. The maximum principle ensures that (u_{D_n}) is bounded in $L^\infty(Q)$. Consequently, there exist $u \in L^\infty(Q\times(0,1))$ such that, up to a subsequence, u_{D_n} tends to u in the nonlinear weak-$\star$ sense.

5.6.2. Compactness in $L^2(Q)$. From discrete estimates obtained in Proposition 5.2 we easily deduce (see, e.g., [EGH00]) the following inequalities on $z_D = \zeta(u_D)-\zeta(\bar u_D)$.

PROPOSITION 5.3 (space translation estimates). *Assume hypotheses* (H1), (H2), (H3), (H4), *and* (H5). *There exists a constant C_1 such that*

$$\forall y \in \mathbb{R}^d, \quad \int_0^T\!\!\int_{\Omega_y} (z_D(t,x+y)-z_D(t,x))^2 dxdt \le C_1|y|(|y|+\mathrm{size}(\mathcal{T})),$$

where $\Omega_y = \{x\in\Omega, [x,x+y]\subset\Omega\}$.

The hypothesis (H5) includes the assumption $\bar u_t \in L^1(Q)$, while the discrete evolution equation (43) relates the discrete time derivative of u_D to its discrete space derivative. Therefore the following time translation estimate on z_D is available.

PROPOSITION 5.4 (time translation estimates). *Assume hypotheses* (H1), (H2), (H3), (H4), *and* (H5). *There exists a constant C_2 such that*

$$\forall s>0, \quad \int_0^{T-s}\int_\Omega (z_D(t+s,x)-z_D(t,x))^2 dxdt \le C_2\, s.$$

Since the function z_D vanishes on Σ, it can be extended by zero out of Q. Then using the Fréchet–Kolmogorov theorem (see, e.g., [Bre83]), we get the existence of a function $z \in L^2(0,T,H^1(\Omega))$ such that, up to a subsequence, $z_{D_n} \to z$ in $L^2(Q)$. Besides, since $z_D = \zeta(u_D)-\zeta(\bar u_D)$ and $\zeta(\bar u_D)$ converges to $\zeta(\bar u)$ in $L^2(Q)$, we get the convergence of $\zeta(u_{D_n})$ in $L^2(Q)$ (to $\zeta(\bar u)+z$). On the other hand, the nonlinear weak-$\star$ convergence of (u_{D_n}) shows that $\zeta(u_{D_n})$ converges also to $\zeta(u)$ weakly in $L^\infty(Q)$, so that $\zeta(\bar u)+z=\zeta(u)$. In particular, $\zeta(u)$ does not depend on the last argument α, and the trace of $\zeta(u)$ on Σ is $\zeta(\bar u)$. See [EGHM02] for more details on this step of the proof.

5.6.3. Conclusion. It remains to pass to the limit in the continuous entropy inequalities to prove that u is an entropy process solution. The uniqueness Theorem 4.1 proves that u does not depend on α and is the unique entropy weak solution of problem (1). Besides, the whole sequence u_{D_n} is convergent (u is the unique possible limit), and by definition of the nonlinear weak-$\star$ convergence, $(u_{D_n})^2$ also converges weakly to $(u)^2$ so that u_{D_n} converges to u in $L^2(Q)$ (strong), and in all $L^p(Q)$, for $1 \leq p < +\infty$. Therefore, we have proved the following theorem.

THEOREM 5.2. *Let D_n be a sequence of discretizations, such that size(D_n) tends to zero. Assume hypotheses* (H1)*,* (H2)*,* (H3)*,* (H4)*,* (H5)*, and* (H6) *(or* (H6bis)*). Then, for every $1 \leq p < +\infty$, (u_{D_n}) converges to the unique entropy solution of problem* (1) *in $L^p(Q)$.*

Acknowledgments. We warmly thank R. Eymard, M. Ohlberger, A. Porretta, and T. Gallouët for wise advice and fruitful discussions.

REFERENCES

[BEK00] R. Bürger, S. Evje, and K. H. Karlsen, *On strongly degenerate convection-diffusion problems modeling sedimentation-consolidation processes*, J. Math. Anal. Appl., 247 (2000), pp. 517–556.

[BGN00] F. Bouchut, F. R. Guarguaglini, and R. Natalini, *Diffusive BGK approximations for nonlinear multidimensional parabolic equations*, Indiana Univ. Math. J., 49 (2000), pp. 723–749.

[BLN79] C. Bardos, A. Y. LeRoux, and J.-C. Nédélec, *First order quasilinear equations with boundary conditions*, Comm. Partial Differential Equations, 4 (1979), pp. 1017–1034.

[Bre83] H. Brezis, *Analyse fonctionnelle*, Collection Mathématiques Appliquées pour la Maîtrise (Master's thesis), Masson, Paris, 1983.

[Car99] J. Carrillo, *Entropy solutions for nonlinear degenerate problems*, Arch. Ration. Mech. Anal., 147 (1999), pp. 269–361.

[CCL95] B. Cockburn, F. Coquel, and P. G. LeFloch, *Convergence of the finite volume method for multidimensional conservation laws*, SIAM J. Numer. Anal., 32 (1995), pp. 687–705.

[CF02] G.-Q. Chen and H. Frid, *Extended divergence-measure fields and the Euler equations for gas dynamics*, Comm. Math. Phys., 236 (2003), pp. 251–280.

[CG99] B. Cockburn and G. Gripenberg, *Continuous dependence on the nonlinearities of solutions of degenerate parabolic equations*, J. Differential Equations, 151 (1999), pp. 231–251.

[CGH93] S. Champier, T. Gallouët, and R. Herbin, *Convergence of an upstream finite volume scheme for a nonlinear hyperbolic equation on a triangular mesh*, Numer. Math., 66 (1993), pp. 139–157.

[CH99] C. Chainais-Hillairet, *Finite volume schemes for a nonlinear hyperbolic equation. Convergence towards the entropy solution and error estimate*, M2AN Math. Model. Numer. Anal., 33 (1999), pp. 129–156.

[DiP85] R.J. DiPerna, *Measure-valued solutions to conservation laws*, Arch. Rational Mech. Anal., 88 (1985), pp. 223–270.

[EGH99] R. Eymard, T. Gallouët, and R. Herbin, *Convergence of finite volume schemes for semilinear convection diffusion equations*, Numer. Math., 82 (1999), pp. 91–116.

[EGH00] R. Eymard, T. Gallouët, and R. Herbin, *Finite volume methods*, in Handbook of Numerical Analysis, Vol. 7, North–Holland, Amsterdam, 2000, pp. 713–1020.

[EGHM02] R. Eymard, T. Gallouet, R. Herbin, and A. Michel, *Convergence of a finite volume scheme for parabolic degenerate equations*, Numer. Math., 92 (2002), pp. 41–82.

[EHM01] R. Eymard, R. Herbin, and A. Michel, *Mathematical study of a petroleum engineering scheme*, RAIRO Modél. Math. Anal. Numér., to appear.

[EK00] S. Evje and K. H. Karlsen, *Discrete approximations of BV solutions to doubly nonlinear degenerate parabolic equations*, Numer. Math., 86 (2000), pp. 377–417.

[GMT96] G. Gagneux and M. Madaune-Tort, *Analyse mathématique de modèles non linéaires de l'ingénierie pétrolière*, Springer-Verlag, Berlin, 1996.

[GMT02] F. R. Guarguaglini, V. Milisic, and A. Terracina, *A discrete BGK approximation for strongly degenerate parabolic problems with boundary conditions*, submitted to J. Differential Equations; also available online at http://www.math.ntnu.no/conservation/2002/029.html.

[KO01] K. Karlsen and M. Ohlberger, *A note on the uniqueness of entropy solutions of nonlinear degenerate parabolic equations*, J. Math. Anal. Appl., 275 (2002), pp. 439–458.

[KR00] K. H. Karlsen and N. H. Risebro, *On the uniqueness and stability of entropy solutions of nonlinear degenerate parabolic equations with rough coefficients*, Discrete Contin. Dyn. Syst. Ser. A, 9 (2003), pp. 1081–1104.

[KR01] K. H. Karlsen and N. H. Risebro, *Convergence of finite difference schemes for viscous and inviscid conservation laws with rough coefficients*, M2AN Math. Model. Numer. Anal., 35 (2001), pp. 239–269.

[Kru70] S. N. Kruzhkov, *First order quasilinear equations with several independent variables*, Mat. Sb. (N.S.), 81 (1970), pp. 228–255.

[KRW96] D. Kröner, M. Rokyta, and M. Wierse, *A Lax–Wendroff type theorem for upwind finite volume schemes in* 2*d*, East-West J. Numer. Math., 4 (1996), pp. 279–292.

[LBS93] A. Lagha-Benadallah and F. Smadhi, *Existence de solutions faibles pour un problème aux limites associé à une équation parabolique dégénérée*, Rev. Maghrébine Math., 2 (1993), pp. 201–221.

[Mic01] A. Michel, *A finite volume scheme for two-phase immiscible flow in porous media*, SIAM J. Numer. Anal., 41 (2003), pp. 1301–1317.

[MPT02] C. Mascia, A. Porretta, and A. Terracina, *Nonhomogeneous Dirichlet problems for degenerate parabolic-hyperbolic equations*, Arch. Ration. Mech. Anal., 163 (2002), pp. 87–124.

[Ohl01] M. Ohlberger, *A posteriori error estimates for vertex centered finite volume approximations of convection-diffusion-reaction equations*, M2AN Math. Model. Numer. Anal., 35 (2001), pp. 355–387.

[Ott96] F. Otto, *Initial-boundary value problem for a scalar conservation law*, C. R. Acad. Sci. Paris Sér. I Math., 322 (1996), pp. 729–734.

[RG99] E. Rouvre and G. Gagneux, *Solution forte entropique de lois scalaires hyperboliques-paraboliques dégénérées*, C. R. Acad. Sci. Paris Sér. I Math., 329 (1999), pp. 599–602.

[Ser96] D. Serre, *Systèmes de lois de conservation.* II, in Structures Géométriques, Oscillation et Problèmes Mixtes, Diderot, Paris, 1996, pp. 264–275.

[Sze91] A. Szepessy, *Convergence of a streamline diffusion finite element method for scalar conservation laws with boundary conditions*, RAIRO Modél. Math. Anal. Numér., 25 (1991), pp. 749–782.

[VH69] A. I. Vol′pert and S. I. Hudjaev, *The Cauchy problem for second order quasilinear degenerate parabolic equations*, Mat. Sb. (N.S.), 78 (1969), pp. 374–396.

[Vil94] J.-P. Vila, *Convergence and error estimates in finite volume schemes for general multidimensional scalar conservation laws.* I. *Explicit monotone schemes*, RAIRO Modél. Math. Anal. Numér., 28 (1994), pp. 267–295.

[Vov02] J. Vovelle, *Convergence of finite volume monotone schemes for scalar conservation laws on bounded domains*, Numer. Math., 90 (2002), pp. 563–596.

SIAM J. Numer. Anal.
Vol. 41, No. 6, pp. 2294–2312

ASYMPTOTICALLY EXACT A POSTERIORI ERROR ESTIMATORS, PART I: GRIDS WITH SUPERCONVERGENCE*

RANDOLPH E. BANK† AND JINCHAO XU‡

Abstract. In Part I of this work, we develop superconvergence estimates for piecewise linear finite element approximations on quasi-uniform triangular meshes where most pairs of triangles sharing a common edge form approximate parallelograms. In particular, we first show a superconvergence of the gradient of the finite element solution u_h and to the gradient of the interpolant u_I. We then analyze a postprocessing gradient recovery scheme, showing that $Q_h \nabla u_h$ is a superconvergent approximation to ∇u. Here Q_h is the global L^2 projection. In Part II, we analyze a superconvergent gradient recovery scheme for general unstructured, shape regular triangulations. This is the foundation for an a posteriori error estimate and local error indicators.

Key words. superconvergence, gradient recovery

AMS subject classifications. 65N50, 65N30

DOI. 10.1137/S003614290139874X

1. Introduction.

The study of superconvergence and a posteriori error estimates has been an area of active research; see the monographs by Verfürth [17], Chen and Huang [8], Wahlbin [18], Lin and Yan [16], and Babuška and Strouboulis [3] and a recent article by Lakhany, Marek, and Whiteman [13] for overviews of the field. In this two-part work we study some new superconvergence results. In Part I, we develop some superconvergence results for finite element approximations of a general class of elliptic partial differential equations (PDEs), based mainly on the geometry of the underlying triangular mesh. In Part II, we develop a gradient recovery technique that can force superconvergence on general shape regular meshes. Patch recovery techniques have been studied by Zienkiewicz and Zhu and this subject has itself evolved into an active subfield of research [25, 14, 23, 24, 9, 22]. Although our algorithm in some respects resembles this and other similar schemes [12, 19, 4, 6, 2, 10], it draws much of its motivation from multilevel iterative methods.

Let $\Omega \subset \mathbb{R}^2$ be a bounded domain with Lipschitz boundary $\partial\Omega$. For simplicity of exposition, we assume that Ω is a polygon. We assume that Ω is partitioned by a shape regular triangulation $\mathcal{T}_h$ of mesh size $h \in (0,1)$. Let $\mathcal{V}_h \subset H^1(\Omega)$ be the corresponding continuous piecewise linear finite element space associated with this triangulation $\mathcal{T}_h$, and $u_h \in \mathcal{V}_h$ be a finite element approximation to a second order elliptic boundary value problem.

Our development has three main steps. In the first step, we prove a superconvergence result for $|u_h - u_I|_{1,\Omega}$, where u_I is the piecewise linear interpolant for u. In

*Received by the editors November 26, 2001; accepted for publication (in revised form) April 30, 2003; published electronically December 17, 2003.
http://www.siam.org/journals/sinum/41-6/39874.html

†Department of Mathematics, University of California, San Diego, La Jolla, CA 92093-0112 (rbank@ucsd.edu). The work of this author was supported by the National Science Foundation under contracts DMS-9973276 and DMS-0208449.

‡Center for Computational Mathematics and Applications, Department of Mathematics, Pennsylvania State University, University Park, PA 16802 (xu@math.psu.edu). The work of this author was supported by the National Science Foundation under contract DMS-0074299 and by the Center for Computational Mathematics and Applications, Penn State University.

particular, we show in Theorem 3.1 that

$$|u_h - u_I|_{1,\Omega} \lesssim h^{1+\min(1,\sigma)}|\log h|^{1/2}\|u\|_{3,\infty,\Omega}. \tag{1.1}$$

Estimate (1.1) holds on quasi-uniform meshes, where an $O(h^2)$ approximate parallelogram property is satisfied for pairs of adjacent triangles in most parts of Ω except for a region of size $O(h^{2\sigma})$; see section 2 for details.

The estimate (1.1) is well known in the literature for the special case $\sigma = \infty$; namely, the $O(h^2)$ approximate parallelogram property is satisfied for all pairs of adjacent triangles, and it is also known for cases when the $O(h^2)$ approximate parallelogram property is satisfied except for triangles along a few lines (see Xu [20] and Lin and Xu [15]) or triangles along the domain boundary (see Lin and Yan [16], Hlaváček and Křížek [11]). Lakhany, Marek, and Whiteman [13] consider a less restrictive $O(h^{1+\alpha})$ approximate parallelogram property. Our new estimate (1.1) is a significant generalization of these known results. First, our analysis is based on local identities for each element that simplify existing techniques. For example, our result can be extended in a straightforward fashion to the mesh in which an $O(h^{1+\alpha})$ (instead of $O(h^2)$) approximate parallelogram property holds for most pairs of triangles (see [13]). Second, the assumptions that we make are weaker than existing ones and should hold for many practical grids for some $\sigma > 0$, although in some cases σ could be very small.

One important case that our theory does not cover in this paper is locally refined grids. Lakhany, Marek, and Whiteman [13] have some results on this topic for piecewise uniform grids (see also Lin and Xu [15]). Because of the local nature of our analysis, our technique can be extended to this type of grid. We will report this type of extension in future work.

Superconvergence results typically depend on delicate estimates involving cancellation of the lowest order terms in some asymptotic expansion of the local error. When one derives elementwise expressions using continuous finite element spaces, often one encounters boundary integrals involving the normal component of the gradient of the test function. Thus, although one can determine that some cancellation takes place between certain error local components, it is difficult to combine elementwise statements because the normal components of the gradient of $v_h \in \mathcal{V}_h$ are discontinuous. On the other hand, *tangential* components of ∇v_h along element edges are continuous. Thus our approach is to derive some expressions for the element error that involve only the tangential derivative of the test function on the element boundary. The key identity of this type is Lemma 2.3.

We also note that Lemma 2.3 is an identity rather than an estimate. Thus global versions of this identity give exact characterizations of the error for arbitrary triangulations. In effect, one can see exactly the cancellations that might occur even on completely unstructured meshes. The $O(h^2)$ approximate parallelogram property can be viewed in this context as one set of sufficient conditions for obtaining superconvergent bounds for those terms.

The techniques used in our analysis are related to but much more refined than many existing superconvergence techniques in the literature such as those summarized in [8, 16]. For example, the identity in Lemma 2.3 may be compared with the integral identities for rectangular elements [16]. In fact it was not known how the integral identities for rectangular elements in [16] could be generalized to triangular elements. Lemma 2.3 offers clues for such generalizations, and more work can obviously be done in this direction.

The second major component of our analysis is a superconvergent approximation to ∇u. This approximation is generated by a gradient recovery procedure. In particular, in Theorem 4.2 we show

$$\|\nabla u - Q_h \nabla u_h\|_{0,\Omega} \lesssim h^{1+\min(1,\sigma)} |\log h|^{1/2} \|u\|_{3,\infty,\Omega}, \tag{1.2}$$

where Q_h is the L^2 projection. When the mesh does not satisfy the $O(h^2)$ parallelogram property or σ becomes very close to zero, then the superconvergence demonstrated in (1.2) will be diminished. Intuitively, it appears that this is due mainly to high frequency errors introduced by the small nonuniformities of the mesh. Preferentially attenuating high frequency errors in mesh functions is of course a widely studied problem in multilevel iterative methods. Our proposal here is to apply these ideas in the present context. Thus, to enhance the superconvergence effect on general shape regular meshes, we compute $S^m Q_h \nabla u_h$, where S is an appropriate multigrid-like smoothing operator. In Part II of this manuscript [7], we analyze this procedure and prove superconvergence estimates somewhat like (1.2) for $\|u - S^m Q_h \nabla u_h\|_{0,\Omega}$.

In the third major component of our analysis, we use the recovered gradient to develop an a posteriori error estimate. An obvious choice is to use $(I - S^m Q_h)\nabla u_h$ to approximate the true error $\nabla(u - u_h)$. In [7], we show that this is a good choice and that in many circumstances we can expect the error estimate to be *asymptotically exact;* that is,

$$\lim \frac{\|(I - S^m Q_h)\nabla u_h\|_{0,\Omega}}{\|\nabla(u - u_h)\|_{0,\Omega}} = 1$$

as $h \to 0$ and $m \to \infty$ in an appropriate fashion.

We also use the recovered gradient to construct local approximations of interpolation errors to be used as local error indicators for adaptive meshing algorithms; see [7] for details.

We remark that both our gradient recovery scheme and our a posteriori error estimate are largely independent of the details of the PDE. Indeed, all of the preliminary lemmas in section 2 are also independent of the PDE. The PDE directly enters only in the proof of Theorem 3.1, and there the properties that we assume are standard. This suggests that superconvergence can be expected for a wide variety of problems, as long as the adaptive meshing yields smoothly varying, shape regular meshes.

The rest of this paper is organized as follows: section 2 contains technical identities and estimates that form the basis for the estimate (1.1). In section 3, we prove (1.1) for general linear elliptic boundary value problems under standard assumptions. We also explore an application to nonlinear elliptic problems. In section 4 we develop and analyze the superconvergent gradient recover scheme in the case of $O(h^2)$ parallelogram meshes. In section 5 we present a few numerical examples illustrating the effectiveness of our procedures.

2. Preliminary lemmas. We begin with some geometric identities for a canonical element τ. Let τ have vertices $\boldsymbol{p}_k^t = (x_k, y_k)$, $1 \le k \le 3$, oriented counterclockwise, and corresponding nodal basis functions (barycentric coordinates) $\{\phi_k\}_{k=1}^3$. Let $\{e_k\}_{k=1}^3$ denote the edges of element τ, $\{\theta_k\}_{k=1}^3$ the angles, $\{\boldsymbol{n}_k\}_{k=1}^3$ the unit outward normal vectors, $\{\boldsymbol{t}_k\}_{k=1}^3$ the unit tangent vectors with counterclockwise orientation, $\{\ell_k\}_{k=1}^3$ the edge lengths, and $\{d_k\}_{k=1}^3$ the perpendicular heights (see Figure 2.1). Let $\tilde{\boldsymbol{p}}$ be the point of intersection for the perpendicular bisectors of the three sides of τ. Let $|s_k|$ denote the distance between $\tilde{\boldsymbol{p}}$ and side k. If τ has no obtuse angles, then the

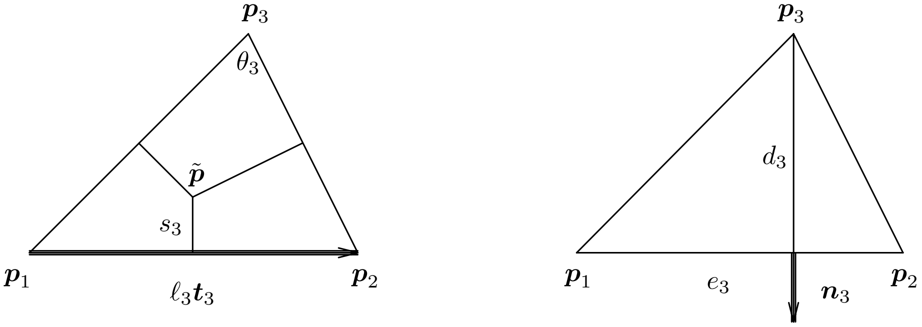


Fig. 2.1. *Parameters associated with the triangle τ.*

s_k will be nonnegative; otherwise, the distance to the side opposite the obtuse angle will be negative.

There are many relationships among these quantities; in particular, we note the following, which hold for $1 \leq k \leq 3$ and $k \pm 1$ permuted cyclically:

$$\begin{gathered} \ell_k d_k = \ell_{k+1}\ell_{k-1}\sin\theta_k = 2|\tau|, \\ 2\ell_{k+1}\ell_{k-1}\cos\theta_k = \ell_{k+1}^2 + \ell_{k-1}^2 - \ell_k^2, \\ \sin\theta_k = \boldsymbol{n}_{k-1}\cdot\boldsymbol{t}_{k+1} = -\boldsymbol{n}_{k+1}\cdot\boldsymbol{t}_{k-1}, \\ \cos\theta_k = -\boldsymbol{t}_{k-1}\cdot\boldsymbol{t}_{k+1} = -\boldsymbol{n}_{k-1}\cdot\boldsymbol{n}_{k+1}, \\ \nabla\phi_k = -\frac{\boldsymbol{n}_k}{d_k}, \\ s_k = -|\tau|\;\ell_k\nabla\phi_{k-1}\cdot\nabla\phi_{k+1} = \frac{\ell_k\cos\theta_k}{2\sin\theta_k}. \end{gathered}$$

Let $\mathcal{D}_\tau$ be a symmetric 2×2 matrix with constant matrix entries. We define

$$\xi_k = -\boldsymbol{n}_{k+1}\cdot\mathcal{D}_\tau\boldsymbol{n}_{k-1}.$$

The important special case $\mathcal{D}_\tau = I$ corresponds to $-\Delta$, and in this case $\xi_k = \cos\theta_k$. Let $q_k = \phi_{k+1}\phi_{k-1}$ denote the quadratic bump function associated with edge e_k, and let $\psi_k = \phi_k(1-\phi_k)$. In Lemma 2.1 we collect several simple identities that are used in the proof of Lemma 2.3.

Lemma 2.1.

$$\sin\theta_k\nabla u\cdot\mathcal{D}_\tau\boldsymbol{n}_k = \xi_{k-1}\frac{\partial u}{\partial\boldsymbol{t}_{k-1}} - \xi_{k+1}\frac{\partial u}{\partial\boldsymbol{t}_{k+1}}, \tag{2.1}$$

$$\frac{\partial u}{\partial\boldsymbol{t}_{k+1}} = -\cos\theta_{k-1}\frac{\partial u}{\partial\boldsymbol{t}_k} - \sin\theta_{k-1}\frac{\partial u}{\partial\boldsymbol{n}_k}, \tag{2.2}$$

$$\frac{\partial u}{\partial\boldsymbol{t}_{k-1}} = -\cos\theta_{k+1}\frac{\partial u}{\partial\boldsymbol{t}_k} + \sin\theta_{k+1}\frac{\partial u}{\partial\boldsymbol{n}_k}, \tag{2.3}$$

$$\int_\tau\frac{\partial u}{\partial\boldsymbol{t}_k} = -\sin\theta_{k+1}\int_{e_{k-1}} u + \sin\theta_{k-1}\int_{e_{k+1}} u, \tag{2.4}$$

$$\sin\theta_k\int_{e_{k-1}} q_{k-1}u = \int_\tau\psi_{k+1}\frac{\partial u}{\partial\boldsymbol{t}_{k+1}} + \sin\theta_{k-1}\int_{e_k} q_k u, \tag{2.5}$$

$$\sin\theta_k \int_{e_{k+1}} q_{k+1} u = -\int_\tau \psi_{k-1} \frac{\partial u}{\partial \boldsymbol{t}_{k-1}} + \sin\theta_{k+1} \int_{e_k} q_k u. \tag{2.6}$$

Proof. We note that (2.1) is an immediate consequence of

$$\mathcal{D}_\tau \boldsymbol{n}_k = \frac{\boldsymbol{n}_{k+1}\cdot \mathcal{D}_\tau \boldsymbol{n}_k}{\boldsymbol{n}_{k+1}\cdot \boldsymbol{t}_{k-1}} \boldsymbol{t}_{k-1} + \frac{\boldsymbol{n}_{k-1}\cdot \mathcal{D}_\tau \boldsymbol{n}_k}{\boldsymbol{n}_{k-1}\cdot \boldsymbol{t}_{k+1}} \boldsymbol{t}_{k+1} = \frac{\xi_{k-1}}{\sin\theta_k}\boldsymbol{t}_{k-1} - \frac{\xi_{k+1}}{\sin\theta_k}\boldsymbol{t}_{k+1}.$$

Proofs for (2.2)–(2.3) follow the same pattern. For (2.4), we note that from Green's identity

$$\int_\tau \nabla u \cdot \boldsymbol{t}_k = \sum_{j=1}^{3} \boldsymbol{n}_j \cdot \boldsymbol{t}_k \cdot \int_{e_k} u.$$

For (2.5)–(2.6), we note that ψ_k is constant along lines parallel to e_k, and $\partial\psi_k/\partial\boldsymbol{t}_k \equiv 0$. Thus

$$\frac{\partial(\psi_k u)}{\partial \boldsymbol{t}_k} = \psi_k \frac{\partial u}{\partial \boldsymbol{t}_k}.$$

Also, on edge e_k we have $q_k = \psi_{k+1} = \psi_{k-1}$. Equations (2.5)–(2.6) follow from these observations and (2.4). □

LEMMA 2.2. *Let $u \in W^{3,\infty}(\Omega)$. Let u_I and u_q be the continuous piecewise linear and piecewise quadratic interpolants, respectively, for u. Then*

$$\int_{e_k} (u - u_I) = \frac{\ell_k^2}{2} \int_{e_k} q_k \frac{\partial^2 u}{\partial \boldsymbol{t}_k^2}, \tag{2.7}$$

$$\int_\tau (u - u_I) = -\frac{1}{24} \int_\tau \sum_{k=1}^{3} \ell_k^2 \frac{\partial^2 u_q}{\partial \boldsymbol{t}_k^2} + \int_\tau (u - u_q). \tag{2.8}$$

Proof. Identity (2.7) is equivalent to the following:

$$\int_a^b u(s)ds - \frac{(b-a)}{2}(u(a) + u(b)) = \frac{1}{2}\int_a^b (s-a)(s-b)u''(s)ds,$$

which follows by an integration by parts. To show (2.8), we note that $u_q - u_I$ is a piecewise quadratic polynomial that is zero at all of the vertices in the mesh and that therefore can be expressed in terms of the quadratic bump functions. A simple calculation shows in a given element τ

$$u_q - u_I = \sum_{k=1}^{3} \ell_k^2 \boldsymbol{t}_k^t M_\tau \boldsymbol{t}_k \, q_k(x,y), \tag{2.9}$$

where

$$M_\tau = -\frac{1}{2}\begin{pmatrix} \partial_{11} u_q & \partial_{12} u_q \\ \partial_{21} u_q & \partial_{22} u_q \end{pmatrix}. \tag{2.10}$$

The matrix M_τ is constant since u_q is quadratic. Let $m_k = (p_{k+1} + p_{k-1})/2$ denote the midpoint of the kth edge. Then

$$\frac{\boldsymbol{t}_k^t M_\tau \boldsymbol{t}_k}{2} = \frac{2u(m_k) - u(p_{k+1}) - u(p_{k-1})}{\ell_k^2}.$$

Identity (2.8) follows from

$$\begin{aligned}\int_\tau (u-u_I) &= \int_\tau (u_q-u_I) + \int_\tau (u-u_q)\\ &= -\frac{1}{2}\sum_{k=1}^{3} \ell_k^2 \frac{\partial^2 u_q}{\partial \boldsymbol{t}_k^2}\int_\tau q_k + \int_\tau (u-u_q)\\ &= -\frac{|\tau|}{24}\sum_{k=1}^{3} \ell_k^2 \frac{\partial^2 u_q}{\partial \boldsymbol{t}_k^2} + \int_\tau (u-u_q)\\ &= -\frac{1}{24}\int_\tau \sum_{k=1}^{3} \ell_k^2 \frac{\partial^2 u_q}{\partial \boldsymbol{t}_k^2} + \int_\tau (u-u_q). \qquad \square\end{aligned}$$

The following is a fundamental identity in our analysis.

Lemma 2.3. *Let $\mathcal{D}_\tau$ be a 2×2 symmetric matrix with constant entries. Then*

$$\begin{aligned}\int_\tau \nabla(u-u_I)\cdot\mathcal{D}_\tau\nabla v_h &= \sum_{k=1}^{3}\int_{e_k}\frac{\xi_k q_k}{2\sin\theta_k}\left\{(\ell_{k+1}^2-\ell_{k-1}^2)\frac{\partial^2 u}{\partial \boldsymbol{t}_k^2} + 4|\tau|\frac{\partial^2 u}{\partial \boldsymbol{t}_k\partial \boldsymbol{n}_k}\right\}\frac{\partial v_h}{\partial \boldsymbol{t}_k}\\ &- \int_\tau \sum_{k=1}^{3}\frac{\ell_k\xi_k}{2\sin^2\theta_k}\left\{\ell_{k+1}\psi_{k-1}\frac{\partial^3 u}{\partial^2 \boldsymbol{t}_{k+1}\partial \boldsymbol{t}_{k-1}} + \ell_{k-1}\psi_{k+1}\frac{\partial^3 u}{\partial^2 \boldsymbol{t}_{k-1}\partial \boldsymbol{t}_{k+1}}\right\}\frac{\partial v_h}{\partial \boldsymbol{t}_k}.\end{aligned}$$

Proof. Using Lemmas 2.1–2.2, we have

$$\begin{aligned}\int_\tau \nabla(u-u_I)\cdot\mathcal{D}_\tau\nabla v_h &= \sum_{k=1}^{3}\int_{e_k}(u-u_I)\nabla v_h v_h\cdot\mathcal{D}_\tau \boldsymbol{n}_k\\ &= \sum_{k=1}^{3}\int_{e_k}(u-u_I)\left\{\frac{\xi_{k-1}}{\sin\theta_k}\frac{\partial v_h}{\partial \boldsymbol{t}_{k-1}} - \frac{\xi_{k+1}}{\sin\theta_k}\frac{\partial v_h}{\partial \boldsymbol{t}_{k+1}}\right\}\\ &= \sum_{k=1}^{3}\left\{\frac{\xi_k}{\sin\theta_{k+1}}\int_{e_{k+1}}(u-u_I)\frac{\partial v_h}{\partial \boldsymbol{t}_k}\right\} - \left\{\frac{\xi_k}{\sin\theta_{k-1}}\int_{e_{k-1}}(u-u_I)\frac{\partial v_h}{\partial \boldsymbol{t}_k}\right\}\\ &= \sum_{k=1}^{3}\left\{\frac{\ell_{k+1}^2\xi_k}{2\sin\theta_{k+1}}\int_{e_{k+1}}q_{k+1}\frac{\partial^2 u}{\partial \boldsymbol{t}_{k+1}^2}\frac{\partial v_h}{\partial \boldsymbol{t}_k}\right\} - \left\{\frac{\ell_{k-1}^2\xi_k}{2\sin\theta_{k-1}}\int_{e_{k-1}}q_{k-1}\frac{\partial^2 u}{\partial \boldsymbol{t}_{k-1}^2}\frac{\partial v_h}{\partial \boldsymbol{t}_k}\right\}\\ &= \sum_{k=1}^{3}\frac{\ell_k\xi_k}{2\sin\theta_k}\left\{\ell_{k+1}\int_{e_{k+1}}q_{k+1}\frac{\partial^2 u}{\partial \boldsymbol{t}_{k+1}^2}\frac{\partial v_h}{\partial \boldsymbol{t}_k} - \ell_{k-1}\int_{e_{k-1}}q_{k-1}\frac{\partial^2 u}{\partial \boldsymbol{t}_{k-1}^2}\frac{\partial v_h}{\partial \boldsymbol{t}_k}\right\}\\ &= \sum_{k=1}^{3}\frac{\xi_k}{2\sin\theta_k}\int_{e_k}q_k\left\{\ell_{k+1}^2\frac{\partial^2 u}{\partial \boldsymbol{t}_{k+1}^2} - \ell_{k-1}^2\frac{\partial^2 u}{\partial \boldsymbol{t}_{k-1}^2}\right\}\frac{\partial v_h}{\partial \boldsymbol{t}_k}\\ &\quad - \int_\tau \sum_{k=1}^{3}\frac{\ell_k\xi_k}{2\sin^2\theta_k}\left\{\ell_{k+1}\psi_{k-1}\frac{\partial^3 u}{\partial \boldsymbol{t}_{k-1}\partial \boldsymbol{t}_{k+1}^2} + \ell_{k-1}\psi_{k+1}\frac{\partial^3 u}{\partial \boldsymbol{t}_{k+1}\partial \boldsymbol{t}_{k-1}^2}+\right\}\frac{\partial v_h}{\partial \boldsymbol{t}_k}.\end{aligned}$$

To complete the proof, we focus attention on the term

$$\ell_{k+1}^2\frac{\partial^2 u}{\partial \boldsymbol{t}_{k+1}^2} - \ell_{k-1}^2\frac{\partial^2 u}{\partial \boldsymbol{t}_{k-1}^2}.$$

Using Lemma 2.1 once again, we have

$$\begin{aligned}\frac{\partial^2 u}{\partial \boldsymbol{t}_{k+1}^2} &= \cos^2\theta_{k-1}\frac{\partial^2 u}{\partial \boldsymbol{t}_k^2} + 2\cos\theta_{k-1}\sin\theta_{k-1}\frac{\partial^2 u}{\partial \boldsymbol{t}_k \partial \boldsymbol{n}_k} + \sin^2\theta_{k-1}\frac{\partial^2 u}{\partial \boldsymbol{n}_k^2},\\ \frac{\partial^2 u}{\partial \boldsymbol{t}_{k-1}^2} &= \cos^2\theta_{k+1}\frac{\partial^2 u}{\partial \boldsymbol{t}_k^2} - 2\cos\theta_{k+1}\sin\theta_{k+1}\frac{\partial^2 u}{\partial \boldsymbol{t}_k \partial \boldsymbol{n}_k} + \sin^2\theta_{k+1}\frac{\partial^2 u}{\partial \boldsymbol{n}_k^2}.\end{aligned}$$

We also need the following identities:

$$\begin{gathered}\ell_{k+1}^2\sin^2\theta_{k-1} - \ell_{k-1}^2\sin^2\theta_{k+1} = 0,\\ \ell_{k+1}^2\cos^2\theta_{k-1} - \ell_{k-1}^2\cos^2\theta_{k+1} = \ell_{k+1}^2 - \ell_{k-1}^2,\\ \ell_{k+1}^2 2\cos\theta_{k-1}\sin\theta_{k-1} + \ell_{k-1}^2 2\cos\theta_{k+1}\sin\theta_{k+1} = 4|\tau|.\end{gathered}$$

Combining these equations leads to

$$\ell_{k+1}^2\frac{\partial^2 u}{\partial \boldsymbol{t}_{k+1}^2} - \ell_{k-1}^2\frac{\partial^2 u}{\partial \boldsymbol{t}_{k-1}^2} = (\ell_{k+1}^2 - \ell_{k-1}^2)\frac{\partial^2 u}{\partial \boldsymbol{t}_k^2} + 4|\tau|\frac{\partial^2 u}{\partial \boldsymbol{t}_k \partial \boldsymbol{n}_k},$$

completing the proof. □

Let e be an interior edge in the triangulation $\mathcal{T}_h$. Let τ and τ' be the two elements sharing e. We say that τ and τ' form an $O(h^2)$ approximate parallelogram if the lengths of any two opposite edges differ only by $O(h^2)$. Let x be a vertex lying on $\partial\Omega$, and let e and e' be the two boundary edges sharing x as an endpoint. Let τ and τ' be the two elements having e and e', respectively, as edges, and let $\boldsymbol{t}$ and and $\boldsymbol{t}'$ be the unit tangents. Take e and e' as one pair of corresponding edges, and make a clockwise traversal of $\partial\tau$ and $\partial\tau'$ to define two additional corresponding edge pairs. In this case, we say that τ and τ' form an $O(h^2)$ approximate parallelogram if $|\boldsymbol{t} - \boldsymbol{t}'| = O(h)$, and the lengths of any two corresponding edges differ only by $O(h^2)$.

Definition 2.4. *The triangulation $\mathcal{T}_h$ is $O(h^{2\sigma})$ irregular if the following hold:*

1. *Let $\mathcal{E} = \mathcal{E}_1 \oplus \mathcal{E}_2$ denote the set of interior edges in $\mathcal{T}_h$. For each $e \in \mathcal{E}_1$, τ and τ' form an $O(h^2)$ approximate parallelogram, while $\sum_{e\in\mathcal{E}_2}|\tau| + |\tau'| = O(h^{2\sigma})$.*
2. *Let $\mathcal{P} = \mathcal{P}_1 \oplus \mathcal{P}_2$ denote the set of boundary vertices. The elements associated with each $x \in \mathcal{P}_1$ form an $O(h^2)$ approximate parallelogram, and $|\mathcal{P}_2| = \kappa$, where κ is fixed independent of h.*

The boundary points $\mathcal{P}$ and the decomposition $\mathcal{P} = \mathcal{P}_1 \oplus \mathcal{P}_2$ are used only in the case of Neumann boundary conditions. Generally speaking, we expect $\mathcal{P}_2$ to consist of the geometric corners of Ω and perhaps a few other isolated points.

We can now state our main lemma.

Lemma 2.5. *Let the triangulation $\mathcal{T}_h$ be $O(h^{2\sigma})$ irregular. Let $\mathcal{D}_\tau$ be a piecewise constant matrix function defined on $\mathcal{T}_h$, whose elements $\mathcal{D}_{\tau ij}$ satisfy*

$$\begin{gathered}|\mathcal{D}_{\tau ij}| \lesssim 1,\\ |\mathcal{D}_{\tau ij} - \mathcal{D}_{\tau' ij}| \lesssim h,\end{gathered}$$

for $i = 1, 2$, $j = 1, 2$. Here τ and τ' are a pair of triangles sharing a common edge. Then

$$\left|\sum_{\tau\in\mathcal{T}_h}\int_\tau \nabla(u - u_I)\cdot\mathcal{D}_\tau\nabla v_h\right| \lesssim h^{1+\min(1,\sigma)}|\log h|^{1/2}\|u\|_{3,\infty,\Omega}|v_h|_{1,\Omega}. \tag{2.11}$$

Proof. Applying Lemma 2.3,

$$\sum_{\tau\in\mathcal{T}_h}\int_\tau \nabla(u-u_I)\cdot\mathcal{D}_\tau\nabla v_h = I_1+I_2, \tag{2.12}$$

where

$$I_1=\sum_{\tau\in\mathcal{T}_h}\sum_{k=1}^{3}\int_{e_k}\frac{\xi_k q_k}{2\sin\theta_k}\left\{(\ell_{k+1}^2-\ell_{k-1}^2)\frac{\partial^2 u}{\partial \boldsymbol{t}_k^2}+4|\tau|\frac{\partial^2 u}{\partial\boldsymbol{t}_k\partial\boldsymbol{n}_k}\right\}\frac{\partial v_h}{\partial\boldsymbol{t}_k},$$

$$I_2=-\sum_{\tau\in\mathcal{T}_h}\int_\tau\sum_{k=1}^{3}\frac{\ell_k\xi_k}{2\sin^2\theta_k}\left\{\ell_{k+1}\psi_{k-1}\frac{\partial^3 u}{\partial^2\boldsymbol{t}_{k+1}\partial\boldsymbol{t}_{k-1}}+\ell_{k-1}\psi_{k+1}\frac{\partial^3 u}{\partial^2\boldsymbol{t}_{k-1}\partial\boldsymbol{t}_{k+1}}\right\}\frac{\partial v_h}{\partial\boldsymbol{t}_k},$$

I_2 is easily estimated by

$$|I_2|\lesssim h^2\|u\|_{3,\Omega}|v_h|_{1,\Omega}. \tag{2.13}$$

To estimate I_1, let $\mathcal{E}=\mathcal{E}_1\oplus\mathcal{E}_2$ denote the set of interior edges. For each $e\in\mathcal{E}$, let τ and τ' share e as a common edge. Denote, with respect to τ,

$$\alpha_e=\frac{\xi_k}{2\sin\theta_k}(\ell_{k+1}^2-\ell_{k-1}^2),\qquad \beta_e=\frac{\xi_k}{2\sin\theta_k}4|\tau|,$$

and with respect to τ',

$$\alpha_e'=\frac{\xi_{k'}}{2\sin\theta_{k'}}(\ell_{k'+1}^2-\ell_{k'-1}^2),\qquad \beta_e'=\frac{\xi_{k'}}{2\sin\theta_{k'}}4|\tau'|.$$

Take $\boldsymbol{n}$ and $\boldsymbol{t}$ to correspond to τ. Then we can write

$$I_1=I_{11}+I_{12}+I_{13},$$

where

$$I_{1j}=\sum_{e\in\mathcal{E}_j}\int_e q_e\left\{(\alpha_e-\alpha_e')\frac{\partial^2 u}{\partial\boldsymbol{t}^2}+(\beta_e-\beta_e')\frac{\partial^2 u}{\partial\boldsymbol{t}\partial\boldsymbol{n}}\right\}\frac{\partial v_h}{\partial\boldsymbol{t}}$$

for $j=1,2$, and

$$I_{13}=\sum_{e\subset\partial\Omega}\int_e q_e\left\{\alpha_e\frac{\partial^2 u}{\partial\boldsymbol{t}^2}+\beta_e\frac{\partial^2 u}{\partial\boldsymbol{t}\partial\boldsymbol{n}}\right\}\frac{\partial v_h}{\partial\boldsymbol{t}}.$$

Using the elementary identity

$$\left|\int_e f\right|\lesssim h^{-1}\int_\tau|f|+\int_\tau|\nabla f|,$$

we obtain (for $\boldsymbol{z}=\boldsymbol{t}$ and $\boldsymbol{z}=\boldsymbol{n}$)

$$\left|\int_e q_e\frac{\partial^2 u}{\partial\boldsymbol{t}\partial\boldsymbol{z}}\frac{\partial v_h}{\partial\boldsymbol{t}}\right|\lesssim h^{-1}\int_\tau|\nabla^2 u||\nabla v_h|+\int_\tau|\nabla^3 u||\nabla v_h|. \tag{2.14}$$

We can estimate this term in a slightly different way:

$$\left|\int_e q_e \frac{\partial^2 u}{\partial \boldsymbol{t} \partial \boldsymbol{z}} \frac{\partial v_h}{\partial \boldsymbol{t}}\right| \lesssim h^{-1}|u|_{2,\infty,\Omega} \int_\tau |\nabla v_h|. \tag{2.15}$$

For $e \in \mathcal{E}_1$,

$$|\alpha_e - \alpha'_e| \lesssim h^3,$$
$$|\beta_e - \beta'_e| \lesssim h^3.$$

Combining this with (2.14), we have

$$|I_{11}| \lesssim h^2 \int_\Omega (|\nabla^2 u| + h\nabla^3 u|)|\nabla v_h| \lesssim h^2 \|u\|_{3,\Omega} |v_h|_{1,\Omega}, \tag{2.16}$$

or, by (2.15), we have

$$|I_{11}| \lesssim h^2 |u|_{2,\infty,\Omega} |v_h|_{1,\Omega}.$$

Now we turn to the estimate for I_{12}. For $e \in \mathcal{E}_2$, we simply estimate

$$|\alpha_e - \alpha'_e| \le |\alpha_e| + |\alpha'_e| \lesssim h^2,$$
$$|\beta_e - \beta'_e| \le |\beta_e| + |\beta'_e| \lesssim h^2.$$

Using (2.15), this leads to

$$|I_{12}| \lesssim h^{1+\sigma} |u|_{2,\infty,\Omega} |v_h|_{1,\Omega}.$$

We now consider I_{13}. It is easy to see that, if $v_h = 0$ on $\partial\Omega$, then $I_{13} = 0$. In the general case, we set

$$B_e(u) = \alpha_e \frac{\partial^2 u}{\partial \boldsymbol{t}^2} + \beta_e \frac{\partial^2 u}{\partial \boldsymbol{t} \partial \boldsymbol{n}}$$

and

$$\overline{B}_e(u) = |e|^{-1} \int_e B_e(u).$$

Then

$$\begin{aligned} I_{13} &= \sum_{e \subset \partial\Omega} \int_e q_e B_e(u) \frac{\partial v_h}{\partial \boldsymbol{t}} \\ &= \sum_{e \subset \partial\Omega} \int_e q_e \overline{B}_e(u) \frac{\partial v_h}{\partial \boldsymbol{t}} + \sum_{e \subset \partial\Omega} \int_e q_e (B_e(u) - \overline{B}_e(u)) \frac{\partial v_h}{\partial \boldsymbol{t}}. \end{aligned}$$

For the second term, we have

$$\begin{aligned} \left| \sum_{e \subset \partial\Omega} \int_e q_e (B_e(u) - \overline{B}_e(u)) \frac{\partial v_h}{\partial \boldsymbol{t}} \right| &\lesssim h^3 |u|_{3,\infty,\Omega} \sum_{e \subset \partial\Omega} \int_e \left| \frac{\partial v_h}{\partial \boldsymbol{t}} \right| \\ &\lesssim h^{5/2} |u|_{3,\infty,\Omega} |v_h|_{1,\Omega}. \end{aligned}$$

We now estimate the first term. Let $\mathcal{P} = \mathcal{P}_1 \oplus \mathcal{P}_2$ denote the set of vertices on $\partial\Omega$. Then we have

$$\begin{aligned}\sum_{e\subset\partial\Omega}\int_e q_e\overline{B}_e(u)\frac{\partial v_h}{\partial \boldsymbol{t}} &= \sum_{e\subset\partial\Omega}\overline{B}_e(u)\frac{\partial v_h}{\partial \boldsymbol{t}}\int_e q_e\\ &= \sum_{e\subset\partial\Omega}\overline{B}_e(u)\frac{\partial v_h}{\partial \boldsymbol{t}}\frac{|e|}{6}\\ &= \frac{1}{6}\sum_{x\in\mathcal{P}}\left(\overline{B}_e(u)-\overline{B}_{e'}(u)\right)v_h(x).\end{aligned}$$

For $x \in \mathcal{P}_1$, we have

$$\begin{aligned}|\alpha_e - \alpha_{e'}| &\lesssim h^3,\\ |\beta_e - \beta_{e'}| &\lesssim h^3.\end{aligned}$$

Thus

$$\left|\overline{B}_e(u)-\overline{B}_{e'}(u)\right| \lesssim h^3|u|_{3,\infty,\Omega}.$$

For $x \in \mathcal{P}_2$, we have

$$\left|\overline{B}_e(u)-\overline{B}_{e'}(u)\right| \le \left|\overline{B}_e(u)\right| + \left|\overline{B}_{e'}(u)\right| \lesssim h^2|u|_{2,\infty,\Omega}.$$

Combining these estimates, we have

$$\begin{aligned}\left|\sum_{x\in\mathcal{P}}\left(\overline{B}_e(u)-\overline{B}_{e'}(u)\right)v_h(x)\right| &\lesssim h^2\left(|u|_{3,\infty,\Omega}+\kappa|u|_{2,\infty,\Omega}\right)\|v_h\|_{\infty,\partial\Omega}\\ &\lesssim h^2|\log h|^{1/2}\|u\|_{3,\infty,\Omega}\|v_h\|_{1,\Omega}.\end{aligned}$$

In the last step, we used the well-known Sobolev inequality

$$\|v_h\|_{\infty,\Omega} \lesssim |\log h|^{1/2}\|v_h\|_{1,\Omega}.$$

Here $\|v_h\|_{1,\Omega}$ can be replaced by $|v_h|_{1,\Omega}$ by a standard argument. Thus our final estimate is

$$|I_{13}| \lesssim h^2|\log h|^{1/2}\|u\|_{3,\infty,\Omega}|v_h|_{1,\Omega}.$$

Consequently

$$|I_1| \lesssim h^{1+\min(1,\sigma)}|\log h|^{1/2}\|u\|_{3,\infty,\Omega}|v_h|_{1,\Omega}. \tag{2.17}$$

Combining (2.12) with (2.13) and (2.17), we obtain (2.11). □

For pure Dirichlet boundary conditions, we have the following better estimate.

COROLLARY 2.6. *Assume the conditions of Lemma* 2.5, *except for the second part of Definition* 2.4 *concerning regularity on the elements near the boundary. Then*

$$\begin{aligned}&\left|\sum_{\tau\in\mathcal{T}_h}\int_\tau \nabla(u-u_I)\cdot\mathcal{D}_\tau\nabla v_h\right|\\ &\qquad\lesssim h^{1+\min(1,\sigma)}(\|u\|_{3,\Omega}+\|u\|_{2,\infty,\Omega})|v_h|_{1,\Omega}, \qquad v_h\in\mathcal{V}_h\cap H_0^1(\Omega).\end{aligned}$$

Proof. Use $I_{13} = 0$ in Lemma 2.5. □

In the general case, without the second part of Definition 2.4, we have this slightly weaker result.

COROLLARY 2.7. *Assume the conditions of Lemma* 2.5*, except for the second part of Definition* 2.4 *concerning regularity on the elements near the boundary. Then*

$$\left|\sum_{\tau\in\mathcal{T}_h}\int_\tau \nabla(u-u_I)\cdot\mathcal{D}_\tau\nabla v_h\right| \lesssim h^{1+\min(1/2,\sigma)}(\|u\|_{3,\Omega}+\|u\|_{2,\infty,\Omega})|v_h|_{1,\Omega}, \qquad v_h\in\mathcal{V}_h.$$

Proof. We always have the following estimate for I_{13}:

$$|I_{13}| \lesssim h^{3/2}|u|_{2,\infty,\partial\Omega}|v_h|_{1,\Omega}. \qquad \square$$

We conclude with a final technical result needed in section 4.

LEMMA 2.8. *Let the triangulation $\mathcal{T}_h$ be $O(h^{2\sigma})$ irregular. Then*

$$\left|\sum_\tau\int_{\partial\tau}\sum_{k=1}^{3}\ell_k^2\frac{\partial^2 u}{\partial \boldsymbol{t}_k^2}\boldsymbol{v}_h\cdot\boldsymbol{n}\right| \lesssim h^{1+\min(1,\sigma)}|\log h|^{1/2}\|u\|_{3,\infty,\Omega}\|\boldsymbol{v}_h\|_{0,\Omega}. \tag{2.18}$$

Proof. Let $e\equiv e_k$ be an arbitrary edge of element τ. We begin with the identity

$$\ell_k^2\frac{\partial^2 u}{\partial \boldsymbol{t}_k^2}+\ell_{k+1}^2\frac{\partial^2 u}{\partial \boldsymbol{t}_{k+1}^2}+\ell_{k-1}^2\frac{\partial^2 u}{\partial \boldsymbol{t}_{k-1}^2} = (\alpha_e-\delta_e)\frac{\partial^2 u}{\partial \boldsymbol{t}_k^2}+\beta_e\frac{\partial^2 u}{\partial \boldsymbol{t}_k\partial \boldsymbol{n}_k}+\delta_e\frac{\partial^2 u}{\partial \boldsymbol{n}_k^2},$$

where

$$\begin{aligned}\alpha_e &= \ell_k^2+\ell_{k+1}^2+\ell_{k-1}^2,\\ \beta_e &= (\ell_{k+1}^2-\ell_{k-1}^2)4|\tau|/\ell_k^2,\\ \delta_e &= 8|\tau|^2/\ell_k^2.\end{aligned}$$

For $e\in\mathcal{E}$, let τ and τ' share e as a common edge. Take $\boldsymbol{n}$ and $\boldsymbol{t}$ to correspond to τ. Then we can write

$$\sum_\tau\int_{\partial\tau}\sum_{k=1}^{3}\ell_k^2\frac{\partial^2 u}{\partial \boldsymbol{t}_k^2}\boldsymbol{v}_h\cdot\boldsymbol{n} = I_1+I_2+I_3,$$

where

$$I_j = \sum_{e\in\mathcal{E}_j}\int_e\left\{(\alpha_e-\alpha_e')\frac{\partial^2 u}{\partial \boldsymbol{t}^2}+(\beta_e-\beta_e')\frac{\partial^2 u}{\partial \boldsymbol{t}\partial \boldsymbol{n}}\right\}\boldsymbol{v}_h\cdot\boldsymbol{n}$$

for $j=1,2$ and

$$I_3 = \sum_{e\in\partial\Omega}\int_e\left\{(\alpha_e-\delta_e)\frac{\partial^2 u}{\partial \boldsymbol{t}^2}+\beta_e\frac{\partial^2 u}{\partial \boldsymbol{t}\partial \boldsymbol{n}}+\delta_e\frac{\partial^2 u}{\partial \boldsymbol{n}^2}\right\}\boldsymbol{v}_h\cdot\boldsymbol{n}.$$

Following the pattern of proof in Lemma 2.5, we estimate

$$\begin{aligned}|I_1| &\lesssim h^2\|u\|_{3,\Omega}\|\boldsymbol{v}_h\|_{0,\Omega},\\ |I_2| &\lesssim h^{1+\sigma}|u|_{2,\infty,\Omega}\|\boldsymbol{v}_h\|_{0,\Omega},\\ |I_3| &\lesssim h^{1+\min(1,\sigma)}|\log h|^{1/2}\|u\|_{3,\infty,\Omega}\|\boldsymbol{v}_h\|_{0,\Omega}.\end{aligned}$$

Equation (2.18) now follows directly from these estimates. □

3. Elliptic boundary value problems. We consider the non–self-adjoint and possibly indefinite problem: find $u \in H^1(\Omega)$ such that

$$B(u,v) = \int_\Omega (\mathcal{D}\nabla u + \boldsymbol{b}u)\cdot\nabla v + cuv\,dx = f(v) \tag{3.1}$$

for all $v \in H^1(\Omega)$. Here $\mathcal{D}$ is a 2×2 symmetric positive definite matrix, $\boldsymbol{b}$ a vector, and c a scalar, and $f(\cdot)$ is a linear functional. We assume that all the coefficient functions are smooth.

In order to insure that (3.1) has a unique solution, we assume that the bilinear form $B(\cdot,\cdot)$ satisfies the continuity condition

$$|B(\phi,\eta)| \le \nu\, \|\phi\|_{1,\Omega}\|\eta\|_{1,\Omega} \tag{3.2}$$

for all $\phi, \eta \in H^1(\Omega)$. We also assume the inf-sup conditions

$$\inf_{\phi\in H^1}\sup_{\eta\in H^1}\frac{B(\phi,\eta)}{\|\phi\|_{1,\Omega}\|\eta\|_{1,\Omega}} = \sup_{\phi\in H^1}\inf_{\eta\in H^1}\frac{B(\phi,\eta)}{\|\phi\|_{1,\Omega}\|\eta\|_{1,\Omega}} \ge \mu > 0. \tag{3.3}$$

Let $\mathcal{V}_h \subset H^1(\Omega)$ be the space of continuous piecewise linear polynomials associated with the triangulation $\mathcal{T}_h$, and consider the approximate problem: find $u_h \in \mathcal{V}_h$ such that

$$B(u_h, v_h) = f(v_h) \tag{3.4}$$

for all $v_h \in \mathcal{V}_h$. To insure a unique solution for (3.4) we assume the inf-sup conditions

$$\inf_{\phi\in \mathcal{V}_h}\sup_{\eta\in \mathcal{V}_h}\frac{B(\phi,\eta)}{\|\phi\|_{1,\Omega}\|\eta\|_{1,\Omega}} = \sup_{\phi\in \mathcal{V}_h}\inf_{\eta\in \mathcal{V}_h}\frac{B(\phi,\eta)}{\|\phi\|_{1,\Omega}\|\eta\|_{1,\Omega}} \ge \mu > 0. \tag{3.5}$$

Xu and Zikatanov [21] have shown that, under these assumptions,

$$\|u-u_h\|_{1,\Omega} \le \frac{\nu}{\mu}\inf_{v_h\in\mathcal{V}_h}\|u-v_h\|_{1,\Omega}.$$

See also Aziz and Babuška [1].

We define the piecewise constant matrix function $\mathcal{D}_\tau$ in terms of the diffusion matrix $\mathcal{D}$ as follows:

$$\mathcal{D}_{\tau ij} = \frac{1}{|\tau|}\int_\tau \mathcal{D}_{ij}\,dx.$$

Note that $\mathcal{D}_\tau$ is symmetric and positive definite.

Theorem 3.1. *Assume that the solution of* (3.1) *satisfies* $u \in W^{3,\infty}(\Omega)$. *Further, assume the hypotheses of Lemma* 2.5, *with* $\mathcal{D}_\tau$ *defined as above. Then*

$$\|u_h - u_I\|_{1,\Omega} \lesssim h^{1+\min(1,\sigma)}|\log h|^{1/2}\|u\|_{3,\infty,\Omega}.$$

Proof. We begin with the identity

$$B(u-u_I, v_h) = \sum_{\tau\in\mathcal{T}_h}\int_\tau \nabla(u-u_I)\cdot\mathcal{D}_\tau\nabla v_h\,dx + \sum_{\tau\in\mathcal{T}_h}\int_\tau \nabla(u-u_I)\cdot(\mathcal{D}-\mathcal{D}_\tau)\nabla v_h\,dx$$
$$+\int_\Omega (u-u_I)(\boldsymbol{b}\cdot\nabla v_h + cv_h)\,dx = I_1 + I_2 + I_3.$$

The first term I_1 is estimated using Lemma 2.5. I_2 and I_3 can be easily estimated by

$$|I_2| + |I_3| \lesssim h^2 \|u\|_{2,\Omega} \|v_h\|_{1,\Omega}.$$

Thus

$$|B(u - u_I, v_h)| \lesssim h^{1+\min(1,\sigma)} |\log h|^{1/2} \|u\|_{3,\infty,\Omega} \|v_h\|_{1,\Omega}.$$

We complete the proof using the inf-sup condition in

$$\begin{aligned} \mu \|u_h - u_I\|_{1,\Omega} &\le \sup_{v_h \in \mathcal{V}_h} \frac{B(u_h - u_I, v_h)}{\|v_h\|_{1,\Omega}} \\ &= \sup_{v_h \in \mathcal{V}_h} \frac{B(u - u_I, v_h)}{\|v_h\|_{1,\Omega}} \\ &\lesssim h^{1+\min(1,\sigma)} |\log h|^{1/2} \|u\|_{3,\infty,\Omega}. \qquad \square \end{aligned}$$

We now consider a more general nonlinear problem: find $u \in H^1(\Omega)$ such that

$$\mathcal{B}(u, v) = f(v) \tag{3.6}$$

for all $v \in H^1(\Omega)$. Here the form $\mathcal{B}(\cdot,\cdot)$ is assumed to be linear in its second argument, but nonlinear in its first. Once again, $f(v)$ is a linear functional. Let u_h be the finite element approximation: find $u_h \in \mathcal{V}_h$ such that

$$\mathcal{B}(u_h, v_h) = f(v_h) \tag{3.7}$$

for all $v_h \in \mathcal{V}_h$. We assume that $\mathcal{B}(\cdot,\cdot)$ is such that its linearization about u is a bilinear form $B(\cdot,\cdot)$ as in (3.1), although the coefficient functions will now generally depend on u. We assume that $B(\cdot,\cdot)$ satisfies the continuity and inf-sup conditions (3.2), (3.3), and (3.5), so that both (3.6) and (3.7) have unique solutions. The linearization process also satisfies

$$\mathcal{B}(u, v_h) - \mathcal{B}(u_h, v_h) = B(u - u_h, v_h) + \mathcal{Q}(u - u_h, v_h) = 0$$

for all $v_h \in \mathcal{V}_h$. The form $\mathcal{Q}(\cdot,\cdot)$ contains higher order truncation terms in the linearization process; as with $\mathcal{B}(\cdot,\cdot)$, it is linear in its second argument. We assume

$$|\mathcal{Q}(u - u_h, v_h)| \lesssim \|u - u_h\|_{1,\Omega}^2 \|v_h\|_{1,\Omega}. \tag{3.8}$$

THEOREM 3.2. *Assume the hypotheses of Theorem* 3.1 *and* (3.8). *Then*

$$\|u_h - u_I\|_{1,\Omega} \lesssim h^{1+\min(1,\sigma)} |\log h|^{1/2} \|u\|_{3,\infty,\Omega} + \|u - u_h\|_{1,\Omega}^2.$$

Proof. As in the proof of Theorem 3.1,

$$\begin{aligned} \mu \|u_h - u_I\|_{1,\Omega} &\le \sup_{v_h \in \mathcal{V}_h} \frac{B(u_h - u_I, v_h)}{\|v_h\|_{1,\Omega}} \\ &\le \sup_{v_h \in \mathcal{V}_h} \frac{B(u - u_I, v_h)}{\|v_h\|_{1,\Omega}} + \frac{\mathcal{Q}(u - u_h, v_h)}{\|v_h\|_{1,\Omega}} \\ &\lesssim h^{1+\min(1,\sigma)} |\log h|^{1/2} \|u\|_{3,\infty,\Omega} + \|u - u_h\|_{1,\Omega}^2. \qquad \square \end{aligned}$$

If $\|u - u_h\|_{1,\Omega}$ is sufficiently small (e.g., $\|u - u_h\|_{1,\Omega} \le C(u)h$), then we will observe superconvergence.

4. A gradient recovery algorithm for $O(h^2)$ approximate parallelogram meshes. In this section, we show that $Q_h \nabla u_I$ can superconverge to ∇u for meshes that are $O(h^{2\sigma})$ irregular.

THEOREM 4.1. *Let $u \in W^{3,\infty}(\Omega)$, and assume the hypotheses of Lemma 2.8. Then*

$$\|\nabla u - Q_h \nabla u_I\|_{0,\Omega} \lesssim h^{1+\min(1,\sigma)} |\log h|^{1/2} \|u\|_{3,\infty,\Omega}.$$

Proof. Given $\boldsymbol{v}_h \in \mathcal{V}_h \times \mathcal{V}_h$, we have

$$(Q_h \nabla(u-u_I), \boldsymbol{v}_h) = (\nabla(u-u_I), \boldsymbol{v}_h) = -((u-u_I), \nabla \cdot \boldsymbol{v}_h) + \int_{\partial\Omega} (u-u_I)\boldsymbol{v}_h \cdot \boldsymbol{n}. \tag{4.1}$$

We estimate the two terms on the right-hand side of (4.1). First,

$$\left| \int_{\partial\Omega} (u-u_I)\boldsymbol{v}_h \cdot \boldsymbol{n} \right| \lesssim h^{3/2} |u|_{2,\infty,\Omega} \|\boldsymbol{v}_h\|_{0,\Omega}.$$

For the other, we use Lemma 2.2 to get

$$\begin{aligned}
\int_\tau (u-u_I)\nabla \cdot \boldsymbol{v}_h &= -\frac{1}{24} \int_\tau \sum_{k=1}^{3} \ell_k^2 \frac{\partial^2 u_q}{\partial \boldsymbol{t}_k^2} \nabla \cdot \boldsymbol{v}_h + \int_\tau (u-u_q)\nabla \cdot \boldsymbol{v}_h \\
&= -\frac{1}{24} \int_\tau \sum_{k=1}^{3} \ell_k^2 \frac{\partial^2 u}{\partial \boldsymbol{t}_k^2} \nabla \cdot \boldsymbol{v}_h \\
&\qquad - \frac{1}{24} \int_\tau \sum_{k=1}^{3} \ell_k^2 \frac{\partial^2 (u_q - u)}{\partial \boldsymbol{t}_k^2} \nabla \cdot \boldsymbol{v}_h + \int_\tau (u-u_q)\nabla \cdot \boldsymbol{v}_h \\
&= -\frac{1}{24} \int_{\partial\tau} \sum_{k=1}^{3} \ell_k^2 \frac{\partial^2 u}{\partial \boldsymbol{t}_k^2} \boldsymbol{v}_h \cdot \boldsymbol{n} + \frac{1}{24} \int_\tau \sum_{k=1}^{3} \ell_k^2 \nabla \frac{\partial^2 u}{\partial \boldsymbol{t}_k^2} \boldsymbol{v}_h \\
&\qquad - \frac{1}{24} \int_\tau \sum_{k=1}^{3} \ell_k^2 \frac{\partial^2 (u_q - u)}{\partial \boldsymbol{t}_k^2} \nabla \cdot \boldsymbol{v}_h + \int_\tau (u-u_q)\nabla \cdot \boldsymbol{v}_h \\
&= I_1 + I_2 + I_3 + I_4.
\end{aligned}$$

Easy estimates show

$$\begin{aligned}
|I_3| + |I_4| &\lesssim h^3 \|u\|_{3,\tau} |\boldsymbol{v}_h|_{1,\tau} \lesssim h^2 \|u\|_{3,\tau} \|\boldsymbol{v}_h\|_{0,\tau}, \\
|I_2| &\lesssim h^2 \|u\|_{3,\tau} \|\boldsymbol{v}_h\|_{0,\tau}.
\end{aligned}$$

$|I_1|$ is estimated using Lemma 2.8. Consequently,

$$|(Q_h \nabla(u-u_I), \boldsymbol{v}_h)| \lesssim h^{1+\min(1,\sigma)} |\log h|^{1/2} \|u\|_{3,\infty,\Omega} \|\boldsymbol{v}_h\|_{0,\Omega}.$$

Taking $\boldsymbol{v}_h = Q_h \nabla(u-u_I)$, it follows that

$$\|Q_h \nabla(u-u_I)\|_{0,\Omega} \lesssim h^{1+\min(1,\sigma)} |\log h|^{1/2} \|u\|_{3,\infty,\Omega}.$$

Theorem 4.1 now follows from the triangle inequality

$$\|\nabla u - Q_h \nabla u_I\|_{0,\Omega} \le \|\nabla u - Q_h \nabla u\|_{0,\Omega} + \|Q_h \nabla(u-u_I)\|_{0,\Omega}. \qquad \square$$

The next result is an immediate consequence of Theorems 3.1 and 4.1.

Theorem 4.2. *Let $u \in W^{3,\infty}(\Omega)$, and assume the hypotheses of Theorems* 3.1 *and* 4.1. *Then*

$$\|\nabla u - Q_h \nabla u_h\|_{0,\Omega} \lesssim h^{1+\min(1,\sigma)} |\log h|^{1/2} \|u\|_{3,\infty,\Omega}.$$

Proof. Using the triangle inequality,

$$\begin{aligned} \|\nabla u - Q_h \nabla u_h\|_{0,\Omega} &\le \|\nabla u - Q_h \nabla u_I\|_{0,\Omega} + \|Q_h \nabla (u_I - u_h)\|_{0,\Omega} \\ &\le \|\nabla u - Q_h \nabla u_I\|_{0,\Omega} + \|\nabla (u_I - u_h)\|_{0,\Omega}. \end{aligned} \tag{4.2}$$

We estimate the two terms on the right-hand side of (4.2) using Theorems 4.1 and 3.1. □

Finally, we would like to point out that many results presented above (such as Theorems 3.1, 3.2, 4.1, and 4.2) can be refined in many ways. Before the end of this section, let us give one such refinement for a piecewise $O(h^{2\sigma})$ irregular grid.

Definition 4.3. *The triangulation $\mathcal{T}_h$ is piecewise $O(h^{2\sigma})$ irregular if Ω can be written as the union of a bounded number of polygonal subdomains and $\mathcal{T}_h$ is $O(h^{2\sigma})$ irregular on each of these subdomains.*

By applying Lemma 2.5 on each subdomain, we can easily get the following result.

Theorem 4.4. *Lemma* 2.5, *Lemma* 2.8, *Theorem* 3.1, *Theorem* 3.2, *Theorem* 4.1, *and Theorem* 4.2 *are all valid for piecewise $O(h^{2\sigma})$ grids.*

The above theorem is related to superconvergence results on piecewise regular (or strongly regular) grids that were discussed in earlier literature (cf. Xu [20] and Lin and Xu [15]). The significance of such an extension will be discussed in the following section.

5. Applications and numerical experiments.

In this section, we develop a few simple applications of our results and present some numerical examples. The numerical experiments were performed using the PLTMG software package [5]. The experiments were done on an Linux PC using double precision arithmetic and the g77 compiler.

We begin our discussion with a very simple example of piecewise uniform grids. As shown in Figure 5.1, we began with a uniform 3×3 mesh with $nt = 8$ elements, and computed a sequence of uniformly refined meshes through regular refinement of each element of a given mesh into four similar triangles in the refined mesh by connecting the midpoints pairwise.

This grid is $O(h)$ irregular ($\sigma = 1/2$) by Definition 2.4, but piecewise $O(h^{2\sigma})$ irregular with $\sigma = \infty$ (namely, piecewise regular) by Definition 4.3. Consequently, for this example, the result claimed by Theorem 4.4 is $O(h^{1/2})$ better than the corresponding result from previous sections. In our first experiment, we consider the problem

$$-\Delta u + u = f, \tag{5.1}$$

$\Omega = (0,1) \times (0,1)$ with either Dirichlet or Neumann boundary conditions. The right-hand side f and the boundary conditions were chosen such that $u = e^{x+y}$ was the exact solution. In this experiment, we begin with the uniform 3×3 mesh with eight triangles described above, and make seven levels of uniform regiment. The results are

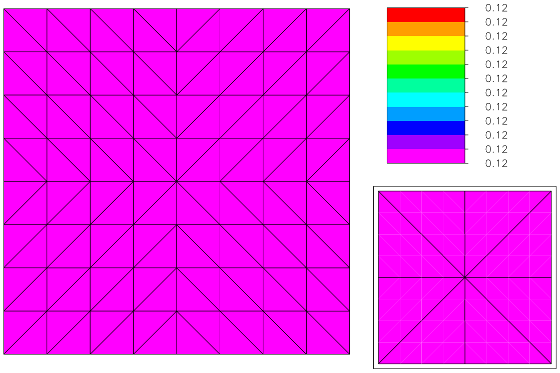


FIG. 5.1. *A (globally) $O(h^{2\sigma})$ irregular grid with $\sigma = 1/2$, but piecewise $O(h^{2\sigma})$ irregular grid with $\sigma = \infty$.*

TABLE 5.1
Results for a square domain, uniform refinement.

	Dirichlet problem			Neumann problem		
nt	H_1	$\widetilde{H}_1$	$\overline{H}_1$	H_1	$\widetilde{H}_1$	$\overline{H}_1$
8	1.2e 0	2.7e-1	6.0e-1	9.5e-1	7.2e-1	6.7e-1
32	6.0e-1	8.1e-2	2.4e-1	5.5e-1	2.4e-1	3.0e-1
128	3.0e-1	2.3e-2	8.8e-2	2.9e-1	7.5e-2	1.1e-1
512	1.5e-1	6.1e-3	3.2e-2	1.5e-1	2.2e-2	3.7e-2
2048	7.5e-2	1.6e-3	1.1e-2	7.5e-2	6.1e-3	1.3e-2
8192	3.8e-2	4.4e-4	4.0e-3	3.8e-2	1.7e-3	4.3e-3
32768	1.9e-2	1.2e-4	1.4e-3	1.9e-2	4.5e-4	1.5e-3
131072	9.4e-3	3.0e-5	5.1e-4	9.4e-3	1.2e-4	5.2e-4
order	1.01	1.95	1.51	1.01	1.91	1.55

reported in Table 5.1. In Table 5.1 and subsequent tables,

$$H_1 = \|\nabla(u - u_h)\|_{0,\Omega},$$
$$\widetilde{H}_1 = \|\nabla(u_I - u_h)\|_{0,\Omega},$$
$$\overline{H}_1 = \|\nabla u - Q_h \nabla u_h)\|_{0,\Omega},$$

where u_I is the linear interpolant of u. In the last line, the order of convergence was estimated from the reported data using a least squares technique.

In Table 5.1, we see quite clearly the first order convergence of $\|\nabla(u - u_h)\|_{0,\Omega}$ and the superconvergence of $\|\nabla(u_I - u_h)\|_{0,\Omega}$. In the latter case, the rate is nearly second order, which is consistent with Theorem 4.4. We also note superconvergence of $\|\nabla u - Q_h \nabla u_h)\|_{0,\Omega}$, with order close to 3/2. This is perhaps the result of most practical significance.

We then repeated the experiment, replacing uniform refinement with the adaptive refinement procedure in PLTMG. This adaptive refinement procedure is based on longest-edge bisection and also includes a mesh smoothing phase that allows the vertices in the mesh to move. The result was a sequence of unstructured, nonuniform, nonnested, shape regular meshes. The target values for the adaptive procedure were selected to produce a sequence of meshes with approximately the same number of elements as for the uniform refinement case. The results are shown in Table 5.2.

For the adaptive meshes, the story is a quite different; $\|\nabla(u_I - u_h)\|_{0,\Omega}$ and $\|\nabla u -$

TABLE 5.2
Results for a square domain, adaptive refinement.

	Dirichlet problem				Neumann problem		
nt	H_1	$\widetilde{H}_1$	$\overline{H}_1$	nt	H_1	$\widetilde{H}_1$	$\overline{H}_1$
8	1.2e 0	2.7e-1	6.0e-1	8	9.5e-1	7.2e-1	6.7e-1
34	5.8e-1	1.0e-1	2.3e-1	36	4.6e-1	2.7e-1	2.3e-1
136	2.2e-1	7.2e-2	7.7e-2	134	2.5e-1	9.3e-2	1.0e-1
528	1.2e-1	3.4e-2	3.8e-2	526	1.2e-1	4.0e-2	3.8e-2
2079	6.0e-2	1.7e-2	1.7e-2	2080	6.0e-2	1.8e-2	1.6e-2
8254	2.9e-2	6.9e-3	7.1e-3	8257	2.9e-2	7.9e-3	7.2e-3
32888	1.4e-2	3.2e-3	3.0e-3	32890	1.4e-2	3.7e-3	3.2e-3
131301	6.9e-3	1.5e-3	1.4e-3	131311	7.0e-3	1.8e-3	1.5e-3
order	1.05	1.10	1.17		1.04	1.11	1.13

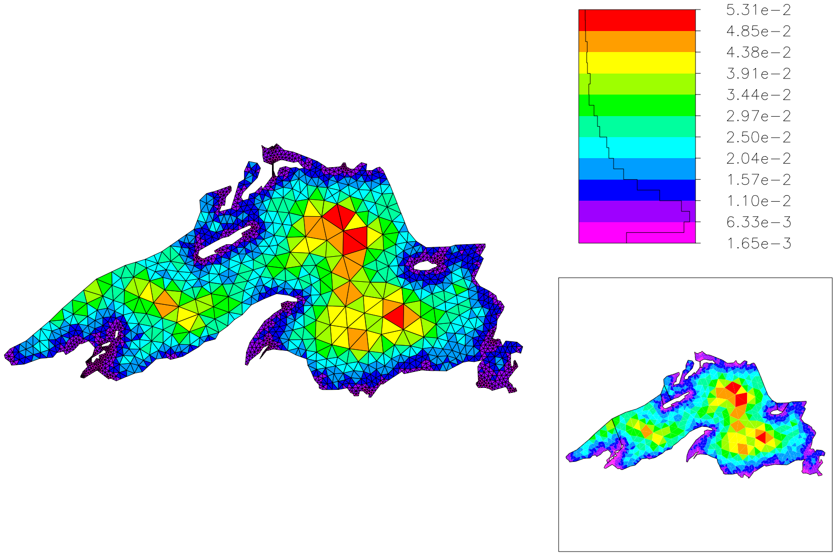


FIG. 5.2. *Lake Superior mesh with* $nt = 2765$. *Elements are colored according to size.*

$Q_h\nabla u_h)\|_{0,\Omega}$ show less superconvergence. In this case $\sigma > 0$, but it is clearly much smaller than in the uniform refinement case. In Part II of this work [7], we show how to obtain strong superconvergence for such meshes using $S^m Q_h \nabla u_h$ as the recovered gradient. Here S is a multigrid-like smoothing operator, and m is a small integer ($m = 1$ or $m = 2$ is usually satisfactory). Analysis and a complete description are deferred to Part II of this work.

In our second experiment, we solved (5.1) on a domain Ω in the shape of Lake Superior. The true solution u in this case was chosen to be $u = \sin x \sin y$. In this case, the initial mesh with $nt = 2765$ elements was unstructured and nonuniform, but shape regular. This mesh is shown in Figure 5.2. As in the first example, we first computed a sequence of uniformly refined meshes through regular refinement of each element of a given mesh into four similar triangles. The results are shown in Table 5.3.

In Table 5.3, we see quite clearly the first order convergence of $\|\nabla(u-u_h)\|_{0,\Omega}$ and the superconvergence of $\|\nabla(u_I - u_h)\|_{0,\Omega}$. In the latter case, the rate is nearly second order, which is again consistent with Theorem 4.4. Evidently, the many small uniform

Table 5.3
Lake Superior domain, uniform refinement.

	Dirichlet problem			Neumann problem		
nt	H_1	$\widetilde{H}_1$	$\overline{H}_1$	H_1	$\widetilde{H}_1$	$\overline{H}_1$
2765	9.2e-1	1.5e-1	2.5e-1	9.1e-1	1.6e-1	2.6e-1
11060	4.6e-1	4.5e-2	1.0e-1	4.6e-1	4.8e-2	1.0e-1
44240	2.3e-1	1.3e-2	3.5e-2	2.3e-1	1.4e-2	3.5e-2
176960	1.2e-1	3.6e-3	1.2e-2	1.2e-1	3.8e-3	1.2e-2
order	1.02	1.88	1.54	1.02	1.89	1.54

Table 5.4
Lake Superior domain, adaptive refinement.

	Dirichlet problem				Neumann problem		
nt	H_1	$\widetilde{H}_1$	$\overline{H}_1$	nt	H_1	$\widetilde{H}_1$	$\overline{H}_1$
2765	9.2e-1	1.5e-1	2.5e-1	2765	9.1e-1	1.6e-1	2.6e-1
11565	2.5e-1	4.0e-2	5.3e-2	11560	2.5e-1	4.3e-2	5.3e-2
45524	1.2e-1	1.8e-2	2.2e-2	45521	1.2e-1	1.9e-2	2.2e-2
179655	6.1e-2	8.3e-3	9.9e-3	179666	6.1e-2	8.7e-3	9.9e-3
order	1.12	1.22	1.32		1.12	1.26	1.32

patches were sufficient to produce a very strong superconvergence effect. We also see superconvergence of $Q_h \nabla u_h$ to ∇u, in a way similar to that of the first example.

We then repeated the experiment, replacing uniform with adaptive refinement. The target values for the adaptive procedure once again were selected to produce a sequence of meshes with approximately the same numbers of elements as in the uniform refinement case. The results are shown in Table 5.4.

The results here are qualitatively similar to those for the first example. In Table 5.4 we note slightly elevated estimates for the estimated order of convergence of $\|\nabla(u - u_h)\|_{0,\Omega}$. This is an artifact of the least squares procedure. Notice that in the first adaptive step there was an unusually large decrease in $\|\nabla(u - u_h)\|_{0,\Omega}$. This was because the initial nonuniform mesh was adapted mainly to the complex geometry of Ω and not to the character of the solution. In subsequent adaptive refinement steps, the error is rapidly approaching first order behavior. The orders for H_1 and $\overline{H}_1$ are also slightly elevated by unusually large decreases in the first adaptive step.

Acknowledgment. The ideas for this manuscript germinated while the authors were attending the workshop on *A Posteriori Error Estimation and Adaptive Approaches in the Finite Element Method,* held at the Mathematical Sciences Research Institute, University of California, Berkeley, April 3–14, 2000. We are grateful for the opportunity to attend this workshop and for the stimulating scientific environment at MSRI. We also thank the reviewers for their careful reviews and their suggestions for improvements.

REFERENCES

[1] A. K. Aziz and I. Babuška, *Survey lectures on the mathematical foundations of the finite element method*, Part I in The Mathematical Foundations of the Finite Element Method with Applications to Partial Differential Equations, Academic Press, New York, 1972, pp. 1–362.

[2] I. Babuška and W. C. Rheinboldt, *A posteriori error estimates for the finite element method*, Internat. J. Numer. Methods Engrg., 12 (1978), pp. 1597–1615.

[3] I. Babuška and T. Strouboulis, *The Finite Element Method and Its Reliability*, Clarendon Press, New York, 2001.

[4] I. Babuška, T. Strouboulis, and C. S. Upadhyay, *η%-superconvergence of finite element approximations in the interior of general meshes of triangles*, Comput. Methods Appl. Mech. Engrg., 122 (1995), pp. 273–305.

[5] R. E. Bank, *PLTMG: A Software Package for Solving Elliptic Partial Differential Equations: Users' Guide* 8.0, Software Environ. Tools 5, SIAM, Philadelphia, 1998.

[6] R. E. Bank and A. Weiser, *Some a posteriori error estimators for elliptic partial differential equations*, Math. Comp., 44 (1985), pp. 283–301.

[7] R. E. Bank and J. Xu, *Asymptotically exact a posteriori error estimators, Part* II: *General unstructured grids*, SIAM J. Numer. Anal., 41 (2003), pp. 2313–2332.

[8] C. Chen and Y. Huang, *High Accuracy Theory of Finite Element Methods*, Hunan Science Press, Hunan, China, 1995 (in Chinese).

[9] L. Du and N. Yan, *Gradient recovery type a posteriori error estimate for finite element approximation on non-uniform meshes*, Adv. Comput. Math., 14 (2001), pp. 175–193.

[10] R. Durán, M. A. Muschietti, and R. Rodríguez, *On the asymptotic exactness of error estimators for linear triangular finite elements*, Numer. Math., 59 (1991), pp. 107–127.

[11] I. Hlaváček and M. Křížek, *On a superconvergent finite element scheme for elliptic systems.* I. *Dirichlet boundary condition*, Apl. Mat., 32 (1987), pp. 131–154.

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

[13] A. M. Lakhany, I. Marek, and J. R. Whiteman, *Superconvergence results on mildly structured triangulations*, Comput. Methods Appl. Mech. Engrg., 189 (2000), pp. 1–75.

[14] B. Li and Z. Zhang, *Analysis of a class of superconvergence patch recovery techniques for linear and bilinear finite elements*, Numer. Methods Partial Differential Equations, 15 (1999), pp. 151–167.

[15] Q. Lin and J. Xu, *Linear finite elements with high accuracy*, J. Comput. Math., 3 (1985), pp. 115–133.

[16] Q. Lin and N. Yan, *The Construction and Analysis of High Efficiency Finite Elements*, Hebei University Press, Hunan, China, 1996 (in Chinese).

[17] R. Verfürth, *A Posteriori Error Estimation and Adaptive Mesh Refinement Techniques*, Teubner Skripten zur Numerik, B. G. Teubner, Stuttgart, 1995.

[18] L. B. Wahlbin, *Superconvergence in Galerkin Finite Element Methods*, Springer-Verlag, Berlin, 1995.

[19] L. B. Wahlbin, *General principles of superconvergence in Galerkin finite element methods*, in Finite Element Methods (Jyväskylä, 1997), Lecture Notes in Pure and Appl. Math. 196, Dekker, New York, 1998, pp. 269–285.

[20] J. Xu, *The error analysis and the improved algorithms for the infinite element method*, in Proceedings of the 1984 Beijing Symposium on Differential Geometry and Differential Equations, Beijing, China, 1985, Science Press, Hoboken, NJ, pp. 326–331.

[21] J. Xu and L. Zikatanov, *Some observations on Babuška and Brezzi theories*, Numer. Math., 94 (2003), pp. 195–202.

[22] N. Yan and A. Zhou, *Gradient recovery type a posteriori error estimates for finite element approximations on irregular meshes*, Comput. Methods Appl. Mech. Engrg., 190 (2001), pp. 4289–4299.

[23] Z. Zhang and H. D. Victory, Jr., *Mathematical analysis of Zienkiewicz-Zhu's derivative patch recovery technique*, Numer. Methods Partial Differential Equations, 12 (1996), pp. 507–524.

[24] Z. Zhang and J. Z. Zhu, *Superconvergence of the derivative patch recovery technique and a posteriori error estimation*, in Modeling, Mesh Generation, and Adaptive Numerical Methods for Partial Differential Equations (Minneapolis, MN, 1993), Springer, New York, 1995, pp. 431–450.

[25] J. Z. Zhu and O. C. Zienkiewicz, *Superconvergence recovery technique and a posteriori error estimators*, Internat. J. Numer. Methods Engrg., 30 (1990), pp. 1321–1339.

SIAM J. NUMER. ANAL.
Vol. 41, No. 6, pp. 2313–2332

ASYMPTOTICALLY EXACT A POSTERIORI ERROR ESTIMATORS, PART II: GENERAL UNSTRUCTURED GRIDS*

RANDOLPH E. BANK† AND JINCHAO XU‡

Abstract. In Part I of this work [*SIAM J. Numer. Anal.*, 41 (2003), pp. 2294–2312], we analyzed superconvergence for piecewise linear finite element approximations on triangular meshes where most pairs of triangles sharing a common edge form approximate parallelograms. In this work, we consider superconvergence for general unstructured but shape regular meshes. We develop a postprocessing gradient recovery scheme for the finite element solution u_h, inspired in part by the smoothing iteration of the multigrid method. This recovered gradient superconverges to the gradient of the true solution and becomes the basis of a global a posteriori error estimate that is often asymptotically exact. Next, we use the superconvergent gradient to approximate the Hessian matrix of the true solution and form local error indicators for adaptive meshing algorithms. We provide several numerical examples illustrating the effectiveness of our procedures.

Key words. superconvergence, gradient recovery, a posteriori error estimates

AMS subject classifications. 65N50, 65N30

DOI. 10.1137/S0036142901398751

1. Introduction.

In Part I of this work [10], we developed some superconvergence estimates and a gradient recovery algorithm appropriate for piecewise linear finite element approximations of elliptic boundary problems. In that work, we restricted attention to triangular meshes that are $O(h^{2\sigma})$ irregular [10]. In this work, we extend the gradient recovery scheme to more general meshes and develop an a posteriori error estimate and local error indicator for use in adaptive meshing algorithms. See [15, 23, 24, 25, 14, 22, 12, 16, 20, 5, 13, 1] for related work, and in particular the monographs by Verfürth [18], Babuška and Strouboulis [4], Chen and Huang [11], Lin and Yan [17], and Wahlbin [19] for recent surveys of the field as a whole.

Our overall development has three major steps. Let $\mathcal{V}_h \subset H^1(\Omega)$ be the finite element subspace consisting of continuous piecewise linear polynomials associated with a shape regular triangulation $\mathcal{T}_h$. Let $u_h \in \mathcal{V}_h$ be the finite element solution of an appropriate linear or nonlinear elliptic boundary value problem. In the first component of our development, we prove a superconvergence result for $|u_h - u_I|_{1,\Omega}$, where u_I is the piecewise linear interpolant for u. In particular, in Part I of this manuscript [10], we prove that

$$|u_h - u_I|_{1,\Omega} \lesssim h^{1+\min(1,\sigma)}|\log h|^{1/2}\|u\|_{3,\infty,\Omega}. \tag{1.1}$$

Estimate (1.1) holds on nonuniform meshes, where most pairs of adjacent triangles satisfy an $O(h^2)$ approximate parallelogram property. $\sigma > 0$ in some sense measures the extent to which this condition is violated; see [10] for details.

*Received by the editors November 26, 2001; accepted for publication (in revised form) April 30, 2003; published electronically December 17, 2003.
http://www.siam.org/journals/sinum/41-6/39875.html

†Department of Mathematics, University of California, San Diego, La Jolla, CA 92093-0112 (rbank@ucsd.edu). The work of this author was supported by the National Science Foundation under contracts DMS-9973276 and DMS-0208449.

‡Center for Computational Mathematics and Applications, Department of Mathematics, Pennsylvania State University, University Park, PA 16802 (xu@math.psu.edu). The work of this author was supported by the National Science Foundation under contract DMS-0074299 and by the Center for Computational Mathematics and Applications, Pennsylvania State University.

The second major component is a superconvergent approximation to ∇u. This approximation is generated by a gradient recovery procedure. In particular, in section 2 of this manuscript we compute $S^m Q_h \nabla u_h$, where S is an appropriate smoothing operator and Q_h is the L^2 projection operator. In words, the discontinuous, piecewise constant gradient ∇u_h is projected into the space of continuous piecewise linear polynomials, and then smoothed, using a multigrid-like smoothing operator. Although the L^2 projection operator is global, the overall work estimate is still $O(N)$ for a mesh with N vertices. In the case of a small number of smoothing steps (the most interesting case), Theorem 2.7 shows that

$$\|\nabla u - S^m Q_h \nabla u_h\|_{0,\Omega} \lesssim h \left\{ \min\left(h^{\min(1,\sigma)} |\log h|, \left[\frac{\kappa - 1}{\kappa}\right]^m \right) + m\, h^{1/2} \right\} \|u\|_{3,\infty,\Omega}. \tag{1.2}$$

Here $\kappa > 1$ is a constant independent of h and u. The term $(1 - \kappa^{-1})^m$ illustrates the well-known effectiveness of a few smoothing steps and is reminiscent of terms arising in connection with multigrid convergence analysis [7]. If σ is sufficiently large, then the L^2 projection itself ($m = 0$) is sufficient to produce superconvergence. The purpose of smoothing is to improve the performance when $\sigma \approx 0$ and the mesh is shape regular.

In the third major component of our analysis, presented in section 3 of this manuscript, we use the recovered gradient to develop an a posteriori error estimate. An obvious choice is to use $(I - S^m Q_h)\nabla u_h$ to approximate the true error $\nabla(u - u_h)$. In Theorem 3.1 we show this is a good choice, and that in many circumstances we can expect the error estimate to be *asymptotically exact*; that is,

$$\lim \frac{\|(I - S^m Q_h)\nabla u_h\|_{0,\Omega}}{\|\nabla(u - u_h)\|_{0,\Omega}} = 1$$

as $h \to 0$ and $m \to \infty$ in an appropriate fashion.

We also use the recovered gradient to construct local approximations of interpolation errors to be used as local error indicators for adaptive meshing algorithms. This is motivated by noting that, under certain circumstances, $|u_q - u_I|_{1,\Omega}$ is an asymptotically exact estimate of $|u - u_h|_{1,\Omega}$. Here u_q is the piecewise quadratic interpolant for u. Thus $u_q - u_I$ is a locally defined quadratic polynomial with value zero at all vertices of the mesh. On a given element τ, $u_q - u_I$ can be expressed as a linear combination of quadratic "bump functions" q_k associated with the edge midpoints of τ,

$$u_q - u_I = \sum_{k=1}^{3} \ell_k^2 \boldsymbol{t}_k^t M_\tau \boldsymbol{t}_k \, q_k(x, y), \tag{1.3}$$

where ℓ_k is the length of edge k, $\boldsymbol{t}_k$ is the unit tangent, and

$$M_\tau = -\frac{1}{2} \begin{pmatrix} \partial_{11} u_q & \partial_{12} u_q \\ \partial_{21} u_q & \partial_{22} u_q \end{pmatrix} \tag{1.4}$$

is the Hessian matrix (for details, see section 3). For convenience in notation, we let $\partial_i u$ denote the partial derivative $\partial u / \partial x_i$. All terms on the right-hand side of (1.3) are known except for the second derivatives appearing in the Hessian matrix M_τ. In our local error indicator, we simply replace $\partial_{ij} u_q$ by $\partial_i S^m Q_h \partial_j u_h$. Let ϵ_τ

denote this locally defined a posteriori error estimate. In Theorem 3.2, we prove the superconvergence estimate

$$\|\partial_i(\partial_k u - S^m Q_h \partial_k u_h)\|_{0,\Omega} \lesssim \left\{ \min\left(h^{\min(1,\sigma)} |\log h|, \left[\frac{\kappa - 1}{\kappa}\right]^m \right) + m\, h^{1/2} \right\} \|u\|_{3,\infty,\Omega}. \tag{1.5}$$

We remark that both our gradient recovery scheme and our a posteriori error estimate are largely independent of the details of the PDE. This suggests that superconvergence can be expected for a wide variety of problems, as long as the adaptive meshing yields smoothly varying, shape regular meshes.

It also is interesting to note that the superconvergent global approximation to ∇u emphasizes once again a classic dilemma in error estimation. On the one hand, generally it seems quite advantageous to take the superconvergent approximation $S^m Q_h \nabla u_h$ as the "accepted" approximation to ∇u. Not only is it of higher order than ∇u_h, but also it is globally continuous and differentiable, often desirable properties. On the other hand, the a posteriori error estimates and resulting adaptive meshing algorithms use $S^m Q_h \nabla u_h$ to estimate the error in ∇u_h. In some respects, the situation is analogous to adaptive time step selection schemes for initial value problems where order p and $p+1$ approximations are computed to estimate the local error in the order p approximation, which is then used to control the time step.

Finally, as a point of practical interest, since the gradient recovery and a posteriori error estimates are independent of the PDE, a single implementation can be used across a broad spectrum of problems. There is no need to have special implementations for each problem class, as is typical of schemes that involve the solution of local problems in each element or patch of elements [3, 9].

The rest of this paper is organized as follows: In section 2, we first provide some notation, describe our gradient recovery scheme, and summarize the main superconvergence estimates of [10] for the case of $O(h^{2\sigma})$ irregular meshes. We then extend the gradient recovery scheme to more general meshes through the use of a multigrid smoother. In section 3, we develop and analyze our a posteriori error estimate and prove (1.5). Finally, in section 4, we present several numerical examples, involving both uniform and nonuniform (adaptive) meshes, with some solutions that satisfy our smoothness assumptions and some that do not. In the latter cases, we observe superconvergence away from singularities for adaptive meshes, although this effect is not covered by our current analysis.

2. A gradient recovery algorithm for shape regular triangulations. Let $\Omega \subset \mathbb{R}^2$ be a bounded domain with Lipschitz boundary $\partial\Omega$. For simplicity of exposition, we assume that Ω is a polyhedron. We assume that Ω is partitioned by a shape regular triangulation $\mathcal{T}_h$ of mesh size h. Let $\mathcal{V}_h \subset H^1(\Omega)$ be the corresponding continuous piecewise linear finite element space associated with this triangulation $\mathcal{T}_h$.

We consider the non–self-adjoint and possibly indefinite problem: Find $u \in H^1(\Omega)$ such that

$$B(u,v) = \int_\Omega (\mathcal{D}\nabla u + \boldsymbol{b}u)\cdot\nabla v + cuv\,dx = f(v) \tag{2.1}$$

for all $v \in H^1(\Omega)$. Here $\mathcal{D}$ is a 2×2 symmetric positive definite matrix, $\boldsymbol{b}$ a vector, and c a scalar, and $f(\cdot)$ is a linear functional. We assume that all the coefficient functions are smooth. Choosing $H^1(\Omega)$ as trial space implies Neumann boundary conditions, a choice made for convenience. In [10], we also analyzed more general nonlinear PDEs and boundary conditions. However, since the details of the PDE do not strongly influence our gradient recovery scheme, here we consider only the most simple case.

In order to insure that (2.1) has a unique solution, we assume that the bilinear form $B(\cdot,\cdot)$ satisfies the continuity condition

$$|B(\phi,\eta)| \le \nu \, \|\phi\|_{1,\Omega}\|\eta\|_{1,\Omega} \tag{2.2}$$

for all $\phi,\eta \in H^1(\Omega)$. We also assume the inf-sup conditions

$$\inf_{\phi\in H^1}\sup_{\eta\in H^1}\frac{B(\phi,\eta)}{\|\phi\|_{1,\Omega}\|\eta\|_{1,\Omega}} = \sup_{\phi\in H^1}\inf_{\eta\in H^1}\frac{B(\phi,\eta)}{\|\phi\|_{1,\Omega}\|\eta\|_{1,\Omega}} \ge \mu > 0. \tag{2.3}$$

For simplicity, we assume that μ and ν are such that the standard Galerkin finite element approximation is an appropriate discretization. Let $\mathcal{V}_h \subset H^1(\Omega)$ be the space of continuous piecewise linear polynomials associated with the triangulation $\mathcal{T}_h$, and consider the approximate problem: Find $u_h \in \mathcal{V}_h$ such that

$$B(u_h, v_h) = f(v_h) \tag{2.4}$$

for all $v_h \in \mathcal{V}_h$. To insure a unique solution for (2.4), we assume the inf-sup conditions

$$\inf_{\phi\in \mathcal{V}_h}\sup_{\eta\in \mathcal{V}_h}\frac{B(\phi,\eta)}{\|\phi\|_{1,\Omega}\|\eta\|_{1,\Omega}} = \sup_{\phi\in \mathcal{V}_h}\inf_{\eta\in \mathcal{V}_h}\frac{B(\phi,\eta)}{\|\phi\|_{1,\Omega}\|\eta\|_{1,\Omega}} \ge \mu > 0. \tag{2.5}$$

Xu and Zikatanov [21] have shown that, under these assumptions,

$$\|u-u_h\|_{1,\Omega} \le \frac{\nu}{\mu}\inf_{v_h\in\mathcal{V}_h}\|u-v_h\|_{1,\Omega}.$$

See also Babuška and Aziz [2]. In this situation, we have standard a priori estimates of the form

$$\|u-u_h\|_{\alpha,\Omega} \lesssim h^{2-\alpha}\|u\|_{2,\Omega}$$

for $0 \le \alpha \le 1$.

We define the piecewise constant matrix function $\mathcal{D}_\tau$ in terms of the diffusion matrix $\mathcal{D}$ as follows:

$$\mathcal{D}_{\tau ij} = \frac{1}{|\tau|}\int_\tau \mathcal{D}_{ij}\,dx.$$

Note that $\mathcal{D}_\tau$ is symmetric and positive definite. The following results are proved in [10].

THEOREM 2.1. *Let the triangulation $\mathcal{T}_h$ be $O(h^{2\sigma})$ irregular* [10]. *Assume that $\mathcal{D}_\tau$ defined above satisfies*

$$|\mathcal{D}_{\tau ij}| \lesssim 1,$$
$$|\mathcal{D}_{\tau ij} - \mathcal{D}_{\tau' ij}| \lesssim h$$

for $i=1,2$, $j=1,2$. Here τ and τ' are a pair of triangles sharing a common edge. Assume that the solution of (2.1) *satisfies $u \in W^{3,\infty}(\Omega)$ and that $u_h \in \mathcal{V}_h$ is the solution of* (2.4). *Then*

$$\|\nabla u_h - \nabla u_I\|_{0,\Omega} \lesssim h^{1+\min(1,\sigma)}|\log h|^{1/2}\|u\|_{3,\infty,\Omega},$$
$$\|\nabla u - Q_h\nabla u_I\|_{0,\Omega} \lesssim h^{1+\min(1,\sigma)}|\log h|^{1/2}\|u\|_{3,\infty,\Omega},$$
$$\|\nabla u - Q_h\nabla u_h\|_{0,\Omega} \lesssim h^{1+\min(1,\sigma)}|\log h|^{1/2}\|u\|_{3,\infty,\Omega}.$$

When the mesh is not $O(h^{2\sigma})$ irregular or when σ becomes very close to zero, the superconvergence demonstrated in Theorem 2.1 will be diminished. Intuitively, it appears that superconvergence of $Q_h \nabla u_h$ is diminished mainly because of high frequency errors introduced by the small nonuniformities of the mesh. Preferentially attenuating high frequency errors in mesh functions is of course a widely studied problem in multilevel iterative methods. Our proposal here is to apply these ideas in the present context. In particular, we construct a multigrid smoother S and take $S^m Q_h \nabla u_h$ as our recovered gradient. As with multigrid methods, we expect that a very small number of smoothing steps will suffice; in our code, we take $m = 2$ as default.

Our postprocessing gradient recovery scheme is based on the following bilinear form:

$$a(u, v) = (\nabla u, \nabla v) + (u, v). \tag{2.6}$$

We introduce the discrete operator $A : v_h \mapsto \mathcal{V}_h$ defined by

$$(Au_h, v_h) = a(u_h, v_h) \quad \forall u_h, v_h \in \mathcal{V}_h.$$

We note that A is symmetric positive definite on $\mathcal{V}_h$ and

$$\lambda \equiv \rho(A) \eqsim h^{-2}. \tag{2.7}$$

Using A, we introduce the smoothing operator S defined by

$$S = I - \lambda^{-1} A.$$

The usual multigrid convergence function

$$f(\alpha, \beta) = \frac{\alpha^\alpha \beta^\beta}{(\alpha + \beta)^{(\alpha+\beta)}},$$

$\alpha, \beta > 0$, plays an important role. Here we summarize some standard properties of $f(\alpha, \beta)$. Let $p, \alpha, \beta > 0$. Then

$$\begin{aligned} \sup_{x \in [0,1]} x^\alpha (1 - x)^\beta &= f(\alpha, \beta), \\ f(\alpha, \beta)^p &= f(p\alpha, p\beta), \\ f(\alpha, \beta) &= f(\beta, \alpha). \end{aligned}$$

For convenience in notation, we let $\partial_i u$ denote the partial derivative $\partial u / \partial x_i$. We now state and prove some preliminary lemmas leading up to the main Theorem 2.7 in this section.

LEMMA 2.2. *For any $z \in \mathcal{V}_h$,*

$$\|(I - S^m) z\|_{0,\Omega} \lesssim mh \big(\|z - \partial_i u\|_{1,\Omega} + h \|u\|_{3,\Omega} + h^{1/2} |u|_{2,\infty,\partial\Omega} \big).$$

Proof. We note, from the definition of S,

$$\begin{aligned} \|(I - S^m) z\| &= \lambda^{-1} \|(I - S^m)(I - S)^{-1} A z\| \\ &\le \lambda^{-1} \max_{s \in [0,1]} [(1 - s^m)(1 - s)^{-1}] \|Az\| \\ &\le \lambda^{-1} m \|Az\| \\ &\lesssim mh^2 \|Az\|. \end{aligned}$$

Let $w = Az$. By definition,

$$(w, \phi) = (\nabla z, \nabla \phi) + (z, \phi) \tag{2.8}$$

for all $\phi \in \mathcal{V}_h$. We take $\phi = w$ in (2.8) and estimate the terms on the right-hand side. The critical term is $(\nabla z, \nabla w)$, where we have

$$\begin{aligned}(\nabla z, \nabla w) &= (\nabla(z - \partial_i u), \nabla w) + (\nabla \partial_i u, \nabla w)\\ &\lesssim \|\nabla(z - \partial_i u)\| \|\nabla w\| - (\Delta \partial_i u, w) + \int_{\partial\Omega} \nabla \partial_i u \cdot \boldsymbol{n}\, w\, ds\\ &\lesssim (h^{-1}\|z - \partial_i u\|_{1,\Omega} + \|u\|_{3,\Omega})\|w\|_{0,\Omega} + |u|_{2,\infty,\partial\Omega} \int_{\partial\Omega} |w|\, ds\\ &\lesssim \left(h^{-1}\|z - \partial_i u\|_{1,\Omega} + \|u\|_{3,\Omega} + h^{-1/2}|u|_{2,\infty,\partial\Omega}\right)\|w\|_{0,\Omega}.\end{aligned}$$

Also

$$(z, w) = (z - \partial_i u, w) + (\partial_i u, w) \lesssim (h^{-1}\|z - \partial_i u\|_{1,\Omega} + \|u\|_{3,\Omega})\|w\|_{0,\Omega}.$$

Thus for $z \in \mathcal{V}_h$,

$$\|Az\| \lesssim h^{-1}\|z - \partial_i u\|_{1,\Omega} + \|u\|_{3,\Omega} + h^{-1/2}|u|_{2,\infty,\partial\Omega},$$

completing the proof. □

LEMMA 2.3. *Suppose that for $v \in \mathcal{V}_h$ and some $0 < \alpha \le 1$ we have*

$$\begin{aligned}\|v\| &\le \omega(h, v),\\ \|v\|_{-\alpha} \equiv \|A^{-\alpha/2}v\| &\le (\mathcal{C}h)^{\alpha}\omega(h, v).\end{aligned}$$

Then

$$\|S^m v\| \le \varepsilon_m\, \omega(h, v),$$

where

$$\varepsilon_m = \begin{cases} \kappa^{\alpha/2} f(m, \alpha/2) \lesssim m^{-\alpha/2} & \text{for } m > (\kappa - 1)\alpha/2,\\ [(\kappa - 1)/\kappa]^m & \text{for } m \le (\kappa - 1)\alpha/2,\end{cases}$$

and $\kappa = (\mathcal{C}h)^2\lambda$.

Proof. Let $0 \le \beta \le \alpha$. Then from the Hölder inequality

$$\|v\|_{-\beta} \le \|v\|_{-\alpha}^{\beta/\alpha}\|v\|^{1-\beta/\alpha}$$

and the hypotheses of the lemma, it follows that

$$\|v\|_{-\beta} \le (\mathcal{C}h)^{\beta}\omega(h, v)$$

for $0 \le \beta \le \alpha$.

Now,

$$\begin{aligned}\|S^m v\| &= \lambda^{\beta/2}\|S^m(I - S)^{\beta/2}A^{-\beta/2}v\|\\ &\le \lambda^{\beta/2} \max_{s\in[0,1]}[s^m(1 - s)^{\beta/2}]\|A^{-\beta/2}v\|\\ &\le \lambda^{\beta/2} f(m, \beta/2)(\mathcal{C}h)^{\beta}\omega(h, v)\\ &\le \kappa^{\beta/2} f(m, \beta/2)\omega(h, v),\end{aligned}$$

where $\kappa = (\mathcal{C}h)^2\lambda$. We now minimize this bound with respect to β on the interval $0 \le \beta \le \alpha$.

$$\begin{aligned}\frac{\partial \kappa^{\beta/2} f(m,\beta/2)}{\partial\beta} &= \frac{1}{2}\log\left\{\frac{(\kappa\beta)}{(2m+\beta)}\right\}\cdot\kappa^{\beta/2} f\left(m,\frac{\beta}{2}\right) = 0\\ &\Leftrightarrow \frac{\kappa\beta}{(2m+\beta)} = 1\\ &\Rightarrow \beta = \frac{2m}{(\kappa-1)}.\end{aligned}$$

There are two cases: The first is when $2m/(\kappa-1) > \alpha$. Here the minimum occurs at $\beta = \alpha$. Hence, for $m > (\kappa-1)\alpha/2$,

$$\varepsilon_m = \kappa^{\alpha/2} f(m,\alpha/2).$$

The second case is when $2m/(\kappa-1) \le \alpha$. Here $\beta = 2m/(\kappa-1)$ and

$$\varepsilon_m = \left(\frac{\kappa-1}{\kappa}\right)^m. \qquad \square$$

LEMMA 2.4. *Let $w \in H^1(\Omega)$. Then, for $1/2 < \alpha \le 1$,*

$$\|S^m Q_h \partial_i w\|_{0,\Omega} \lesssim \varepsilon_m\left(h^{-1}\|w\|_{0,\Omega} + \|w\|_{1,\Omega} + h^{-\alpha}\|w\|_{0,\infty,\partial\Omega}\right),$$

with ε_m defined as in Lemma 2.3.

Proof. Our plan is to apply Lemma 2.3 to $v = Q_h\partial_i w$. Note that

$$\|v\|_{-\alpha} = \|Q_h\partial_i w\|_{-\alpha} = \sup_{\phi\in\mathcal{V}_h}\frac{(Q_h\partial_i w,\phi)}{\|\phi\|_\alpha} = \sup_{\phi\in\mathcal{V}_h}\frac{(\partial_i w,\phi)}{\|\phi\|_\alpha}.$$

Using integration by parts,

$$\begin{aligned}(\partial_i w,\phi) &= -(w,\partial_i\phi) + \int_{\partial\Omega} w\phi n_i\, ds\\ &\lesssim \|w\|_{0,\Omega}\|\phi\|_{1,\Omega} + \|w\|_{0,\infty,\partial\Omega}\int_{\partial\Omega}|\phi|\, ds\\ &\lesssim h^{\alpha-1}\|w\|_{0,\Omega}\|\phi\|_{\alpha,\Omega} + \|w\|_{0,\infty,\partial\Omega}\|\phi\|_{\alpha,\Omega}\\ &\lesssim (h^{\alpha-1}\|w\|_{0,\Omega} + \|w\|_{0,\infty,\partial\Omega})\|\phi\|_{\alpha,\Omega}.\end{aligned}$$

Thus

$$\|v\|_{-\alpha,\Omega} \lesssim h^\alpha\omega(h,v)$$

with

$$\omega(h,v) = h^{-1}\|w\|_{0,\Omega} + \|w\|_{1,\Omega} + h^{-\alpha}\|w\|_{0,\infty,\partial\Omega}.$$

Since

$$\|v\|_{0,\Omega} = \|Q_h\partial_i w\|_{0,\Omega} \le \omega(h,v),$$

the desired estimate now follows from Lemma 2.3. $\square$

Lemma 2.5. *Let $u \in H^3(\Omega) \cap W^{2,\infty}(\Omega)$. Then for any $v_h \in \mathcal{V}_h$ and $1/2 < \alpha \leq 1$ we have*

$$\|\nabla u - S^m Q_h \nabla v_h\|_{0,\Omega} \lesssim m h^{3/2}\big(h^{1/2}\|u\|_{3,\Omega} + |u|_{2,\infty,\partial\Omega}\big) + \varepsilon_m \left(h^{-1}\|u - v_h\|_{0,\Omega} + \|u - v_h\|_{1,\Omega} + h^{-\alpha}\|u - v_h\|_{0,\infty,\partial\Omega}\right),$$

with ε_m defined as in Lemma 2.3.

Proof. By the triangle inequality,

$$\|\partial_i u - S^m Q_h \partial_i v_h\|_{0,\Omega} \leq \|(I - Q_h)\partial_i u\|_{0,\Omega} + \|(I - S^m)Q_h \partial_i u\|_{0,\Omega} + \|S^m Q_h \partial_i (u - v_h)\|_{0,\Omega}.$$

We now estimate these three terms. The first term is easy; by standard arguments,

$$\|(I - Q_h)\partial_i u\|_{0,\Omega} \lesssim h^2 \|u\|_{3,\Omega}.$$

The second term is estimated by Lemma 2.2 with $z = Q_h \partial_i u$. For the third, we apply Lemma 2.4 with $w = u - v_h$. □

In the case in which $v_h = u_h \in \mathcal{V}_h \cap H_0^1(\Omega)$ is the finite element approximation to $u \in H_0^1(\Omega)$, the boundary terms vanish and

$$\|\nabla u - S^m Q_h \nabla v_h\|_{0,\Omega} \lesssim h(mh + \varepsilon_m)\|u\|_{3,\Omega}.$$

In the more general case, if $v_h = u_h \in \mathcal{V}_h$ and $1/2 < \alpha < 1$, we use the well-known L^∞ norm estimate for the linear finite element approximation to obtain

$$h^{-\alpha}\|u - u_h\|_{0,\infty,\partial\Omega} \lesssim h^{1-\alpha}|\log h| h |u|_{2,\infty,\Omega} \lesssim h|u|_{2,\infty,\Omega}$$

and, hence,

$$\|\nabla u - S^m Q_h \nabla v_h\|_{0,\Omega} \lesssim h(mh^{1/2} + \varepsilon_m)(\|u\|_{3,\Omega} + |u|_{2,\infty,\Omega}).$$

Similar estimates hold for the case $v = u_I$. We now turn to the main theorems in this section. The next theorem is based only on the results developed in this section and summarizes the above discussion.

Theorem 2.6. *Let $u \in H^3(\Omega) \cap W^{2,\infty}(\Omega)$ and $u_h \in \mathcal{V}_h$ be an approximation of u satisfying*

$$\|u - u_h\|_{k,\Omega} \lesssim h^{2-k}|u|_{2,\Omega}, \quad k = 0, 1,$$
$$\|u - u_h\|_{0,\infty\Omega} \lesssim h^2 |\log h||u|_{2,\infty\Omega}.$$

Then

$$\|\nabla u - S^m Q_h \nabla u_h\|_{0,\Omega} \lesssim h(mh^{1/2} + \varepsilon_m)\left(\|u\|_{3,\Omega} + |u|_{2,\infty,\Omega}\right),$$

where ε_m is defined as in Lemma 2.3 *and $1/2 < \alpha < 1$.*

The following theorem combines results from this section with our earlier superconvergence results.

Theorem 2.7. *Let $u \in W^{3,\infty}(\Omega)$, and assume the hypotheses of Theorem* 2.1. *Then*

$$\|\nabla u - S^m Q_h \nabla u_I\|_{0,\Omega} \lesssim h\big(\min(h^{\min(1,\sigma)}|\log h|, \varepsilon_m) + m\, h^{1/2}\big)\|u\|_{3,\infty,\Omega}, \tag{2.9}$$

$$\|\nabla u - S^m Q_h \nabla u_h\|_{0,\Omega} \lesssim h\big(\min(h^{\min(1,\sigma)}|\log h|, \varepsilon_m) + m\, h^{1/2}\big)\|u\|_{3,\infty,\Omega}, \tag{2.10}$$

where ε_m is defined as in Lemma 2.3 *and $1/2 < \alpha < 1$.*

Proof. Our proof combines Lemma 2.5 and Theorem 2.1. We first use the triangle inequality

$$\|\partial_i u - S^m Q_h \partial_i u_I\|_{0,\Omega} \le \|(I - Q_h)\partial_i u\|_{0,\Omega} + \|(I - S^m)Q_h \partial_i u\|_{0,\Omega} + \|S^m Q_h \partial_i (u - u_I)\|_{0,\Omega}.$$

The first two terms are estimated as in Lemma 2.5. For the third term, we can first use Theorem 2.1 as

$$\begin{aligned}\|S^m Q_h \partial_i (u - u_I)\|_{0,\Omega} &\lesssim \|Q_h \partial_i (u - u_I)\|_{0,\Omega} \\ &\lesssim \|\partial_i u - Q_h \partial_i u_I\|_{0,\Omega} + \|(I - Q_h)\partial_i u\|_{0,\Omega} \\ &\lesssim h^{1+\min(1,\sigma)} |\log h|^{1/2} \|u\|_{3,\infty,\Omega} + h^2 \|u\|_{3,\Omega}.\end{aligned}$$

The third term can also be estimated as in Lemma 2.5. Taken together, these estimates establish (2.9). The proof of (2.10) is identical. □

We conclude with a few implementation details. First, with respect to the selection of the critical parameter m: Balancing the terms

$$\left(\frac{\kappa - 1}{\kappa}\right)^m \approx m h^{-1/2}$$

suggests that m should grow in a logarithmic-like fashion as the mesh is refined. On the other hand, in our empirical investigations, we have found that taking $m \le 2$ has been adequate for scalar PDE equations involving $O(10^5)$ unknowns, which suggests that a simple fixed strategy is good enough for most purposes.

Second, with respect to the L^2 projection: This linear system is solved approximately by an iterative method, in our case, a symmetric Gauss–Seidel method with conjugate gradient acceleration (SGSCG). The mass matrix is assembled in the standard nodal basis and is sparse and diagonally dominant, so convergence is very rapid; typically 4–6 iterations are sufficient. In the context of an adaptive refinement feedback loop, the initial guess is taken as zero for the first (coarsest) mesh and interpolated from the previous mesh at all subsequent refinement steps. The overall complexity of this step is thus $O(N)$ for a mesh with N vertices. If necessary, this step could be made more efficient (in terms of the size of the constant, not the order of complexity) by using some standard mass lumping scheme to construct a diagonal approximation to the mass matrix. This would also make the calculation *local* rather than global.

Third, with respect to the smoothing steps: We do not compute the constant λ exactly. In fact, we use a Jacobi conjugate gradient (JCG) iteration. The stiffness matrix A corresponding to the operator $-\Delta$ is assembled in the nodal basis; this matrix is symmetric positive semidefinite with a one dimensional kernel corresponding the constant function. (We used the complete H^1 inner product in our analysis to avoid the technical complications introduced by a nontrivial kernel.) Then m JCG steps are applied to the linear system $Ax = 0$, with initial guess corresponding to the finite element function $Q_h \partial_i u_h$. Our default choice is $m = 2$ iterations; thus this step also has complexity $O(N)$.

3. An a posteriori error estimator. In this section we use the recovered gradient to develop an a posteriori error estimator. The obvious choice for a global a posteriori error estimator is to approximate $\|\nabla(u - u_h)\|_{0,\Omega}$ by $\|(I - S^m Q_h)\nabla u_h\|_{0,\Omega}$. In Theorem 3.1, we show that this is indeed a good approximation.

THEOREM 3.1. *Assume the hypotheses of Theorem* 2.7. *Then*

$$\|\nabla(u-u_h)\|_{0,\Omega} \le \|(I-S^mQ_h)\nabla u_h\|_{0,\Omega} + Ch\big(\min(h^{\min(1,\sigma)}|\log h|, \varepsilon_m) + m\,h^{1/2}\big)\|u\|_{3,\infty,\Omega}, \tag{3.1}$$

$$\|(I-S^mQ_h)\nabla u_h\|_{0,\Omega} \le \|\nabla(u-u_h)\|_{0,\Omega} + Ch\big(\min(h^{\min(1,\sigma)}|\log h|, \varepsilon_m) + m\,h^{1/2}\big)\|u\|_{3,\infty,\Omega}, \tag{3.2}$$

where ε_m *is defined as in Lemma* 2.3 *for* $1/2<\alpha<1$. *Furthermore, if there exists a positive constant* $c_0(u)$ *independent of* h *such that*

$$\|\nabla(u-u_h)\|_{0,\Omega} \ge c_0(u)h, \tag{3.3}$$

then

$$\left|\frac{\|(I-S^mQ_h)\nabla u_h\|_{0,\Omega}}{\|\nabla(u-u_h)\|_{0,\Omega}} - 1\right| \lesssim \min(h^{\min(1,\sigma)}|\log h|, \varepsilon_m) + m\,h^{1/2}. \tag{3.4}$$

Proof. The proof of (3.1)–(3.2) is just a simple application of the triangle inequalities

$$\|\nabla(u-u_h)\|_{0,\Omega} \le \|(I-S^mQ_h)\nabla u_h\|_{0,\Omega} + \|\nabla u - S^mQ_h\nabla u_h\|_{0,\Omega},$$
$$\|(I-S^mQ_h)\nabla u_h\|_{0,\Omega} \le \|\nabla(u-u_h)\|_{0,\Omega} + \|\nabla u - S^mQ_h\nabla u_h\|_{0,\Omega}$$

and Theorem 2.7. Estimate (3.4) follows from (3.1)–(3.2) and from the assumption (3.3). □

Taken together, (3.1)–(3.2) show that if the true error is first order, $\|\nabla(u-u_h)\|_{0,\Omega} = O(h)$, then the a posteriori error estimate $\|(I-S^mQ_h)\nabla u_h\|_{0,\Omega}$ will also be $O(h)$. In particular, given a superconvergent approximation to ∇u, we can expect the effectivity ratio $\|(I-S^mQ_h)\nabla u_h\|_{0,\Omega}/\|\nabla(u-u_h)\|_{0,\Omega}$ to be close to unity. Furthermore, in this case Theorem 3.1 shows that the a posteriori error estimate will be asymptotically exact.

In terms of local error indicators, an obvious choice would be to estimate the local error in a given element τ by $\|(I-S^mQ_h)\nabla u_h\|_{0,\tau}$. For practical reasons discussed below, we prefer an alternative approach where the recovered gradient is used to approximate the Hessian matrix of second derivatives of u. By way of motivation, we note that for an $O(h^{2\sigma})$ irregular mesh

$$\begin{aligned}\|\nabla(u-u_h)\|_{0,\Omega} &\le \|\nabla(u-u_q)\|_{0,\Omega} + \|\nabla(u_q-u_I)\|_{0,\Omega} + \|\nabla(u_I-u_h)\|_{0,\Omega}\\ &\le C(u)h^{1+\min(1,\sigma)}|\log h| + \|\nabla(u_q-u_I)\|_{0,\Omega},\\ \|\nabla(u_q-u_I)\|_{0,\Omega} &\le \|\nabla(u_q-u)\|_{0,\Omega} + \|\nabla(u-u_h)\|_{0,\Omega} + \|\nabla(u_h-u_I)\|_{0,\Omega}\\ &\le C(u)h^{1+\min(1,\sigma)}|\log h| + \|\nabla(u-u_h)\|_{0,\Omega}.\end{aligned}$$

From this pair of estimates, it follows that $\|\nabla(u_q-u_I)\|_{0,\Omega} = O(h)$ if and only if $\|\nabla(u-u_h)\|_{0,\Omega} = O(h)$, and that in this case $\|\nabla(u_q-u_I)\|_{0,\Omega}$ is asymptotically exact.

The function u_q-u_I is a locally defined, piecewise quadratic polynomial with value zero at all vertices of the mesh. Let a canonical element $\tau\in\mathcal{T}_h$ have vertices $\boldsymbol{p}_k^t=(x_k,y_k)$, $1\le k\le 3$, oriented counterclockwise, and corresponding nodal basis functions (barycentric coordinates) $\{\psi_k\}_{k=1}^3$. Let $\{e_k\}_{k=1}^3$ denote the edges of element

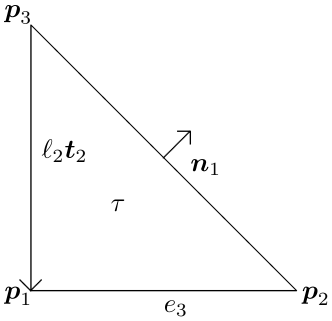


FIG. 3.1. *Parameters associated with the triangle τ.*

τ, $\{\boldsymbol{n}_k\}_{k=1}^3$ the unit outward normal vectors, $\{\boldsymbol{t}_k\}_{k=1}^3$ the unit tangent vectors with counterclockwise orientation, and $\{\ell_k\}_{k=1}^3$ the edge lengths (see Figure 3.1). Let $q_k = \psi_{k+1}\psi_{k-1}$ denote the quadratic bump function associated with edge k of τ, where $(k-1,k,k+1)$ is a cyclic permutation of $(1,2,3)$. Thus in element τ, $u_q - u_I$ is a linear combination of the quadratic bump functions associated with the edge midpoints of the element,

$$u_q - u_I = \sum_{k=1}^{3} \ell_k^2 \boldsymbol{t}_k^t M_\tau \boldsymbol{t}_k \, q_k(x,y),$$

where

$$M_\tau = -\frac{1}{2}\begin{pmatrix} \partial_{11}u_q & \partial_{12}u_q \\ \partial_{21}u_q & \partial_{22}u_q \end{pmatrix}.$$

In our local error indicator, we simply approximate the second derivatives in the Hessian matrix M_τ using gradients of $S^m Q_h \partial_i u_h$. In particular, let

$$\tilde{M}_\tau = -\frac{1}{2}\begin{pmatrix} \partial_1 S^m Q_h \partial_1 u_h & \partial_1 S^m Q_h \partial_2 u_h \\ \partial_2 S^m Q_h \partial_1 u_h & \partial_2 S^m Q_h \partial_2 u_h \end{pmatrix},$$
$$\bar{M}_\tau = \frac{\alpha_\tau}{2}(\tilde{M}_\tau + \tilde{M}_\tau^t),$$

where $\alpha_\tau > 0$ is a constant described below. For the case of meshes that are $O(h^{2\sigma})$ irregular, we can have $m = 0$, but for general shape regular meshes, we have $m > 0$. In either case, the local error estimate ϵ_τ is given by

$$\epsilon_\tau = \sum_{k=1}^{3} \ell_k^2 \boldsymbol{t}_k^t \bar{M}_\tau \boldsymbol{t}_k \, q_k(x,y). \tag{3.5}$$

The normalization constant α_τ is chosen such that the local error indicator η_τ satisfies

$$\eta_\tau \equiv \|\nabla \epsilon_\tau\|_{0,\tau} = \|(I - S^m Q_h)\nabla u_h\|_{0,\tau}.$$

Normally we expect that $\alpha_\tau \approx 1$, which is likely to be the case in regions where the Hessian matrix for the true solution is well defined. Near singularities, u is not smooth, and we anticipate difficulties in estimating the Hessian. For elements near such singularities, α_τ provides a heuristic for partly compensating for poor approximation.

The form of our a posteriori error estimate (3.5) is quite useful in practice. It explicitly shows the dependence on the shape, size, and orientation of the elements, as well as the dependence on the second derivatives of u. This leads to many interesting algorithms for adaptive mesh smoothing and topology modification (e.g., "edge flipping") [8]. For example, if $\bar{M}_\tau$ is assumed to be constant, then ϵ_τ and η_τ^2 are rational functions of vertex locations, and derivatives with respect to the vertex locations are easily computed.

Using ϵ_τ also provides a simple, robust, and elegant solution to an important practical problem for adaptive mesh refinement schemes: how to provide error estimates for the refined elements without immediately resolving the global problem. In the past, most schemes were based on crude (or not-so-crude) extrapolation ideas, using the error indicator of the parent element as a basis. In most such schemes it is difficult to take into account the details of the geometry, and they tend to become very inaccurate after only a few levels of refinement. On the other hand, with our error estimator, the children elements inherit only the Hessian matrix from the parent, and all the geometrical information is derived from the refined elements themselves. Thus it is possible to have many levels of refinement before the approximation breaks down and a new global solution is required. This has proved to be very effective in the PLTMG 8.0 package [6], which employs this scheme, but uses a different a posteriori error estimate to compute the approximate Hessian matrix [8]. In Theorem 3.2, we show that our recovered gradients can provide reasonable approximation to the Hessian.

THEOREM 3.2. *Assume the hypotheses of Theorem* 2.7. *Then*

$$\|\partial_i(\partial_k u - S^m Q_h \partial_k u_h)\|_{0,\Omega} \lesssim \left(\min(h^{\min(1,\sigma)}|\log h|, \varepsilon_m) + m\, h^{1/2}\right)\|u\|_{3,\infty,\Omega}, \tag{3.6}$$

where ε_m *is defined as in Lemma* 2.3 *for* $1/2 \leq \alpha < 1$.

Proof. Let $z = I_h \partial_k u \in \mathcal{V}_h$. Then

$$\begin{aligned}
\|\partial_i(\partial_k u - S^m Q_h \partial_k u_h)\|_{0,\Omega} &\leq \|\partial_i(\partial_k u - z)\|_{0,\Omega} + \|\partial_i(z - S^m Q_h \partial_k u_h)\|_{0,\Omega} \\
&\lesssim h\|u\|_{3,\Omega} + h^{-1}\|z - S^m Q_h \partial_k u_h\|_{0,\Omega} \\
&\lesssim h\|u\|_{3,\Omega} + h^{-1}\left(\|z - \partial_k u\|_{0,\Omega} + \|\partial_k u - S^m Q_h \partial_k u_h\|_{0,\Omega}\right) \\
&\lesssim \left(\min(h^{\min(1,\sigma)}|\log h|, \varepsilon_m) + m\, h^{1/2}\right)\|u\|_{3,\infty,\Omega}. \qquad \square
\end{aligned}$$

4. Numerical experiments.

In this section, we present some numerical illustrations of our recovery scheme in the cases of uniform and adaptively refined (nonuniform) meshes. Our gradient recovery scheme and a posteriori error estimate were implemented in the PLTMG package [6], which was then used for our numerical experiments. The experiments were done on an SGI Octane using double precision arithmetic.

In our first example, we consider the solution of the problem

$$\begin{aligned}
-\Delta u &= f \qquad \text{in } \Omega = (0,1) \times (0,1), \\
u &= g \qquad \text{on } \partial\Omega,
\end{aligned}$$

where f and g are chosen such that $u = e^{x+y}$ is the exact solution. This is a very smooth solution that satisfies all the assumptions of our theory. Here we will compare the recovery scheme with $m = 2$ smoothing steps, for the case of uniform and adaptive meshes. We begin with a uniform 3×3 mesh consisting of eight right triangles, as shown in Figure 4.1. Elements in Figure 4.1 are shaded according to size; this allows

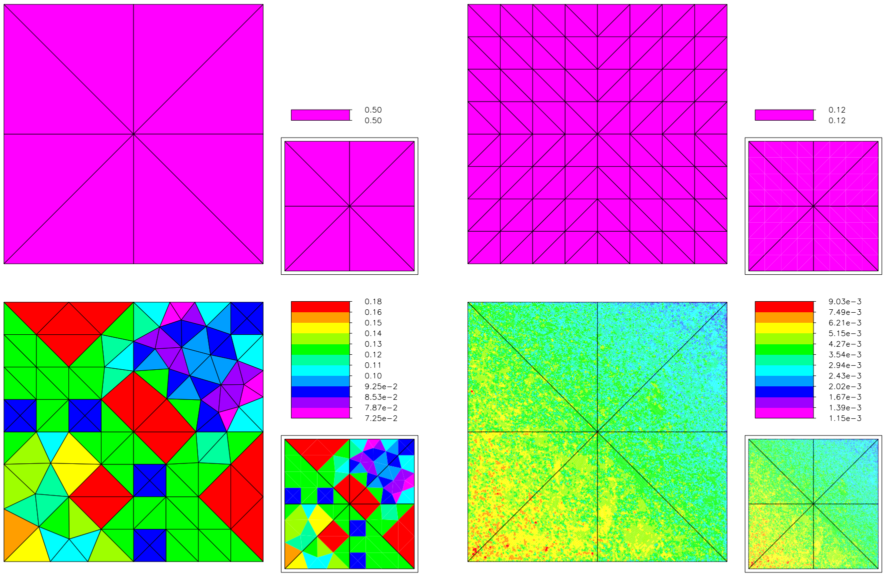


FIG. 4.1. *Top left:* 3×3 *initial mesh. Top right: uniform refinement with* $nt = 128$*. Bottom left: adaptive refinement with* $nt = 134$*. Bottom right: adaptive refinement with* $nt = 130961$*. Elements are shaded according to size.*

TABLE 4.1
Error estimates for the case $m = 2$*.*

Adaptive meshes				Uniform meshes			
nt	$L2$	$H1$	Ef	nt	$L2$	$H1$	Ef
8	1.5e-1	1.7e 0	1.68	8	1.5e-1	1.7e 0	1.68
34	4.9e-2	5.9e-1	1.36	32	3.8e-2	8.7e-1	1.74
134	1.3e-2	2.9e-1	1.54	128	9.6e-3	3.6e-1	1.50
510	2.4e-3	7.1e-2	1.17	512	2.4e-3	1.6e-1	1.41
2037	5.2e-4	2.4e-2	1.09	2048	6.0e-4	6.7e-2	1.30
8148	1.1e-4	7.1e-3	1.04	8192	1.5e-4	2.6e-2	1.20
32683	2.7e-5	2.0e-3	1.01	32768	3.8e-5	1.0e-2	1.12
130961	7.0e-6	6.2e-4	1.00	131072	9.4e-6	3.7e-3	1.07
	$L2$	$H1$	$\widetilde{H1}$		$L2$	$H1$	$\widetilde{H1}$
Order	2.06	1.76	1.04		2.03	1.42	1.01

one to obtain some impression of the structure of highly refined meshes with many elements, even if individual elements can no longer be resolved.

In Table 4.1, we record the results of the computation. We give the error as a function of the number of elements, choosing targets for the adaptive refinement procedure to produce adaptive meshes with numbers of elements similar to those for the uniform refinement case. The values are defined as follows:

$$\begin{aligned} L2 &= \|u - u_h\|_{0,\Omega}, \\ H1 &= \|\nabla u - S^m Q_h \nabla u_h\|_{0,\Omega}, \\ \widetilde{H1} &= \|\nabla(u - u_h)\|_{0,\Omega}, \\ Ef &= \frac{\|(I - S^m Q_h)\nabla u_h\|_{0,\Omega}}{\|\nabla(u - u_h)\|_{0,\Omega}}. \end{aligned}$$

Table 4.2
Order of convergence as a function of m for adaptive meshes.

	$m=0$		$m=1$		$m=2$		$m=3$		$m=10$	
K	$H1$	Ef	$H1$	Ef	$H1$	Ef	$H1$	Ef	$H1$	Ef
1	6.1e 0	0.83	7.8e-1	1.16	1.7e 0	1.68	1.7e 0	1.70	1.7e 0	1.70
2	2.3e-1	0.92	3.2e-1	1.00	5.9e-1	1.36	1.3e 0	2.34	1.8e 0	3.67
3	7.8e-2	0.94	1.1e-1	1.06	2.9e-1	1.54	3.8e-1	1.64	1.8e 0	6.75
4	3.8e-2	0.95	3.4e-2	1.02	7.1e-2	1.17	1.5e-1	1.49	1.1e 0	7.09
5	1.7e-2	0.96	1.1e-2	1.01	2.4e-2	1.09	4.8e-2	1.26	3.2e-1	4.05
6	7.2e-3	0.97	3.7e-3	1.00	7.1e-3	1.04	1.3e-2	1.09	1.2e-1	3.63
7	3.0e-3	0.98	1.4e-3	1.00	2.0e-3	1.01	3.4e-3	1.03	3.2e-2	2.27
8	1.5e-3	0.98	5.7e-4	1.00	6.2e-4	1.00	8.9e-4	1.01	7.0e-3	1.34
	$H1$	$\widetilde{H1}$	$H1$	$\widetilde{H1}$	$H1$	$\widetilde{H1}$	$H1$	$\widetilde{H1}$	$H1$	$\widetilde{H1}$
Order	1.16	1.04	1.39	1.03	1.76	1.04	1.92	1.07	2.01	1.08

For the cases of $L2$, $H1$, and $\widetilde{H1}$, we made a least squares fit of the data to a function of the form $F(N) = CN^{-p/2}$ to estimate the order of convergence p. All integrals were approximated using a 12-point order 7 quadrature formula applied to each triangle.

What is most striking is the similarity in the data. $L2$ is approximately the same for both uniform and adaptive refinement, while $H1$ is slightly better in the adaptive case. This is consistent with our strategy, which adaptively refines with respect to $\|\nabla\epsilon_\tau\|_{0,\tau}$. Nonetheless, both cases exhibit some superconvergence for the recovered gradients. This is further supported by noting that the effectivity ratios Ef suggest asymptotic exactness of the a posteriori error estimates.

In Table 4.2, we show the effect of varying the number of smoothing steps. To reduce the amount of data, we report only the case of adaptive meshes. Since the a posteriori error estimates are used to create the meshes, the meshes differ for each value of m but at level K have $nt \approx 2^{2K+1}$ elements. For $m = 0$, we note only slight superconvergence; thus although the meshes are shape regular and quasi-uniform, apparently $\sigma \approx 0$. In contrast, uniform meshes for $m = 0$ have a computed order of convergence for $H1$ of 1.52, essentially that predicted by our theory. However, the data show that the situation for adaptive meshes improves dramatically for $m = 1, 2$. For $m = 10$, one can see the effects of "too many" smoothings; Ef becomes more erratic, and $H1$ *increases* for some of the coarser refinement steps. But even in this case, for more refined meshes (e.g., $K = 8$), Ef again appears to be converging towards 1. This is likely due to the well-known (and in this case extremely useful) effect of the smoothing iteration "slowing down" quickly as h becomes smaller.

In Table 4.3, we explore the effect of "lumping" the mass matrix in the L^2 projection step. In particular, the mass matrix was replaced by a diagonal matrix, with diagonal entries given by the sum of all nonzero entries of the corresponding row of the mass matrix. In Table 4.3, we see results that are quite comparable to those of Table 4.1, although the gradient errors are generally slightly larger. Nonetheless, these results suggest that our gradient recovery algorithm could be modified to use only local calculations without much loss in effectiveness.

In our second example, we consider the nonlinear problem

$$\begin{aligned} -\nabla\cdot(a\nabla u) + e^u = f \qquad & \text{in } \Omega = (0,1)\times(0,1), \\ u = 0 \qquad & \text{on } \partial\Omega, \end{aligned}$$

Table 4.3
The effect of a lumped *mass matrix.*

Adaptive meshes				Uniform meshes			
nt	$L2$	$H1$	Ef	nt	$L2$	$H1$	Ef
8	1.5e-1	1.7e 0	1.69	8	1.5e-1	1.7e 0	1.69
34	4.9e-2	8.2e-1	1.70	32	3.8e-2	1.1e 0	2.02
134	1.3e-2	3.0e-1	1.57	128	9.6e-3	5.2e-1	1.96
514	2.8e-3	8.9e-2	1.23	512	2.4e-3	2.2e-1	1.74
2036	5.5e-4	2.7e-2	1.11	2048	6.0e-4	9.3e-2	1.55
8148	1.2e-4	7.7e-3	1.04	8192	1.5e-4	3.7e-2	1.37
32676	2.8e-5	2.3e-3	1.01	32768	3.8e-5	1.4e-2	1.23
130904	6.8e-6	8.2e-4	1.01	131072	9.4e-6	5.1e-3	1.13
	$L2$	$H1$	$\widetilde{H1}$		$L2$	$H1$	$\widetilde{H1}$
Order	2.10	1.64	1.05		2.03	1.43	1.01

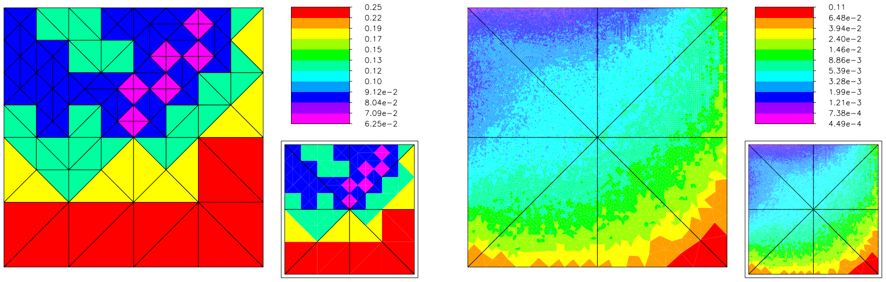


Fig. 4.2. *Left: adaptive refinement with* $nt = 138$. *Right: adaptive refinement with* $nt = 131112$. *Elements are shaded according to size.*

Table 4.4
Error estimates for the case $m = 2$.

Adaptive meshes				Uniform meshes			
nt	$L2$	$H1$	Ef	nt	$L2$	$H1$	Ef
8	1.9e-3	1.7e-2	0.30	8	1.9e-3	1.7e-2	0.30
32	1.0e-3	1.6e-2	0.84	32	1.0e-3	1.6e-2	0.84
138	2.7e-4	1.1e-2	1.46	128	3.8e-4	1.2e-2	1.42
531	2.4e-3	3.7e-3	1.67	512	1.1e-4	7.8e-3	1.89
2060	4.4e-5	1.0e-3	1.32	2048	3.0e-5	4.1e-3	2.09
8203	1.2e-5	2.6e-4	1.09	8192	7.7e-6	1.9e-3	2.01
32736	8.6e-7	7.6e-5	1.00	32768	1.9e-6	7.7e-4	1.79
131112	2.5e-7	3.0e-5	0.98	131072	4.9e-7	3.0e-4	1.53
	$L2$	$H1$	$\widetilde{H1}$		$L2$	$H1$	$\widetilde{H1}$
Order	1.83	1.58	1.05		2.01	1.30	1.02

where a is the 2×2 diagonal matrix

$$a = \begin{pmatrix} .01 & \\ & 1 \end{pmatrix}.$$

The function f is chosen such that $u = x(1-x)^3 y^5 (1-y)$ is the exact solution. We repeat the same computations as in the first example, with uniform and adaptive meshes. The uniform meshes are identical to those of the first example. Some of the adaptive meshes are shown in Figure 4.2. The numerical results are summarized in Table 4.4.

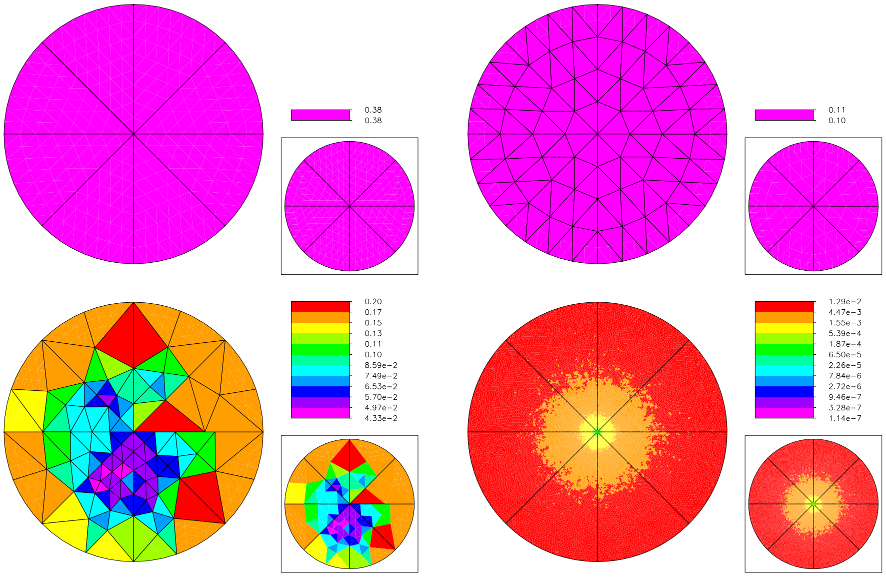


FIG. 4.3. *Top left: the initial mesh. Top right: uniform refinement with* $nt = 128$. *Bottom left: adaptive refinement with* $nt = 138$. *Bottom right: adaptive refinement with* $nt = 131105$. *Elements are shaded according to size.*

This problem is more difficult than the first in several respects. The diffusion is anisotropic, and the operator is nonlinear. The solution is smooth but generally has larger derivatives than the first example. Nonetheless, we see a similar behavior of the gradient recovery scheme and a posteriori error estimate. In this example, the adaptive meshes are more strongly graded than in the first example, suggesting that localization and equilibration of the error are more important effects for superconvergence of our gradient recovery procedure than geometric uniformity in the mesh.

In our third example, we consider the problem

$$\begin{aligned} -\Delta u &= 0 && \text{in } \Omega, \\ u &= g && \text{on } \partial\Omega_1, \\ u_n &= 0 && \text{on } \partial\Omega_2, \end{aligned}$$

where Ω is a circle of radius one centered at the origin, and with a crack along the positive x-axis, $0 \le x \le 1$. $\partial\Omega_2$ is the bottom edge of the crack, and $\partial\Omega_1 = \partial\Omega - \partial\Omega_2$. The function g is chosen such that the exact solution is $u = r^{1/4}\sin(\theta/4)$, the leading term of the singularity associated with the interior angle of 2π and change in boundary conditions at the origin. In Figure 4.3 we illustrate the initial mesh and several of the uniformly and adaptively refined meshes.

Convergence results for uniform and adaptive refinement are reported in Table 4.5. The solution u is not smooth in this case ($u \in H^{5/4-\epsilon}(\Omega)$), and this is reflected in the results. For the case of uniform refinement, the 0.25 order of convergence of the gradient coincides with the smoothness of the solution. For the adaptive meshes, the order of convergence improves and seems to be approaching order one for the gradient. This sort of behavior is typical of a reasonable adaptive refinement procedure. However, even in this case, there is no apparent superconvergence.

Table 4.5
Error estimates for the case $m = 2$.

Adaptive meshes				Uniform meshes			
nt	$L2$	$H1$	Ef	nt	$L2$	$H1$	Ef
8	3.0e-1	7.1e-1	1.32	8	3.0e-1	7.1e-1	1.32
31	1.9e-1	6.0e-1	1.14	32	1.8e-1	5.9e-1	1.18
138	8.3e-2	5.3e-1	1.18	128	1.1e-1	5.4e-1	1.16
535	3.3e-2	3.9e-1	1.26	512	7.3e-2	4.6e-1	1.14
2091	1.1e-2	2.3e-1	1.22	2048	4.9e-2	3.9e-1	1.12
8242	2.9e-3	1.2e-1	1.28	8192	3.4e-2	3.3e-1	1.11
32832	6.8e-4	4.9e-2	1.08	32768	2.3e-2	2.8e-1	1.10
131105	1.3e-4	2.2e-2	1.11	131072	1.6e-2	2.3e-1	1.10
	$L2$	$H1$	$\widetilde{H1}$		$L2$	$H1$	$\widetilde{H1}$
Order	2.23	1.15	1.10		0.54	0.25	0.27

Table 4.6
The effect of the singularity.

Adaptive meshes				Uniform meshes			
nt	$L2'$	$H1'$	Ef'	nt	$L2'$	$H1'$	Ef'
8	3.0e-1	7.1e-1	1.32	8	3.0e-1	7.1e-1	1.32
31	1.9e-1	6.0e-1	1.14	32	1.8e-1	5.9e-1	1.18
138	8.3e-2	5.3e-1	1.18	128	1.1e-1	5.4e-1	1.16
535	3.0e-2	2.4e-1	0.92	512	7.3e-2	4.6e-1	1.14
2091	1.0e-2	4.3e-2	0.78	2048	4.6e-2	1.9e-1	0.69
8242	2.7e-3	1.2e-2	0.94	8192	3.1e-2	1.1e-1	0.28
32832	6.2e-4	3.0e-3	0.99	32768	2.1e-2	7.8e-2	0.19
131105	1.2e-4	7.9e-4	1.00	131072	1.5e-2	5.3e-2	0.13
	$L2'$	$H1'$	$\widetilde{H1}'$		$L2'$	$H1'$	$\widetilde{H1}'$
Order	2.24	1.95	1.12		0.56	0.60	0.61

Let Ω' now denote the union of all triangles in Ω with a least one vertex outside a circle of $r = 0.1$; note that this definition of Ω' is mesh dependent, but eventually Ω' excludes small triangles close to the singularity. In Table 4.6, we report results for the same computations but with the error calculation restricted to Ω'. For $L2'$, the results do not change much. However, for the gradients ($H1'$), the results are quite striking. For the case of adaptive refinement, the improvement is quite dramatic, in that, away from the singularity, the gradient recovery scheme exhibits the same sort of behavior as for a smooth problem. For the case of uniform refinement, there is also some improvement in order of convergence, but the recovered gradient does not appear significantly more accurate than ∇u_h. We also note the quite different behavior of Ef' compared with Ef for the case of uniform refinement. Taken as a whole, it seems likely that the error in the uniform refinement case is not localized, and pollution effects are still dominant even on the most refined meshes.

In our fourth example, we consider a problem with discontinuous coefficients:

$$\begin{aligned} -\nabla \cdot (a \nabla u) &= f \qquad \text{in } \Omega, \\ u &= 0 \qquad \text{on } \partial\Omega, \end{aligned}$$

where Ω is once again the unit square $(0,1) \times (0,1)$. The scalar coefficient function $a(p) = 1$ for $p \in \Omega_1 \equiv (0,1/2) \times (0,1/2) \cup (1/2,1) \times (1/2,1)$, and $a(p) = 10^{-2}$ for $p = \Omega - \Omega_1$. The function f is given by $f = 8\pi^2 \sin(2\pi x) \sin(2\pi y)$, and the exact solution u is given by $u = a^{-1} \sin(2\pi x) \sin(2\pi y)$. The initial mesh is the same as that in the first two examples. In Table 4.7, we give the results for both uniform and

Table 4.7
Error estimates for the case $m = 2$.

Adaptive meshes				Uniform meshes			
nt	$L2$	$H1$	Ef	nt	$L2$	$H1$	Ef
8	3.5e 1	3.2e 2	0.00	8	3.5e 1	3.2e 2	0.00
32	1.7e 1	2.9e 2	1.97	32	1.7e 1	3.0e 2	1.07
136	2.6e 0	2.2e 2	2.72	128	5.6e 0	2.5e 2	1.96
528	7.4e-1	1.1e 2	2.81	512	1.5e 0	1.7e 2	2.76
2071	2.2e-1	5.8e 1	2.67	2048	3.8e-1	9.0e 1	3.04
8143	6.9e-2	2.8e 1	2.42	8192	9.5e-2	5.4e 1	3.64
33102	2.0e-2	1.4e 1	2.36	32768	2.4e-2	3.7e 1	4.85
130809	6.1e-3	7.3e 0	2.31	131072	6.0e-3	2.6e 1	6.72
	$L2$	$H1$	$\widetilde{H1}$		$L2$	$H1$	$\widetilde{H1}$
Order	1.78	1.00	0.94		2.04	0.58	1.02

Table 4.8
The effect of the discontinuity.

Adaptive meshes				Uniform meshes			
nt	$L2'$	$H1'$	Ef'	nt	$L2'$	$H1'$	Ef'
8	3.5e 1	3.2e 2	0.00	8	3.5e 1	3.2e 2	0.00
32	7.1e 0	3.1e 2	2.11	32	1.7e 1	3.1e 2	1.11
136	3.5e 0	1.7e 2	1.83	128	5.6e 0	2.3e 2	1.81
532	9.1e-1	7.6e 1	1.85	512	1.5e 0	1.5e 2	2.49
2071	1.6e-1	2.4e 1	1.57	2048	3.8e-1	7.2e 1	2.47
8092	4.2e-2	6.3e 0	1.20	8192	9.5e-2	2.4e 1	1.84
32486	9.4e-3	1.3e 0	1.05	32768	2.4e-2	7.3e 0	1.37
130586	2.4e-3	3.0e-1	1.01	131072	6.0e-3	2.2e 0	1.15
	$L2$	$H1$	$\widetilde{H1}$		$L2$	$H1$	$\widetilde{H1}$
Order	2.10	2.16	1.09		2.04	1.68	1.02

adaptive meshes. Here we note no superconvergence for $H1$ in either case; indeed, in the uniform mesh case, $S^m Q_h \nabla u_h$ is much less accurate than ∇u_h in terms of order of convergence. The gradient of the exact solution is discontinuous along the lines $x = 1/2$ and $y = 1/2$. Since the mesh is aligned with the discontinuity, ∇u_h is able to capture this discontinuity with no problem. Since $Q_h \nabla u_h$ is continuous, the lack of superconvergence is due to the global L^2 projection; making local L^2 projections in each of the four subregions where $a(p)$ is constant would allow the projected gradient to remain discontinuous along the interfaces. In Table 4.8, we show the results for such a calculation. We computed $S^m \hat{Q}_h \nabla u_u$, where $\hat{Q}_h$ corresponds to the four local L^2 projections. Since we used a different projection, the adaptive meshes changed slightly. Nonetheless, we see quite clearly that allowing for the discontinuity in ∇u corrects the problems. We observe superconvergence in $H1'$ for both the uniform and adaptive meshes, as well as significant improvement in the effectivity ratios Ef'. We note in passing that, away from the discontinuities, the approximation $S^m Q_h \nabla u_h$ is superconvergent. In this respect, the behavior here is similar to that in the third example.

These numerical examples show that the effectiveness of our adaptive scheme does not necessarily depend critically on either the quasi-uniformity of the mesh or on the global regularity of the solution. However, theoretically there is still much work to do to fill in the gaps. We believe that our quasi-uniformity assumption on the triangular meshes can be removed with some extra effort. Our results are very local in the sense that the domain Ω in our theorems could be any subdomain of the actual physical domain. It will be much more challenging to obtain theoretical results with more realistic assumptions on a continuous solution's regularity. However, heuristically speaking, at places where the solution is singular, the solution will have large (or infinite) W^3_∞ norm, and as a result, our error indicator will be large in those regions and hence the grid will be refined there.

Acknowledgments. The ideas for this manuscript germinated while the authors were attending the workshop on *A Posteriori Error Estimation and Adaptive Approaches in the Finite Element Method,* held at the Mathematical Sciences Research Institute, University of California, Berkeley, April 3–14, 2000. We are grateful for the opportunity to attend this workshop and for the stimulating scientific environment at MSRI. We also thank Lars B. Wahlbin for helpful discussions during the aforementioned workshop, and the reviewers for their careful reviews and suggestions for improvements.

REFERENCES

[1] M. Ainsworth and A. Craig, *A posteriori error estimators in the finite element method*, Numer. Math., 60 (1992), pp. 429–463.

[2] I. Babuška and A. K. Aziz, *Survey lectures on the mathematical foundations of the finite element method*, in The Mathematical Foundations of the Finite Element Method with Applications to Partial Differential Equations (Baltimore, 1972), Academic Press, New York, 1972, pp. 1–362.

[3] I. Babuška and W. C. Rheinboldt, *A posteriori error estimates for the finite element method*, Internat. J. Numer. Methods Engrg., 12 (1978), pp. 1597–1615.

[4] I. Babuška and T. Strouboulis, *The Finite Element Method and Its Reliability*, Numer. Math. Sci. Comput., Oxford Science Publications, New York, 2001.

[5] I. Babuška, T. Strouboulis, and C. S. Upadhyay, *η%-superconvergence of finite element approximations in the interior of general meshes of triangles*, Comput. Methods Appl. Mech. Engrg., 122 (1995), pp. 273–305.

[6] R. E. Bank, *PLTMG: A Software Package for Solving Elliptic Partial Differential Equations: Users' Guide* 8.0, Software Environ. Tools 5, SIAM, Philadelphia, 1998.

[7] R. E. Bank and C. C. Douglas, *Sharp estimates for multigrid rates of convergence with general smoothing and acceleration*, SIAM J. Numer. Anal., 22 (1985), pp. 617–633.

[8] R. E. Bank and R. K. Smith, *Mesh smoothing using a posteriori error estimates*, SIAM J. Numer. Anal., 34 (1997), pp. 979–997.

[9] R. E. Bank and A. Weiser, *Some a posteriori error estimators for elliptic partial differential equations*, Math. Comp., 44 (1985), pp. 283–301.

[10] R. E. Bank and J. Xu, *Asymptotically exact a posteriori error estimators, Part* I*: Grids with superconvergence*, SIAM J. Numerical Analysis, 41 (2003), pp. 2294–2312.

[11] C. Chen and Y. Huang, *High Accuracy Theory of Finite Element Methods*, Hunan Science Press, Hunan, China, 1995 (in Chinese).

[12] L. Du and N. Yan, *Gradient recovery type a posteriori error estimate for finite element approximation on non-uniform meshes*, Adv. Comput. Math., 14 (2001), pp. 175–193.

[13] R. Durán, M. A. Muschietti, and R. Rodríguez, *On the asymptotic exactness of error estimators for linear triangular finite elements*, Numer. Math., 59 (1991), pp. 107–127.

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

[15] B. Li and Z. Zhang, *Analysis of a class of superconvergence patch recovery techniques for linear and bilinear finite elements*, Numer. Methods Partial Differential Equations, 15 (1999), pp. 151–167.

[16] J. Li, *Convergence and superconvergence analysis of finite element methods on highly nonuniform anisotropic meshes for singularly perturbed reaction-diffusion problems*, Appl. Numer. Math., 36 (2001), pp. 129–154.

[17] Q. Lin and N. Yan, *The Construction and Analysis of High Efficiency Finite Elements*, Hebei University Press, Hunan, China, 1996 (in Chinese).

[18] R. Verfürth, *A Posteriori Error Estimation and Adaptive Mesh Refinement Techniques*, Teubner Skr. Numer., Teubner, Stuttgart, 1995.

[19] L. B. Wahlbin, *Superconvergence in Galerkin Finite Element Methods*, Springer-Verlag, Berlin, 1995.

[20] L. B. Wahlbin, *General principles of superconvergence in Galerkin finite element methods*, in Finite Element Methods (Jyväskylä, 1997), Dekker, New York, 1998, pp. 269–285.

[21] J. Xu and L. Zikatanov, *Some observations on Babuška and Brezzi theories*, Numer. Math., 94 (2003), pp. 195–202.

[22] N. Yan and A. Zhou, *Gradient recovery type a posteriori error estimates for finite element approximations on irregular meshes*, Comput. Methods Appl. Mech. Engrg., 190 (2001), pp. 4289–4299.

[23] Z. Zhang and H. D. Victory, Jr., *Mathematical analysis of Zienkiewicz-Zhu's derivative patch recovery technique*, Numer. Methods Partial Differential Equations, 12 (1996), pp. 507–524.

[24] Z. Zhang and J. Z. Zhu, *Superconvergence of the derivative patch recovery technique and a posteriori error estimation*, in Modeling, Mesh Generation, and Adaptive Numerical Methods for Partial Differential Equations (Minneapolis, MN, 1993), Springer, New York, 1995, pp. 431–450.

[25] J. Z. Zhu and O. C. Zienkiewicz, *Superconvergence recovery technique and a posteriori error estimators*, Internat. J. Numer. Methods Engrg., 30 (1990), pp. 1321–1339.

SIAM J. NUMER. ANAL.
Vol. 41, No. 6, pp. 2333–2349

CONVERGENCE ANALYSIS OF SPECTRAL COLLOCATION METHODS FOR A SINGULAR DIFFERENTIAL EQUATION[*]

WEIZHANG HUANG[†], HEPING MA[‡], AND WEIWEI SUN[§]

Abstract. Solutions of partial differential equations with coordinate singularities often have special behavior near the singularities, which forces them to be smooth. Special treatment for these coordinate singularities is necessary in spectral approximations in order to avoid degradation of accuracy and efficiency. It has been observed numerically in the past that, for a scheme to attain high accuracy, it is unnecessary to impose all the pole conditions, the constraints representing the special solution behavior near singularities. In this paper we provide a theoretical justification for this observation. Specifically, we consider an existing approach, which uses a pole condition as the boundary condition at a singularity and solves the reformulated boundary value problem with a commonly used Gauss–Lobatto collocation scheme. Spectral convergence of the Legendre and Chebyshev collocation methods is obtained for a singular differential equation arising from polar and cylindrical geometries.

Key words. coordinate singularity, convergence, spectral collocation method

AMS subject classifications. 65N35, 65N12, 65L10

DOI. 10.1137/S0036142902381024

1. Introduction.

Physical problems in polar, cylindrical, or spherical geometries often give rise to mathematical models involving singular partial differential equations (PDEs) with smooth solutions. A common feature of these PDEs is that their solutions have special behavior near coordinate singularities, which forces the solutions to be smooth. For the spectral solution of this type of equation, special treatment for the coordinate singularities is needed, since a traditional spectral scheme either does not fully capture the special solution behavior or is ill-suited to fast transform techniques; e.g., see [6, 7, 11].

A number of spectral approaches have been developed in the past in attempts to capture the solution behavior near coordinate singularities. They include those expanding the solution in specially designed basis functions, such as spherical harmonics, parity-modified Fourier series, modified Robert functions, and eigenfunctions of singular Sturm–Liouville problems [6, 9, 11, 20, 21, 22, 27]; approaches using inherent symmetries of the solution [9, 10]; and methods using pole conditions (i.e., compatibility conditions at the center of polar coordinates) as boundary conditions in the collocation context [12] and the Galerkin context [24, 25]. Many of these approaches have been successfully applied to steady state and time dependent problems including the Navier–Stokes equations; e.g., see [13, 14, 20, 23, 26, 28].

[*] Received by the editors August 12, 2002; accepted for publication (in revised form) June 11, 2003; published electronically December 17, 2003. This work was supported in part by the NSF (U.S.A.) through grant DMS–0074240, the Research Grants Council of the Hong Kong Special Administrative Region, China, through grant CityU-1141/01P, and the Special Funds for State Major Basic Research Projects of China through grant G1999032804.

http://www.siam.org/journals/sinum/41-6/38102.html

[†] Department of Mathematics, University of Kansas, Lawrence, KS 66045 (huang@math.ukans.edu).

[‡] Department of Mathematics, Shanghai University, Shanghai 200436, China (hpma@mail.shu.edu.cn).

[§] Department of Mathematics, City University of Hong Kong, Kowloon, Hong Kong, China (maweiw@math.cityu.edu.hk).

The solution behavior of PDEs near coordinate singularities can be described by an infinite number of pole conditions derived either from the underlying differential equation, by assuming some kind of smoothness of the solution, or more generally from the analyticity of the solution at singularities. Most of the existing methods are developed more or less to accommodate this behavior. However, it is observed numerically first by Orszag [21] and then by many other researchers (e.g., [6]) that it is unnecessary to impose all of the pole conditions in order for a numerical scheme to attain high accuracy. In fact, Huang and Sloan [12], using some of these pole conditions as the boundary conditions at the coordinate singularities, show numerically that the spectral collocation approximation of the Helmholtz equation on the unit disk has a spectral convergence rate.

Three Legendre-type pseudospectral schemes and their convergence analysis have been developed for axisymmetric domains by Bernardi, Dauge, and Maday in their recent book [1]. The basic idea behind these schemes is to incorporate the natural measure rdr of the coordinate singularity into the quadrature formula defining the spectral approximation. In the radial direction the formula reads as

$$\int_0^1 v(r)rdr = \sum_{j=0}^{N} v(r_j)\omega_j \qquad \forall v \in \mathbf{P}, \tag{1.1}$$

where $\mathbf{P}$ is a polynomial space, $r_j = (1+\rho_j)/2$, and the ω_j's are the corresponding weights. Three sets of ρ_j's and $\mathbf{P}$ are chosen, one for each scheme:

Method A (Gauss–Radau): $\mathbf{P} = \mathbf{P}_{2N}$, $\rho_N = 1$,
ρ_j $(0 \le j \le N-1)$ are the roots of $L'_{N+1}(\rho)$;
Method B (Gauss–Lobatto): $\mathbf{P} = \mathbf{P}_{2N-1}$, $\rho_0 = -1$, $\rho_N = 1$,
ρ_j $(1 \le j \le N-1)$ are the roots of $\left(\frac{L_N(\rho)+L_{N+1}(\rho)}{1+\rho}\right)'$;
Method C (Gauss–Radau): $\mathbf{P} = \mathbf{P}_{2N-1}$, $\rho_N = 1$,
ρ_j $(0 \le j \le N-1)$ are the roots of $L_{N+1}(\rho) - L_N(\rho)$,

where L_N is the Legendre polynomial of degree N. The pseudospectral approximations are then defined through the boundary condition(s) and the Galerkin formulation of the underlying problem in the discrete inner product $((u,v))_N \equiv \sum_{j=0}^{N} u(r_j)v(r_j)\omega_j$ induced from the quadrature formula (1.1). It is noted that the nodes used in these schemes are different from those in a traditional (unweighted) spectral collocation method. Moreover, among these three schemes, only Method C is equivalent to a collocation system. Furthermore, the authors of the book suggest that two boundary conditions $u(0) = 0$ (which is a pole condition, cf. (2.6)) and $u(1) = g$ be used for a reduced equation (see (2.1)–(2.2)) with $n \neq 0$. Thus, only Method B, which uses the Gauss–Lobatto nodes but cannot be interpreted as a collocation scheme, can be applied to the case $n \neq 0$.

The objective of this paper is to provide a theoretical justification for the method developed in [12], which uses a pole condition as the boundary condition at the coordinate singularity and solves the reformulated boundary value problem with a Gauss–Lobatto collocation scheme. The method corresponds to the standard quadrature formula

$$\int_0^1 v(r)\omega(r)dr = \sum_{j=0}^{N} w_j v(r_j) \qquad \text{for } v \in \mathbf{P}_{2N-1}, \tag{1.2}$$

where $\omega(r)$ is the weight function and $r_j = (1+\rho_j)/2$. For example, $\omega(r) = 1$ and ρ_j $(0 \le j \le N)$ are the roots of $(1-\rho^2)L_N^{'}(\rho)$ for the Legendre collocation method, and $\omega(r) = (r-r^2)^{-1/2}$ and $\rho_j = \cos\frac{\pi(N-j)}{N}$ $(0 \le j \le N)$ for the Chebyshev collocation method. We emphasize that the method of [12], which is no more than a traditional collocation method, is different from those considered by Bernardi, Dauge, and Maday [1]. The method shares with many existing methods the common feature of explicitly using pole conditions, and has been successfully applied to practical problems including the Navier–Stokes equations; e.g., see [13, 14]. Our analysis is given for both the Legendre and Chebyshev schemes. A Chebyshev collocation scheme is often desirable in practical computation because the fast Fourier transformation (FFT) can be utilized. We find that for the current situation with coordinate singularities the corresponding bilinear form lacks the coercive property which is often crucial to the convergence analysis of a Chebyshev scheme. Because of this, the error estimate of the Chebyshev scheme is obtained in the weighted energy norm $\|\cdot\|_{E_n,\omega}$ with $\omega(r)$ being the Chebyshev weight function for the reduced equation with $n > 0$, but in the unweighted norm $\|\cdot\|_{E_n}$ for the case $n = 0$.

An outline of this paper is as follows. The method of [12] is briefly described in section 2. The convergence analysis of the Legendre and Chebyshev methods is given in section 3. In section 4 we present numerical results to verify the theoretical findings. Finally, section 5 contains conclusions and further comments.

2. Pole conditions and spectral collocation approximation.

In this section we briefly describe the spectral collocation method of [12] for a model problem

$$-\frac{d^2u}{dr^2} - \frac{1}{r}\frac{du}{dr} + \frac{n^2}{r^2}u = f, \quad 0 < r < 1, \tag{2.1}$$

$$u(1) = g, \tag{2.2}$$

where $n \ge 0$ is a given integer. This problem is obtained using separation of variables for the Poisson equation on the unit disk.

2.1. Pole conditions.

Equation (2.1) has a coordinate singularity at $r = 0$. Assume that both f and u are sufficiently smooth. A Taylor series expansion of u about $r = 0$ yields the pole conditions

$$O\left(\frac{1}{r^2}\right) : n^2 u(0) = 0, \tag{2.3}$$

$$O\left(\frac{1}{r}\right) : (n^2-1)\frac{du}{dr}(0) = 0, \tag{2.4}$$

$$O(1) : \left(\frac{n^2}{2} - 2\right)\frac{d^2u}{dr^2}(0) = f(0), \tag{2.5}$$

$$\cdots\cdots$$

These conditions contain full information about the solution behavior near $r = 0$. It was observed first by Orszag [21] and later by many other researchers (see [6]) that it is unnecessary to impose all of these pole conditions in order for a numerical scheme to obtain high accuracy. In fact, using one constraint

$$\begin{cases} u(0) = 0 & \text{for } n \ne 0, \\ \frac{du}{dr}(0) = 0 & \text{for } n = 0 \end{cases} \tag{2.6}$$

as the boundary condition at $r = 0$, Huang and Sloan [12] obtain spectrally accurate solutions; also see [24, 25] for the spectral Galerkin approximation. Once a boundary condition has been defined at $r = 0$, it is straightforward to apply a traditional spectral collocation scheme to the singular problem (2.1) and (2.2).

2.2. Legendre and Chebyshev collocation approximations. Hereafter, the weight functions $\omega(r) = 1$ and $\omega(r) = (r-r^2)^{-1/2}$ will be associated with the Legendre and Chebyshev methods, respectively. For simplicity, we use subscript ω in common notation for both methods and for those which apply only to the Chebyshev method, and suppress the subscript for the Legendre method.

For a given integer $N > 0$, let $\{\rho_{j,\omega}\}_{j=0}^{N}$ be a set of Gauss–Lobatto points associated with the weight function $\omega(r)$. Define

$$r_{j,\omega} = \frac{1+\rho_{j,\omega}}{2}, \qquad j = 0, 1, \dots, N. \tag{2.7}$$

The solution $u(r)$ is approximated by

$$u^N(r) = \sum_{j=0}^{N} u_{j,\omega} l_{j,\omega}(r), \tag{2.8}$$

where $u_{j,\omega}$ denotes the approximation of $u(r_{j,\omega})$ and $l_{j,\omega}(r)$ is the Lagrangian interpolation polynomial

$$l_{j,\omega}(r) = \prod_{\substack{i=0\\ i\neq j}}^{N} \frac{r - r_{i,\omega}}{r_{j,\omega} - r_{i,\omega}}. \tag{2.9}$$

A collocation approximation to (2.1), (2.2), and (2.6) is then defined by the collocation equations

$$-\frac{d^2u^N}{dr^2}(r_{j,\omega}) - \frac{1}{r_{j,\omega}}\frac{du^N}{dr}(r_{j,\omega}) + \frac{n^2}{r_{j,\omega}^2}u^N(r_{j,\omega}) = f(r_{j,\omega}), \qquad j = 1, \dots, N-1, \tag{2.10}$$

$$u^N(1) = g, \tag{2.11}$$

$$\begin{cases} u^N(0) = 0 & \text{for } n \neq 0, \\ \frac{du^N}{dr}(0) = 0 & \text{for } n = 0. \end{cases} \tag{2.12}$$

Recall that the transformed Gauss–Lobatto quadrature rule satisfies

$$\int_0^1 v(r)\omega(r)dr = \sum_{j=0}^{N} w_{j,\omega} v(r_{j,\omega}) \qquad \text{for } v \in \mathbf{P}_{2N-1}, \tag{2.13}$$

where the $w_{j,\omega}$'s are the corresponding weights and $\mathbf{P}_{2N-1}$ is the space of real polynomials (in r) of degree no more than $2N-1$. The associated interpolation operator $I^N : C[0,1] \to \mathbf{P}_N$ is defined as

$$I^N v \in \mathbf{P}_N : \quad (I^N v)(r_{j,\omega}) = v(r_{j,\omega}), \qquad j = 0, 1, \dots, N. \tag{2.14}$$

We use the notation

$$\langle u, v\rangle_\omega = \int_0^1 uv\omega dr, \quad \|u\|_\omega = \langle u, u\rangle_\omega^{1/2}, \tag{2.15}$$

$$\|u\|_{m,\omega} = \left(\sum_{k=0}^{m} \left\|\frac{d^k u}{dr^k}\right\|_\omega^2\right)^{1/2} \qquad \text{for } u \in H_\omega^m, \tag{2.16}$$

$$\langle u, v\rangle_{\omega,N} = \sum_{j=0}^{N} w_{j,\omega} u(r_{j,\omega}) v(r_{j,\omega}), \quad \|u\|_{\omega,N} = \langle u, u\rangle_{\omega,N}^{1/2}, \tag{2.17}$$

where H_ω^m $(m \geq 0)$ is a (weighted) Sobolev space on [0, 1].

For the Legendre collocation scheme, the set of Legendre–Gauss–Lobatto points is defined by $\rho_0 = -1$, $\rho_N = 1$, and ρ_j $(j = 1, 2, \ldots, N-1)$ being the roots of $L_N^{'}$, the first derivative of the Legendre polynomial L_N of degree N. We have

$$l_j(r) = \frac{2}{N(N+1)} \frac{r(1-r)\bar{L}_N^{'}(r)}{(r-r_j)\bar{L}_N(r_j)}, \qquad w_j = \frac{1}{N(N+1)} \frac{1}{\bar{L}_N^2(r_j)},$$

with $\bar{L}_N(r)$ being the transformed Legendre polynomial $L_N(2r-1)$.

For the Chebyshev approximation, the set of Gauss–Lobatto points is defined by $\rho_{0,\omega} = -1$, $\rho_{N,\omega} = 1$, and $\rho_{j,\omega}$ $(j = 1, 2, \ldots, N-1)$ being the roots of $T_N^{'}$, the derivative of the Chebyshev polynomial T_N of degree N. We have

$$l_{j,\omega}(r) = (-1)^{j+1} \frac{2}{c_j N^2} \frac{r(1-r)\bar{T}_N^{'}(r)}{(r - r_{j,\omega})}, \qquad w_{j,\omega} = \frac{\pi}{c_j N},$$

where $\bar{T}_k(r)$ is the transformed Chebyshev polynomial $T_k(2r-1)$ and

$$c_j = \begin{cases} 2, & j = 0, N, \\ 1, & j = 1, \ldots, N-1. \end{cases}$$

3. Convergence analysis.

3.1. Preliminary approximation results. To start with, we introduce some preliminary results. Hereafter, C is used to denote the generic constant. We shall assume that $N \gg m$ (the smoothness order of functions); otherwise, the estimates given below, especially those involving seminorms, will not be true.

LEMMA 3.1. *Let P^N denote the Legendre (or Chebyshev) truncated operator; i.e., $P^N v$ is the truncated Legendre (or Chebyshev) series of v. Then, for $m \geq 0$ and for $\omega(r) = 1$ (the Legendre case) or $\omega(r) = (r - r^2)^{-1/2}$ (the Chebyshev case),*

$$\|v - P^N v\|_\omega \leq CN^{-m} \left\|(r - r^2)^{m/2} \frac{d^m v}{dr^m}\right\|_\omega \qquad \forall v \in H_\omega^m. \tag{3.1}$$

LEMMA 3.2. *For $\omega(r) = 1$ (the Legendre case) or $\omega(r) = (r - r^2)^{-1/2}$ (the Chebyshev case) and for $m \geq 1$,*

$$\begin{aligned} &\left\|(r - r^2)^{-1/2}(v - I^N v)\right\|_\omega + N^{-1}\|v - I^N v\|_{1,\omega} \\ &\qquad \leq CN^{-m} \left\|(r - r^2)^{\frac{m-1}{2}} \frac{d^m v}{dr^m}\right\|_\omega \qquad \forall v \in H_\omega^m. \end{aligned} \tag{3.2}$$

Lemma 3.3. *Let $\omega(r) = 1$ (for the Legendre case) or $\omega(r) = (r - r^2)^{-1/2}$ (for the Chebyshev case). There exists a positive constant C, independent of N and M, such that for all $\phi \in \mathbf{P}_M$ with M being any nonnegative integer,*

$$\|\phi\|_{\omega,N} \le C\left(1 + \frac{M}{N}\right)\|\phi\|_{\omega}, \tag{3.3}$$

$$\|\phi\|_{\omega} \le \ \|\phi\|_{\omega,N} \ \le C\|\phi\|_{\omega}. \tag{3.4}$$

The interested reader is referred to [2, 3, 4, 7, 16, 18, 19] for the proofs of these Lemmas. Lemmas 3.1 and 3.2 are the improvements of existing results in terms of the weight and can be obtained by the method in the aforementioned references.

3.2. Convergence analysis of the Legendre method. We now proceed to the convergence analysis for the Legendre approximation (2.10)–(2.12). Let ϕ be an arbitrary polynomial in $\mathbf{P}_N$ satisfying

$$\begin{cases} \phi(1) = \phi(0) = 0 & \text{if } n \neq 0, \\ \phi(1) = 0 & \text{if } n = 0. \end{cases} \tag{3.5}$$

Multiplying (2.10) by $r_j w_j \phi(r_j)$ and summing over the range of j from 1 to $N-1$, we have

$$\sum_{j=1}^{N-1} w_j\phi(r_j)\left[-r_j\frac{d^2u^N}{dr^2}(r_j) - \frac{du^N}{dr}(r_j) + \frac{n^2}{r_j}u^N(r_j)\right] = \sum_{j=1}^{N-1} r_j w_j \phi(r_j) f(r_j). \tag{3.6}$$

It is not difficult to see from (3.5) that

$$\begin{aligned} & w_N\phi(r_N)\left[-r_N\frac{d^2u^N}{dr^2}(r_N) - \frac{du^N}{dr}(r_N) + \frac{n^2}{r_N}u^N(r_N)\right] \\ & = r_N w_N \phi(r_N) f(r_N) \\ & = 0. \end{aligned} \tag{3.7}$$

Noticing that $u^N(r)/r$ is a polynomial of degree not greater than $N-1$, that $\phi(r_0) = 0$ (see (3.5)) when $n \neq 0$, and that $(du^N/dr)(r_0) = 0$ when $n = 0$, we have

$$\begin{aligned} & w_0\phi(r_0)\left[-r_0\frac{d^2u^N}{dr^2}(r_0) - \frac{du^N}{dr}(r_0) + \frac{n^2}{r_0}u^N(r_0)\right] \\ & = r_0 w_0 \phi(r_0) f(r_0) \\ & = 0. \end{aligned} \tag{3.8}$$

Thus, (3.6)–(3.8) imply that

$$\left\langle -r\frac{d^2u^N}{dr^2} - \frac{du^N}{dr} + \frac{n^2}{r}u^N, \phi\right\rangle_N = \langle rf, \phi\rangle_N. \tag{3.9}$$

Since $r\phi(d^2u^N/dr^2)$, $\phi(du^N/dr)$, and $n^2u^N\phi/r$ are in $\mathbf{P}_{2N-1}$, (2.13) and (3.9) lead to

$$\left\langle -r\frac{d^2u^N}{dr^2} - \frac{du^N}{dr} + \frac{n^2}{r}u^N, \phi\right\rangle = \langle rf, \phi\rangle_N \tag{3.10}$$

or, taking integration by parts,

$$\left\langle \frac{du^N}{dr}, r\frac{d\phi}{dr}\right\rangle + n^2\left\langle \frac{u^N}{r}, \phi\right\rangle \;=\; \langle rf, \phi\rangle_N. \tag{3.11}$$

Multiplying the continuous equation (2.1) by $r\phi$ and integrating from $r = 0$ to 1, we obtain

$$\left\langle \frac{du}{dr}, r\frac{d\phi}{dr}\right\rangle + n^2\left\langle \frac{u}{r}, \phi\right\rangle \;=\; \langle rf, \phi\rangle. \tag{3.12}$$

Then, subtracting (3.11) from (3.12) gives the error equation

$$\left\langle \frac{d(u-u^N)}{dr}, r\frac{d\phi}{dr}\right\rangle + n^2\left\langle \frac{u-u^N}{r}, \phi\right\rangle \;=\; \langle rf, \phi\rangle - \langle rf, \phi\rangle_N, \tag{3.13}$$

which can be written in a simpler form as

$$a_{r,n}(u-u^N, \phi) = F_r(\phi) \quad \forall \phi \in \mathbf{P}_N, \tag{3.14}$$

where

$$\begin{aligned} a_{r,n}(u,v) \;&=\; \left\langle \frac{du}{dr}, r\frac{dv}{dr}\right\rangle + n^2\left\langle u, \frac{v}{r}\right\rangle, \\ \|v\|_{E_n} \;&=\; a_{r,n}(v,v)^{1/2}, \\ F_r(\phi) \;&=\; \langle rf, \phi\rangle - \langle rf, \phi\rangle_N. \end{aligned} \tag{3.15}$$

We first consider the case $n \neq 0$. Recall that we have $u^N(0) = u(0) = 0$ (cf. (2.6)). Let

$$V_{0g} = \{v \in H^1(I) : v(0) = 0, v(1) = g\}, \quad V_{0g}^N = V_{0g} \cap \mathbf{P}_N.$$

Lemma 3.4. *Let u and u^N be the solutions of the problem* (2.1)–(2.2) ($n \neq 0, u(0) = 0$) *and the approximation* (2.10)–(2.11) ($u^N(0) = 0$), *respectively. We have*

$$\|u - u^N\|_{E_n} \leq \sup_{\varphi \in V_{00}^N} \frac{F_r(\varphi)}{\|\varphi\|_{E_n}} + 2 \inf_{v \in V_{0g}^N} \|v - u\|_{E_n}. \tag{3.16}$$

Proof. Equation (3.14) can be rewritten as

$$a_{r,n}(v-u^N, \phi) = F_r(\phi) + a_{r,n}(v-u, \phi) \qquad \forall \phi \in V_{00}^N \text{ and } v \in V_{0g}^N.$$

Taking $\phi = v - u^N \in V_{00}^N$ results in

$$\|v-u^N\|_{E_n}^2 = a_{r,n}(v-u^N, \phi) \leq \sup_{\varphi \in V_{00}^N} \frac{F_r(\varphi)}{\|\varphi\|_{E_n}} \|\phi\|_{E_n} + a_{r,n}(v-u, \phi)\,.$$

Since

$$\begin{aligned} &a_{r,n}(v-u, \phi) \\ &\quad \leq \|r^{1/2}(v-u)_r\| \cdot \|r^{1/2}\phi_r\| + n^2\|r^{-1/2}(v-u)\| \cdot \|r^{-1/2}\phi\| \\ &\quad \leq \left(\|r^{1/2}(v-u)_r\|^2 + n^2\|r^{-1/2}(v-u)\|^2\right)^{1/2} \left(\|r^{1/2}\phi_r\|^2 + n^2\|r^{-1/2}\phi\|^2\right)^{1/2} \\ &\quad \leq C a_{r,n}(v-u, v-u)^{1/2} a_{r,n}(\phi, \phi)^{1/2} \\ &\quad = C\|v-u\|_{E_n} \cdot \|v-u^N\|_{E_n}, \end{aligned}$$

we have

$$\|v-u^N\|_{E_n}^2 \le C\left(\sup_{\varphi\in V_{00}^N}\frac{F_r(\varphi)}{\|\varphi\|_{E_n}}+\|v-u\|_{E_n}\right)\|v-u^N\|_{E_n}.$$

Then the desired result follows from

$$\|u-u^N\|_{E_n}\le\|u-v\|_{E_n}+\|v-u^N\|_{E_n}. \qquad \square$$

We now use Lemma 3.4 to obtain the estimate of $\|u-u^N\|_{E_n}$. For the first term on the right-hand side of (3.16), we have from the Cauchy–Schwarz inequality and Lemmas 3.1–3.3 that, for any $\phi\in\mathbf{P}_N$,

$$\begin{aligned}|F_r(\phi)| &= |\langle rf,\phi\rangle-\langle rf,\phi\rangle_N|\\ &= |\langle rf,\phi\rangle-\langle P^{N-1}(rf),\phi\rangle\\ &\quad+\langle P^{N-1}(rf),\phi\rangle-\langle I^N(rf),\phi\rangle_N|\\ &\le|\langle rf-P^{N-1}(rf),\phi\rangle|+|\langle P^{N-1}(rf)-I^N(rf),\phi\rangle_N|\\ &\le\|rf-P^{N-1}(rf)\|\,\|\phi\|+C\|P^{N-1}(rf)-I^N(rf)\|\,\|\phi\|_N\\ &\le\|rf-P^{N-1}(rf)\|\,\|\phi\|+C\left(\|rf-P^{N-1}(rf)\|+\|rf-I^N(rf)\|\right)\|\phi\|\\ &= C\left(\|rf-P^{N-1}(rf)\|+\|rf-I^N(rf)\|\right)\|\phi\|.\end{aligned} \tag{3.17}$$

By Lemmas 3.1 and 3.2 (and taking $v=rf\in H^{\bar m}$, $\bar m:=\max\{m-1,1\}$), we obtain

$$\begin{aligned}|F_r(\phi)| &\le CN^{1-m}\left\|(r-r^2)^{\frac{\bar m-1}{2}}\frac{d^{\bar m}(rf)}{dr^{\bar m}}\right\|\|\phi\|\\ &\le CN^{1-m}\left\|(r-r^2)^{\frac{\bar m-1}{2}}\frac{d^{\bar m}(rf)}{dr^{\bar m}}\right\|\|\phi\|_{E_n}.\end{aligned} \tag{3.18}$$

For the second term on the right-hand side of (3.16), taking $v=I^Nu$ and using the definition of the energy norm leads to

$$\begin{aligned}\|v-u\|_{E_n} &\le |v-u|_1+\|[r(1-r)]^{-1/2}(v-u)\|\\ &\le CN^{1-m}\left\|(r-r^2)^{\frac{m-1}{2}}\frac{d^m u}{dr^m}\right\|.\end{aligned} \tag{3.19}$$

Substituting (3.18) and (3.19) into (3.16), we obtain the estimate

$$\|u-u^N\|_{E_n}\le CN^{1-m}\left(\left\|(r-r^2)^{\frac{m-1}{2}}\frac{d^m u}{dr^m}\right\|+\left\|(r-r^2)^{\frac{\bar m-1}{2}}\frac{d^{\bar m}(rf)}{dr^{\bar m}}\right\|\right). \tag{3.20}$$

We now consider the case $n=0$. Recall again that we have $\frac{du^N}{dr}(0)=\frac{du}{dr}(0)=0$. Define

$$W_g=\{v\in H^1(I):v(1)=g\},\qquad W_g^N=W_g\cap\mathbf{P}_N.$$

LEMMA 3.5. *Let u and u^N be the solutions of the problem* (2.1)–(2.2) ($n=0$, $\frac{du}{dr}(0)=0$) *and the approximation* (2.10)–(2.11) ($\frac{du^N}{dr}(0)=0$), *respectively. We have*

$$\|u-u^N\|_{E_0}\le 2\sup_{\varphi\in W_0^N}\frac{F_r(\varphi)}{\|\varphi\|_{E_0}}+2\inf_{v\in W_g^N}\|v-u\|_{E_0}.$$

Proof. Equation (3.14) can be written as

$$a_{r,0}(v-u^N,\phi)=F_r(\phi)+a_{r,0}(v-u,\phi) \qquad \forall\phi\in W_0^N \text{ and } v\in W_g^N.$$

Taking $\phi=v-u^N\in W_0^N$, we have

$$\begin{aligned}\|v-u^N\|_{E_0}^2&=a_{r,0}(v-u^N,\phi)\\&\le\left(\sup_{\varphi\in W_0^N}\frac{F_r(\varphi)}{\|\varphi\|_{E_0}}+\|v-u\|_{E_0}\right)\|v-u^N\|_{E_0}.\end{aligned}$$

Then the conclusion follows. □

From the Hardy-type inequality

$$\|v\|\le\left\|\sqrt{r}\frac{dv}{dr}\right\|=\|v\|_{E_0} \qquad \forall v\in W_0,$$

Lemma 3.5 leads to the same result as in (3.20). Hence, we have proved the following theorem.

THEOREM 3.1. *For any integer $m\ge1$, the Legendre-collocation approximation u^N defined by the scheme* (2.10)–(2.12) *for the problem* (2.1) *and* (2.2) *satisfies*

$$\|u-u^N\|_{E_n}\le CN^{1-m}\left(\left\|(r-r^2)^{\frac{m-1}{2}}\frac{d^mu}{dr^m}\right\|+\left\|(r-r^2)^{\frac{\bar m-1}{2}}\frac{d^{\bar m}(rf)}{dr^{\bar m}}\right\|\right), \tag{3.21}$$

where u is the exact solution of (2.1) *and* (2.2) *and* $\bar m=\max\{m-1,1\}$.

This theorem shows that the Legendre collocation approximation is convergent and the error decays faster than algebraically, provided that the right-hand-side term f and the solution u are infinitely differentiable. As shown in [1], the correct regularity requirement for u and f should be considered in a weighted Sobolev space

$$H_r^s=\left\{v\;\middle|\;\sum_{l=0}^{s}\left\|\sqrt{r}\frac{d^lv}{dr^l}\right\|^2<\infty\right\} \tag{3.22}$$

for some integer s. It is not difficult to see that the terms in the bracket on the right-hand side of (3.21) are bounded for $u\in H_r^m$ and $f\in H_r^{\bar m}$ with $m\ge2$. In this sense, the result of Theorem 3.1 is optimal.

3.3. Convergence analysis of the Chebyshev method. We now consider the convergence of the Chebyshev method (2.10)–(2.12). Let ϕ be the same as in (3.5). As for (3.10), we have

$$\left\langle -r\frac{d^2u^N}{dr^2}-\frac{du^N}{dr}+\frac{n^2}{r}u^N,\phi\right\rangle_\omega = \langle rf,\phi\rangle_{\omega,N} \tag{3.23}$$

or, taking integration by parts,

$$\left\langle\frac{du^N}{dr},r\frac{d(\phi\omega)}{dr}\right\rangle+n^2\left\langle\frac{u^N}{r},\phi\right\rangle_\omega = \langle rf,\phi\rangle_{\omega,N}. \tag{3.24}$$

The error equation reads as

$$\begin{aligned}&\left\langle\frac{d(u-u^N)}{dr},r\frac{d(\phi\omega)}{dr}\right\rangle+n^2\left\langle\frac{u-u^N}{r},\phi\right\rangle_\omega\\&=\langle rf,\phi\rangle_\omega-\langle rf,\phi\rangle_{\omega,N}.\end{aligned} \tag{3.25}$$

We also write it in a simpler form

$$a_{r,n,\omega}(u-u^N,\phi)=F_{r,\omega}(\phi)\qquad \forall\phi\in\mathbf{P}_N, \tag{3.26}$$

where

$$\begin{aligned} a_{r,n,\omega}(u,v) &= b_{r,\omega}(u,v)+n^2\left\langle u,\frac{v}{r}\right\rangle_\omega,\\ b_{r,\omega}(u,v) &= \left\langle \frac{du}{dr}, r\frac{d(v\omega)}{dr}\right\rangle,\\ F_{r,\omega}(\phi) &= \langle rf,\phi\rangle_\omega-\langle rf,\phi\rangle_{\omega,N}. \end{aligned} \tag{3.27}$$

It is known that the nonsymmetric bilinear form $b_{r,\omega}(\cdot,\cdot)$, without the factor r, is coercive (see [3, 7, 8, 17, 15]). On the other hand, in the current situation, $b_{r,\omega}(v,v)$ can become negative for some polynomials subject to the boundary conditions $v(-1)=v(1)=0$ or $\frac{dv}{dr}(-1)=v(1)=0$. We have the following Gårding-type inequality.

Lemma 3.6. *For all $u,v\in H^1_{\omega,0}$ we have*

$$\frac{1}{4}\left\|\sqrt{r}\frac{dv}{dr}\right\|_\omega^2+\frac{3}{8}\left\|\frac{v}{\sqrt{1-r}}\right\|_\omega^2-\frac{1}{8}\left\|\frac{v}{\sqrt{r}}\right\|_\omega^2\le b_{r,\omega}(v,v)\le\left\|\sqrt{r}\frac{dv}{dr}\right\|_\omega^2, \tag{3.28}$$

$$|b_{r,\omega}(u,v)|\le 3\left\|\sqrt{r}\frac{du}{dr}\right\|_\omega\left\|\sqrt{r}\frac{dv}{dr}\right\|_\omega. \tag{3.29}$$

Proof. For notational simplicity, define

$$I_1(v)=\int_0^1\left(\frac{dv}{dr}\right)^2 r\omega\,dr. \tag{3.30}$$

We have from integrating by parts

$$\begin{aligned} b_{r,\omega}(v,v) &= I_1(v)-\int_0^1\frac{dv}{dr}v\frac{1-2r}{2(r-r^2)}r\omega\,dr\\ &= I_1(v)-\frac{1}{8}\int_0^1 v^2r(1-2r+4r^2)\omega^5\,dr\\ &= I_1(v)-\frac{3}{8}\int_0^1 v^2r^3\omega^5\,dr-\frac{1}{8}\int_0^1 v^2r(1-r)^2\omega^5\,dr. \end{aligned} \tag{3.31}$$

Thus $b_{r,\omega}(v,v)\le I_1(v)$. On the other hand,

$$\begin{aligned} 0 &\le \int_0^1\left(\frac{dv}{dr}+vr\omega^2\right)^2 r\omega\,dr\\ &= I_1(v)+\int_0^1 v^2r^3\omega^5\,dr+\int_0^1\frac{d(v^2)}{dr}r^2\omega^3\,dr\\ &= I_1(v)-\frac{1}{2}\int_0^1 v^2r^2\omega^5\,dr, \end{aligned} \tag{3.32}$$

which gives

$$b_{r,\omega}(v,v)\ge\frac{1}{4}I_1(v)+\frac{3}{8}\int_0^1 v^2r^2(1-r)\omega^5\,dr-\frac{1}{8}\int_0^1\frac{v^2}{r}\omega\,dr. \tag{3.33}$$

The result (3.28) follows. To prove (3.29), we estimate $b_{r,\omega}(u,v)$ by

$$\begin{aligned} |b_{r,\omega}(u,v)| &\le \left|\int_0^1 \frac{du}{dr}\frac{dv}{dr} r\omega\, dr\right| + \left|\int_0^1 \frac{du}{dr} v \frac{1-2r}{2(r-r^2)} r\omega\, dr\right| \\ &\le [I_1(u)]^{1/2}[I_1(v)]^{1/2} + I_2(u,v), \end{aligned} \tag{3.34}$$

where

$$I_2(u,v) = \left|\int_0^1 \frac{du}{dr} v \frac{1-2r}{2(r-r^2)} r\omega\, dr\right| \le [I_1(u)]^{1/2}[I_3(v)]^{1/2} \tag{3.35}$$

with

$$I_3(v) = \int_0^1 v^2 \frac{(1-2r)^2}{4(r-r^2)^2} r\omega\, dr = \frac{1}{4}\int_0^1 v^2 r(1-2r)^2\omega^5\, dr. \tag{3.36}$$

On the other hand, from integrating by parts,

$$I_2(v,v) = \frac{1}{8}\int_0^1 v^2(2r^2 + r(1-2r)^2)\omega^5\, dr \ge \frac{1}{2} I_3(v). \tag{3.37}$$

Thus we get from (3.37) and (3.35)

$$I_3(v) \le 2I_2(v,v) \le 2[I_1(v)]^{1/2}[I_3(v)]^{1/2}, \tag{3.38}$$

which gives $I_3(v) \le 4I_1(v)$ and

$$\begin{aligned} |b_{r,\omega}(u,v)| &\le [I_1(u)]^{1/2}([I_1(v)]^{1/2} + [I_3(v)]^{1/2}) \\ &\le 3[I_1(u)]^{1/2}[I_1(v)]^{1/2}. \qquad \square \end{aligned} \tag{3.39}$$

We first consider the case $n \neq 0$. Let V_{0g} and V_{0g}^N be the same as before and

$$\|v\|_{E_n,\omega} = \left(\left\| \sqrt{r}\frac{dv}{dr} \right\|_\omega^2 + n^2 \left\| \frac{v}{\sqrt{r}} \right\|_\omega^2 \right)^{1/2}.$$

LEMMA 3.7. *Let u and u^N be the solutions of the problem* (2.1)–(2.2) ($n \neq 0, u(0) = 0$) *and the Chebyshev approximation* ($u^N(0) = 0$)*, respectively. We have*

$$\|u - u^N\|_{E_n,\omega} \le C \sup_{\varphi \in V_{00}^N} \frac{F_{r,\omega}(\varphi)}{\|\varphi\|_{E_n,\omega}} + C \inf_{v \in V_{0g}^N} \|v - u\|_{E_n,\omega}. \tag{3.40}$$

Proof. Equation (3.26) can be rewritten as

$$a_{r,n,\omega}(v - u^N, \phi) = F_{r,\omega}(\phi) + a_{r,n,\omega}(v - u, \phi) \qquad \forall \phi \in V_{00}^N \text{ and } v \in V_{0g}^N.$$

Taking $\phi = v - u^N \in V_{00}^N$ and using the inequality (3.28) of Lemma 3.6 yields

$$\|v - u^N\|_{E_n,\omega}^2 \le Ca_{r,n,\omega}(v - u^N, \phi) \le C \sup_{\varphi \in V_{00}^N} \frac{F_{r,\omega}(\varphi)}{\|\varphi\|_{E_n,\omega}} \|\phi\|_{E_n,\omega} + Ca_{r,n,\omega}(v - u, \phi)\,.$$

From (3.29) of Lemma 3.6 we have

$$\begin{aligned}
&a_{r,n,\omega}(v-u,\phi)\\
&\quad\le C\left\|\sqrt{r}\frac{d(v-u)}{dr}\right\|_\omega\left\|\sqrt{r}\frac{d\phi}{dr}\right\|_\omega+n^2\left\|\frac{v-u}{\sqrt{r}}\right\|_\omega\left\|\frac{\phi}{\sqrt{r}}\right\|_\omega\\
&\quad\le C\left(\left\|\sqrt{r}\frac{d(v-u)}{dr}\right\|_\omega^2+n^2\left\|\frac{v-u}{\sqrt{r}}\right\|_\omega^2\right)^{1/2}\left(\left\|\sqrt{r}\frac{d(\phi)}{dr}\right\|_\omega^2+n^2\left\|\frac{\phi}{\sqrt{r}}\right\|_\omega^2\right)^{1/2}\\
&\quad= C\|v-u\|_{E_n,\omega}\cdot\|v-u^N\|_{E_n,\omega},
\end{aligned}$$

and therefore

$$\|v-u^N\|_{E_n,\omega}^2\le C\left(\sup_{w\in V_{00}^N}\frac{F_{r,\omega}(w)}{\|w\|_{E_n,\omega}}+\|v-u\|_{E_n,\omega}\right)\|v-u^N\|_{E_n,\omega}.$$

Then, the desired result follows from the triangular inequality:

$$\|u-u^N\|_{E_n,\omega}\le\|u-v\|_{E_n,\omega}+\|v-u^N\|_{E_n,\omega}. \qquad \square$$

We now use Lemma 3.7 to obtain the estimate of $\|u-u^N\|_{E_n,\omega}$. For the first term on the right-hand side of (3.40), we have from the Cauchy–Schwarz inequality and Lemmas 3.1–3.3 that, for any $\phi\in\mathbf{P}_N$,

$$\begin{aligned}
|F_{r,\omega}(\phi)|&=|\langle rf,\phi\rangle_\omega-\langle rf,\phi\rangle_{\omega,N}|\\
&=|\langle rf,\phi\rangle_\omega-\langle P^{N-1}(rf),\phi\rangle_\omega+\langle P^{N-1}(rf),\phi\rangle_\omega-\langle I^N(rf),\phi\rangle_{\omega,N}|\\
&=C\left(\|rf-P^{N-1}(rf)\|_\omega+\|rf-I^N(rf)\|_\omega\right)\|\phi\|_\omega.
\end{aligned}$$

Then, by Lemmas 3.1 and 3.2,

$$\begin{aligned}
|F_{r,\omega}(\phi)|&\le CN^{1-m}\left\|(r-r^2)^{\frac{\bar m-1}{2}}\frac{d^{\bar m}(rf)}{dr^{\bar m}}\right\|_\omega\|\phi\|_\omega\\
&\le CN^{1-m}\left\|(r-r^2)^{\frac{\bar m-1}{2}}\frac{d^{\bar m}(rf)}{dr^{\bar m}}\right\|_\omega\|\phi\|_{E_n,\omega}. \qquad (3.41)
\end{aligned}$$

For the second term on the right-hand side of (3.16), taking $v=I^Nu$ leads to

$$\begin{aligned}
\|v-u\|_{E_n,\omega}&\le\left\|\sqrt{r}\frac{d(v-u)}{dr}\right\|_\omega+\left\|\frac{v-u}{\sqrt{r(1-r)}}\right\|_\omega\\
&\le CN^{1-m}\left\|(r-r^2)^{\frac{m-1}{2}}\frac{d^mu}{dr^m}\right\|_\omega. \qquad (3.42)
\end{aligned}$$

Substituting (3.41) and (3.42) into (3.16), we obtain the estimate

$$\begin{aligned}
&\|u-u^N\|_{E_n,\omega}\\
&\quad\le CN^{1-m}\left(\left\|(r-r^2)^{\frac{m-1}{2}}\frac{d^mu}{dr^m}\right\|_\omega+\left\|(r-r^2)^{\frac{\bar m-1}{2}}\frac{d^{\bar m}(rf)}{dr^{\bar m}}\right\|_\omega\right). \qquad (3.43)
\end{aligned}$$

We now consider the case $n=0$. In this case, $a_{r,0,\omega}(\cdot,\cdot)=b_{r,\omega}(\cdot,\cdot)$ is not coercive. Thus it does not seem likely to us that an error bound can be obtained in the weighted energy norm $\|\cdot\|_{E_0,\omega}$. For this reason, we conduct the estimation in the energy norm $\|\cdot\|_{E_0}$ without the Chebyshev weight. We first note that the polar condition $\frac{du^N}{dr}(0)=0$ allows us to extend the collocation equations

$$-r_{j,\omega}\frac{d^2u^N}{dr^2}(r_{j,\omega})-\frac{du^N}{dr}(r_{j,\omega})=r_{j,\omega}f(r_{j,\omega}),\qquad 1\le j\le N-1, \tag{3.44}$$

to the point $r=r_{j,\omega}=0$. Since the left-hand side of (3.44) is a polynomial of degree $N-1$, (3.44) holds for all $r\in[0,1]$, provided that $f\in\mathbf{P}_{N-2}$. We introduce an auxiliary interpolation operator $\tilde{I}^{N-2}$: $C([0,1])\to\mathbf{P}_{N-2}$ defined by

$$\tilde{I}^{N-2}v(r_j)=v(r_j),\qquad 1\le j\le N-1. \tag{3.45}$$

Thus we are able to rewrite (3.44) as

$$-r\frac{d^2u^N}{dr^2}(r)-\frac{du^N}{dr}(r)=r\tilde{I}^{N-2}f(r),\qquad 0\le r\le 1. \tag{3.46}$$

Let W_g and W_g^N be the same as before.

Lemma 3.8. *Let u and u^N be the solutions of the problem* (2.1)–(2.2) ($n=0$, $\frac{du}{dr}(0)=0$) *and the Chebyshev collocation approximation* ($\frac{du^N}{dr}(0)=0$), *respectively. We have*

$$\|u-u^N\|_{E_0}\le\sup_{\varphi\in W_0^N}\frac{\tilde{F}_r(\varphi)}{\|\varphi\|_{E_0}}+2\inf_{v\in W_g^N}\|v-u\|_{E_0},$$

where $\tilde{F}_r(\varphi)=\langle rf-r\tilde{I}^{N-2}f,\varphi\rangle$.

Proof. We have from (2.1) and (3.46)

$$a_{r,0}(v-u^N,\phi)=\tilde{F}_r(\phi)+a_{r,0}(v-u,\phi)\qquad\forall\phi\in W_0^N\text{ and }v\in W_g^N.$$

Taking $\phi=v-u^N\in W_0^N$, we have

$$\begin{aligned}\|v-u^N\|_{E_0}^2&=a_{r,0}(v-u^N,\phi)\\&\le\left(\sup_{\varphi\in W_0^N}\frac{\tilde{F}_r(\varphi)}{\|\varphi\|_{E_0}}+\|v-u\|_{E_0}\right)\|v-u^N\|_{E_0},\end{aligned}$$

which gives the desired result. □

We need further to estimate the term $\tilde{F}_r(\varphi)$. According to the definition, it is easy to see that $(r-r^2)\tilde{I}^{N-2}f=I^N((r-r^2)f)$. Therefore, we have from the Cauchy–Schwarz inequality and Lemma 3.2 that, for any $\varphi\in W_0^N$,

$$\begin{aligned}|\tilde{F}_r(\varphi)|&=|\langle(r-r^2)f-I^N((r-r^2)f),(1-r)^{-1}\varphi\rangle|\\&\le\|(r-r^2)f-I^N((r-r^2)f)\|\,\|(1-r)^{-1}\varphi\|\\&\le CN^{1-m}\left\|(r-r^2)^{\frac{\bar{m}-1}{2}}\frac{d^{\bar{m}}((r-r^2)f)}{dr^{\bar{m}}}\right\|_{\omega}\|\varphi\|_{E_0},\end{aligned} \tag{3.47}$$

where we have used

$$\|(1-r)^{-1}\varphi\|\le 2\left\|\sqrt{r}\frac{d\varphi}{dr}\right\|\qquad\forall\varphi\in W_0^N, \tag{3.48}$$

which can be derived from

$$\begin{aligned}0 &\le \int_0^1 \left(\sqrt{r}\frac{d\varphi}{dr} - \frac{1}{2}(1-r)^{-1}\varphi\right)^2 dr \\ &= \left\|\sqrt{r}\frac{d\varphi}{dr}\right\|^2 + \frac{1}{4}\|(1-r)^{-1}\varphi\|^2 - \frac{1}{2}\int_0^1 \frac{d(\varphi)^2}{dr} r^{1/2}(1-r)^{-1}\, dr \\ &\le \left\|\sqrt{r}\frac{d\varphi}{dr}\right\|^2 - \frac{1}{4}\|(1-r)^{-1}\varphi\|^2. \qquad (3.49)\end{aligned}$$

Hence, we have proved the following theorem.

THEOREM 3.2. *For any integer $m \ge 1$, the Chebyshev-collocation approximation u^N defined by the scheme* (2.10)–(2.12) *for the problem* (2.1) *and* (2.2) *satisfies*

$$\begin{aligned}\|u - u^N\|_{E_n} \le CN^{1-m} &\left(\left\|(r-r^2)^{\frac{m}{2}-\frac{3}{4}} \frac{d^m u}{dr^m}\right\| + \left\|(r-r^2)^{\frac{\bar m}{2}-\frac{3}{4}} \frac{d^{\bar m}(rf)}{dr^{\bar m}}\right\| \right. \\ &\left. + \left\|(r-r^2)^{\frac{\bar m}{2}-\frac{3}{4}} \frac{d^{\bar m}((r-r^2)f)}{dr^{\bar m}}\right\| \right), \qquad (3.50)\end{aligned}$$

where u is the exact solution of (2.1) *and* (2.2) *and* $\bar m = \max\{m-1, 1\}$. *For the case $n \ne 0$,* (3.50) *also holds in the stronger norm* $\|\cdot\|_{E_n,\omega}$.

Thus, we obtain a convergence result similar to that of the Legendre collocation method. As in Theorem 3.1, when $u \in H_r^m$ and $f \in H_r^{\bar m}$ with $m \ge 5/2$, the terms in the bracket on the right-hand side of (3.50) are bounded.

4. Numerical experiments. In this section we present some numerical results to demonstrate the accuracy of the Legendre and Chebyshev collocation methods (2.10)–(2.12) for the model problem (2.1)–(2.2).

Example 1. The function $f(r)$ and the Dirichlet boundary condition at $r = 1$ are chosen such that the exact solution of (2.1) and (2.2) is

$$u(r) = r^2 \cos(10\pi r), \qquad 0 < r < 1. \qquad (4.1)$$

We note that the energy norm $\|\cdot\|_{E_n}$ is stronger than the L^∞ norm for $n > 0$. (This is not true for $n = 0$.) For this reason, we use the maximum norm to measure the error for the case $n > 0$, but use the energy norm for the case $n = 0$. These norms are numerically approximated in the computations, viz.,

$$E_{0,N} = \left\{ \sum_{j=0}^N r_j \left| \frac{du}{dr}(r_j) - \frac{du^N}{dr}(r_j)\right|^2 w_j \right\}^{1/2} \qquad \text{for } n = 0, \qquad (4.2)$$

$$E_{1,N} = \max_{0 \le j \le N} |u(r_{j,\omega}) - u^N(r_{j,\omega})| \qquad \text{for } n = 1, \qquad (4.3)$$

where the Legendre points $\{r_j\}$ are used in (4.2) and both the Legendre and Chebyshev points $\{r_{j,\omega}\}$ are used in (4.3). The numerical results for the cases $n = 0$ and $n = 1$ are listed in Table 1. The spectral convergence of the methods is clearly shown in the table. One may also notice that the Legendre and Chebyshev collocation methods produce very comparable results.

Example 2. The function $f(r)$ and the boundary condition at $r = 1$ are chosen such that the problem has a less regular solution

$$u(r) = r^{5/2}, \qquad 0 < r < 1. \qquad (4.4)$$

Table 1

Numerical results obtained with the Legendre (LC) and Chebyshev (CC) collocation methods for Example 1*. The convergence order is* N^{-Order}.

	$n = 0$				$n = 1$			
	LC-method		CC-method		LC-method		CC-method	
N	$E_{0,N}$	Order	$E_{0,N}$	Order	$E_{1,N}$	Order	$E_{1,N}$	Order
10	6.9e+01		1.5e+02		1.7e+01		4.1e+01	
20	2.1e+01	1.8	2.4e+01	2.7	3.9e+00	2.2	5.0e+00	3.1
30	1.0e+00	7.4	1.3e+00	7.2	5.3e−02	10.6	6.4e−02	10.7
40	2.0e−03	21.7	2.9e−03	21.2	7.3e−05	22.9	9.2e−05	22.8
50	1.8e−07	41.8	3.5e−07	40.5	6.7e−09	41.7	7.3e−09	42.3
60	1.6e−12	63.9	1.9e−12	66.3	7.8e−14	62.3	1.0e−13	61.3
70	2.2e−13		4.2e−12		1.5e−14		1.2e−13	

Table 2

Numerical results obtained with the Legendre collocation method for Example 2*. The convergence order is* N^{-Order}.

	$n = 0$		$n = 1$	
N	$E_{0,N}$	Order	$E_{1,N}$	Order
40	3.2e−07		1.7e−08	
80	1.2e−08	4.75	5.6e−10	4.96
120	1.7e−09	4.79	7.4e−11	4.97
160	4.2e−10	4.81	1.8e−11	4.98
200	1.4e−10	4.82	5.8e−12	4.96
240	5.9e−11	4.83	2.4e−12	4.91

The computation is done with the Legendre collocation method for $n = 0$ and $n = 1$. The solution error and the convergence order are listed in Table 2. One can easily see that $E_{0,N} \approx \mathrm{O}\left(N^{-5}\right)$ and $E_{1,N} \approx \mathrm{O}\left(N^{-5}\right)$. That is, the rate of convergence is nearly twice the exponent of r in (4.4), 5/2. On the other hand, it is not difficult to show that the first term on the right-hand side of (3.21) is bounded for $m < 5$, while the second term is bounded for $\bar{m} < 3$ or $m < 4$ for the current example. Thus, the right-hand-side terms are bounded for $m < 4$. From Theorem 3.1, we have $\|u - u^N\|_{E_n} \approx \mathrm{O}(N^{-3})$. This indicates that the convergence rate predicted by (3.21) is not sharp, although the estimate is optimal according to the regularity requirement ([1]; also cf. (3.22)). Such an order loss seems typical in the convergence analysis of collocation schemes, especially for problems involving force terms; e.g., see [4] for comparison of typical estimates for the Legendre Galerkin method ((8.7) on p. 274) and the Legendre collocation method ((15.15) on p. 310). It is interesting to note that sharp estimates have been obtained for the p-version finite element method (which is of Galerkin type); e.g., see Babuska and Suri [5]. Finally, we mention that the Chebyshev collocation method leads to very comparable results.

5. Conclusions and comments.

In the previous sections we have proved that the Legendre and Chebyshev collocation approximations presented in [12] are convergent and that the error decays faster than algebraically when f and u are infinitely differentiable for the singular problem (2.1) and (2.2). Our main results are given in Theorems 3.1 and 3.2.

The key feature of the spectral collocation approximation is that it uses a pole condition as the boundary condition at the singularity and employs a commonly used collocation scheme. Thus, the convergence result provides a theoretical justification for the well-known fact that it is unnecessary to impose all the pole conditions in order for numerical schemes to obtain high accuracy. Because most of the existing spectral approaches for singular problems use more or less the pole conditions, we expect that our result can also be regarded as a theoretical justification for these methods.

Finally we make a few comments on the method we analyzed. The method has been successfully applied to solving steady-state Navier–Stokes equations in [13, 14]. However, since it uses the Chebyshev or Legendre type of collocation methods in the r interval (0,1), one may suspect that the clustering of grid points near $r = 0$ leads to a very severe restriction on time steps for time dependent problems. To see this, let us consider the time dependent problem on the unit disk

$$u_t = \Delta u + au_x + bu_y. \tag{5.1}$$

In polar coordinates the equation becomes

$$\begin{aligned} u_t = {} & \frac{\partial^2 u}{\partial r^2} + \frac{1}{r}\frac{\partial u}{\partial r} + \frac{1}{r^2}\frac{\partial^2 u}{\partial \theta^2} \\ & + (a\cos\theta + b\sin\theta)\frac{\partial u}{\partial r} + (-a\sin\theta + b\cos\theta)\frac{1}{r}\frac{\partial u}{\partial \theta}. \end{aligned} \tag{5.2}$$

Assume that (5.2) is approximated in r using the Legendre or Chebyshev collocation, and in θ using Fourier collocation. Then it is not difficult to see that for the diffusion term the time step restriction for an explicit integration scheme is

$$\Delta t_{\max} \approx \min\{(\Delta r_0)^2, r_1 \Delta r_0\} \approx (r_1)^2 = O\left(\frac{1}{N^4}\right) \tag{5.3}$$

at $r \approx 0$ and

$$\Delta t_{\max} \approx \min\{(\Delta r_{N-1})^2, r_N \Delta r_{N-1}\} \approx \left(\frac{1}{N^4}\right) \tag{5.4}$$

at $r \approx 1$. Obviously these two time scales are the same. For the convection term we have at $r \approx 0$

$$\Delta t_{\max} \approx \min\{\Delta r_0, r_1\} \approx O\left(\frac{1}{N^2}\right) \tag{5.5}$$

and at $r \approx 1$

$$\Delta t_{\max} \approx \min\{\Delta r_{N-1}, r_N\} \approx O\left(\frac{1}{N^2}\right). \tag{5.6}$$

Once again they are the same. Thus, the above simple analysis tells us that the clustering of grid points near the singularity does not result in a time restriction worse than that near the outer boundary. Of course, just like spectral methods applied to nonsingular problems, a restriction $O(1/N^4)$ on time steps is too severe. Implicit or semi-implicit time integrators should be used. The resultant algebraic systems can be solved using either iterative methods with effective preconditioners [7, 12] or fast direct solvers [24].

For problems in spheric geometries the method can be applied straightforwardly. However, the severe restriction on time steps at the north and south poles could be a potential problem for the method (see discussion in [6, pp. 480–482]). This issue deserves further investigation.

Acknowledgment. The authors would like to thank the anonymous referees for their valuable comments.

REFERENCES

[1] C. Bernardi, M. Dauge, and Y. Maday, *Spectral Methods for Axisymmetric Domains*, Gauthier-Villars, Éditions Scientifiques et Médicales Elsevier, Paris, 1999.

[2] C. Bernardi and Y. Maday, *Properties of some weighted Sobolev spaces and application to spectral approximations*, SIAM J. Numer. Anal., 26 (1989), pp. 769–829.

[3] C. Bernardi and Y. Maday, *Polynomial interpolation results in Sobolev space*, J. Comput. Appl. Math., 43 (1992), pp. 53–82.

[4] C. Bernardi and Y. Maday, *Spectral Methods, Techniques of Scientific Computing (Part 2)*, Handb. Numer. Anal. 5, P. Ciarlet and J. L. Lions, eds., Elsevier, Amsterdam, 1997, pp. 209–486.

[5] I. Babuška and M. Suri, *The p and h-p versions of the finite element method, Basic principles and properties*, SIAM Rev., 36 (1994), pp. 578–632.

[6] J. P. Boyd, *Chebyshev and Fourier Spectral Methods*, Springer-Verlag, Berlin, 1989.

[7] C. Canuto, M. Y. Hussaini, A. Quarteroni, and T. A. Zang, *Spectral Methods in Fluid Dynamics*, Springer-Verlag, Berlin, 1988.

[8] C. Canuto and A. Quarteroni, *Variational methods in the theoretical analysis of spectral approximations*, in Spectral Methods for Partial Differential Equations (NASA Langley Research Center, 1982), R. G. Voigt, D. Gottlieb, and M. Y. Hussaini, eds., SIAM, Philadelphia, PA, 1984, pp. 55–78.

[9] H. Eisen, W. Heinrichs, and K. Witsch, *Spectral collocation methods and polar coordinate singularities*, J. Comput. Phys., 96 (1991), pp. 241–257.

[10] B. Fornberg, *A pseudospectral approach for polar and spherical geometries*, SIAM J. Sci. Comput., 16 (1995), pp. 1071–1081.

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

[12] W. Huang and D. M. Sloan, *Pole condition for singular problems: The pseudospectral approximation*, J. Comput. Phys., 107 (1993), pp. 254–261.

[13] W. Huang and T. Tang, *Pseudospectral solutions for steady motion of a viscous fluid inside a circular boundary*, Appl. Numer. Math., 33 (2000), pp. 167–173.

[14] A. Karageorghis and T. Tang, *A spectral domain decomposition approach for steady Navier–Stokes problems in circular geometries*, Comput. Fluids, 25 (1996), pp. 541–549.

[15] S. D. Kim and S. V. Parter, *Preconditioning Chebyshev spectral collocation by finite-difference operators*, SIAM J. Numer. Anal., 34 (1997), pp. 939–958.

[16] J. Li, H.-P. Ma, and W.-W. Sun, *Error analysis for solving the Korteweg–de Vries equation by a Legendre pseudo-spectral method*, Numer. Methods Partial Differential Equations, 16 (2000), pp. 513–534.

[17] H.-P. Ma and B.-Y. Guo, *The Chebyshev spectral method for Burgers-like equations*, J. Comput. Math., 6 (1988), pp. 48–53.

[18] H.-P. Ma and W.-W. Sun, *A Legendre–Petrov–Galerkin and Chebyshev collocation method for third-order differential equations*, SIAM J. Numer. Anal., 38 (2000), pp. 1425–1438.

[19] H.-P. Ma and W.-W. Sun, *Optimal error estimates of the Legendre–Petrov–Galerkin method for the Korteweg–de Vries equation*, SIAM J. Numer. Anal., 39 (2001), pp. 1380–1394.

[20] T. Matsushima and P. S. Marcus, *A spectral method for polar coordinates*, J. Comput. Phys., 120 (1995), pp. 365–374.

[21] S. A. Orszag, *Fourier series on spheres*, Mon. Weather Rev., 102 (1974), pp. 56–75.

[22] V. G. Priymak, *Pseudospectral algorithms for Navier–Stokes simulation of turbulent flows in cylindrical geometry with coordinate singularities*, J. Comput. Phys., 118 (1995), pp. 366–379.

[23] V. G. Priymak and T. Miyazaki, *Accurate Navier–Stokes investigation of transitional and turbulent flows in a circular pipe*, J. Comput. Phys., 142 (1998), pp. 370–411.

[24] J. Shen, *Efficient spectral-Galerkin methods* III: *Polar and cylindrical geometries*, SIAM J. Sci. Comput., 18 (1997), pp. 1583–1604.

[25] J. Shen, *Efficient spectral-Galerkin methods* IV: *Spherical geometries*, SIAM J. Sci. Comput., 20 (1999), pp. 1438–1455.

[26] D. J. Torres and E. A. Coutsias, *Pseudospectral solution of the two-dimensional Navier–Stokes equations in a disk*, SIAM J. Sci. Comput., 21 (1999), pp. 378–403.

[27] W. T. Verkley, *A pseudo-spectral model for two-dimensional incompressible flow in a circular basin,* I. *Mathematical formulation*, J. Comput. Phys., 136 (1997), pp. 100–114.

[28] W. T. Verkley, *A pseudo-spectral model for two-dimensional incompressible flow in a circular basin,* II. *Numerical examples*, J. Comput. Phys., 136 (1997), pp. 115–131.

SIAM J. NUMER. ANAL.
Vol. 41, No. 6, pp. 2350–2373

EXISTENCE VERIFICATION FOR HIGHER DEGREE SINGULAR ZEROS OF NONLINEAR SYSTEMS*

R. BAKER KEARFOTT[†] AND JIANWEI DIAN[‡]

Abstract. Finding approximate solutions to systems of n nonlinear equations in n real variables is a much studied problem in numerical analysis. Somewhat more recently, researchers have developed numerical methods to provide mathematically rigorous error bounds on such solutions. (We say that we "verify" existence of the solution within those bounds on the variables.) However, when the Jacobi matrix is singular at the solution, no computational techniques to verify existence can handle the general case. Nonetheless, computational verification that one or more solutions exists within a region in complex space containing the real bounds is possible by computing the topological degree. In a previous paper, we presented theory and algorithms for the simplest case, when the rank-defect of the Jacobian matrix at the solution is 1 and the topological index is 2. Here, we generalize that result to arbitrary topological index $d \geq 2$: We present theory, algorithms, and experimental results. We also present a heuristic for determining the degree, obtaining a value that we can subsequently verify with our algorithms. Although execution times are slow compared to corresponding bound verification processes for nonsingular systems, the order with respect to system size is still cubic.

Key words. complex nonlinear systems, interval computations, verified computations, singularities, topological degree

AMS subject classifications. 65G10, 65H10

DOI. 10.1137/S0036142901386057

1. Introduction.

Solution of linear and nonlinear systems of equations is a fundamental problem in numerical analysis, underlying much, if not most, of modern scientific computation. A system of n equations in n unknowns, where the expressions defining the system are defined in some closed, bounded subset $\mathbf{D}$ of n-dimensional space, may be expressed mathematically by

$$F(x) = 0, \quad F : \mathbf{D} \subset \mathbb{R}^n \to \mathbb{R}^n. \tag{1.1}$$

Throughout scientific computing, floating point arithmetic is used to solve equations (1.1) approximately. If F is linear, for example, then various direct (Gaussian elimination–based) methods, or iterative methods such as the preconditioned conjugate gradient method, are used. If F is nonlinear, then numerical solution of (1.1) involves various iterative methods, and the corresponding computer code can be sophisticated or involve numerous heuristics. In both the linear and nonlinear cases, the result of the computation is an approximate solution vector $\check{x} \in \mathbb{R}^n$, $F(\check{x}) \approx 0$. Hopefully, $\check{x}$ is near an exact or "true" solution x^*, $F(x^*) = 0$, such that $\|\check{x} - x^*\|$ is small.[1] However, with a few exceptions, the computation that produces the approximate solution $\check{x}$ does not give a bound on $\|\check{x} - x^*\|$. Indeed, it is not hard to find

*Received by the editors March 2, 2001; accepted for publication (in revised form) June 16, 2003; published electronically December 17, 2003. This work was partially supported by National Science Foundation grant DMS-9701540.

http://www.siam.org/journals/sinum/41-6/38605.html

[†]Department of Mathematics, University of Louisiana at Lafayette, Lafayette, LA 70504 (rbk@louisiana.edu).

[‡]Hewlett–Packard Company, 3000 Waterview Parkway, Richardson, TX 75080 (jianwei_dian@hp.com).

[1] Except when explicitly noted, we assume only that the norm is some fixed norm (independent of x), since the norm is discussed in terms of the order $\mathcal{O}(\|\cdot\|)$ and since we are working in finite-dimensional spaces.

instances of practical problems for which the output vector $\check{x}$ of an algorithm to solve a nonlinear system of the form (1.1) is not near a true solution at all, and for which the modeler does not recognize this fact; see, for example, [5].

On the other hand, efficient methods have been available for some time to construct bounds about such approximate solutions $\check{x}$ at which a true solution is known to exist. Specifically, an interval vector

$$\boldsymbol{x} = ([\underline{x}_1, \overline{x}_1], [\underline{x}_2, \overline{x}_2], \ldots, [\underline{x}_n, \overline{x}_n])^T \tag{1.2}$$

is found such that each width $\mathrm{w}([\underline{x}_i, \overline{x}_i]) = \overline{x}_i - \underline{x}_i$ is small (a small multiple of the machine precision, depending on the problem), and such that the computational process has proven mathematically (with no uncertainty due to roundoff error) that there is an exact solution $x^* \in \boldsymbol{x}$. Although it is not universally recognized within the general numerical analysis community, such methods can be developed to be practical more often than not and can give rigorous bounds that both are tighter than heuristic error estimates and are obtained with less effort; see, for example, [15]. An explanation of these methods appears in [6, 11, 17] and in numerous other works. The mathematical assumptions under which such verification methods can be expected to be successful are basically that the Jacobi matrix for the system is continuous and nonsingular at the solution; see the aforementioned references for a precise statement of the assumptions. For a practical implementation of such methods (with interval arithmetic), the function residuals and Jacobi matrix need to be representable as a computer program.

Although these verification methods involve interval arithmetic, notorious for impracticality due to overestimation when naively used, the intervals (the coordinates of $\boldsymbol{x}$) in a posteriori verification computations are small. It is known, from both theory and practice, that the overestimation in such small intervals is asymptotically insignificant, making such methods more generally applicable.

In this work, we consider not finding an approximate solution $\check{x}$ but constructing and verifying bounds $\boldsymbol{x}$ about such a point $\check{x}$ (however found) such that an exact solution x^* lies within $\boldsymbol{x}$. Specifically, we address the following problem.

(1.3)

> Given $F : \mathbf{D} \to \mathbb{R}^n$, where $\mathbf{D}$ is some closed, bounded subset of $\mathbb{R}^n$ with nonempty interior, and given an approximate solution $\check{x} \in \mathbf{D}$, construct bounds $\boldsymbol{x} \in \mathbb{IR}^n$, $\check{x} \in \boldsymbol{x}$, with $\boldsymbol{x}$ as in (1.2), for which we *rigorously* verify
>
> - there exists an $x^* \in \boldsymbol{x}$ such that $F(x^*) = 0$.

Throughout this paper, by "rigorous" we mean "with the same standard as for a traditional mathematical proof." Our algorithms for such verification will employ techniques derived from traditional floating point computations but will use directed roundings to take the finite nature of floating point arithmetic into account.

As is seen in [6, 11, 17] and elsewhere, when the Jacobian matrix $F'(x^*)$ is well-conditioned and not too quickly varying, interval computations have no trouble proving that there is a *unique* solution within small boxes with x^* reasonably near the center. (Various techniques, such as those in [16], can be used to initially construct the bounds over which the verification algorithm proceeds.) However, when $F'(x^*)$ is ill-conditioned or singular, in general, no computational techniques can verify the existence of a solution within a given region $\boldsymbol{x}$ of $\mathbb{R}^n$. Indeed, common thinking among researchers in such verification methods has been that verification is not possible in the singular case. Nonetheless, in [14] we introduced an algorithm for computational but rigorous verification, in the singular case, that a given number of true solutions exists within a region in complex space containing $\boldsymbol{x}$. There we studied the simplest case, when the rank-defect of the Jacobian matrix at the solution is one, and we developed and experimentally validated algorithms for the case when the topological index is 2. There, we also proved the special case of Theorem 3.1 (see section 3 below) when $d = 2$. Under the same assumptions as those in section 2 below, we developed specialized versions of the algorithms in section 4 below, and we presented varying-dimensional experimental results in [14].

We were surprised and pleased that the results in [14] could be generalized so easily. In particular, we developed an alternate simple, general proof for Theorem 3.1 below. Furthermore, the algorithms in section 4 below, although not taking advantage of special efficiencies in the degree-2 case as in [14], are similar in structure and have the same computational complexity as the algorithms in [14].

The developments below proceed by thinking of the function F in terms of a model of the form

$$F(x) = M(x) + R(x), \tag{1.4}$$

where $M(x)$ is a Taylor approximation to F about x^* and $R(x)$ is the error term. The number of solutions to $F(x) = 0$ is determined according to the topological degree (reviewed in section 1.2 below) of F. In Theorem 3.1 (see section 3 below) we show that, if $F(x) = M(x)$, where $M(x)$ has some verifiable properties, then the topological degree of F must equal d. (This proves existence, since the topological degree over a region in complex space is equal to the number of solutions in the region, counting multiplicities.) Basing the computations on the structure of M, we use a heuristic test to guess the integer d. Speaking roughly, we then take account of both roundoff error and the error term $R(x)$ with interval computations. In particular, we use the structure of M to efficiently arrange an exhaustive search that rigorously verifies that the topological degree actually is d. Even though the search is exhaustive, completion of the search requires only the same order of magnitude of computational work as a step of Newton's method on the system; this is due to the postulated structure of M and the way we have arranged the search.

As explained in [14, section 1.4], if d is even, it is meaningless to discuss the existence of a solution in $\mathbb{R}^n$ within the framework of errors in the data, model, and floating point system, since the topological degree in real space in such cases may be equal to 0. The even d case is a generalization of the situation with $f(x) = x^2$ at $x = 0$: The function f itself has a unique solution at $x = 0$, yet perturbations of f result in either no solutions or two solutions near $x = 0$. In contrast, $f(z) = (1+\epsilon_1)z^2+\epsilon_2 z+\epsilon_3$, $|\epsilon_i|$ small for $i = 1, 2, 3$, has two solutions, counting multiplicities, in all sufficiently large (but with diameters that can be chosen to be $\mathcal{O}(\epsilon)$ as $\epsilon \to 0$) open sets in $\mathbb{C}$ containing $z = 0$. This illustrates a general phenomenon: Whereas small perturbations of the data change the existence of (one or more) solutions near a particular point in $\mathbb{R}^n$, the solutions vary continuously with perturbations of complex extensions. We have presented one precise statement of this in Theorem 3.1 of [4]: Under the assumptions in that theorem (essentially, that the Jacobi matrix have rank defect 1 at the solution, and that certain derivative tensors up to order d vanish), $\mathrm{d}(F, \boldsymbol{x}, 0) = 0$ for a box $\boldsymbol{x} \in \mathbb{R}^n$ whenever d is even (and $\mathrm{d}(F, \boldsymbol{x}, 0) = \pm 1$ over such a box when d is odd). Thus, in that case when d is even (and, we believe, in many fairly general cases) verifying the value of the topological degree within the real context cannot verify existence of the solution. In contrast, $\mathrm{d}(F, \boldsymbol{z}, 0)$ must always be nonzero if there is a $z \in \boldsymbol{z}$ with $F(\boldsymbol{z}) = 0$ and $0 \notin \boldsymbol{\partial z}$, for $\boldsymbol{z} \subset \mathbb{C}^n$. Also, the *number* of solutions counting multiplicities can change under perturbations in $\mathbb{R}^n$, even for odd-order functions such as x^3, for which we can use techniques such as those in [4] to prove existence; in contrast, $\mathrm{d}(F, \boldsymbol{z}, 0)$ gives the exact number of solutions within $\boldsymbol{z}$, counting multiplicities, for complex-valued functions F of complex variables $\boldsymbol{z}$.

In this paper, we consider the case of general d, to verify the existence of solutions in small neighborhoods of $\mathbb{C}^n$, as illustrated in problem (1.5) below. Our hope is that such verification will be useful in analysis of systems having even-order roots, even though the validation is in a different space; in any case, rigorous validation of such systems in the original space may not be possible and may not be meaningful if the system was derived from measurements with errors. (We present special theory, analysis, and algorithms for the real case and odd-order roots in [4].)

(1.5) Given F, $\mathbf{D}$, and $\check{x}$ as in problem (1.3), consider an analytic extension $\tilde{F}$ of F to a domain $\tilde{\mathbf{D}} \subseteq \mathbb{C}^n$, $\mathbf{D} \subset \tilde{\mathbf{D}}$. Construct bounds $\boldsymbol{x}$ as in problem (1.3) and $\boldsymbol{y} = ([\underline{y}_1, \overline{y}_1], \ldots, [\underline{y}_n, \overline{y}_n])$, $0 \in \boldsymbol{y}$, for which we *rigorously* verify the following:

- there exists a $z^* \in \boldsymbol{z}$ such that $\tilde{F}(z^*) = 0$, where
- $\boldsymbol{z} = \big\{(x_1 + iy_1, x_2 + iy_2, \ldots, x_n + iy_n)^T \in \mathbb{C}^n \;\big|\; x_j \in \boldsymbol{x}_j,\, y_j \in \boldsymbol{y}_j,\, 1 \le j \le n\big\}$.

Hiding detail and revealing overall ideas, we have simplified the notation in this work, compared to that in [14].

After introducing our notation in section 1.1, we briefly review the relevant portions of topological degree theory in sections 1.2 and 1.3. We introduce our use of the structure of the model $M(x)$ in section 2. We present our scheme for setting the coordinate bounds $\boldsymbol{x}$ within which we prove the existence of solutions in section 2.2; though related, this scheme is improved and works more generally than that in [14]. In section 3, we show that the degree must be equal to d if $F(x) = M(x)$ (i.e., $R(x) = 0$), within the context introduced in section 2. In section 4, we present the algorithm that verifies that the degree is d for nonzero $R(x)$, within the framework introduced in section 2. In section 5, we present an easily implemented heuristic computation for guessing the value of d, necessary for the verification algorithm in section 4. Finally, in section 6.3 we present results of trying the computations on several examples; these results illustrate that the algorithm can be practical for a variety of problems, and that the computation does not necessarily increase rapidly with the dimension of the problem.

1.1. Notation.

We assume familiarity with the fundamentals of interval arithmetic; see [1, 6, 11, 17, 19] for introductory material.

Throughout, scalars and vectors will be denoted by lower case, while matrices will be denoted by upper case. Intervals, interval vectors (also called "boxes"), and interval matrices will be denoted by boldface. For instance, $\boldsymbol{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ denotes an interval vector, $A = (a_{i,j})$ denotes a point matrix, and $\boldsymbol{A} = (\boldsymbol{a}_{i,j})$ denotes an interval matrix. The midpoint of an interval or interval vector $\boldsymbol{x}$ will be denoted by $\mathrm{m}(\boldsymbol{x})$. As in section 1, $\mathrm{w}(\boldsymbol{x})$ denotes the width of an interval $\boldsymbol{x} = [\underline{x}, \overline{x}]$, that is, $\mathrm{w}(\boldsymbol{x}) = \overline{x} - \underline{x}$; if $\boldsymbol{x}$ represents an interval vector, then the midpoint $\mathrm{m}(\boldsymbol{x})$ and width $\mathrm{w}(\boldsymbol{x})$ will be real vectors, understood componentwise. Real n-space will be denoted by $\mathbb{R}^n$, while complex n-space will be denoted by $\mathbb{C}^n$. The set of real interval vectors will be denoted by $\mathbb{IR}^n$, while the set of complex interval vectors will be denoted by $\mathbb{IC}^n$.

Suppose $\boldsymbol{x} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)$ is an n-dimensional real box, where $\boldsymbol{x}_k = [\underline{x}_k, \overline{x}_k]$. The nonoriented boundary of $\boldsymbol{x}$, denoted by $\boldsymbol{\partial x}$, consists of $2n$ $(n-1)$-dimensional real boxes

$$\boldsymbol{x}_{\underline{k}} \equiv (\boldsymbol{x}_1, \dots, \boldsymbol{x}_{k-1}, \underline{x}_k, \boldsymbol{x}_{k+1}, \dots, \boldsymbol{x}_n) \quad \text{and} \quad \boldsymbol{x}_{\overline{k}} \equiv (\boldsymbol{x}_1, \dots, \boldsymbol{x}_{k-1}, \overline{x}_k, \boldsymbol{x}_{k+1}, \dots, \boldsymbol{x}_n),$$

where $k = 1, \dots, n$.

The *orientation* of a region $\mathbf{D} \subset \mathbb{R}^n$ and of its boundary $\boldsymbol{\partial}\mathbf{D}$ is a generalization of the concept of orientation of a region and its boundary (counterclockwise being positive orientation) in complex analysis, or of the concepts of orientation of a region and its boundary when applying Green's theorem or Stokes' theorem; see [3, pp. 4–10] or [2], for example, for a detailed formal definition. In particular, a simplex $\langle a^{(0)}, a^{(1)}, \dots, a^{(n)} \rangle$, $a^{(k)} \in \mathbb{R}^n$, $0 \le k \le n$, is positively oriented, provided that a certain determinant formed from the coordinates of the points $a^{(k)}$ is positive, and is negatively oriented if that determinant is negative; polygonal regions formed by juxtaposing such oriented simplexes have a positive orientation, provided that each component simplex is positively oriented.

To explain the algorithms in this paper, we need concern ourselves only with the derived orientation of the boundary of an interval vector (i.e., of a box) $\boldsymbol{x}$. The following "definition" can be derived as a theorem from the general definition of a positively oriented polygonal region. (For a more detailed presentation, see our technical report [13, pp. 7–8].)

Definition 1.1. *Suppose that a box $\boldsymbol{x}$ as in* (1.2) *is positively oriented. Then the positively oriented boundary $b(\boldsymbol{x})$ is given by the formal sum*

$$\sum_{k=1}^{n} \left\{ (-1)^k \boldsymbol{x}_{\underline{k}} + (-1)^{k+1} \boldsymbol{x}_{\overline{k}} \right\}$$

of the $2n$ $(n-1)$-dimensional boxes $\boldsymbol{x}_{\underline{k}}$ and $\boldsymbol{x}_{\overline{k}}$.

Our model $M(x)$ of $F(x)$ as in (1.4) is a multivariate Taylor polynomial. In particular we will write a component f_i of F as

$$f_i(x) = f_i(\check{x}) + \sum_{j=1}^{d} \frac{1}{j!} D^j f_i(\check{x})[x-\check{x}, \dots, x-\check{x}] + \mathcal{O}\left(\|x-\check{x}\|\right)^{d+1}, \tag{1.6}$$

where

$$D^j f_i(\check{x})[x-\check{x}, \dots, x-\check{x}] = \sum_{k_1=1}^{n} \cdots \sum_{k_j=1}^{n} \frac{\partial^j f_i}{\partial x_{k_1} \cdots \partial x_{k_j}}(\check{x})(x_{k_1} - \check{x}_{k_1}) \cdots (x_{k_j} - \check{x}_{k_j}) \tag{1.7}$$

is the jth derivative tensor.

In our verification algorithms, the domains will be interval vectors, i.e., rectangular boxes $\boldsymbol{x}$. However, we state some of the known topological degree theory results more generally, in terms of the closed, bounded set $\mathbf{D}$ with nonempty interior that we introduced above.

1.2. Formulas from degree theory. In [14], we reviewed the topological degree in the context of this paper. Also see [2, 3, 8, 9, 18, 20]. Here, we repeat several properties used in the proofs in subsequent sections.

Although a formal definition of the topological degree is somewhat cumbersome, one obtains an intuitive understanding of the topological degree from its properties. In particular, for $n = 1$, the topological degree of F at 0 over an interval $\boldsymbol{x}$, denoted $\mathrm{d}(F, \boldsymbol{x}, 0)$, is the number of times the graph of F crosses the x-axis in the positive direction, minus the number of times the graph of F crosses the x-axis in the negative direction. If $F : \mathbb{C} \to \mathbb{C}$ and $\mathbf{D}$ is a simply connected region (a region without holes, such as a disk) containing the origin in $\mathbb{C}$, then the topological degree $\mathrm{d}(F, \mathbf{D}, 0)$ is equal to the winding number of F with respect to the curve bounding $\mathbf{D}$. Because of this fact, $\mathrm{d}(p_d, \mathbf{D}, 0) = d$, where p_d is any polynomial of degree d and $\mathbf{D}$ is any sufficiently large simply connected domain in $\mathbb{C}$ with $0 \in \mathbf{D}$ (where the size depends on the particular p_d). Thus, in $\mathbb{C}$, the topological degree roughly corresponds to the notion of algebraic degree, which is the same as the number of solutions, counting multiplicity. If we think of $\mathbf{D}$ as being the closure of a very small region containing a solution z^*, $F(z^*) = 0$, then $\mathrm{d}(F, \mathbf{D}, 0)$ is termed the *topological index* of z^*; the topological index corresponds to the multiplicity of z^*. For example, the topological index of z^d at $z^* = 0$ is equal to d. In this paper, we prove the existence of solutions within small domains $\mathbf{D}$ by verifying, essentially, that the topological index is nonzero.

Formal definitions of the topological degree can be given analytically (in terms of an integral) as in [18, Chapter 6], or in terms of fundamental concepts of algebraic topology, as in [2]. In either case, either definition can be obtained as a theorem, starting with the other one as the definition. We can actually think of the degree in terms of the following.

THEOREM 1.2 (see [18, p. 150]). *Suppose that F is continuous, and suppose that the Jacobian matrix $F'(x)$ is defined and nonsingular at each zero of F within a domain $\mathbf{D}$, which is the closure of an open region in $\mathbb{R}^n$, and suppose that $F(x) \neq 0$ when $x \in \boldsymbol{\partial}\mathbf{D}$. Then, the degree* $\mathrm{d}(F, \mathbf{D}, 0)$ *is equal to the number of zeros of F at which the determinant of the Jacobian matrix $F'(x)$ is positive, minus the number of zeros of F at which the determinant of the Jacobian matrix $F'(x)$ is negative.*

Basically, Theorem 1.2 states that the degree is an algebraic number of zeros of F in $\mathbf{D}$ when the Jacobian matrix is nonsingular at each zero. However, the degree does not change as F is perturbed, and we can imagine the degree remaining defined as two or more zeros of F coalesce into a single zero at which the Jacobian matrix is singular (important in our context here). Similarly, F need only be continuous (not necessarily differentiable) for $\mathrm{d}(F, \mathbf{D}, 0)$ to be defined. To define the degree for arbitrary continuous functions that do not vanish on the boundary $\boldsymbol{\partial}\mathbf{D}$, the analytic definition as in [18, Chapter 6] uses an integral and mollifying functions, whereas the topological definition approximates the image of the boundary $\boldsymbol{\partial}\mathbf{D}$ with a piecewise-linear simplicial complex (similar to how engineers approximate an object with triangles for the finite element method, except that the topologist's simplicial complex is oriented).

Starting either from the analytical definition of [18] or from the algebraic-topological definition of [2], we obtain the following properties of the degree. These properties are what will concern us in our verification procedures.

THEOREM 1.3 (see [18, p. 150]). *Let F, $G : \mathbf{D} \subset \mathbb{R}^n \to \mathbb{R}^n$ be two continuous functions that do not vanish on $\boldsymbol{\partial}D$. If $F(x) = G(x)$ for $x \in \boldsymbol{\partial}\mathbf{D}$, then* $\mathrm{d}(F, \mathbf{D}, 0) = \mathrm{d}(G, \mathbf{D}, 0)$.

Theorem 1.3 states one of the most important properties of degree: The degree depends only on the function values on the boundary.

THEOREM 1.4 (see [18, p. 157]). *Let F, $G : \mathbf{D} \subset \mathbb{R}^n \to \mathbb{R}^n$ be two continuous functions. If*

$$0 \notin \{tF(x) + (1-t)G(x) | x \in \boldsymbol{\partial}\mathbf{D} \text{ and } t \in [0, 1]\},$$

then

$$\mathrm{d}(F, \mathbf{D}, 0) = \mathrm{d}(G, \mathbf{D}, 0).$$

Theorem 1.4 is the famous Poincaré–Bohl theorem. It is a particular case of the homotopy invariant property of the topological degree. Since $\mathbf{D}$ is compact, this homotopy invariance implies, without too much argument, that the degree is a continuous function of F.

COROLLARY 1.5. *Suppose $F : \mathbf{D} \subset \mathbb{R}^n \to \mathbb{R}^n$ is continuous and $\mathrm{d}(F, \mathbf{D}, 0) = d$. Then there is an $\epsilon > 0$ such that, for all continuous $G : \mathbf{D} \to \mathbb{R}^n$ with $|F(x) - G(x)| < \epsilon$ for $x \in \mathbf{D}$, $\mathrm{d}(F, \mathbf{D}, 0) = \mathrm{d}(G, \mathbf{D}, 0)$.*

Suppose $F : \mathbf{D} \subset \mathbb{C}^n \to \mathbb{C}^n$ is analytic, and view the real and imaginary components of F and its argument $z \in \mathbb{C}^n$ as real components in $\mathbb{R}^{2n}$. Let $z = x + iy$ and $F(z) = u(x, y) + iv(x, y)$, where $x = (x_1, \ldots, x_n)$, $y = (y_1, \ldots, y_n)$, $u(x, y) = (u_1(x, y), \ldots, u_n(x, y))$, and $v(x, y) = (v_1(x, y), \ldots, v_n(x, y))$. We define $\tilde{\mathbf{D}}$ by

$$\tilde{\mathbf{D}} \equiv \{(x_1, y_1, \ldots, x_n, y_n) | (x_1 + iy_1, \ldots, x_n + iy_n) \in \mathbf{D}\}$$

and $\tilde{F} : \tilde{\mathbf{D}} \subset \mathbb{R}^{2n} \to \mathbb{R}^{2n}$ by $\tilde{F} = (u_1, v_1, \ldots, u_n, v_n)$. We then have the following properties.

THEOREM 1.6 (see [14]). *Suppose that $F : \mathbf{D} \subset \mathbb{C}^n \to \mathbb{C}^n$ is analytic, with $F(z) \neq 0$ for any $z \in \boldsymbol{\partial}\mathbf{D}$, and suppose that $\tilde{\mathbf{D}}$ and $\tilde{F} : \tilde{\mathbf{D}} \to \mathbb{R}^{2n}$ are defined as above. Then $\mathrm{d}(\tilde{F}, \tilde{\mathbf{D}}, 0)$ is nonnegative and is equal to the number of solutions $z^* \in \mathbf{D}$, $F(z^*) = 0$, counting multiplicities.*

1.3. A basic degree computation formula. Theorem 1.7 below relates the basic theory of the topological degree to the computational verification procedures in section 4 below. Theorem 1.7 is similar to Theorem 2.5 of [14]. We can obtain Theorem 1.7 from formulas (4.12) and (4.14) in [20], by taking into account the orientations of the faces of $\boldsymbol{x}$.

Theorem 1.7 characterizes $\mathrm{d}(F, \boldsymbol{x}, 0)$ in terms of certain components of F on $\boldsymbol{\partial x}$. In particular, set

$$F_{\neg k}(\boldsymbol{x}) \equiv \big(f_1(\boldsymbol{x}), \ldots, f_{k-1}(\boldsymbol{x}), f_{k+1}(\boldsymbol{x}), \ldots, f_n(\boldsymbol{x})\big).$$

Then we have the following result.

THEOREM 1.7. *Let $s \in \{-1, 1\}$ be fixed arbitrarily, suppose $F \neq 0$ on $\boldsymbol{\partial x}$, and suppose that there is a p, $1 \leq p \leq n$, such that*

1. *$F_{\neg p} \equiv (f_1, \ldots, f_{p-1}, f_{p+1}, \ldots, f_n) \neq 0$ on $\boldsymbol{\partial x}_{\underline{k}}$ or $\boldsymbol{\partial x}_{\overline{k}}$, $k = 1, \ldots, n$; and*
2. *the Jacobi matrices of $F_{\neg p}$ are nonsingular at all solutions of $F_{\neg p} = 0$ on $\boldsymbol{\partial x}$ and are continuous in a neighborhood of such solutions.*

Then

$$\begin{aligned}\mathrm{d}(F, \boldsymbol{x}, 0) \ &= (-1)^{p-1} s \left\{ \sum_{k=1}^{n} (-1)^k \sum_{\substack{x \in \boldsymbol{x}_{\underline{k}} \\ F_{\neg p}(x) = 0 \\ \mathrm{sgn}(f_p(x)) = s}} \mathrm{sgn} \left| \frac{\partial F_{\neg p}}{\partial x_1 x_2 \cdots x_{k-1} x_{k+1} \cdots x_n}(x) \right| \right. \\ &\qquad \left. + \sum_{k=1}^{n} (-1)^{k+1} \sum_{\substack{x \in \boldsymbol{x}_{\overline{k}} \\ F_{\neg p}(x) = 0 \\ \mathrm{sgn}(f_p(x)) = s}} \mathrm{sgn} \left| \frac{\partial F_{\neg p}}{\partial x_1 x_2 \cdots x_{k-1} x_{k+1} \cdots x_n}(x) \right| \right\}.\end{aligned}$$

2. Assumptions and choice of box. In this section, we present the basic assumptions. We also introduce how we choose the coordinate bounds $\boldsymbol{x}_i = [\underline{x}_i, \overline{x}_i]$ to satisfy the assumptions and enable more efficient algorithms. When the rank of $F'(x^*)$ is $n-p$ for some $p>0$, an appropriate preconditioner can be used to reduce $\boldsymbol{F'}(\boldsymbol{x})$ to approximately the pattern shown in Figure 2.1. (See [11] and [14] for details on preconditioning.)

$$Y\boldsymbol{F'}(\boldsymbol{x}) \approx \begin{pmatrix} 1 & 0 & \cdots & 0 & \overbrace{*\cdots*}^{p} \\ 0 & 1 & 0\cdots & 0 & *\cdots* \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & 0 & 1 & *\cdots* \\ 0 & \cdots & 0 & 0 & 0\cdots 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & 0 & 0\cdots 0 \end{pmatrix}.$$

FIG. 2.1. *An approximate form for a preconditioned singular interval system of approximate rank $n-p$, where "$*$" represents a nonzero element.*

In the analysis to follow, we assume that the system has already been preconditioned, so that it is, to within second-order terms with respect to $\mathrm{w}(\boldsymbol{x})$, of the form in Figure 2.1. That is, we assume that the preconditioned system is of the form seen in Figure 2.1 if we interpret "$*$" to represent any interval, "1" to represent intervals of the form $[1-\mathcal{O}(\|x-x^*\|), 1+\mathcal{O}(\|x-x^*\|)]$, and "0" to represent intervals of the form $[-\mathcal{O}(\|x-x^*\|), \mathcal{O}(\|x-x^*\|)]$. Here as in [14], we concentrate on the case $p=1$.

2.1. The basic assumptions. As in the special case $d=2$ of [14], we assume

1. $F : \mathbf{D} \subset \mathbb{R}^n \to \mathbb{R}^n$ can be extended to an analytic function in $\mathbb{C}^n$.
2. $\boldsymbol{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) = ([\underline{x}_1, \overline{x}_1], \ldots, [\underline{x}_n, \overline{x}_n])$ is a small box constructed to be centered at an approximate solution $\check{x}$, i.e., $\mathrm{m}(\boldsymbol{x}) = (\check{x}_1, \ldots, \check{x}_n)$.
3. $\check{x}$ is near a point x^* with $F(x^*)=0$ such that $\|\check{x}-x^*\|$ is much smaller than the norm of the width of the box $\boldsymbol{x}$, and the width of the box $\boldsymbol{x}$ is small enough that mean value interval extensions lead, after preconditioning, to a system like Figure 2.1, with small intervals replacing the zeros.
4. F has been preconditioned as in Figure 2.1, and $F'(x^*)$ has null space of dimension 1.

Define

$$\begin{aligned} \alpha_k &\equiv \frac{\partial f_k}{\partial x_n}(\check{x}), \qquad 1 \le k \le n-1, \\ \alpha_n &\equiv -1, \\ \Delta_1 &\equiv \left| \frac{\partial F}{\partial x_1 \cdots \partial x_n}(\check{x}) \right| \\ \Delta_l &\equiv \sum_{k_1=1}^{n} \cdots \sum_{k_l=1}^{n} \frac{\partial^l f_n}{\partial x_{k_1} \cdots \partial x_{k_l}}(\check{x}) \alpha_{k_1} \cdots \alpha_{k_l}, \qquad 2 \le l. \end{aligned}$$

The following representation of $F(x)$ near $\check{x}$ is appropriate under these assumptions:

$$\begin{aligned}
&(2.1) \qquad f_k(x) = (x_k - \check{x}_k) + \alpha_k(x_n - \check{x}_n) + \mathcal{O}\left(\|x - \check{x}\|\right)^2 \\
&\qquad\qquad \text{for } 1 \le k \le n-1, \\
&(2.2) \qquad f_n(x) = \sum_{\ell=2}^{d} \frac{1}{\ell!} D^\ell f_n(\check{x})[x-\check{x}, \ldots, x-\check{x}] + \mathcal{O}\left(\|x-\check{x}\|\right)^{d+1}.
\end{aligned}$$

Here and below, "d" is a fixed constant that represents the postulated topological index (obtained, say, with the heuristic in section 5 below); the index d will be verified with our proposed algorithms (in section 4 below).

We now introduce additional notation to describe the complex extensions. For $F : \mathbb{R}^n \to \mathbb{R}^n$, extend F to complex space: $x+iy$, with y in a small box $\boldsymbol{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n) = ([\underline{y}_1, \overline{y}_1], \ldots, [\underline{y}_n, \overline{y}_n])$, where $\boldsymbol{y}$ is centered at $(0, \ldots, 0)$. As in Theorem 1.6 above, define $\tilde{\boldsymbol{x}} \equiv (\boldsymbol{x}, \boldsymbol{y}) \equiv (\boldsymbol{x}_1, \boldsymbol{y}_1, \ldots, \boldsymbol{x}_n, \boldsymbol{y}_n) = ([\underline{x}_1, \overline{x}_1], [\underline{y}_1, \overline{y}_1], \ldots, [\underline{x}_n, \overline{x}_n], [\underline{y}_n, \overline{y}_n])$, $u_k(x,y) \equiv \Re(f_k(x+iy))$ and $v_k(x,y) \equiv \Im(f_k(x+iy))$. With this, define

$$\tilde{F}(x,y) \equiv (u_1(x,y), v_1(x,y), \ldots, u_n(x,y), v_n(x,y)) : \mathbb{R}^{2n} \to \mathbb{R}^{2n}.$$

Also define

$$\tilde{F}_{\neg u_n}(x,y) \equiv \big(u_1(x,y), v_1(x,y), \ldots, u_{n-1}(x,y), v_{n-1}(x,y), v_n(x,y)\big).$$

Then, based on (2.1) and (2.2), for $1 \le k \le (n-1)$,

$$(2.3) \qquad \left.\begin{array}{rcl} u_k(x,y) & = & (x_k - \check{x}_k) + \alpha_k(x_n - \check{x}_n) \\ & & + \, \mathcal{O}\left(\|(x-\check{x}, y)\|\right)^2, \\ v_k(x,y) & = & y_k + \alpha_k y_n + \mathcal{O}\left(\|(x-\check{x}, y)\|\right)^2, \end{array}\right\}$$

or

$$(2.4) \qquad \left.\begin{array}{rcl} u_k(x,y) & \approx & (x_k - \check{x}_k) + \alpha_k(x_n - \check{x}_n), \\ v_k(x,y) & \approx & y_k + \alpha_k y_n. \end{array}\right\}$$

2.2. Choosing the coordinate bounds. In our verification algorithms below, we drastically reduce the amount of computation required by astutely choosing the ratios of coordinate widths of the boxes $\boldsymbol{x}$ and $\boldsymbol{y}$. We will use a scheme similar to that of section 5 of [14]. In particular, having defined $\boldsymbol{x}_{\underline{k}}$ and $\boldsymbol{x}_{\overline{k}}$ in section 1.1, we define $\boldsymbol{y}_{\underline{k}}$ and $\boldsymbol{y}_{\overline{k}}$ similarly:

$$\begin{aligned}
&\boldsymbol{y}_{\underline{k}} \equiv (\boldsymbol{x}_1, \boldsymbol{y}_1, \ldots, \boldsymbol{x}_{k-1}, \boldsymbol{y}_{k-1}, \boldsymbol{x}_k, \underline{y}_k, \boldsymbol{x}_{k+1}, \boldsymbol{y}_{k+1}, \ldots, \boldsymbol{x}_n, \boldsymbol{y}_n) \quad \text{and} \\
&\boldsymbol{y}_{\overline{k}} \equiv (\boldsymbol{x}_1, \boldsymbol{y}_1, \ldots, \boldsymbol{x}_{k-1}, \boldsymbol{y}_{k-1}, \boldsymbol{x}_k, \overline{y}_k, \boldsymbol{x}_{k+1}, \boldsymbol{y}_{k+1}, \ldots, \boldsymbol{x}_n, \boldsymbol{y}_n).
\end{aligned}$$

To compute the degree $\mathrm{d}(\tilde{F}, \tilde{\boldsymbol{x}}, 0)$, we will consider $\tilde{F}_{\neg u_n}$ on the boundary of $\tilde{\boldsymbol{x}}$. This boundary consists of the $4n$ faces $\boldsymbol{x}_{\underline{1}}$, $\boldsymbol{x}_{\overline{1}}$, $\boldsymbol{y}_{\underline{1}}$, $\boldsymbol{y}_{\overline{1}}$, …, $\boldsymbol{x}_{\underline{n}}$, $\boldsymbol{x}_{\overline{n}}$, $\boldsymbol{y}_{\underline{n}}$, $\boldsymbol{y}_{\overline{n}}$. We will set $\boldsymbol{x}_n$ and $\boldsymbol{y}_n$ so that the coordinate widths $\mathrm{w}(\boldsymbol{x}_k)$ obey

$$(2.5) \qquad \mathrm{w}(\boldsymbol{x}_n) \le \frac{1}{2} \min_{1 \le k \le n-1} \left\{ \frac{\mathrm{w}(\boldsymbol{x}_k)}{|\alpha_k|} \right\} \quad \text{and} \quad \mathrm{w}(\boldsymbol{y}_n) \le \frac{1}{2} \min_{1 \le k \le n-1} \left\{ \frac{\mathrm{w}(\boldsymbol{y}_k)}{|\alpha_k|} \right\}.$$

In the above two relationships, when $\alpha_k = 0$ for some k, that particular k can be ignored in obtaining the minima, and $\mathrm{w}(\boldsymbol{x}_k)$ and $\mathrm{w}(\boldsymbol{y}_k)$ can be set to any small positive values as long as the assumptions in section 2.1 are met. If $\alpha_k = 0$ for $k = 1, \ldots, n-1$, then $\mathrm{w}(\boldsymbol{x}_k)$ and $\mathrm{w}(\boldsymbol{y}_k)$, $k = 1, \ldots, n$, can independently be set to any small positive values, as long as the assumptions in section 2.1 are met.

Constructing the box widths this way will make it unlikely that $u_k(x,y) = 0$ on either $\boldsymbol{x}_{\underline{k}}$ or $\boldsymbol{x}_{\overline{k}}$ and unlikely that $v_k(x,y) = 0$ on either $\boldsymbol{y}_{\underline{k}}$ or $\boldsymbol{y}_{\overline{k}}$, for $k = 1, \ldots, n-1$. This, in turn, will allow us to replace searches on $4n-4$ of the $4n$ faces of $\boldsymbol{\partial\tilde{x}}$ by simple interval evaluations, reducing the total computational cost dramatically. See [14] for details.

A difference between the scheme used here and that of [14] is the way the ratio $\mathrm{w}(\boldsymbol{y}_n)/\mathrm{w}(\boldsymbol{x}_n)$ is chosen. In [14], $\mathrm{w}(\boldsymbol{y}_n)$ was chosen large relative to $\boldsymbol{x}_n$, to arrange no solutions of $u_n = 0$ on $\boldsymbol{y}_{\underline{n}}$ and $\boldsymbol{y}_{\overline{n}}$. When the degree is odd, that is not possible, and we have found the strategy represented by formula (4.1) below, implying $\mathrm{w}(\boldsymbol{y}_n)$ small relative to $\mathrm{w}(\boldsymbol{x}_n)$, as in Figure 4.1 below, to be more convenient.

3. When the polynomial model is exact.

In [14] we proved that, under the assumptions in section 2, if the $\mathcal{O}\left(\|x-\check{x}\|\right)^2$ term is absent in (2.1) and the $\mathcal{O}\left(\|x-\check{x}\|\right)^{d+1}$ term is absent in (2.2) with $d = 2$, and if $\Delta_1 = 0$ but $\Delta_2 \neq 0$, then $\mathrm{d}(\tilde{F}, \tilde{\boldsymbol{x}}, 0) = 2$. Here, we generalize that result to $\Delta_1 = \cdots = \Delta_{d-1} = 0$, $\Delta_d \neq 0$. Since the degree doesn't change under small perturbations of the function $\tilde{F}$ (see Theorem 1.5 above), the conclusion in Theorem 3.1 below also holds for more general continuous functions for which the $\mathcal{O}\left(\|x-\check{x}\|\right)^2$ and $\mathcal{O}\left(\|x-\check{x}\|\right)^{d+1}$ terms are not absent but are sufficiently small. In our computational existence verification algorithm in the next section, we use interval arithmetic to rigorously encompass the $\mathcal{O}\left(\|x-\check{x}\|\right)^2$ and $\mathcal{O}\left(\|x-\check{x}\|\right)^{d+1}$ terms. In this way, Theorem 3.1 below provides guidance for construction of our general algorithm.

THEOREM 3.1. *Suppose that*

1. $\tilde{\boldsymbol{x}}$ *is a nondegenerate box in* $\mathbb{R}^{2n}$ *as defined in section* 2;
2. $(\check{x}, \check{y}) = (\check{x}_1, \check{y}_1, \ldots, \check{x}_n, \check{y}_n)$ *is the midpoint of* $\tilde{\boldsymbol{x}}$*;*
3. F *and* $\tilde{F}$ *are as in section* 2;
4. F *is such that the* $\mathcal{O}\left(\|x-\check{x}\|\right)^2$ *and* $\mathcal{O}\left(\|x-\check{x}\|\right)^{d+1}$ *terms in* (2.1) *and* (2.2) *are absent; and*
5. $\Delta_1 = \cdots = \Delta_{d-1} = 0$, $\Delta_d \neq 0$, *where* $2 \leq d$.

Then $\mathrm{d}(\tilde{F}, \tilde{\boldsymbol{x}}, 0) = d$.

In contrast to the proof in [14], we use a homotopy argument to prove Theorem 3.1.

Proof. Let $z = (z_1, \ldots, z_n) = (x_1 + iy_1, \ldots, x_n + iy_n)$. Then

$$F(z) = (f_1(z), \ldots, f_{n-1}(z), f_n(z)),$$

where

$$\begin{aligned} f_k(z) &= (z_k - \check{z}_k) + \frac{\partial f_k}{\partial x_n}(\check{x})(z_n - \check{z}_n) \\ &= (z_k - \check{z}_k) + \alpha_k(z_n - \check{z}_n) \\ &\qquad \text{for } 1 \leq k \leq n-1, \end{aligned}$$

$$f_n(z) = \sum_{\ell=2}^{d} \frac{1}{\ell!} D^\ell f_n(\check{x})[z - \check{z}, \ldots, z - \check{z}]. \tag{3.1}$$

We construct $A : \mathbb{C}^n \to \mathbb{C}^n$ by

$$A(z) = (a_1(z), \ldots, a_{n-1}(z), a_n(z)),$$

where

$$\begin{aligned} a_k(z) &= (z_k - \check{z}_k) + \alpha_k(z_n - \check{z}_n) \quad \text{for } 1 \le k \le n-1, \\ a_n(z) &= \frac{(-1)^d \Delta_d}{d!}(z_n - \check{z}_n)^d. \end{aligned} \tag{3.2}$$

Let $r_k(x, y) \equiv \Re(a_k(x + iy))$ and $s_k(x, y) \equiv \Im(a_k(x + iy))$. With this, define $\tilde{A} : \mathbb{R}^{2n} \to \mathbb{R}^{2n}$ by

$$\tilde{A}(x, y) \equiv (r_1(x, y), s_1(x, y), \ldots, r_n(x, y), s_n(x, y)).$$

We construct $G : \mathbb{C}^n \to \mathbb{C}^n$ by

$$G(z) = (g_1(z), \ldots, g_{n-1}(z), g_n(z)),$$

where

$$\begin{aligned} g_k(z) &= (z_k - \check{z}_k) \quad \text{for } 1 \le k \le n-1, \\ g_n(z) &= \frac{(-1)^d \Delta_d}{d!}(z_n - \check{z}_n)^d. \end{aligned} \tag{3.3}$$

Let $p_k(x, y) \equiv \Re(g_k(x + iy))$ and $q_k(x, y) \equiv \Im(g_k(x + iy))$. With this, define $\tilde{G} : \mathbb{R}^{2n} \to \mathbb{R}^{2n}$ by

$$\tilde{G}(x, y) \equiv (p_1(x, y), q_1(x, y), \ldots, p_n(x, y), q_n(x, y)).$$

We will prove $\mathrm{d}(\tilde{F}, \tilde{\boldsymbol{x}}, 0) = \mathrm{d}(\tilde{A}, \tilde{\boldsymbol{x}}, 0) = \mathrm{d}(\tilde{G}, \tilde{\boldsymbol{x}}, 0)$. First, we prove $\mathrm{d}(\tilde{F}, \tilde{\boldsymbol{x}}, 0) = \mathrm{d}(\tilde{A}, \tilde{\boldsymbol{x}}, 0)$.

Define

$$\begin{aligned} \tilde{H}_1((x, y), t) &\equiv t\tilde{F}(x, y) + (1 - t)\tilde{A}(x, y) \\ \text{and} \qquad H_1(z, t) &\equiv tF(z) + (1 - t)A(z). \end{aligned}$$

We will prove that $\tilde{H}_1((x, y), t) \ne 0$ when $(x, y) \in \partial\tilde{\boldsymbol{x}}$ and $t \in [0, 1]$. It is clear that $\tilde{H}_1((x, y), t) = 0$ is equivalent to $H_1(z, t) = 0$, so we consider $H_1(z, t)$. The definition of H_1 and some rearrangement of terms give

$$\begin{aligned} H_1(z, t) = \Big(&(z_1 - \check{z}_1) + \alpha_1(z_n - \check{z}_n), \ldots, (z_{n-1} - \check{z}_{n-1}) + \alpha_{n-1}(z_n - \check{z}_n), \\ &tf_n(z) + (1 - t)\frac{(-1)^d \Delta_d}{d!}(z_n - \check{z}_n)^d\Big). \end{aligned}$$

Thus, $H_1(z, t) = 0$ implies $z_k = \check{z}_k - \alpha_k(z_n - \check{z}_n)$ for $k = 1, \ldots, n - 1$. By definition, $\alpha_n = -1$, and thus $z_n = \check{z}_n - \alpha_n(z_n - \check{z}_n)$. Substituting $z_k - \check{z}_k = -\alpha_k(z_n - \check{z}_n)$ for each such k ($k = 1, 2, \ldots, n$) in the derivative tensor evaluation $D^\ell f_n(\check{x})[z - \check{z}, \ldots, z - \check{z}]$ in (3.1), we obtain

$$D^\ell f_n(\check{x})[z - \check{z}, \ldots, z - \check{z}] = (-1)^\ell \Delta_\ell (z_n - \check{z}_n)^\ell, \quad 2 \le \ell \le d. \tag{3.4}$$

Since we are assuming that Δ_ℓ, $\ell < d$, vanish, (3.4) and (3.1) give

$$f_n(z) = \frac{(-1)^d \Delta_d}{d!}(z_n - \check{z}_n)^d. \tag{3.5}$$

Thus, the last component of $H_1(z,t)$ is

$$\begin{aligned} t f_n(z) &+ (1-t)\frac{(-1)^d \Delta_d}{d!}(z_n - \check{z}_n)^d \\ &= t\frac{(-1)^d \Delta_d}{d!}(z_n - \check{z}_n)^d + (1-t)\frac{(-1)^d \Delta_d}{d!}(z_n - \check{z}_n)^d \\ &= \frac{(-1)^d \Delta_d}{d!}(z_n - \check{z}_n)^d. \end{aligned}$$

Then, $H_1(z,t) = 0$ implies $(z_n - \check{z}_n)^d = 0$, and consequently, $z_n - \check{z}_n = 0$ or $z_n = \check{z}_n$. This implies $z_k = \check{z}_k - \alpha_k(z_n - \check{z}_n) = \check{z}_k$ for $k = 1, \dots, n-1$.

Now we know that $H_1(z,t)$ has a unique zero at $(\check{z}_1, \dots, \check{z}_{n-1}, \check{z}_n)$. This is saying that $\tilde{H}_1((x,y),t)$ has a unique zero at $(\check{x}, \check{y})$, which is the midpoint of nondegenerate box $\tilde{\boldsymbol{x}}$. Thus, $\tilde{H}_1((x,y),t) \neq 0$ for $(x,y) \in \boldsymbol{\partial}\tilde{\boldsymbol{x}}$ and $t \in [0,1]$. Then, by Theorem 1.4,

$$\mathrm{d}(\tilde{F}, \tilde{\boldsymbol{x}}, 0) = \mathrm{d}(\tilde{A}, \tilde{\boldsymbol{x}}, 0).$$

Next, we prove $\mathrm{d}(\tilde{A}, \tilde{\boldsymbol{x}}, 0) = \mathrm{d}(\tilde{G}, \tilde{\boldsymbol{x}}, 0)$. Define

$$\begin{aligned} \tilde{H}_2((x,y),t) &\equiv t\tilde{A}(x,y) + (1-t)\tilde{G}(x,y) \\ \text{and} \qquad H_2(z,t) &\equiv tA(z) + (1-t)G(z). \end{aligned}$$

We will prove that $\tilde{H}_2((x,y),t) \neq 0$ when $(x,y) \in \partial\tilde{\boldsymbol{x}}$ and $t \in [0,1]$. It is clear that $\tilde{H}_2((x,y),t) = 0$ is equivalent to $H_2(z,t) = 0$, so we consider $H_2(z,t)$. The definition of H_2 and some rearrangement of terms give

$$\begin{aligned} H_2(z,t) = \Big((z_1 - \check{z}_1) &+ t\alpha_1(z_n - \check{z}_n), \dots, (z_{n-1} - \check{z}_{n-1}) + t\alpha_{n-1}(z_n - \check{z}_n), \\ &\frac{(-1)^d \Delta_d}{d!}(z_n - \check{z}_n)^d\Big). \end{aligned}$$

Because of the last component of $H_2(z,t)$, $H_2(z,t) = 0$ implies $z_n = \check{z}_n$. Then, from the first $n-1$ components of $H_2(z,t)$, $H_2(z,t) = 0$ implies $z_k = \check{z}_k - t\alpha_k(z_n - \check{z}_n) = \check{z}_k$ for $k = 1, \dots, n-1$. Thus, $H_2(z,t)$ has a unique zero at $(\check{z}_1, \dots, \check{z}_{n-1}, \check{z}_n)$. This is saying that $\tilde{H}_2((x,y),t)$ has a unique zero at $(\check{x}, \check{y})$, which is the midpoint of the nondegenerate box $\tilde{\boldsymbol{x}}$. Thus, $\tilde{H}_2((x,y),t) \neq 0$ for $(x,y) \in \boldsymbol{\partial}\tilde{\boldsymbol{x}}$ and $t \in [0,1]$. Then, by Theorem 1.4,

$$\mathrm{d}(\tilde{A}, \tilde{\boldsymbol{x}}, 0) = \mathrm{d}(\tilde{G}, \tilde{\boldsymbol{x}}, 0).$$

Next, we prove $\mathrm{d}(\tilde{G}, \tilde{\boldsymbol{x}}, 0) = d$. Perturb $G(z)$ by an arbitrary small ϵ to define

$$G_\epsilon(z) = (g_{1\epsilon}(z), \dots, g_{(n-1)\epsilon}(z), g_{n\epsilon}(z)),$$

where

$$\begin{aligned} g_{k\epsilon}(z) &= g_k(z) = (z_k - \check{z}_k) \quad \text{for } 1 \le k \le n-1, \\ g_{n\epsilon}(z) &= g_n(z) + \epsilon = \frac{(-1)^d \Delta_d}{d!}(z_n - \check{z}_n)^d + \epsilon. \end{aligned} \tag{3.6}$$

Let $p_{k\epsilon}(x,y) \equiv \Re(g_{k\epsilon}(x+iy))$ and $q_{k\epsilon}(x,y) \equiv \Im(g_{k\epsilon}(x+iy))$. With this, define

$$\tilde{G}_\epsilon(x,y) \equiv (p_{1\epsilon}(x,y), q_{1\epsilon}(x,y), \ldots, p_{n\epsilon}(x,y), q_{n\epsilon}(x,y)).$$

It is obvious that $p_{k\epsilon}(x,y) = x_k - \check{x}_k$ and $q_{k\epsilon}(x,y) = y_k - \check{y}_k$ for $k = 1, \ldots, n-1$. Assume that ϵ is small enough. Then $G_\epsilon(z)$, and thus $\tilde{G}_\epsilon(x,y)$, have d zeros $\tilde{z} = (\tilde{z}_1, \ldots, \tilde{z}_{n-1}, \tilde{z}_n)$, or $\tilde{x} = (\tilde{x}_1, \tilde{y}_1, \ldots, \tilde{x}_{n-1}, \tilde{y}_{n-1}, \tilde{x}_n, \tilde{x}_n)$ in $\tilde{\boldsymbol{x}}$, with $\tilde{z}_k - \check{z}_k = 0$, or $\tilde{x}_k - \check{x}_k = 0$ and $\tilde{y}_k - \check{y}_k = 0$ for $k = 1, \ldots, n-1$, and $(\tilde{z}_n - \check{z}_n)^d = \frac{d!\epsilon}{(-1)^{d+1}\Delta_d} \neq 0$. $\frac{\partial g_{n\epsilon}}{\partial z_n}(\tilde{z}) = \frac{(-1)^d \Delta_d}{(d-1)!}(\tilde{z}_n - \check{z}_n)^{d-1} \neq 0$.

$$\begin{aligned}
\left|\frac{\partial \tilde{G}_\epsilon}{\partial x_1 \partial y_1 \ldots \partial x_n \partial y_n}(\tilde{x})\right| &= \begin{vmatrix} 1 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \\ 0 & 0 & \cdots & 1 & 0 & 0 \\ 0 & 0 & \cdots & 0 & \frac{\partial p_{n\epsilon}}{\partial x_n} & \frac{\partial p_{n\epsilon}}{\partial y_n} \\ 0 & 0 & \cdots & 0 & \frac{\partial q_{n\epsilon}}{\partial x_n} & \frac{\partial q_{n\epsilon}}{\partial y_n} \end{vmatrix} \\
&= \begin{vmatrix} \frac{\partial p_{n\epsilon}}{\partial x_n} & \frac{\partial p_{n\epsilon}}{\partial y_n} \\ \frac{\partial q_{n\epsilon}}{\partial x_n} & \frac{\partial q_{n\epsilon}}{\partial y_n} \end{vmatrix} = \begin{vmatrix} \frac{\partial p_{n\epsilon}}{\partial x_n} & \frac{\partial p_{n\epsilon}}{\partial y_n} \\ -\frac{\partial p_{n\epsilon}}{\partial y_n} & \frac{\partial p_{n\epsilon}}{\partial x_n} \end{vmatrix} \qquad (3.7) \\
&= \left(\frac{\partial p_{n\epsilon}}{\partial x_n}\right)^2 + \left(\frac{\partial p_{n\epsilon}}{\partial y_n}\right)^2 = \left|\frac{\partial g_{n\epsilon}}{\partial z_n}(\tilde{z})\right|^2 > 0.
\end{aligned}$$

Thus, by Theorem 1.2, $\mathrm{d}(\tilde{G}_\epsilon, \tilde{\boldsymbol{x}}, 0) = d$, and then $\mathrm{d}(\tilde{G}, \tilde{\boldsymbol{x}}, 0) = d$ by Theorem 1.5. Finally,

$$\mathrm{d}(\tilde{F}, \tilde{\boldsymbol{x}}, 0) = \mathrm{d}(\tilde{G}, \tilde{\boldsymbol{x}}, 0) = d. \qquad \square$$

Unless the components of F are exactly linear and degree-d polynomials, the $\mathcal{O}(\|x - \check{x}\|)^2$ and $\mathcal{O}(\|x - \check{x}\|)^{d+1}$ terms in (2.1) and (2.2) are not absent. However, since $\mathrm{d}(F, \boldsymbol{z}, 0)$ is a continuous function of F, $\mathrm{d}(F, \boldsymbol{z}, 0)$ will still be equal to d if the widths $\mathrm{w}(\boldsymbol{x}_k - \check{x}_k)$ (and hence $\|x_k - \check{x}_k\|$) are small, for $1 \leq k \leq n$. Nonetheless, the proof of Theorem 3.1 does not lead to a practical computational verification technique that the degree is d for such more general F: If we try to verify $H(z,t) \neq 0$ or $\tilde{H}((x,y),t) \neq 0$ when $(x,y) \in \partial\tilde{\boldsymbol{x}}$ and $t \in [0,1]$, then it would require an inordinate amount of work for a verification process that would normally require only a single step of an interval Newton method in the nonsingular case. First, we would need to compute Δ_d, which involves all partial derivatives of order 1 and order d. This is expensive when both n and d are large. Second, we would need to know where the solutions of $u_n(x) = 0$ and $v_n(x) = 0$ are on $\boldsymbol{x}_{\underline{n}}$, $\boldsymbol{x}_{\overline{n}}$, $\boldsymbol{y}_{\underline{n}}$, and $\boldsymbol{y}_{\overline{n}}$ when $z_k = \check{z}_k - t\alpha_k(z_n - \check{z}_n)$, and the search process for such solutions is expensive.

We could try to verify $H(z,t) \neq 0$ when $(x,y) \in \partial\tilde{\boldsymbol{x}}$ and $t \in [0,1]$ in another way: verify that $H(z,t) = 0$ has a unique solution in the interior of $\tilde{\boldsymbol{x}}$ when $t \in [0,1]$. However, we will run into the singular situation again if we do that.

In fact, there is an alternative algorithm to verify that the degree is d. That will be the subject of the next section.

4. Algorithm to verify a nonzero topological degree.

The algorithm we present here is similar to the algorithm in [14]. Based on Theorem 1.7 in section 1.2, the following theorem underlies our algorithm.

THEOREM 4.1. *Suppose that*

1. $u_k \neq 0$ *on* $\boldsymbol{x}_{\underline{k}}$ *and* $\boldsymbol{x}_{\overline{k}}$, *and* $v_k \neq 0$ *on* $\boldsymbol{y}_{\underline{k}}$ *and* $\boldsymbol{y}_{\overline{k}}$, $k = 1, \dots, n-1$;
2. $\tilde{F}_{\neg u_n} = 0$ *has solutions, if there are any, on* $\boldsymbol{x}_{\underline{n}}$ *and* $\boldsymbol{x}_{\overline{n}}$ *with* y_n *in the interior of* $\boldsymbol{y}_n$, *and* $\tilde{F}_{\neg u_n} = 0$ *has solutions, if there are any, on* $\boldsymbol{y}_{\underline{n}}$ *and* $\boldsymbol{y}_{\overline{n}}$ *with* x_n *in the interior of* $\boldsymbol{x}_n$;
3. $u_n \neq 0$ *at the solutions of* $\tilde{F}_{\neg u_n} = 0$ *in condition* 2; *and*
4. *the Jacobi matrices of* $\tilde{F}_{\neg u_n}$ *are nonsingular at the solutions of* $\tilde{F}_{\neg u_n} = 0$ *in condition* 2.

Then, for a fixed $s \in \{-1, 1\}$,

$$\begin{aligned}
\mathrm{d}(\tilde{F}, \tilde{\boldsymbol{x}}, 0) = -s & \sum_{\substack{x_n = \underline{x}_n \\ \tilde{F}_{\neg u_n}(x,y)=0 \\ \mathrm{sgn}(u_n(x,y))=s}} \mathrm{sgn} \left| \frac{\partial \tilde{F}_{\neg u_n}}{\partial x_1 y_1 \cdots x_{n-1} y_{n-1} y_n}(x,y) \right| \\
+ s & \sum_{\substack{x_n = \overline{x}_n \\ \tilde{F}_{\neg u_n}(x,y)=0 \\ \mathrm{sgn}(u_n(x,y))=s}} \mathrm{sgn} \left| \frac{\partial \tilde{F}_{\neg u_n}}{\partial x_1 y_1 \cdots x_{n-1} y_{n-1} y_n}(x,y) \right| \\
+ s & \sum_{\substack{y_n = \underline{y}_n \\ \tilde{F}_{\neg u_n}(x,y)=0 \\ \mathrm{sgn}(u_n(x,y))=s}} \mathrm{sgn} \left| \frac{\partial \tilde{F}_{\neg u_n}}{\partial x_1 y_1 \cdots x_{n-1} y_{n-1} x_n}(x,y) \right| \\
- s & \sum_{\substack{y_n = \overline{y}_n \\ \tilde{F}_{\neg u_n}(x,y)=0 \\ \mathrm{sgn}(u_n(x,y))=s}} \mathrm{sgn} \left| \frac{\partial \tilde{F}_{\neg u_n}}{\partial x_1 y_1 \cdots x_{n-1} y_{n-1} x_n}(x,y) \right| .
\end{aligned}$$

Proof. Condition 1 implies $\tilde{F} \neq 0$ on $\boldsymbol{x}_{\underline{k}}$, $\boldsymbol{x}_{\overline{k}}$, $\boldsymbol{y}_{\underline{k}}$, and $\boldsymbol{y}_{\overline{k}}$, $k = 1, \dots, n-1$, and conditions 2 and 3 imply $\tilde{F} \neq 0$ on $\boldsymbol{x}_{\underline{n}}$, $\boldsymbol{x}_{\overline{n}}$, $\boldsymbol{y}_{\underline{n}}$, and $\boldsymbol{y}_{\overline{n}}$. Thus, $\tilde{F} \neq 0$ on $\boldsymbol{\partial} \tilde{\boldsymbol{x}}$. Now, condition 1 implies $\tilde{F}_{\neg u_n} \neq 0$ on $\boldsymbol{\partial x}_{\underline{k}}$, $\boldsymbol{\partial x}_{\overline{k}}$, $\boldsymbol{\partial y}_{\underline{k}}$, and $\boldsymbol{\partial y}_{\overline{k}}$, $k = 1, \dots, n-1$. $\boldsymbol{\partial x}_{\underline{n}}$ consists of $2(n-1)$ $(2n-2)$-dimensional boxes, each of which is either embedded in some $\boldsymbol{x}_{\underline{k}}$, $\boldsymbol{x}_{\overline{k}}$, $\boldsymbol{y}_{\underline{k}}$, or $\boldsymbol{y}_{\overline{k}}$, $1 \leq k \leq n-1$, or is embedded in $\boldsymbol{\partial y}_{\underline{n}}$ or $\boldsymbol{\partial y}_{\overline{n}}$. Thus, by 1 and 2, $\tilde{F}_{\neg u_n} \neq 0$ on $\boldsymbol{\partial x}_{\underline{n}}$. Similarly, $\tilde{F}_{\neg u_n} \neq 0$ on $\boldsymbol{\partial x}_{\overline{n}}$, $\boldsymbol{\partial y}_{\underline{n}}$, and $\boldsymbol{\partial y}_{\overline{n}}$. Thus, condition 1 in Theorem 1.7 is satisfied. Finally, with condition 4, all the conditions of Theorem 1.7 are satisfied. The formula is thus obtained. □

By constructing the box $\tilde{\boldsymbol{x}}$ according to (2.5), we can verify $u_k \neq 0$ on $\boldsymbol{x}_{\underline{k}}$ and $\boldsymbol{x}_{\overline{k}}$, and $v_k \neq 0$ on $\boldsymbol{y}_{\underline{k}}$ and $\boldsymbol{y}_{\overline{k}}$, $k = 1, \dots, n-1$, since $u_k(x,y) \approx (x_k - \check{x}_k) + \alpha_k(x_n - \check{x}_n) \neq 0$ on $\boldsymbol{x}_{\underline{k}}$ and $\boldsymbol{x}_{\overline{k}}$, and $v_k(x,y) \approx y_k + \alpha_k y_n \neq 0$ on $\boldsymbol{y}_{\underline{k}}$ and $\boldsymbol{y}_{\overline{k}}$. This needs only $4n-4$ interval evaluations. Then, we need to search only the four faces $\boldsymbol{x}_{\underline{n}}, \boldsymbol{x}_{\overline{n}}, \boldsymbol{y}_{\underline{n}}$, and $\boldsymbol{y}_{\overline{n}}$ for solutions of $\tilde{F}_{\neg u_n}(x,y) = 0$, regardless of how large n is. The four faces $\boldsymbol{x}_{\underline{n}}$, $\boldsymbol{x}_{\overline{n}}$, $\boldsymbol{y}_{\underline{n}}$, and $\boldsymbol{y}_{\overline{n}}$ remaining to be searched are $(2n-1)$-dimensional boxes. However, exploitation of (2.3) will reduce the search for solutions of $\tilde{F}_{\neg u_n}(x,y) = 0$ on the $(2n-1)$-dimensional boxes to a one-dimensional search. We use $\boldsymbol{x}_{\underline{n}}$ as an example to explain this.

On $\boldsymbol{x}_{\underline{n}}$, $x_n = \underline{x}_n$. We know from (2.3) that if x_n is known precisely, formally solving $\boldsymbol{u}_k(\boldsymbol{x}, \boldsymbol{y}) = 0$ for x_k gives sharper bounds $\tilde{\boldsymbol{x}}_k$ with $\mathrm{w}(\tilde{\boldsymbol{x}}_k) = \mathcal{O}\left(\|(\boldsymbol{x} - \check{x}, \boldsymbol{y})\|\right)^2$,

$1 \le k \le n-1$. Then, we can divide $\boldsymbol{y}_n$ into smaller subintervals. For a small subinterval $\boldsymbol{y}_n^0$ of $\boldsymbol{y}_n$, we can formally solve $\boldsymbol{v}_k(\boldsymbol{x}, \boldsymbol{y}) = 0$ for y_k to get sharper bounds $\tilde{\boldsymbol{y}}_k$ with $\mathrm{w}(\tilde{\boldsymbol{y}}_k) = \mathcal{O}(\max(\|(\boldsymbol{x} - \check{x}, \boldsymbol{y})\|^2, \|\boldsymbol{y}_n^0\|))$, $1 \le k \le n-1$. Thus, we have reduced the search to searching the one-dimensional interval $\boldsymbol{y}_n$, much less costly than searching a $(2n-1)$-dimensional box when n is large. Furthermore, if we know approximately where the solutions of $\tilde{F}_{\neg u_n}(x,y) = 0$ are, we can reduce even the cost of the one-dimensional search. To this end, we will next analyze the solutions of $\tilde{F}_{\neg u_n}(x,y) = 0$ on the four faces $\boldsymbol{x}_{\underline{n}}$, $\boldsymbol{x}_{\overline{n}}$, $\boldsymbol{y}_{\underline{n}}$, and $\boldsymbol{y}_{\overline{n}}$.

To expedite the search, we obtain approximate locations of the places on $\boldsymbol{x}_{\underline{n}}$, $\boldsymbol{x}_{\overline{n}}$, $\boldsymbol{y}_{\underline{n}}$, and $\boldsymbol{y}_{\overline{n}}$ where $\tilde{F}_{\neg u_n}(x,y) = 0$. To obtain these locations, we assume that the $\mathcal{O}\left(\|x - \check{x}\|\right)^2$ terms in (2.1) and the $\mathcal{O}\left(\|x - \check{x}\|\right)^{d+1}$ terms in (2.2) are absent. Proceeding as in the proof of Theorem 3.1, we plug $z_k - \check{z}_k = -\alpha_k(z_n - \check{z}_n)$, $k = 1, \ldots, n-1$, into $f_n(z)$ to obtain

$$f_n(z) = \frac{(-1)^d \Delta_d}{d!}(z_n - \check{z}_n)^d$$

as before. Thus, $u_n(x,y) = \Re(f_n(z)) = \left\{(-1)^d \Delta_d / d!\right\} \Re((z_n - \check{z}_n)^d)$ and $v_n(x,y) = \Im(f_n(z)) = \left\{(-1)^d \Delta_d / d!\right\} \Im((z_n - \check{z}_n)^d)$. Setting $z_n - \check{z}_n = r(\cos(\theta) + i\sin(\theta))$, we obtain $u_n(x,y) = \left\{(-1)^d \Delta_d / d!\right\} r\cos(d\theta)$ and $v_n(x,y) = \left\{(-1)^d \Delta_d / d!\right\} r \sin(d\theta)$, so $u_n(x,y) = 0$ is equivalent to $\cos(d\theta) = 0$ and $v_n(x,y) = 0$ is equivalent to $\sin(d\theta) = 0$. If we choose $\boldsymbol{x}_n$ and $\boldsymbol{y}_n$ such that

$$\frac{\mathrm{w}(\boldsymbol{y}_n)}{\mathrm{w}(\boldsymbol{x}_n)} = \tan\left(\frac{\pi}{4d}\right), \qquad \text{that is,} \qquad \mathrm{w}(\boldsymbol{y}_n) = \tan\left(\frac{\pi}{4d}\right)\mathrm{w}(\boldsymbol{x}_n), \tag{4.1}$$

then all solutions of $v_n(x,y) = 0$, and consequently all solutions of $\tilde{F}_{\neg u_n}(x,y) = 0$, are arranged in a known pattern on $\boldsymbol{x}_{\underline{n}}$, $\boldsymbol{x}_{\overline{n}}$, $\boldsymbol{y}_{\underline{n}}$, and $\boldsymbol{y}_{\overline{n}}$. In particular, on $\boldsymbol{x}_{\underline{n}}$, $\tilde{x}_n = \underline{x}_n$. $v_n(x,y) = 0$ has a unique solution $\tilde{y}_n = 0$. Substituting these into the conditions

$$\left.\begin{array}{rcl} x_k &=& \check{x}_k - \alpha_k(x_n - \check{x}_n), \\ y_k &=& -\alpha_k y_n, \end{array}\right\} \qquad 1 \le k \le n-1, \tag{4.2}$$

we get the unique solution of $\tilde{F}_{\neg u_n}(x,y) = 0$ with

$$(\tilde{x}, \tilde{y}) = (\check{x}_1 - \alpha_1(\underline{x}_n - \check{x}_n), 0, \ldots, \check{x}_{n-1} - \alpha_{n-1}(\underline{x}_n - \check{x}_n), 0, \underline{x}_n, 0)\,.$$

Similarly, $\tilde{F}_{\neg u_n}(x,y) = 0$ has a unique solution on $\boldsymbol{x}_{\overline{n}}$ with

$$(\tilde{x}, \tilde{y}) = (\check{x}_1 - \alpha_1(\overline{x}_n - \check{x}_n), 0, \ldots, \check{x}_{n-1} - \alpha_{n-1}(\overline{x}_n - \check{x}_n), 0, \overline{x}_n, 0).$$

On $\boldsymbol{y}_{\underline{n}}$, $\tilde{y}_n = \underline{y}_n$. $v_n(x,y) = 0$ has $d-1$ solutions with

$$\tilde{x}_n = \check{x}_n + \frac{\mathrm{w}(\boldsymbol{y}_n)}{2\tan\left(\frac{m\pi}{d}\right)}, \quad m = d-1, d-2, \ldots, 1. \tag{4.3}$$

Substituting these into (4.2) gives the $d-1$ solutions $(\tilde{x}, \tilde{y})$ of $\tilde{F}_{\neg u_n}(x,y) = 0$ with

$$\begin{aligned} (\tilde{x}, \tilde{y}) = \big(&\check{x}_1 - \alpha_1\left(\tilde{x}_n - \check{x}_n\right), \alpha_1 \underline{y}_n, \ldots, \check{x}_{n-1} - \alpha_{n-1}\left(\tilde{x}_n - \check{x}_n\right), \\ &- \alpha_{n-1}\underline{y}_n, \tilde{x}_n, \underline{y}_n\big). \end{aligned}$$

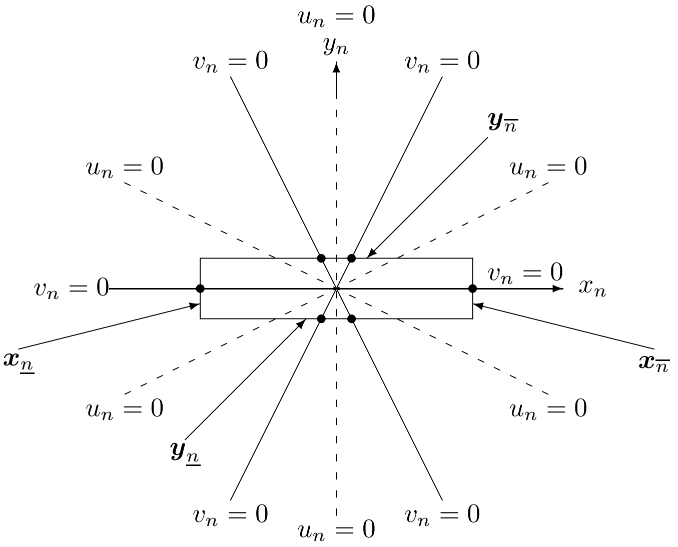

$u_n = 0$
y_n
$v_n = 0$
$v_n = 0$
$\boldsymbol{y}_{\overline{n}}$
$u_n = 0$
$u_n = 0$
$v_n = 0$
$v_n = 0$
x_n
$\boldsymbol{x}_{\underline{n}}$
$\boldsymbol{x}_{\overline{n}}$
$u_n = 0$
$u_n = 0$
$\boldsymbol{y}_{\underline{n}}$
$v_n = 0$
$v_n = 0$
$u_n = 0$

FIG. 4.1. *The zero structure when d is odd. Here,* $d = 3$. $v_n = 0$ *on solid lines, and* $u_n = 0$ *on dashed lines. The thick dots are the solutions of* $\tilde{F}_{\neg u_n}(x, y) = 0$ *on* $\boldsymbol{\partial\tilde{x}}$.

Similarly, $\tilde{F}_{\neg u_n}(x, y) = 0$ has $d - 1$ solutions on $\boldsymbol{y}_{\overline{n}}$ with

$$\begin{aligned}(\tilde{x}, \tilde{y}) = (&\check{x}_1 - \alpha_1 (\tilde{x}_n - \check{x}_n), \dots, \check{x}_{n-1} - \alpha_{n-1} (\tilde{x}_n - \check{x}_n), \\ &- \alpha_{n-1}\overline{y}_n, \tilde{x}_n, \overline{y}_n).\end{aligned}$$

For example, Figure 4.1 gives the solutions of $v_n(x, y) = 0$ on the four faces $\boldsymbol{x}_{\underline{n}}$, $\boldsymbol{x}_{\overline{n}}$, $\boldsymbol{y}_{\underline{n}}$, and $\boldsymbol{y}_{\overline{n}}$ when $d = 3$.

To use the above analysis to find approximations to the solutions of $\tilde{F}_{\neg u_n} = 0$ on the faces we search, we need to know d; we present a heuristic for d in section 5 below.

Now, we present our algorithm. The algorithm consists of three phases:

1. the box-construction phase, where we set $\tilde{\boldsymbol{x}}$,
2. the elimination phase, where we use interval evaluations to verify that $u_k \neq 0$ on $\boldsymbol{x}_{\underline{k}}$ and $\boldsymbol{x}_{\overline{k}}$, and $v_k \neq 0$ on $\boldsymbol{y}_{\underline{k}}$ and $\boldsymbol{y}_{\overline{k}}$, where $1 \leq k \leq n - 1$, and thus eliminate those $4n - 4$ faces, and
3. the search phase, where we
   (a) search $\boldsymbol{x}_{\underline{n}}$, $\boldsymbol{x}_{\overline{n}}$, $\boldsymbol{y}_{\underline{n}}$, and $\boldsymbol{y}_{\overline{n}}$ to locate the solutions of $\tilde{F}_{\neg u_n}(x, y) = 0$,
   (b) compute the signs of u_n and determinants of the Jacobi matrices of $\tilde{F}_{\neg u_n}$ at those solutions,
   (c) compute the degree contributions of each of the four faces $\boldsymbol{x}_{\underline{n}}$, $\boldsymbol{x}_{\overline{n}}$, $\boldsymbol{y}_{\underline{n}}$, and $\boldsymbol{y}_{\overline{n}}$ according to Theorem 4.1, and
   (d) finally sum up to get the degree.

ALGORITHM 1.
INPUT: An approximate solution $\check{x} \in \mathbf{D} \subseteq \mathbb{R}^n$ and a heuristically derived guess d for the topological index of the solution to $\tilde{F}(z) = 0$ near $\check{x}$. (See section 5 below.)
OUTPUT: Either "`A solution is verified`" or "`Verification failed`." If a solution is verified, then also output real bounds $\boldsymbol{x} \subset \mathbb{R}^n$, $\check{x} \in \boldsymbol{x}$, and imaginary bounds $\boldsymbol{y} \in \mathbb{R}^n$, $0 \in \boldsymbol{y}$, such that a solution of $\tilde{F}(z) = 0$ must lie in $(\boldsymbol{x}_1 + i\boldsymbol{y}_1, \dots, \boldsymbol{x}_n + i\boldsymbol{y}_n) \in \mathbb{IC}^n$.

```
Box-setting phase.
  1. Compute the preconditioner of the original system, using Gaussian elimination with full pivoting.
  2. Set the widths of $\boldsymbol{x}_k$ and $\boldsymbol{y}_k$ (see explanation below), for $1 \le k \le n-1$.
  3. Set the width of $\boldsymbol{x}_n$ as in (2.5).
  4. Set the width of $\boldsymbol{y}_n$ to be the minimum of that obtained from conditions (2.5) and (4.1).
Elimination phase.
Do for $1 \le k \le n-1$
  1. DO for $\boldsymbol{x}_{\underline{k}}$ and $\boldsymbol{x}_{\overline{k}}$
       (a) Compute the mean-value extension of $\boldsymbol{u}_k$ over that face.
       (b) IF $0 \in \boldsymbol{u}_k$, THEN STOP and signal failure.
     END DO
  2. DO for $\boldsymbol{y}_{\underline{k}}$ and $\boldsymbol{y}_{\overline{k}}$
       (a) Compute the mean-value extension of $\boldsymbol{v}_k$ over that face.
       (b) IF $0 \in \boldsymbol{v}_k$, THEN STOP and signal failure.
     END DO
Search phase.
  1. Set the value of $s \in \{+1, -1\}$.
       (a) Initialize $s$ to be $+1$. Initialize search_lower and search_upper to be false.
           (See the second note below.)
       (b) DO for $\boldsymbol{x}_{\underline{n}}$ and $\boldsymbol{x}_{\overline{n}}$
             i. Use mean-value extensions for $\boldsymbol{u}_k(\boldsymbol{x}, \boldsymbol{y}) = 0$ to solve for $x_k$ to get
                sharper bounds $\tilde{\boldsymbol{x}}_k$ with width $\mathcal{O}\left(\|(\boldsymbol{x} - \check{x}, \boldsymbol{y})\|\right)^2$, $1 \le k \le n-1$,
                and thus to get a subface $\boldsymbol{x}_{\underline{n}}^0$ (or $\boldsymbol{x}_{\overline{n}}^0$) of $\boldsymbol{x}_{\underline{n}}$ (or $\boldsymbol{x}_{\overline{n}}$.).
             ii. IF $\tilde{\boldsymbol{x}}_k \cap \boldsymbol{x}_k = \emptyset$, THEN CYCLE.
             iii. Compute the mean-value extension $\boldsymbol{u}_n$ over $\boldsymbol{x}_{\underline{n}}^0$ (or $\boldsymbol{x}_{\overline{n}}^0$).
             iv. IF $\boldsymbol{u}_n$ contains 0, THEN
                   A. set search_lower (or search_upper) to be true
                   B. CYCLE.
                 END IF
             v. IF $\boldsymbol{u}_n$ does not contain 0, THEN set $s = -\mathrm{sgn}(\boldsymbol{u}_n)$.
           END DO
       (c) IF $\boldsymbol{u}_n$ does not contain 0 on both $\boldsymbol{x}_{\underline{n}}$ and $\boldsymbol{x}_{\overline{n}}$,
           THEN set $s$ to be the opposite sign to the sign of $\boldsymbol{u}_n$ on $\boldsymbol{x}_{\overline{n}}$, and
           IF $\boldsymbol{u}_n$ has different signs on $\boldsymbol{x}_{\underline{n}}$ and $\boldsymbol{x}_{\overline{n}}$,
           THEN set search_lower to be true.
  2. For $\boldsymbol{x}_{\underline{n}}$ (or $\boldsymbol{x}_{\overline{n}}$), IF search_lower (or search_upper) is true,
     THEN apply Algorithm 2 with $\boldsymbol{x}_{\underline{n}}$ (or $\boldsymbol{x}_{\overline{n}}$) and 0 as input, to compute the
     degree contribution of $\boldsymbol{x}_{\underline{n}}$ (or $\boldsymbol{x}_{\overline{n}}$).
  3. For $\boldsymbol{y}_{\underline{n}}$ (or $\boldsymbol{y}_{\overline{n}}$)
       (a) Use (4.3) to compute the $\tilde{x}_n^m$, $m = d-1, d-2, \ldots, 1$, $\tilde{x}_n^{d-1} < \tilde{x}_n^{d-2} <$
           $\cdots < \tilde{x}_n^1$, corresponding to the $d-1$ approximate solutions of $\tilde{F}_{\neg u_n} = 0$
           on $\boldsymbol{y}_{\underline{n}}$.
       (b) Divide $\boldsymbol{x}_n$ into $d-1$ parts $\boldsymbol{x}_n^m$, $m = 1, \ldots, d-1$, as follows:
           $\boldsymbol{x}_n^1 = [\underline{x}_n, (\tilde{x}_n^{d-1} + \tilde{x}_n^{d-2})/2]$,
           $\boldsymbol{x}_n^m = [(\tilde{x}_n^{d-(m-1)} + \tilde{x}_n^{d-m})/2, (\tilde{x}_n^{d-m} + \tilde{x}_n^{d-(m+1)})/2]$
           for $m = 2, \ldots, d-2$, and $\boldsymbol{x}_n^{d-1} = [(\tilde{x}_n^2 + \tilde{x}_n^1)/2, \overline{x}_n]$.
       (c) DO for $m = 1, \ldots, d-1$
             i. Set a subface $\boldsymbol{y}_{\underline{n}}^m$ of $\boldsymbol{y}_{\underline{n}}$ (or $\boldsymbol{y}_{\overline{n}}^m$ of $\boldsymbol{y}_{\overline{n}}$) by replacing $\boldsymbol{x}_n$ by $\boldsymbol{x}_n^m$.
```

```
            ii. Apply Algorithm 3 with $\boldsymbol{y}^m_{\underline{n}}$ and $\tilde{x}^m_n$ as inputs, to compute the degree
                contribution of $\boldsymbol{y}^m_{\underline{n}}$ (or $\boldsymbol{y}^m_{\overline{n}}$).
            END DO
        (d) Add the degree contributions in the last step to get the degree contribution of $\boldsymbol{y}_{\underline{n}}$ (or $\boldsymbol{y}_{\overline{n}}$).
    4. Add the degree contributions of $\boldsymbol{x}_{\underline{n}}$, $\boldsymbol{x}_{\overline{n}}$, $\boldsymbol{y}_{\underline{n}}$, and $\boldsymbol{y}_{\overline{n}}$ to get the overall degree.
```

Notes for Algorithm 1.

1. In step 3 of the box-setting phase, the width $\mathrm{w}(\boldsymbol{x}_n)$ of $\boldsymbol{x}_n$ depends on the accuracy of the approximate solution $\check{x}$ of the system $F(x) = 0$: $\mathrm{w}(\boldsymbol{x}_n)$ should be much larger than $|\check{x}_k - x^*{}_k|$, but also should be small enough to make a quadratic model accurate over the box.
2. We may set s to minimize the amount of work required to evaluate the sum in Theorem 4.1. In particular, if we know $\mathrm{sgn}(u_n) = \sigma$ on a large number of faces, then setting $s = -\sigma$ will eliminate the need to search those faces.

Algorithm 2.

```
INPUT: $\boldsymbol{x}_{\underline{n}}$ (or $\boldsymbol{x}_{\overline{n}}$) and $\boldsymbol{y}$ from Algorithm 1.
OUTPUT: The contribution of $\boldsymbol{x}_{\underline{n}}$ (or $\boldsymbol{x}_{\overline{n}}$) to the degree in Algorithm 1.
    1. (a) Use mean-value extensions for $\boldsymbol{u}_k(\boldsymbol{x}, \boldsymbol{y}) = 0$ to solve for $x_k$ to get sharper
           bounds $\tilde{\boldsymbol{x}}_k$ with width $\mathcal{O}\left(\|(\boldsymbol{x} - \check{x}, \boldsymbol{y})\|\right)^2$, $1 \le k \le n - 1$.
       (b) IF $\tilde{\boldsymbol{x}}_k \cap \boldsymbol{x}_k = \emptyset$,
           THEN RETURN the degree contribution of that face as 0.
       (c) Update $\boldsymbol{x}_k$.
    2. (a) Compute the mean-value extension $\boldsymbol{u}_n$ over that face.
       (b) IF $s \times \mathrm{sgn}(\boldsymbol{u}_n) < 0$,
           THEN RETURN the degree contribution of that face as 0.
    3. Construct a small subinterval $\boldsymbol{y}^0_n$ of $\boldsymbol{y}_n$ centered at $\check{y}_n$.
    4. (Steps 4 to 9 are identical to steps 1(d) to 1(i), respectively, of the search phase
       in the algorithm in [14]. These steps are repeated here for completeness.)
       (a) Use mean-value extensions for $\boldsymbol{v}_k(\boldsymbol{x}, \boldsymbol{y}) = 0$ to solve for $y_k$ to get sharper
           bounds $\tilde{\boldsymbol{y}}_k$ with width $\mathcal{O}(\max(\|(\boldsymbol{x} - \check{x}, \boldsymbol{y})\|^2, \|\boldsymbol{y}^0_n\|))$, $1 \le k \le n-1$, thus
           getting a subface $\boldsymbol{x}^0_{\underline{n}}$ (or $\boldsymbol{x}^0_{\overline{n}}$) of $\boldsymbol{x}_{\underline{n}}$ (or $\boldsymbol{x}_{\overline{n}}$.)
       (b) IF $\tilde{\boldsymbol{y}}_k \cap \boldsymbol{y}_k = \emptyset$,
           THEN STOP and signal failure.
    5. (a) Set up an interval Newton method for $\tilde{F}_{\neg u_n}$ to verify existence and
           uniqueness of a zero in the subface $\boldsymbol{x}^0_{\underline{n}}$ (or $\boldsymbol{x}^0_{\overline{n}}$).
       (b) IF the zero cannot be verified,
           THEN STOP and signal failure.
    6. Inflate $\boldsymbol{y}^0_n$ as much as possible subject to verification of existence and uniqueness
       of the zero of $\tilde{F}_{\neg u_n}$ over the corresponding subface, and thus get a subinterval
       $\boldsymbol{y}^1_n$ of $\boldsymbol{y}_n$.
    7. In this step, we verify that $\tilde{F}_{\neg u_n} = 0$ has no solutions when $y_n \in \boldsymbol{y}_n \setminus \boldsymbol{y}^1_n$.
       $\boldsymbol{y}_n \setminus \boldsymbol{y}^1_n$ has two separate parts; we denote the lower part by $\boldsymbol{y}^l_n$ and the upper
       part by $\boldsymbol{y}^u_n$. We present the processing of only the lower part. The upper
       part can be processed similarly.
       (a) DO
           i. Use mean-value extensions for $\boldsymbol{v}_k(\boldsymbol{x}, \boldsymbol{y}) = 0$ to solve for $y_k$ to get
              sharper bounds for $y_k$, $1 \le k \le n - 1$, and thus to get a subface of
              $\boldsymbol{x}_{\underline{n}}$ (or $\boldsymbol{x}_{\overline{n}}$).
           ii. Compute the mean-value extensions $\tilde{\boldsymbol{F}}_{\neg u_n}$ over the subface obtained
               in the last step.
```

```
            iii. IF $0 \in \tilde{\boldsymbol{F}}_{\neg u_n}$, THEN
                 A. bisect $\boldsymbol{y}_n^l$, update the lower part as a new $\boldsymbol{y}_n^l$;
                 B. CYCLE.
                 END IF
                 IF $0 \notin \tilde{\boldsymbol{F}}_{\neg u_n}$, THEN EXIT the loop.
            END DO
        (b) DO
            i. IF $\underline{y}_n^1 \le \overline{y}_n^l$, THEN EXIT the loop.
            ii. $\boldsymbol{y}_n^l \longleftarrow [\overline{y}_n^l, \overline{y}_n^l + \mathrm{w}(\boldsymbol{y}_n^l)]$.
            iii. Use mean-value extensions for $\boldsymbol{v}_k(\boldsymbol{x}, \boldsymbol{y}) = 0$ to solve for $y_k$ to get
                 sharper bounds for $y_k$, $1 \le k \le n-1$, and thus to get a subface of
                 $\boldsymbol{x}_{\underline{n}}$ (or $\boldsymbol{x}_{\overline{n}}$).
            iv. Compute the mean-value extensions $\tilde{\boldsymbol{F}}_{\neg u_n}$ over the subface obtained
                in the last step.
            v. IF $0 \notin \tilde{\boldsymbol{F}}_{\neg u_n}$, THEN CYCLE.
               IF $0 \in \tilde{\boldsymbol{F}}_{\neg u_n}$, THEN
                 A. $\boldsymbol{y}_n^l \longleftarrow [\underline{y}_n^l, \mathrm{mid}(\boldsymbol{y}_n^l)]$;
                 B. CYCLE.
               END IF
            END DO
    8.  (a) Compute the mean-value extension of $\boldsymbol{u}_n$ over $\boldsymbol{x}_{\underline{n}}^0$ (or $\boldsymbol{x}_{\overline{n}}^0$).
        (b) IF $\boldsymbol{u}_n < 0$,
            THEN RETURN the degree contribution of that face as 0.
    9.  (a) Compute $|\frac{\partial \tilde{F}_{\neg u_n}}{\partial x_1 y_1 \ldots x_{n-1} y_{n-1} y_n}(\boldsymbol{x}_{\underline{n}}^0)|$ (or $|\frac{\partial \tilde{F}_{\neg u_n}}{\partial x_1 y_1 \ldots x_{n-1} y_{n-1} y_n}(\boldsymbol{x}_{\overline{n}}^0)|$).
        (b) IF $0 \in |\frac{\partial \tilde{F}_{\neg u_n}}{\partial x_1 y_1 \ldots x_{n-1} y_{n-1} y_n}(\boldsymbol{x}_{\underline{n}}^0)|$ (or $0 \in |\frac{\partial \tilde{F}_{\neg u_n}}{\partial x_1 y_1 \ldots x_{n-1} y_{n-1} y_n}(\boldsymbol{x}_{\overline{n}}^0)|$),
            THEN STOP and signal failure.
    10. Apply Theorem 4.1 to compute the degree contribution of $\boldsymbol{x}_{\underline{n}}$ or $\boldsymbol{x}_{\overline{n}}$.
```

ALGORITHM 3.

```
INPUT: $\boldsymbol{y}_{\underline{n}}$ (or $\boldsymbol{y}_{\overline{n}}$) and $\boldsymbol{x}$.
OUTPUT: The contribution of $\boldsymbol{y}_{\underline{n}}$ (or $\boldsymbol{y}_{\overline{n}}$) to the degree in Algorithm 1.
    1. (a) Use mean-value extensions for $\boldsymbol{v}_k(\boldsymbol{x}, \boldsymbol{y}) = 0$ to solve for $y_k$ to get sharper
           bounds $\tilde{\boldsymbol{y}}_k$ with width $\mathcal{O}\left(\|(\boldsymbol{x} - \check{x}, \boldsymbol{y})\|\right)^2$, $1 \le k \le n-1$.
       (b) IF $\tilde{\boldsymbol{y}}_k \cap \boldsymbol{y}_k = \emptyset$,
           THEN RETURN the degree contribution of that face as 0.
       (c) Update $\boldsymbol{y}_k$.
    2. (a) Compute the mean-value extension $\boldsymbol{u}_n$ over that face.
       (b) IF $s \times \mathrm{sgn}(\boldsymbol{u}_n) < 0$,
           THEN RETURN the degree contribution of that face as 0.
    3. Construct a small subinterval $\boldsymbol{x}_n^0$ of $\boldsymbol{x}_n$ which is centered at $\check{x}_n$.
    4. (Steps 4 to 9 are identical to steps 2(d) to 2(i), respectively, of the search
       phase in the algorithm in [14], but are included here for completeness.) Same
       as step 4 of Algorithm 2, except change $y_k$ to $x_k$, $\tilde{\boldsymbol{y}}_k$ to $\tilde{\boldsymbol{x}}_k$, $\boldsymbol{y}_k$ to $\boldsymbol{x}_k$, $\boldsymbol{x}_{\underline{n}}^0$ to
       $\boldsymbol{y}_{\underline{n}}^0$, $\boldsymbol{x}_{\overline{n}}^0$ to $\boldsymbol{y}_{\overline{n}}^0$, $\boldsymbol{x}_{\underline{n}}$ to $\boldsymbol{y}_{\underline{n}}$, and $\boldsymbol{x}_{\overline{n}}$ to $\boldsymbol{y}_{\overline{n}}$.
    5. Same as step 5 of Algorithm 2, except change $\boldsymbol{x}_{\underline{n}}^0$ to $\boldsymbol{y}_{\underline{n}}^0$ and $\boldsymbol{x}_{\overline{n}}^0$ to $\boldsymbol{y}_{\overline{n}}^0$.
    6. Same as step 6 of Algorithm 2, except change $\boldsymbol{y}_n^0$ to $\boldsymbol{x}_n^0$, $\boldsymbol{y}_n^1$ to $\boldsymbol{x}_n^1$, and $\boldsymbol{y}_n$
       to $\boldsymbol{x}_n$.
    7. Same as step 7 of Algorithm 2, except change $\boldsymbol{y}_n \setminus \boldsymbol{y}_n^1$ to $\boldsymbol{x}_n \setminus \boldsymbol{x}_n^1$.
```

8. Same as step 8 of Algorithm 2, except change $\boldsymbol{x}^0_{\underline{n}}$ to $\boldsymbol{y}^0_{\underline{n}}$ and $\boldsymbol{x}^0_{\overline{n}}$ to $\boldsymbol{y}^0_{\overline{n}}$.
9. Same as step 9 of Algorithm 2, except change $|\frac{\partial \tilde{F}_{\neg u_n}}{\partial x_1 y_1 \ldots x_{n-1} y_{n-1} y_n}(\boldsymbol{x}^0_{\underline{n}})|$ to $|\frac{\partial \tilde{F}_{\neg u_n}}{\partial x_1 y_1 \ldots x_{n-1} y_{n-1} x_n}(\boldsymbol{y}^0_{\underline{n}})|$ and $|\frac{\partial \tilde{F}_{\neg u_n}}{\partial x_1 y_1 \ldots x_{n-1} y_{n-1} y_n}(\boldsymbol{x}^0_{\overline{n}})|$ to $|\frac{\partial \tilde{F}_{\neg u_n}}{\partial x_1 y_1 \ldots x_{n-1} y_{n-1} x_n}(\boldsymbol{y}^0_{\overline{n}})|$.
10. Same as step 10 of Algorithm 2.

Notes for Algorithms 2 *and* 3.

1. Algorithms 2 and 3 are identical to steps 1 and 2 of the search phase of the algorithm in [14], except, in Algorithm 2, $\check{y}_n$ can be any interior point of $\boldsymbol{y}_n$, while $\check{y}_n$ is assumed to equal zero in step 1 of the search phase in the algorithm in [14]. Similarly, in Algorithm 3, $\check{x}_n$ can be any interior point of $\boldsymbol{x}_n$, whereas $\check{x}_n$ is assumed to equal the center of $\boldsymbol{x}_n$ in step 2 of the search phase in the algorithm in [14].
2. In the overall algorithm, Algorithm 1, the actual inputs are $\boldsymbol{y}^m_{\underline{n}}$ and $\tilde{x}^m_n$ when Algorithm 3 is applied. However, for notational simplicity, we use $\boldsymbol{y}_{\underline{n}}$ and $\check{x}_n$ as inputs in the presentation of Algorithm 3.

In a certain sense, the computational complexity of Algorithms 1, 2, and 3 is $\mathcal{O}(n^3)$. (See [14] for detailed analysis.) Thus, the computational complexity of the overall algorithm, Algorithm 1, is $\mathcal{O}(n^3)$. This is the best possible order, since computing preconditioners of the original system and the system $\tilde{F}_{\neg u_n}$ is necessary and computing each preconditioner is of order $\mathcal{O}(n^3)$.

5. A heuristic for the degree.

The algorithms in section 4 require a value for d to locate the approximate positions of solutions of $\tilde{F}_{\neg u_n} = 0$ on the faces we search. Here, we present a practical heuristic for the value of d.

Proceeding as in the proof of Theorem 3.1, we assume that the $\mathcal{O}\left(\|x - \check{x}\|\right)^2$ terms in (2.1) and the $\mathcal{O}\left(\|x - \check{x}\|\right)^{d+1}$ terms in (2.2) are absent, and we substitute $x_k - \check{x}_k = -\alpha_k(x_n - \check{x}_n)$, $k = 1, \ldots, n-1$, into f_n to enable us to define the univariate function

$$g(x_n - \check{x}_n) = \frac{(-1)^d \Delta_d}{d!}(x_n - \check{x}_n)^d = \frac{\Delta_d}{d!}(\check{x}_n - x_n)^d. \tag{5.1}$$

Setting

$$K(r, x_n - \check{x}_n) \equiv \frac{g(x_n - \check{x}_n)}{(x_n - \check{x}_n)^r} = \frac{\Delta_d}{d!}(\check{x}_n - x_n)^{d-r},$$

it is clear that $K(d, x_n - \check{x}_n) = \Delta_d/d!$ is independent of x_n, while $K(r, x_n - \check{x}_n)$ depends on x_n for any other r value. Letting δ be a heuristically chosen constant, we have the following ratios:

$$\frac{K(d, \delta(x_n - \check{x}_n))}{K(d, x_n - \check{x}_n)} = \frac{\frac{\Delta_d}{d!}}{\frac{\Delta_d}{d!}} = 1, \quad \text{while}$$

$$R(r) = \frac{K(r, \delta(x_n - \check{x}_n))}{K(r, x_n - \check{x}_n)} = \frac{\frac{\Delta_d}{d!}(\delta(\check{x}_n - x_n))^{d-r}}{\frac{\Delta_d}{d!}(\check{x}_n - x_n)^{d-r}} = \delta^{d-r}$$

for any other r value. The first ratio $R(d)$ always equals 1, but $R(r)$, $r \neq d$, depends on the δ value. We can choose δ to distinguish d from other r values. For example, if we choose $\delta = 100$, then $R(r)$ is not smaller than 100 when r is smaller than d, and is not larger than 0.01 when r is larger than d. Both values are sufficiently different from 1. We can also vary the δ value to check our detection of d. Thus, $R(r)$ is a good heuristic to determine the value of d.

The above discussion is based on the assumptions in section 2. However, unless the first $n-1$ components of F are exactly linear and the last component is a homogeneous degree-d polynomial of n variables, those assumptions are only approximately true. In practice, if $g(x_n - \check{x}_n) \approx \frac{\Delta_d}{d!}(\check{x}_n - x_n)^d$ is an accurate approximation, then $(\check{x}_n - x_n)^d$ should dominate the value of $g(x_n - \check{x}_n)$. Actually,

$$g(x_n - \check{x}_n) = \sum_{k=1}^{d-1} c_k \Delta_k (x_n - \check{x}_n)^k + c_d \Delta_d (x_n - \check{x}_n)^d + \sum_{k=d+1}^{\infty} c_k \Delta_k (x_n - \check{x}_n)^k,$$

where, approximately, $\Delta_1 = \cdots = \Delta_{d-1} = 0$, $\Delta_d \neq 0$. Thus, $x_n - \check{x}_n$ and $\delta(x_n - \check{x}_n)$ should not be too small, since $\sum_{k=1}^{d-1} c_k \Delta_k (x_n - \check{x}_n)^k$ could dominate otherwise. They should not be too big either, since $\sum_{k=d+1}^{\infty} c_k \Delta_k (x_n - \check{x}_n)^k$ could dominate otherwise. If $\Delta_k \approx 0, k = 1, \ldots, d-1$, are quite accurate, then we can choose $x_n - \check{x}_n$ very small, so both $\sum_{k=1}^{d-1} c_k \Delta_k (x_n - \check{x}_n)^k$ and $\sum_{k=d+1}^{\infty} c_k \Delta_k (x_n - \check{x}_n)^k$ can be ignored in the detection of d.

The choice of $x_n - \check{x}_n$ is independent of the settings of $\boldsymbol{x}_k$, $k = 1, \ldots, n$, since we only want to know what d is at that stage.

An alternative choice for detecting d is to compute the values of Δ_k, $k = 1, 2, \ldots$, by interval evaluations until we get some Δ_{k_0} that is sufficiently different from 0. Then, we can decide $d = k_0$. The obvious disadvantage of this method is that it is too expensive for just detecting the value of d, since computation of Δ_k involves computations of all kth-order derivatives. Furthermore, even if we actually evaluate Δ_k, $k = 1, 2, \ldots$, spending much time in the process, we still can not detect the value of d if the magnitudes of Δ_k, $k = 1, \ldots, d-1, d$, are not sufficiently different either due to the problem itself or due to the range overestimation in interval computations.

6. Numerical results.

In this section, we present numerical results for the algorithm in section 4.

The testing described in this section is not meant to be exhaustive, but is meant to illustrate that the algorithms are programmable and do succeed for a variety of problems, as well as to illustrate that the technique can be practical for higher-dimensional problems. We emphasize that, unless there are programming blunders, the implementation can never give an incorrect result. (That is, the degree can never be incorrectly verified to be d.) The only ways that the algorithms can fail are by either asserting that they cannot verify that the degree is d or by running out of computer resources (typically, CPU time limits).

6.1. Test problems.

Our test problems are represented in Examples 1 through 5 below. This set includes both simple problems, such as Example 1, and slightly more realistic problems, such as Example 4. There are both lower degree problems, like Example 2, and slightly higher degree problems, like Example 5.

Consistent with the analysis and algorithms in this paper, the null-space of the Jacobi matrix at the solution has dimension 1 in all of these examples. (We discuss the higher-order rank defect case in [12].)

Examples 2, 3, and 4 are variable-dimension examples coming from finite difference discretization of a bifurcation problem. In choosing these three problems, we looked for a simple way to vary both the actual topological index at the solution and the dimension of the problem. Actual verification procedures for differential equation models should differ somewhat from what is seen here, since the discretization error should also be taken into account, to be able to assert properties about the solutions to the differential equation itself, rather than just properties about solutions of the discretization.

Although Examples 2, 3, and 4 have a special structure (tridiagonal systems), this actual structure was not used in the present algorithms; that is, dense linear algebra was used throughout. In this sense, the observed dependence of computational time on dimension is representative, although the precise form of the nonlinearity conceivably could make a difference.

Example 1.

$$\begin{aligned} f_1(x_1, x_2) &= x_1^2 - x_2, \\ f_2(x_1, x_2) &= x_1^2 + x_2. \end{aligned}$$

Example 2 (the same as Example 3 from [14], motivated from considerations in [7]). *Set $F(x) = H(x,t) = (1-t)(Ax - x^2) - tx$, where $A \in \mathbb{R}^{n\times n}$ is the matrix corresponding to central difference discretization of the boundary value problem $-u'' = 0$, $u(0) = u(1) = 0$, and $x^2 = (x_1^2, \ldots, x_n^2)^T$. t was chosen to be equal to $t_1 = \lambda_1/(1+\lambda_1)$, where λ_1 is the largest eigenvalue of A.*

In Example 2, if we change the exponent of x from 2 to 3 and 4, then we get Examples 3 and 4.

Example 3. This example is identical to Example 2, except that we set $F(x) = H(x,t) = (1-t)(Ax - x^3) - tx$.

Example 4. This example is identical to Example 2, except that we set $F(x) = H(x,t) = (1-t)(Ax - x^4) - tx$.

We tested with $n = 5$, 10, 20, 40, 80, and 160 for Examples 2, 3, and 4.

Example 5.

$$\begin{aligned} f_1(x_1, x_2, x_3) &= x_1^5 + x_2 + x_2^6 + 3x_3, \\ f_2(x_1, x_2, x_3) &= 4x_1^5 + 5x_2 - 4x_2^6 + 5x_3 - x_3^6, \\ f_3(x_1, x_2, x_3) &= 7x_1^5 + 8x_2 - 100x_2^7 + 10x_3 + 50x_3^6. \end{aligned}$$

For each test problem, we used $(0, 0, \ldots, 0)$, the exact solution to $F(x) = 0$, as the approximate solution to the problem $F(x) = 0$. For each problem except Example 4, we set the widths $\mathrm{w}(\boldsymbol{x}_k)$ and $\mathrm{w}(\boldsymbol{y}_k)$ to 10^{-2} for $1 \le k \le n-1$; then the algorithm automatically computed $\mathrm{w}(\boldsymbol{x}_n)$ and $\mathrm{w}(\boldsymbol{y}_n)$. For Example 4, we set the widths $\mathrm{w}(\boldsymbol{x}_k)$ and $\mathrm{w}(\boldsymbol{y}_k)$ to 10^{-1}, instead of 10^{-2}, for $1 \le k \le n-1$. The reason for this setting for Example 4 is that the system $F(x)$ is flatter near the singular solution, since the degree is higher. Because of the flatness, the condition number of the Jacobian matrix of the system $\tilde{F}_{\neg u_n}$ is larger. Then, because of this ill-conditioning, the interval Newton method to verify the unique solutions of $\tilde{F}_{\neg u_n}$ in step 5 of Algorithm 2 and step 5 of Algorithm 3 is less efficient: More iterations can be expected. We tried 10^{-2} first, but the interval Newton method was not able to verify the solutions when the maximum allowed number of iterations was set to be the same as for Examples 2 and 3.

6.2. Test environment. We programmed the algorithms in section 4 in the Fortran 90 environment developed and described in [10, 11]. Similarly, the test functions were programmed using the same Fortran 90 system, which generated internal symbolic representations of the functions. In the actual tests, generic routines then interpreted the internal representations to obtain both floating point and interval values.

The Sun Fortran 95 compiler, version 6.0, was used on a Sparc Ultra-1 model 140 (with a 140 megaHertz clock) with optimization level 0 (that is, with no optimization).

Table 6.1
Numerical results.

		Heuristic		Verified		
Problem	n	degree	Success	degree	CPU time	Time ratio
Example 1	2	2	Yes	2	0.13	-
Example 2	5	2	Yes	2	1.13	-
Example 2	10	2	Yes	2	5.99	5.30
Example 2	20	2	Yes	2	38.40	6.41
Example 2	40	2	Yes	2	273.61	7.13
Example 2	80	2	Yes	2	2198.14	8.03
Example 2	160	2	Yes	2	13033.22	5.93
Example 3	5	3	Yes	3	39.27	-
Example 3	10	3	Yes	3	10.31	0.26
Example 3	20	3	Yes	3	74.32	7.21
Example 3	40	3	Yes	3	481.23	6.48
Example 3	80	3	Yes	3	3805.06	7.91
Example 3	160	3	Yes	3	33944.20	8.92
Example 4	5	4	Yes	4	23.02	-
Example 4	10	4	Yes	4	154.00	6.69
Example 4	20	4	Yes	4	115.55	0.75
Example 4	40	4	Yes	4	3867.51	33.47
Example 4	80	4	Yes	4	6671.20	1.72
Example 4	160	4	-	-	-	-
Example 5	3	5	Yes	5	16.43	-

Execution times were measured with the Port library routine `ETIME`. All times are given in CPU seconds.

6.3. Test results. We present the numerical results in Table 6.1. The column labels of the table are as follows:

Problem: names of the problems identified in section 6.1,
n: number of independent variables,
Heuristic degree: the heuristic value of the degree computed by the heuristic described in section 5,
Success: whether the algorithm was successful,
Verified degree: topological degree verified by the algorithm,
CPU time: CPU time in seconds of the algorithm,
Time ratio: the ratio of two successive CPU times. This column is only meaningful for Examples 2, 3, and 4.

The algorithm, that is, existence verification, succeeded for all problems except Example 4 when $n = 160$. For that problem, we aborted the program after it ran for 36 hours.

We can see from the CPU time ratios that the algorithm is approximately of order $\mathcal{O}(n^3)$ for Examples 2 and 3. However, as we pointed out at the end of section 6.1, when the degree is higher, the system $F(x)$ is flatter near the singular solution. Because of this ill-conditioning, the interval Newton method to verify the unique solutions of $\tilde{F}_{\neg u_n}$ in step 5 of Algorithm 2 and step 5 of Algorithm 3 will be less efficient: More iterations should be expected, and more irregularity in timing could occur. We can see this from the timing results of Example 4. The experimental results are consistent with our expectations.

In certain preliminary experiments, the heuristic failed to compute the correct value of d. The subsequent verification then returned fairly rapidly with "failure to verify" (generally due to failure to verify that there were no solutions to $u_k = 0$ on $\boldsymbol{x}_{\underline{k}}$ or $\boldsymbol{x}_{\overline{k}}$ or to $v_k = 0$ on $\boldsymbol{y}_{\underline{k}}$ or $\boldsymbol{y}_{\overline{k}}$). The heuristic is the weakest part of the verification process.

Although we arranged $\check{x}$ to be exactly the solution x^*, this should not be crucial to the functioning of the algorithm, as long as the box center $\check{x}$ is a sufficiently accurate approximation to an actual root x^* to allow us to choose a box that is large in relationship to this accuracy but small enough to satisfy our other criteria.

Finally, we expect that additional tuning (selection of initial box size, maximum number of inner iterations in the interval Gauss–Seidel method, etc.) could significantly change timing and success for particular problems. The actual times could improve significantly with a more efficient interval arithmetic environment than that of [10, 11], such as direct use of Sun's interval data type in Fortran.

Acknowledgments. We wish to thank both referees for their careful reading and their valuable suggestions to make the exposition more lucid.

REFERENCES

[1] G. Alefeld and J. Herzberger, *Introduction to Interval Computations*, Academic Press, New York, 1983.

[2] P. S. Alexandrov and H. Hopf, *Topologie*, Springer, Berlin, 1935.

[3] J. Cronin, *Fixed Points and Topological Degree in Nonlinear Analysis*, American Mathematical Society, Providence, RI, 1964.

[4] J. Dian and R. B. Kearfott, *Existence verification for singular and non-smooth zeros of real nonlinear systems*, Math. Comp., 72 (2003), pp. 757–766.

[5] C.-Y. Gau, J. F. Brennecke, and M. A. Stadtherr, *Reliable parameter estimation in VLE modeling,* Fluid Phase Equilib., 168 (2000), pp. 1–18.

[6] E. R. Hansen, *Global Optimization Using Interval Analysis*, Marcel Dekker, New York, 1992.

[7] H. Jürgens, H.-O. Peitgen, and D. Saupe, *Topological perturbations in the numerical nonlinear eigenvalue and bifurcation problems*, in Analysis and Computation of Fixed Points, S. M. Robinson, ed., Academic Press, New York, 1980, pp. 139–181.

[8] R. B. Kearfott, *Computing the Degree of Maps and a Generalized Method of Bisection*, Ph.D. thesis, University of Utah, Salt Lake City, UT, 1977.

[9] R. B. Kearfott, *An efficient degree-computation method for a generalized method of bisection*, Numer. Math., 32 (1979), pp. 109–127.

[10] R. B. Kearfott, *A Fortran* 90 *environment for research and prototyping of enclosure algorithms for nonlinear equations and global optimization*, ACM Trans. Math. Software, 21 (1995), pp. 63–78.

[11] R. B. Kearfott, *Rigorous Global Search: Continuous Problems*, Kluwer, Dordrecht, The Netherlands, 1996.

[12] R. B. Kearfott and J. Dian, *Verifying topological indices for higher-order rank deficiencies*, J. Complexity, 18 (2002), pp. 589–611.

[13] R. B. Kearfott, J. Dian, and A. Neumaier, *Existence verification for singular zeros of nonlinear systems*, Technical report, University of Louisiana at Lafayette, Lafayette, LA, 1999; available online at http://interval.louisiana.edu/preprints/singular_existence.ps.

[14] R. B. Kearfott, J. Dian, and A. Neumaier, *Existence verification for singular zeros of complex nonlinear systems*, SIAM. J. Numer. Anal., 38 (2000), pp. 360–379.

[15] C. F. Korn and Ch. Ullrich, *Extending LINPACK by verification routines for linear systems*, Math. Comput. Simulation, 39 (1995), pp. 21–37.

[16] G. Mayer, *Epsilon-inflation in verification algorithms*, J. Comput. Appl. Math., 60 (1994), pp. 147–169.

[17] A. Neumaier, *Interval Methods for Systems of Equations*, Cambridge University Press, Cambridge, UK, 1990.

[18] J. M. Ortega and W. C. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, 1970.

[19] H. Ratschek and J. Rokne, *New Computer Methods for Global Optimization*, Wiley, New York, 1988.

[20] F. Stenger, *Computing the topological degree of a mapping in* $\mathbb{R}^n$, Numer. Math., 25 (1975), pp. 23–38.

SIAM J. NUMER. ANAL.
Vol. 41, No. 6, pp. 2374–2399

A POSTERIORI ERROR ESTIMATES FOR A DISCONTINUOUS GALERKIN APPROXIMATION OF SECOND-ORDER ELLIPTIC PROBLEMS*

OHANNES A. KARAKASHIAN† AND FREDERIC PASCAL‡

Abstract. Several a posteriori error estimators are introduced and analyzed for a discontinuous Galerkin formulation of a model second-order elliptic problem. In addition to residual-type estimators, we introduce some estimators that are couched in the ideas and techniques of domain decomposition. Results of numerical experiments are presented.

Key words. discontinuous Galerkin methods, a posteriori estimates

AMS subject classifications. 65N55, 65F10

DOI. 10.1137/S0036142902405217

1. Introduction.

One of the important objectives of the numerical approximation of differential equations has been to obtain approximations whose error, as measured in some norm, falls in a given range, preferably as narrow as possible. In the finite element method, and specifically for elliptic boundary value problems, such a goal became possible with the advent of a posteriori estimates pioneered by Babuška and Rheinboldt [4, 5]. Acting on an approximation u_h calculated on a given mesh, such a posteriori estimates give lower and upper bounds on the error expressed in terms of contributions from individual triangles and interfaces. This makes it possible to calculate a new mesh by means of refinement and coarsening. For a survey of the vast amount of work spurred by the above two references, we refer the reader to the book by Verfürth [18]. More recently, attention has increasingly focused on the important issue of convergence, whereby a given tolerance is achieved after a finite number of refinement steps; cf., e.g., [10, 17, 15].

Our aim is to present a posteriori error estimates in the energy norm for a discontinuous Galerkin formulation of a simple second-order elliptic problem. In contrast to standard Galerkin methods, such work is still very rare. Indeed, we are aware only of [8] as taking the a posteriori approach; also, only the estimator (3.1) is treated, and with a different proof which relies on a Helmholtz-type decomposition of the gradient of the error, thus following a technique first used in the context of a posteriori estimates for nonconforming methods. See also [16] for a posteriori estimates in the L^2 norm. Recall that in discontinuous Galerkin methods the trial and test spaces consist of piecewise totally discontinuous polynomials. That is, no continuity constraints are explicitly imposed on the trial and test functions across the element interfaces. As a consequence, weak formulations must include jump terms across interfaces, and typically penalty terms are (artificially) added to control the jump terms. Several variants of this approach exist; cf., e.g., [2, 9, 19]. For a nice survey of various discontinuous Galerkin methods, see [3].

*Received by the editors April 8, 2002; accepted for publication (in revised form) April 3, 2003; published electronically December 17, 2003.
http://www.siam.org/journals/sinum/41-6/40521.html

†Department of Mathematics, The University of Tennessee, Knoxville, TN 37996 (ohannes@math.utk.edu). The research of this author was supported by the French Ministère de la Recherche for a "séjour scientifique de haut niveau" at the University of Paris-Sud.

‡Laboratoire de Mathématiques, Université de Paris-Sud, and Centre National de la Recherche Scientifique 91405, Orsay, France (Frederic.Pascal@math.u-psud.fr).

Discontinuous Galerkin methods have several advantages over other types of finite element methods. For example, the trial and test spaces are very easy to construct; they can naturally handle inhomogeneous boundary conditions and curved boundaries; and they allow the use of highly nonuniform and unstructured meshes. In addition, the fact that the mass matrices are block diagonal is an attractive feature in the context of time-dependent problems, especially if explicit time discretizations are used.

In this paper, we will concentrate on the construction and analysis of error estimators, postponing to a subsequent work the study of other important issues such as convergence of the adaptive scheme. In section 3 we present residual-type estimators whose form and analysis follow traditional lines, with the exception of some technical issues caused by the discontinuous nature of the finite element spaces.

In section 4 we present estimators requiring the solution of local problems. In a departure from more traditional techniques, ours flow from the ideas and techniques of domain decomposition, and specifically those expounded in [11]. In a nutshell, we view the computed solution u_h corresponding to a mesh $\mathcal{T}_h$ as a "coarse-mesh" approximation to a more accurate approximation u_h' to u, with an eye towards using $u_h' - u_h$ to estimate $u - u_h$. Obviously computing u_h' would prove too costly; instead, a good approximation thereof is obtained by adding to u_h the solutions of local problems, the supports of these local contributions playing the role of the subdomains. Indeed, our technique offers the tightest coupling yet known between a posteriori error estimation and domain decomposition, to the extent that the matrices involved in the solution of the local problems consist of the diagonal blocks of the global stiffness matrix that correspond to the individual triangles. A somewhat similar idea is found in [20] in the context of mortar finite elements. There are, however, substantial differences between the two approaches.

In section 5, we present results of numerical experiments focusing on the behavior of the effectivity indices as well as other characteristics of the various estimators.

2. Preliminaries. Let $\Omega \subset \mathbf{R}^d,\ d = 1, 2, 3$, be a bounded domain. We consider the following model problem:

$$-\Delta u = f \quad \text{in } \Omega, \tag{2.1}$$

$$u = 0 \quad \text{on } \partial\Omega. \tag{2.2}$$

The treatment of second-order elliptic problems with more general coefficients and boundary conditions will be contained in a parallel work [13].

Throughout this paper, the standard space, norm, and inner product notation are adopted. Their definitions can be found in [1]. Also, c is used to denote a generic positive mesh-independent constant.

The discontinuous Galerkin method considered in this paper for discretizing problem (2.1)–(2.2) is the one proposed in [6] and [7, 12], where the biharmonic and Stokes problems, respectively, were considered.

Let $\mathcal{T}_h = \{K_i : \ i = 1, 2, \dots, m_h\}$ be a family of star-like partitions (triangulations) of the domain Ω parametrized by $0 < h \le 1$. We assume the following:

(i) The elements of $\mathcal{T}_h$ satisfy the minimal angle condition. Specifically, there is a constant $\theta_0 > 0$ such that $h_K/\rho_K \ge \theta_0\ \forall K \in \mathcal{T}_h$, where h_K and ρ_K denote, respectively, the diameters of the circumscribed and inscribed balls to K.

(ii) $\mathcal{T}_h$ is locally quasi-uniform; that is, if two elements K_j and K_ℓ are adjacent in the sense that $\mu_{d-1}(\partial K_j \cap \partial K_\ell) > 0$, then $\mathrm{diam(K_j)} \approx \mathrm{diam(K_\ell)}$.

Here μ_{d-1} denotes the $(d-1)$-dimensional Lebesgue measure. On $\mathcal{T}_h$ we define the "energy space" $E_h = \Pi_{K\in\mathcal{T}_h} H^2(K) \subset L^2(\Omega)$. For $r \geq 2$, we define the finite element space $V_h^r \subset E_h$ by $V_h^r = \Pi_{K\in\mathcal{T}_h} P_{r-1}(K)$, where $P_{r-1}(K)$ denotes the space of polynomials of total degree $r-1$.

Given the discontinuous nature of the piecewise polynomial functions, we define $\mathcal{E}^I$ and $\mathcal{E}^B$ to be the set of all interior and boundary edges (faces in the case $d=3$), respectively:

$$\begin{aligned}\mathcal{E}^I &= \{e = \partial K_j \cap \partial K_\ell, \quad \mu_{d-1}(\partial K_j \cap \partial K_\ell) > 0\},\\ \mathcal{E}^B &= \{e = \partial K \cap \partial\Omega, \quad \mu_{d-1}(\partial K \cap \partial\Omega) > 0\}.\end{aligned}$$

We also set $\mathcal{E} = \mathcal{E}^I \cup \mathcal{E}^B$. We note that elements of $\mathcal{E}^B$ may be curved. Also, if $e \in \mathcal{E}^I$, then $e = \partial K^+ \cap \partial K^-$ for $K^+, K^- \in \mathcal{T}_h$. We may designate as K^+ the triangle with the higher of the two indices.

Note that elements of the energy space E_h are not functions in the proper sense, and care must be applied in defining their values on $\mathcal{E}$. This is done in the sense of trace.

In order to construct a weak formulation for the problem (2.1)–(2.2), we introduce the bilinear form $a_h^\gamma : E_h \times E_h \to \mathbf{R}$:

$$\begin{aligned} a_h^\gamma(u,v) = \sum_{K\in\mathcal{T}_h} (\nabla u, \nabla v)_K - \sum_{e\in\mathcal{E}^I} \Big[\langle\{\partial_n u\}, [v]\rangle_e + \langle\{\partial_n v\}, [u]\rangle_e - \gamma h_e^{-1} \langle [u], [v]\rangle_e \Big] \\ - \sum_{e\in\mathcal{E}^B} \Big[\langle \partial_n u, v\rangle_e + \langle \partial_n v, u\rangle_e - \gamma h_e^{-1} \langle u, v\rangle_e \Big], \end{aligned} \tag{2.3}$$

where $h_e =$ diam(e) and

$$\begin{aligned} (u,v)_D &= \int_D u\cdot v\, dx,\\ \langle u,v\rangle_\Gamma &= \int_\Gamma uv\, ds, \qquad \text{edge/surface integrals}, \qquad |v|_\Gamma = \langle v,v\rangle_\Gamma^{1/2},\\ [v]\,|_e &= v^+|_e - v^-|_e, \quad v^+ = v|_{K^+}, \quad v^- = v|_{K^-}, \quad e\in\mathcal{E}^I,\\ \{\partial_n v\}\big|_e &= \frac{\partial v^+}{\partial n^+}\bigg|_e, \quad \big[\partial_n v\big]\big|_e = \frac{\partial v^+}{\partial n^+}\bigg|_e - \frac{\partial v^-}{\partial n^+}\bigg|_e, \quad e\in\mathcal{E}^I,\\ \partial_n v\big|_e &= \frac{\partial v^+}{\partial n^+}\bigg|_e, \quad e\in\mathcal{E}^B. \end{aligned}$$

Some further comments on the nature of the form a_h^γ are in order:

(a) The third and sixth terms have been added to symmetrize a_h^γ. Note that the former is zero for smooth u, while the latter is a known quantity since $u\,\big|_{\partial\Omega}$ is given. The a priori estimates remain valid if these terms are removed.

(b) γ is a positive (penalty) parameter that must be chosen appropriately in order for a_h^γ to be coercive.

Remark 2.1. There is an alternative formulation due to Arnold [2], which is obtained by setting $\{\partial_n v\}\big|_e = \frac{1}{2}(\frac{\partial v^+}{\partial n^+} + \frac{\partial v^-}{\partial n^+})|_e$. The results presented in this paper should be valid for Arnold's formulation as well.

The form $a_h^\gamma(\cdot,\cdot)$ is consistent with the Laplacian in the sense that if $u\in H^2(\Omega)$, then

$$a_h^\gamma(u,v) = -(\Delta u, v) - \sum_{e\in\mathcal{E}^B} \big\langle u, \partial_n v - \gamma h_e^{-1} v\big\rangle_e \quad \forall v\in E_h. \tag{2.4}$$

Thus, we define the discontinuous Galerkin approximation of u to be the element u_h^γ in V_h^r that satisfies

$$a_h^\gamma(u_h^\gamma, v) = F(v) := (f, v) \quad \forall v \in V_h^r. \tag{2.5}$$

The existence of a unique u_h^γ follows from Lemma 2.1.

We define the "energy" norm on E_h by

$$\|v\|_{1,h} = \left(\sum_{K\in\mathcal{T}_h} ||\nabla v||_K^2 + \sum_{e\in\mathcal{E}^I} \left[h_e \left|\{\partial_n v\}\right|_e^2 + h_e^{-1}|[v]|_e^2 \right] \right.$$
$$\left. + \sum_{e\in\mathcal{E}^B} \left[h_e \left|\partial_n v\right|_e^2 + h_e^{-1}|v|_e^2 \right] \right)^{1/2}.$$

Concerning the continuity and coercivity of the form a_h^γ, we have the following result (cf. [7]).

Lemma 2.1. (i)

$$|a_h^\gamma(u, v)| \le (1+\gamma)\|u\|_{1,h}\|v\|_{1,h} \qquad \forall u, v \in E_h. \tag{2.6}$$

(ii) *There exist positive constants γ_0 and c_a such that for $\gamma \ge \gamma_0$*

$$a_h^\gamma(v, v) \ge c_a \|v\|_{1,h}^2 \qquad \forall v \in V_h^r. \tag{2.7}$$

Here γ_0 depends only on r and the (aspect) ratios h_K/ρ_K of the elements. In view of condition (i) on the mesh, γ_0 can grow only as a function of r. Numerical experiments reveal that $\gamma_0 \approx 5$ for $r = 2$ and $\gamma_0 \approx 15$ for $r = 3$.

The proofs of the above rely on the following *trace* and *inverse* inequalities. Let D be a regular and starlike domain, and let $\mu =$ diam (D). Then

$$|v|_{\partial D}^2 \le c_{tr}(\mu^{-1}\|v\|_D^2 + \mu\|\nabla v\|_D^2) \quad \forall v \in H^1(D). \tag{2.8}$$

Let $|\cdot|_{j,D}$ denote the seminorm of $H^j(D)$. Then

$$|v|_{j,D} \le c_{inv}\mu^{i-j}|v|_{i,D} \quad \forall v \in P_r,\ 0 \le i \le j \le 2, \tag{2.9}$$

the constant c_{inv} in (2.9) depending only on r.

We shall assume that the following approximation property holds: Let $0 \le m \le r$. Then there exists a constant $c > 0$, independent of $\mathcal{T}_h$, such that for any $u \in H^m(\Omega)$ and $K \in \mathcal{T}_h$ there exists $\chi \in P_{r-1}(K)$ satisfying

$$|u - \chi|_{j,K} \le c h_K^{m-j}|u|_{m,K}, \quad 0 \le j \le m. \tag{2.10}$$

It can be shown that the following error estimates hold (cf. [7]).

Theorem 2.1. *Let u and u_h^γ be the solutions of* (2.1)–(2.2) *and* (2.5), *respectively, and suppose that $u \in H^r(\Omega) \cap H_0^1(\Omega)$ with $r \ge 2$. Then there exists a positive constant c, which is independent of h and u, such that*

$$\|u - u_h^\gamma\|_{1,h} \le c \left(\sum_{K\in\mathcal{T}_h} h_K^{2(r-1)}|u|_{r,K}^2 \right)^{1/2}, \tag{2.11}$$

$$\|u - u_h^\gamma\| \le c h^r |u|_{r,\Omega}. \tag{2.12}$$

2.1. An approximation result. For our first residual-type a posteriori estimate we will need to see how well an element of V_h^r can be approximated by elements of $\overset{0}{V}{}_h^r = V_h^r \cap H_0^1(\Omega)$. The result, Theorem 2.2 below, can also be found in [14]. The proof we give here is constructive and differs entirely from the one given in [14]. We also consider separately in Theorem 2.3 the case when the mesh is nonconforming, i.e., is characterized by the presence of *hanging nodes*.

LEMMA 2.2. *Given N real numbers $\{\alpha_1, \ldots, \alpha_N\}$ let $\beta = \frac{1}{N}\sum_{j=1}^N \alpha_j$. Then,*

$$\sum_{j=1}^N |\alpha_j - \beta|^2 \le C \sum_{j=1}^{N-1} |\alpha_{j+1} - \alpha_j|^2, \tag{2.13}$$

where C depends only on N.

Proof. For any j, the Cauchy–Schwarz inequality gives

$$|\alpha_j - \beta|^2 = \frac{1}{N^2}\left|\sum_{i=1}^N (\alpha_j - \alpha_i)\right|^2 \le \frac{N-1}{N^2}\sum_{i=1}^N |\alpha_j - \alpha_i|^2.$$

Summing over j, we obtain $\sum_{j=1}^N |\alpha_j - \beta|^2 \le \frac{2(N-1)}{N}\sum_{j>i} |\alpha_j - \alpha_i|^2$. The required result now follows, upon writing $\alpha_j - \alpha_i = \sum_{k=i}^{j-1}(\alpha_{k+1} - \alpha_k)$ and using the arithmetic-geometric mean inequality. □

THEOREM 2.2. *Let $\mathcal{T}_h$ be a conforming mesh consisting of triangles when $d = 2$, and tetrahedra when $d = 3$. Then for any $v_h \in V_h^r$ there exists $\chi \in \overset{0}{V}{}_h^r$ satisfying*

$$\sum_{K \in \mathcal{T}_h} \|\nabla(v_h - \chi)\|_K^2 \le C \left(\sum_{e \in \mathcal{E}^I} h_e^{-1} |[v_h]|_e^2 + \sum_{e \in \mathcal{E}^B} h_e^{-1} |v_h|_e^2 \right) \tag{2.14}$$

for some constant C independent of h and v_h but which may depend on the constant θ_0 in assumption (i) *on the mesh.*

Proof. The main argument is quite natural. Given $v_h \in V_h^r$, we construct a function $\chi \in \overset{0}{V}{}_h^r$ as follows: At every node of the mesh $\mathcal{T}_h$ corresponding to a Lagrangian-type degree of freedom for $\overset{0}{V}{}_h^r$, the value of χ is set to the average of the values of v_h at that node.

For each $K \in \mathcal{T}_h$ let $\mathcal{N}_K = \{x_K^{(j)},\ j = 1, \ldots, m\}$ be the Lagrange nodes (points) of K and $\{\phi_K^{(j)},\ j = 1, \ldots, m\}$ the corresponding (local) basis functions satisfying $\phi_K^{(j)}(x_K^{(i)}) = \delta_{ij}$. Set $\mathcal{N} = \cup_{K \in \mathcal{T}_h} \mathcal{N}_K$. We view $\mathcal{N}$ as the union of three disjoint classes:

$$\begin{aligned}
\mathcal{N}_i &= \{\nu \in \mathcal{N} : \nu \text{ is interior to some element}\},\\
\mathcal{N}_b &= \{\nu \in \mathcal{N} : \nu \in e \in \mathcal{E}^B\},\\
\mathcal{N}_v &= \mathcal{N} \setminus (\mathcal{N}_i \cup \mathcal{N}_b).
\end{aligned}$$

For each $\nu \in \mathcal{N}$, let $\omega_\nu = \{K \in \mathcal{T}_h |\ \nu \in K\}$ and denote its cardinality by $|\omega_\nu|$. If $\nu \in \mathcal{N}_i$, then $|\omega_\nu| = 1$. On the other hand if $\nu \in \mathcal{N}_b \cup \mathcal{N}_v$, then $|\omega_\nu|$ is bounded by a constant depending only on the constant θ_0.

Now let $\overset{0}{\mathcal{N}}$ be the collection of distinct Lagrange nodes ν needed to construct a function $\chi \in \overset{0}{V}{}_h^r$. To each node $\nu \in \overset{0}{\mathcal{N}}$ we associate the basis function $\phi^{(\nu)}$ given by

$$\operatorname{supp} \phi^{(\nu)} = \bigcup_{K \in \omega_\nu} K, \quad \phi^{(\nu)}\big|_K = \phi_K^{(j)}, \quad x_K^{(j)} = \nu.$$

We make it a point to include the boundary nodes $\mathcal{N}_b$ in $\overset{0}{\mathcal{N}}$ even though this is not necessary in view of the vanishing on $\partial\Omega$ of the functions in $\overset{0}{V}{}_h^r$. We then can state the following characterization: $\overset{0}{\mathcal{N}} \subseteq \mathcal{N}$ and the mesh $\mathcal{T}_h$ is conforming if and only if $\overset{0}{\mathcal{N}} = \mathcal{N}$.

Now, given $v_h \in V_h^r$, written $v_h = \sum_{K \in \mathcal{T}_h} \sum_{j=1}^m \alpha_K^{(j)} \phi_K^{(j)}$, we define the function $\chi \in \overset{0}{V}{}_h^r$ by (note that $\overset{0}{\mathcal{N}} = \mathcal{N}$ since the mesh is conforming)

$$\chi = \sum_{\nu \in \overset{0}{\mathcal{N}}} \beta^{(\nu)} \phi^{(\nu)}, \text{ where } \beta^{(\nu)} = \begin{cases} 0 & \text{if } \nu \in \mathcal{N}_b, \\ \frac{1}{|\omega_\nu|} \sum_{x_K^{(j)} = \nu} \alpha_K^{(j)}, & \text{if } \nu \in \overset{0}{\mathcal{N}} \setminus \mathcal{N}_b. \end{cases} \tag{2.15}$$

Now set $\beta_K^{(j)} = \beta^{(\nu)}$ whenever $x_K^{(j)} = \nu$.

A simple scaling argument shows that $\|\nabla \phi_K^{(j)}\|_K^2 \le c h_K^{d-2}$. Hence

$$\begin{aligned} \sum_{K \in \mathcal{T}_h} \|\nabla(v_h - \chi)\|_K^2 &\le c\, m \sum_{K \in \mathcal{T}_h} h_K^{d-2} \sum_{j=1}^m \left|\alpha_K^{(j)} - \beta_K^{(j)}\right|^2 \\ &\le c \sum_{\nu \in \mathcal{N}} h_\nu^{d-2} \sum_{x_K^{(j)} = \nu} \left|\alpha_K^{(j)} - \beta^{(\nu)}\right|^2 \qquad \left(h_\nu = \max_{K \in \omega_\nu} h_K\right) \\ &= c \sum_{\nu \in \mathcal{N}_v} h_\nu^{d-2} \sum_{x_K^{(j)} = \nu} \left|\alpha_K^{(j)} - \beta^{(\nu)}\right|^2 + c \sum_{\nu \in \mathcal{N}_b} h_\nu^{d-2} \sum_{x_K^{(j)} = \nu} |\alpha_K^{(j)}|^2. \end{aligned} \tag{2.16}$$

Note that there are no contributions from $\mathcal{N}_i$. We now temporarily focus on the case $d = 2$. For $\nu \in \mathcal{N}_v$, we enumerate the elements of ω_ν as $\{K_1, \ldots, K_{|\omega_\nu|}\}$ so that any consecutive pair K_i, K_{i+1} in that list share an edge. Then from Lemma 2.2, with some constant c depending only on $|\omega_\nu|$ and thus on θ_0, we have

$$\sum_{x_K^{(j)} = \nu} \left|\alpha_K^{(j)} - \beta^{(\nu)}\right|^2 \le c \sum_{i=1}^{|\omega_\nu|-1} \left|\alpha_{K_i}^{(j_i)} - \alpha_{K_{i+1}}^{(j_{i+1})}\right|^2. \tag{2.17}$$

For $d = 3$, it may not be possible to enumerate ω_ν in such a way. However, by allowing some repetitions of its elements, we can write $\omega_\nu = \{K_{\ell_1}, \ldots, K_{\ell_{n(\nu)}}\}$ for some $n(\nu)$, so that in this case also K_{ℓ_i} and $K_{\ell_{i+1}}$ share a face or an edge. Having done so, by applying Lemma 2.2 to the list obtained by removing all repetitions of elements of ω_ν and then using the arithmetic-geometric mean inequality, we obtain

$$\sum_{x_K^{(j)} = \nu} \left|\alpha_K^{(j)} - \beta^{(\nu)}\right|^2 \le c \sum_{i=1}^{n(\nu)-1} \left|\alpha_{K_{\ell_i}}^{(j_{\ell_i})} - \alpha_{K_{\ell_{i+1}}}^{(j_{\ell_{i+1}})}\right|^2. \tag{2.18}$$

Using (2.17) if $d = 2$, or (2.18) if $d = 3$, from (2.16) we have

$$\begin{aligned}\sum_{K \in \mathcal{T}_h} \|\nabla(v_h - \chi)\|_K^2 \leq c \sum_{e \in \mathcal{E}^I} \sum_{\nu \in e} h_\nu^{d-2} \left| \alpha_{K^+}^{(j_\nu^+)} - \alpha_{K^-}^{(j_\nu^-)} \right|^2 \\ + c \sum_{\nu \in \mathcal{N}_b} \sum_{x_K^{(j)} = \nu} h_\nu^{d-2} \left| \alpha_K^{(j)} \right|^2,\end{aligned} \tag{2.19}$$

with $x_{K^+}^{(j_\nu^+)} = x_{K^-}^{(j_\nu^-)} = \nu$. Note that $\alpha_{K^+}^{(j_\nu^+)} - \alpha_{K^-}^{(j_\nu^-)}$ is the jump in the values of v_h at ν across e. Also, since the mesh $\mathcal{T}_h$ is locally quasi-uniform, it follows that

$$\begin{aligned}\sum_{\nu \in e} h_\nu^{d-2} \left| \alpha_{K^+}^{(j_\nu^+)} - \alpha_{K^-}^{(j_\nu^-)} \right|^2 &\leq c h_e^{d-2} |[v_h]|_{L^\infty(e)}^2 \\ &\leq c h_e^{-1} |[v_h]|_e^2,\end{aligned} \tag{2.20}$$

where the constant c depends on the number of nodes in e.

Similarly, it can be shown that for $\nu \in e \in \mathcal{E}^B$,

$$\sum_{x_K^{(j)} = \nu} h_\nu^{d-2} \left| \alpha_K^{(j)} \right|^2 \leq c h_e^{-1} |v_h|_e^2. \tag{2.21}$$

The required result now follows from (2.19)–(2.21). □

We now consider the case when the mesh is nonconforming. We make the following observations, using the notation established in Theorem 2.2:

(i) The hanging nodes are precisely the members of $\mathcal{N} \setminus \overset{0}{\mathcal{N}}$.

(ii) A hanging node cannot be a member of $\mathcal{N}_b$ or $\mathcal{N}_i$.

(iii) For every hanging node ν there is a nonempty proper subset $\tilde{\omega}_\nu$ of ω_ν such that if $\tilde{K} \in \tilde{\omega}_\nu$, then ν is not a local node of $\tilde{K}$. For $d = 2$, $|\tilde{\omega}_\nu| = 1$.

We shall also require that $\mathcal{T}_h$ be obtained from a conforming mesh $\mathcal{T}_h^0$ via a finite number of refinement/coarsening steps. In particular, we assume that there is a mapping $Level : \mathcal{T}_h \to N$, the set of nonnegative integers, such that

(iv) $Level(K) = 0 \ \forall K \in \mathcal{T}_h^0$ ($\mathcal{T}_h^0 \subseteq \mathcal{T}_h$).

(v) If $K \in \omega_\nu \setminus \tilde{\omega}_\nu$ and $\tilde{K} \in \tilde{\omega}_\nu$ are as in (iii) above, then $Level(K) > Level(\tilde{K})$.

An example of the mapping *Level* can be constructed for $d = 2$ as follows. Suppose that we *refine* a given triangle (the *father*) by cutting it in the usual way (see, e.g., Figure 3.2) into four triangles of equal area (the *sons*). On the other hand, we *coarsen* the mesh by merging four sons of the same father. Then we define $Level(K)$, $K \in \mathcal{T}_h$, by $|K| = |K^0|(\frac{1}{4})^{Level(K)}$, where K^0 is the triangle in $\mathcal{T}_h^0$ that contains K and where $|\cdot|$ denotes area.

We have the following result.

Theorem 2.3. *Let $\mathcal{T}_h$ be a nonconforming mesh consisting of triangles when $d = 2$ and tetrahedra when $d = 3$. We shall also assume that $\mathcal{T}_h$ can be described in terms of the mapping Level as discussed above. Then* (2.14) *holds, but the constant C may also depend on $L_{max} = \max\{Level(K), K \in \mathcal{T}_h\}$.*

Proof. Noting that an element of $\overset{0}{V_h^r}$ is still defined by its values at the nodes in $\overset{0}{\mathcal{N}}$, we define the approximant $\chi \in \overset{0}{V_h^r}$ of v_h via (2.15). This uniquely determines the values of χ at the hanging nodes, so we let $\beta^{(\nu)} = \chi(\nu)$ for $\nu \in \mathcal{N} \setminus \overset{0}{\mathcal{N}}$. In a similar fashion, we introduce the quantities $\tilde{\alpha}_{\tilde{K}}^{(\nu)} = (v_h\big|_{\tilde{K}})(\nu)$, $\tilde{K} \in \tilde{\omega}_\nu$, $\nu \in \mathcal{N} \setminus \overset{0}{\mathcal{N}}$.

Proceeding as in (2.16), we obtain

$$\begin{aligned}\sum_{K\in\mathcal{T}_h}\|\nabla(v_h-\chi)\|_K^2 &\le c\sum_{\ell=0}^{L_{max}}\sum_{K\in\mathcal{T}_h^\ell} h_K^{d-2}\sum_{j=1}^{m}\big|\alpha_K^{(j)}-\beta_K^{(j)}\big|^2\\ &\le c\sum_{\nu\in\overset{0}{\mathcal{N}}} h_\nu^{d-2}\sum_{x_K^{(j)}=\nu}\big|\alpha_K^{(j)}-\beta^{(\nu)}\big|^2 + c\sum_{\nu\in\mathcal{N}\setminus\overset{0}{\mathcal{N}}} h_\nu^{d-2}\sum_{\substack{x_K^{(j)}=\nu\\ K\in\omega_\nu\setminus\tilde{\omega}_\nu}}\big|\alpha_K^{(j)}-\beta^{(\nu)}\big|^2,\end{aligned}\tag{2.22}$$

where $\mathcal{T}_h^\ell=\{K\in\mathcal{T}_h \mid Level(K)=\ell\}$. The first sum on the right-hand side of (2.22) can be handled as in the conforming case using steps (2.17)–(2.21). As for the second sum, for a given $x_K^{(j)}=\nu,\ K\in\omega_\nu\setminus\tilde{\omega}_\nu$, we choose $\tilde{K}\in\tilde{\omega}_\nu$ such that K and $\tilde{K}$ share an edge or a face. A crucial observation is that $Level(\tilde{K})<Level(K)$. Then

$$\big|\alpha_K^{(j)}-\beta^{(\nu)}\big|^2\le 2\big|\alpha_K^{(j)}-\tilde{\alpha}_{\tilde{K}}^{(\nu)}\big|^2+2\big|\tilde{\alpha}_{\tilde{K}}^{(\nu)}-\beta^{(\nu)}\big|^2.$$

Now $|\alpha_K^{(j)}-\tilde{\alpha}_{\tilde{K}}^{(\nu)}|$ is the jump in the values of v_h at the node ν which belongs to the interface e between K and $\tilde{K}$, and thus $h_\nu^{d-2}|\alpha_K^{(j)}-\tilde{\alpha}_{\tilde{K}}^{(\nu)}|^2$ can be bounded by $ch_e^{-1}|[v_h]|_e^2$ as was done in (2.20). On the other hand,

$$\tilde{\alpha}_{\tilde{K}}^{(\nu)}-\beta^{(\nu)}=\big(v_h\big|_{\tilde{K}}\big)(\nu)-\big(\chi\big|_{\tilde{K}}\big)(\nu)=\sum_{j=1}^{m}(\alpha_{\tilde{K}}^{(j)}-\beta_{\tilde{K}}^{(j)})\phi_{\tilde{K}}^{(j)}.$$

Thus,

$$\big|\tilde{\alpha}_{\tilde{K}}^{(\nu)}-\beta^{(\nu)}\big|^2\le\sum_{j=1}^{m}\big|\alpha_{\tilde{K}}^{(j)}-\beta_{\tilde{K}}^{(j)}\big|^2\cdot\sum_{j=1}^{m}\big|\phi_{\tilde{K}}^{(j)}(\nu)\big|^2\le c\sum_{j=1}^{m}\big|\alpha_{\tilde{K}}^{(j)}-\beta_{\tilde{K}}^{(j)}\big|^2$$

for some constant c independent of h. Gathering these results, we have

$$\begin{aligned}\sum_{\nu\in\mathcal{N}\setminus\overset{0}{\mathcal{N}}} h_\nu^{d-2}\sum_{\substack{x_K^{(j)}=\nu\\ K\in\omega_\nu\setminus\tilde{\omega}_\nu}}\big|\alpha_K^{(j)}-\beta^{(\nu)}\big|^2 &= \sum_{\nu\in\mathcal{N}\setminus\overset{0}{\mathcal{N}}} h_\nu^{d-2}\sum_{\ell=0}^{L_{max}}\sum_{K\in\mathcal{T}_h^\ell}\sum_{\substack{x_K^{(j)}=\nu\\ K\in\omega_\nu\setminus\tilde{\omega}_\nu}}\big|\alpha_K^{(j)}-\beta^{(\nu)}\big|^2\\ &\le c\sum_{e\in\mathcal{E}^I} h_e^{-1}|[v_h]|_e^2+c\sum_{\ell=0}^{L}\sum_{K\in\mathcal{T}_h^\ell} h_K^{d-2}\sum_{j=1}^{m}\big|\alpha_K^{(j)}-\beta_K^{(j)}\big|^2\\ &\le c\sum_{e\in\mathcal{E}^I} h_e^{-1}|[v_h]|_e^2+c\sum_{\nu\in\overset{0}{\mathcal{N}}} h_\nu^{d-2}\sum_{x_K^{(j)}=\nu}\big|\alpha_K^{(j)}-\beta^{(\nu)}\big|^2\\ &\quad+c\sum_{\nu\in\mathcal{N}\setminus\overset{0}{\mathcal{N}}} h_\nu^{d-2}\sum_{\ell=0}^{L}\sum_{K\in\mathcal{T}_h^\ell}\sum_{\substack{x_K^{(j)}=\nu\\ K\in\omega_\nu\setminus\tilde{\omega}_\nu}}\big|\alpha_K^{(j)}-\beta^{(\nu)}\big|^2\end{aligned}\tag{2.23}$$

for some L that satisfies $0\le L<L_{max}$. Repeating this argument a finite number of times, the last sum in (2.23) (over the hanging nodes) will be eventually replaced by $c\sum_{e\in\mathcal{E}^I} h_e^{-1}|[v_h]|_e^2+c\sum_{\nu\in\overset{0}{\mathcal{N}}} h_\nu^{d-2}\sum_{x_K^{(j)}=\nu}|\alpha_K^{(j)}-\beta^{(\nu)}|^2$. As we mentioned earlier, the latter term can be bounded by $c\left(\sum_{e\in\mathcal{E}^I} h_e^{-1}|[v_h]|_e^2+\sum_{e\in\mathcal{E}^B} h_e^{-1}|v_h|_e^2\right)$ just as in the conforming case. This concludes the proof. □

3. A posteriori estimates. This section is devoted to residual-type a posteriori estimates. The estimators as well as the exposition follow the lines found in Verfürth [18], with the exception of the technical details stemming from the discontinuous nature of V_h^r. We also note that our estimators (3.11) and (3.12) are entirely local.

Again for the sake of simplifying the exposition, and in this section only, we shall assume that f is a piecewise polynomial function on the mesh $\mathcal{T}_h$. Given that we have decided not to worry about quadrature errors, this is not an unreasonable assumption, since any given quadrature rule used to evaluate $(f|_K, v)$ cannot distinguish between $f|_K$ and the Lagrange interpolant of f at the quadrature points in K.

THEOREM 3.1. *Let $e = u - u_h^\gamma$. Then*

$$\sum_{K\in\mathcal{T}_h} \|\nabla e\|_K^2 \le c\Big\{ \sum_{K\in\mathcal{T}_h} h_K^2 \|f + \Delta u_h^\gamma\|_K^2 + \sum_{e\in\mathcal{E}^I} h_e \big|\big[\partial_n u_h^\gamma\big]\big|_e^2 + \gamma^2 \sum_{e\in\mathcal{E}^I} h_e^{-1} |[u_h^\gamma]|_e^2 + \gamma^2 \sum_{e\in\mathcal{E}^B} h_e^{-1} |u_h^\gamma|_e^2 \Big\}. \tag{3.1}$$

Proof. From (2.4) and (2.5) there follows the orthogonality relation $a_h^\gamma(e, v_h) = 0 \; \forall v_h \in V_h^r$. Now for $v \in E_h$ and $v_h \in V_h^r$, let $\eta = v - v_h$. We have

$$\begin{aligned} a_h^\gamma(e,v) &= a_h^\gamma(e,\eta) = (f,\eta) - a_h^\gamma(u_h^\gamma,\eta) \\ &= (f,\eta) - \Bigg\{ \sum_{K\in\mathcal{T}_h} (\nabla u_h^\gamma, \nabla\eta)_K \\ &\qquad - \sum_{e\in\mathcal{E}^I} \Big[\langle \{\partial_n u_h^\gamma\}, [\eta] \rangle_e + \langle \{\partial_n \eta\}, [u_h^\gamma] \rangle_e - \gamma h_e^{-1} \langle [u_h^\gamma], [\eta] \rangle_e \Big] \\ &\qquad - \sum_{e\in\mathcal{E}^B} \Big[\langle \partial_n u_h^\gamma, \eta \rangle_e + \langle \partial_n \eta, u_h^\gamma \rangle_e - \gamma h_e^{-1} \langle u_h^\gamma, \eta \rangle_e \Big] \Bigg\}. \end{aligned} \tag{3.2}$$

Now, integrating by parts, we see that

$$\begin{aligned} \sum_{K\in\mathcal{T}_h} (\nabla u_h^\gamma, \nabla\eta)_K &= \sum_{K\in\mathcal{T}_h} (-\Delta u_h^\gamma, \eta)_K + \sum_{K\in\mathcal{T}_h} \langle \partial_n u_h^\gamma, \eta \rangle_{\partial K} \\ &= \sum_{K\in\mathcal{T}_h} (-\Delta u_h^\gamma, \eta)_K + \sum_{e\in\mathcal{E}^I} \Big[\langle \{\partial_n u_h^\gamma\}, [\eta] \rangle_e + \langle [\partial_n u_h^\gamma], \eta^* \rangle_e \Big] \\ &\qquad + \sum_{e\in\mathcal{E}^B} \langle \partial_n u_h^\gamma, \eta \rangle_e, \end{aligned} \tag{3.3}$$

where $\eta^* = \eta^-$ for Baker's method and $\eta^* = \frac{1}{2}(\eta^+ + \eta^-)$ for Arnold's method. Now, using (3.3) in (3.2), we obtain

$$\begin{aligned} a_h^\gamma(e,v) = &\sum_{K\in\mathcal{T}_h} (f + \Delta u_h^\gamma, \eta)_K \\ &+ \sum_{e\in\mathcal{E}^I} \Big[\langle \{\partial_n \eta\}, [u_h^\gamma] \rangle_e - \big\langle \big[\partial_n u_h^\gamma\big], \eta^* \big\rangle_e - \gamma h_e^{-1} \langle [u_h^\gamma], [\eta] \rangle_e \Big] \\ &+ \sum_{e\in\mathcal{E}^B} \Big[\langle \partial_n \eta, u_h^\gamma \rangle_e - \gamma h_e^{-1} \langle u_h^\gamma, \eta \rangle_e \Big]. \end{aligned} \tag{3.4}$$

From the definition of $a_h^\gamma(e,v)$ and using (3.4), we get

$$\begin{aligned}
\sum_{K\in\mathcal{T}_h}(\nabla e,\nabla v)_K+\gamma\sum_{e\in\mathcal{E}^I}h_e^{-1}\langle[e],[v]\rangle_e+\gamma\sum_{e\in\mathcal{E}^B}h_e^{-1}\langle e,v\rangle_e &= a_h^\gamma(e,v)\\
&\quad+\sum_{e\in\mathcal{E}^I}\Big[\langle\{\partial_n e\},[v]\rangle_e+\langle\{\partial_n v\},[e]\rangle_e\Big]\\
&\quad+\sum_{e\in\mathcal{E}^B}\Big[\langle\partial_n e,v\rangle_e+\langle\partial_n v,e\rangle_e\Big]\\
&= \sum_{K\in\mathcal{T}_h}(f+\Delta u_h^\gamma,\eta)_K+\sum_{e\in\mathcal{E}^I}\Big[\langle\{\partial_n\eta\},[u_h^\gamma]\rangle_e-\big\langle[\partial_n u_h^\gamma],\eta^*\big\rangle_e\\
&\quad+\langle\{\partial_n e\},[v]\rangle_e+\langle\{\partial_n v\},[e]\rangle_e-\gamma h_e^{-1}\langle[u_h^\gamma],[\eta]\rangle_e\Big]\\
&\quad+\sum_{e\in\mathcal{E}^B}\Big[\langle\partial_n\eta,u_h^\gamma\rangle_e+\langle\partial_n e,v\rangle_e+\langle\partial_n v,e\rangle_e-\gamma h_e^{-1}\langle u_h^\gamma,\eta\rangle_e\Big].
\end{aligned}\tag{3.5}$$

First note that

$$\langle\{\partial_n\eta\},[u_h^\gamma]\rangle_e+\langle\{\partial_n v\},[e]\rangle_e=-\langle\{\partial_n v_h\},[u_h^\gamma]\rangle_e,\quad e\in\mathcal{E}^I,$$

and

$$\langle\partial_n\eta,u_h^\gamma\rangle_e+\langle\partial_n v,e\rangle_e=-\langle\partial_n v_h,u_h^\gamma\rangle_e,\quad e\in\mathcal{E}^B.$$

We will choose v_h to be piecewise constant on $\mathcal{T}_h$. Thus these four terms are zero. Hence (3.5) reduces to

$$\begin{aligned}
\sum_{K\in\mathcal{T}_h}(\nabla e,\nabla v)_K+\gamma\sum_{e\in\mathcal{E}^I}h_e^{-1}\langle[e],[v]\rangle_e+\gamma\sum_{e\in\mathcal{E}^B}h_e^{-1}\langle e,v\rangle_e &= \sum_{K\in\mathcal{T}_h}(f+\Delta u_h^\gamma,\eta)_K\\
&\quad+\sum_{e\in\mathcal{E}^I}\Big[-\big\langle[\partial_n u_h^\gamma],\eta^*\big\rangle_e+\langle\{\partial_n e\},[v]\rangle_e-\gamma h_e^{-1}\langle[u_h^\gamma],[\eta]\rangle_e\Big]\\
&\quad+\sum_{e\in\mathcal{E}^B}\Big[\langle\partial_n e,v\rangle_e-\gamma h_e^{-1}\langle u_h^\gamma,\eta\rangle_e\Big].
\end{aligned}\tag{3.6}$$

At this point, we set $v=e$ and observe that

$$\begin{aligned}
\sum_{e\in\mathcal{E}^I}\langle\{\partial_n e\},[e]\rangle_e+\sum_{e\in\mathcal{E}^B}\langle\partial_n e,e\rangle_e &= -\sum_{e\in\mathcal{E}^I}\langle\{\partial_n e\},[u_h^\gamma]\rangle_e-\sum_{e\in\mathcal{E}^B}\langle\partial_n e,u_h^\gamma\rangle_e\\
&= -\sum_{e\in\mathcal{E}^I}\langle\{\partial_n e\},[u_h^\gamma-\chi]\rangle_e-\sum_{e\in\mathcal{E}^B}\langle\partial_n e,u_h^\gamma-\chi\rangle_e
\end{aligned}\tag{3.7}$$

for any $\chi\in\overset{0}{V}{}_h^r$. Since $a_h^\gamma(e,u_h^\gamma-\chi)=0$, we replace the terms $\sum_{e\in\mathcal{E}^I}\langle\{\partial_n e\},[e]\rangle_e+\sum_{e\in\mathcal{E}^B}\langle\partial_n e,e\rangle_e$ on the right-hand side of (3.6) with

$$\begin{aligned}
-\sum_{K\in\mathcal{T}_h}(\nabla e,\nabla(u_h^\gamma-\chi))_K-\sum_{e\in\mathcal{E}^I}\Big[\langle\{\partial_n(u_h^\gamma-\chi)\},[u_h^\gamma]\rangle_e-\gamma h_e^{-1}|[u_h^\gamma]|_e^2\Big]\\
-\sum_{e\in\mathcal{E}^B}\Big[\langle\partial_n(u_h^\gamma-\chi),u_h^\gamma\rangle_e-\gamma h_e^{-1}|u_h^\gamma|_e^2\Big]
\end{aligned}$$

to obtain

$$\begin{aligned}\sum_{K\in\mathcal{T}_h}\|\nabla e\|_K^2 = {}& \sum_{K\in\mathcal{T}_h}(f+\Delta u_h^\gamma,\eta)_K - \sum_{e\in\mathcal{E}^I}\Big[\left\langle[\partial_n u_h^\gamma],\eta^*\right\rangle_e + \gamma h_e^{-1}\left\langle[u_h^\gamma],[\eta]\right\rangle_e\Big]\\ &- \gamma\sum_{e\in\mathcal{E}^B} h_e^{-1}\left\langle u_h^\gamma,\eta\right\rangle_e - \sum_{K\in\mathcal{T}_h}(\nabla e,\nabla(u_h^\gamma-\chi))_K\\ &- \sum_{e\in\mathcal{E}^I}\left\langle\{\partial_n(u_h^\gamma-\chi)\},[u_h^\gamma]\right\rangle_e - \sum_{e\in\mathcal{E}^B}\left\langle\partial_n(u_h^\gamma-\chi),u_h^\gamma\right\rangle_e. \qquad (3.8)\end{aligned}$$

Here we have used the facts that $\eta = e - v_h$, $[e]|_e = -[u_h]|_e\ \forall e\in\mathcal{E}^I$, and $e|_e = -u_h|_e\,\forall e\in\mathcal{E}^B$.

We now obtain bounds for the terms on the right-hand side of (3.8). Those that contain η are bounded by $\frac{1}{2}$ times

$$\begin{aligned}&\frac{1}{\epsilon_1}\sum_{K\in\mathcal{T}_h}h_K^2\|f+\Delta u_h^\gamma\|_K^2 + \frac{1}{\epsilon_2}\sum_{e\in\mathcal{E}^I}h_e\big|[\partial_n u_h^\gamma]\big|_e^2 + \frac{1}{\epsilon_3}\gamma\sum_{e\in\mathcal{E}^I}h_e^{-1}|[u_h^\gamma]|_e^2\\ &+\frac{1}{\epsilon_4}\gamma\sum_{e\in\mathcal{E}^B}h_e^{-1}|u_h^\gamma|_e^2 + \epsilon_1\sum_{K\in\mathcal{T}_h}h_K^{-2}\|\eta\|_K^2 + \epsilon_2\sum_{e\in\mathcal{E}^I}h_e^{-1}|\eta^*|_e^2\\ &+\epsilon_3\gamma\sum_{e\in\mathcal{E}^I}h_e^{-1}|[\eta]|_e^2 + \epsilon_4\gamma\sum_{e\in\mathcal{E}^B}h_e^{-1}|\eta|_e^2 \qquad (3.9)\end{aligned}$$

for any $\epsilon_i > 0,\ i = 1,\ldots,4$. To estimate the "$\eta$" terms in (3.9) we choose as v_h the best piecewise constant approximation of e. From (2.10) this gives

$$h_K^{-2}\|\eta\|_K^2 = h_K^{-2}\|e-v_h\|_K^2 \le c\|\nabla e\|_K^2.$$

Also, using the trace inequality (2.8) and (2.10), we obtain

$$\begin{aligned}h_e^{-1}\left(|\eta^*|_e^2+|[\eta]|_e^2\right) &\le c\sum_{K=K^+,K^-}h_e^{-1}(h_K^{-1}\|\eta\|_K^2 + h_K\|\nabla\eta\|_K^2)\\ &\le c\sum_{K=K^+,K^-}h_e^{-1}h_K\|\nabla e\|_K^2.\end{aligned}$$

The local quasiuniformity of the mesh implies that $h_e \approx h_{K^+}\approx h_{K^-}$. Thus $h_e^{-1}h_K \le c$. A similar bound holding for $\sum_{e\in\mathcal{E}^B}h_e^{-1}|\eta|_e^2$, we can now hide the "$\eta$" terms in the left-hand side of (3.8) by taking the ϵ's sufficiently small. In particular, we must take $\epsilon_3 \approx 1/\gamma$ and $\epsilon_4\approx 1/\gamma$.

To obtain (3.1), we need to estimate the terms containing $u_h^\gamma-\chi$. Indeed these are bounded by

$$\begin{aligned}&\epsilon\sum_{K\in\mathcal{T}_h}\|\nabla e\|_K^2 + \frac{1}{\epsilon}\sum_{K\in\mathcal{T}_h}\|\nabla(u_h^\gamma-\chi)\|_K^2 + \sum_{e\in\mathcal{E}^I}h_e\big|\{\partial_n(u_h^\gamma-\chi)\}\big|_e^2\\ &+\sum_{e\in\mathcal{E}^I}h_e^{-1}|[u_h^\gamma]|_e^2 + \sum_{e\in\mathcal{E}^B}h_e\big|\partial_n(u_h^\gamma-\chi)\big|_e^2 + \sum_{e\in\mathcal{E}^B}h_e^{-1}|u_h^\gamma|_e^2. \qquad (3.10)\end{aligned}$$

Using the estimates (2.8) and (2.9), we see that the two terms in (3.10) that contain $\partial_n(u_h^\gamma-\chi)$ are bounded by $\sum_{K\in\mathcal{T}_h}\|\nabla(u_h^\gamma-\chi)\|_K^2$. In view of Theorem 2.2, the latter is bounded by $\sum_{e\in\mathcal{E}^I}h_e^{-1}\ |[u_h^\gamma]|_e^2 + \sum_{e\in\mathcal{E}^B}h_e^{-1}|u_h^\gamma|_e^2$. Using this fact completes the proof. □

THEOREM 3.2. *Suppose that f is a piecewise polynomial on $\mathcal{T}_h$. Then*
(i) *for each $K \in \mathcal{T}_h$,*

$$h_K^2 \|f + \Delta u_h^\gamma\|_K^2 \leq c\, \|\nabla e\|_K^2; \tag{3.11}$$

(ii) *for $e = K^+ \cap K^- \in \mathcal{E}^I$,*

$$h_e \big| \big[\partial_n u_h^\gamma \big] \big|_e^2 \leq c(\|\nabla e\|_{K^+}^2 + \|\nabla e\|_{K^-}^2). \tag{3.12}$$

Proof. To estimate $\|f + \Delta u_h^\gamma\|_K$, we set $v_h = 0$ and $v\big|_K = (f + \Delta u_h^\gamma)\, b_K$, where b_K is the "bubble" function $27\lambda_1\lambda_2\lambda_3$ expressed in terms of the barycentric coordinates of K; we extend v to the outside of K by zero. Using this v in (3.6), we obtain

$$\int_K (f + \Delta u_h^\gamma)^2\, b_K\, dx = (\nabla e, \nabla((f + \Delta u_h^\gamma)\, b_K))_K.$$

Now since $b_K > 0$ on $\mathrm{int}(K)$, $(\int_K (\cdot)^2\, b_K\, dx)^{1/2}$ defines a norm on $L^2(K)$, equivalent to the L^2 norm on $P_m(K)$ for any fixed m. Thus, there exists a constant $c > 0$ such that

$$\int_K (f + \Delta u_h^\gamma)^2\, b_K\, dx \geq c\, \|f + \Delta u_h^\gamma\|_K^2. \tag{3.13}$$

Since $\|b_K\|_{L^\infty(K)} = 1$, a scaling argument can be used to show that, while the constant c may depend on r and the degree of f, it is independent of h_K. On the other hand, using the inverse inequality (2.9), we have

$$\begin{aligned}(\nabla e, \nabla((f + \Delta u_h^\gamma)\, b_K))_K &\leq \|\nabla e\|_K \|\nabla((f + \Delta u_h^\gamma)\, b_K)\|_K \\ &\leq c\epsilon \|f + \Delta u_h^\gamma\|_K^2 + \frac{1}{\epsilon} h_K^{-2} \|\nabla e\|_K^2.\end{aligned}$$

This gives (i). We next estimate $h_e|[\partial_n u_h^\gamma]|_e^2$. Let $e = \partial K^+ \cap \partial K^-$ and suppose that e is a full edge of both K^+ and K^-. (See Remark 3.1 below.) Extend $\big[\partial_n u_h^\gamma\big]$ to a function ϕ defined over $\tilde{K} = K^+ \cup K^-$ by extending by constants along lines normal to e; see Figure 3.1. Also, let b denote the bubble function on $\tilde{K}$ given by

$$b\big|_{K^+} = 4\lambda_1^+\lambda_2^+, \quad b\big|_{K^-} = 4\lambda_1^-\lambda_2^-.$$

Let $v = \phi\, b$ and set $v_h = 0$. Using this v in (3.6), we get

$$\begin{aligned} h_e \int_e \big[\partial_n u_h^\gamma\big]^2 b\, ds &= h_e \sum_{K = K^+, K^-} \Big[(f + \Delta u_h^\gamma, \phi\, b)_K - (\nabla e, \nabla(\phi\, b))_K \Big] \\ &\leq h_e \sum_{K = K^+, K^-} \Big[\|f + \Delta u_h^\gamma\|_K \|\phi\, b\|_K + \|\nabla e\|_K \|\nabla(\phi\, b)\|_K \Big] \\ &\leq \frac{1}{2\epsilon} \sum_{K = K^+, K^-} \Big\{ h_K^2 \|f + \Delta u_h^\gamma\|_K^2 + \|\nabla e\|_K^2 \Big\} \qquad (3.14) \\ &\quad + \frac{\epsilon}{2} \sum_{K = K^+, K^-} \Big\{ \|\phi\, b\|_K^2 + h_K^2 \|\nabla(\phi\, b)\|_K^2 \Big\}. \end{aligned}$$

Using arguments similar to those leading to (3.13), we obtain

$$\int_e \big[\partial_n u_h^\gamma\big]^2 b\, ds \geq c \big| \big[\partial_n u_h^\gamma\big] \big|_e^2 \tag{3.15}$$

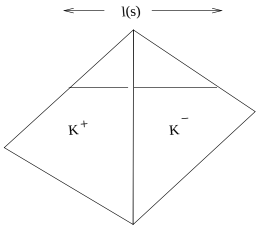


FIG. 3.1.

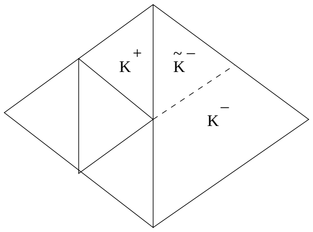


FIG. 3.2.

for some positive constant c depending only on r. Moreover,

$$\|\phi\, b\|^2_{\tilde{K}} \le \|\phi\|^2_{\tilde{K}} = \int_e \left[\partial_n u_h^\gamma\right]^2 l(s) ds \le h_e \big| \left[\partial_n u_h^\gamma\right] \big|^2_e, \tag{3.16}$$

where $l(s)$ is as in Figure 3.1. Using the inverse inequality (2.9), we see that

$$\sum_{K=K^+,K^-} h_K^2 \|\nabla(\phi\, b)\|^2_K \le c\|\phi\, b\|^2_{\tilde{K}} \le ch_e \big| \left[\partial_n u_h^\gamma\right] \big|^2_e. \tag{3.17}$$

The required estimate now follows from (3.14)–(3.17). □

Remark 3.1. If e is not a full edge of one of the triangles, say K^-, then we can work with $\tilde{K}^-$ instead; see Figure 3.2.

4. Estimates based on the solution of local problems. In this section, we shall introduce and analyze a posteriori estimates that are based on domain decomposition techniques proposed in [11]. The approach consists of viewing the computed solution u_h^γ as the coarse-mesh approximation to some function which is arguably a more accurate approximation to u. Before suggesting some choices for this quantity, let us say that there will be no attempt to compute it directly, but rather to approximate it by adding to u_h^γ a function obtained through the solution of "local" problems. For simplicity, we restrict the exposition to $d = 2$ and assume that $\mathcal{T}_h$ is a conforming mesh of triangles. Also, we should note that the results of [11] concern Baker's formulation only; however, we believe that similar results can be obtained for Arnold's formulation.

4.1. A nonoverlapping approach. To begin, let $\mathcal{T}_{h/2}$ be the mesh obtained by cutting every $K \in \mathcal{T}_h$ into four equal triangles. In a similar way, we may define $\mathcal{T}_{h'} := \mathcal{T}_{h/2^p}$, $h' := h/2^p$ by repeating this process p times. On the latter, we define a finite element space $V' := V_{h'}^{r'}$ of discontinuous piecewise polynomial functions of degree less than or equal to $r' - 1$, where $r' \ge r$. This way, V_h^r is a subspace of V'. On $V' \times V'$ we define the bilinear form $a' := a_{h'}^{\gamma'}$ just as in the definition of a_h^γ in (2.3). It is crucial for the analysis that a_h^γ be the restriction of a' to V_h^r in the sense that

$$a_h^\gamma(v, w) = a'(v, w) \quad \forall v, w \in V_h^r. \tag{4.1}$$

Remark 4.1. By comparing the penalty terms in a_h^γ and a', we see that (4.1) requires the condition $\gamma'(h')^{-1} = \gamma h^{-1}$, which, in view of the fact that $h' = h/2^p$, is equivalent to $\gamma' = \gamma 2^{-p}$. Now since the coercivity of the forms a' and a_h^γ can be guaranteed only if $\gamma' \ge \gamma_0'(r')$ and $\gamma \ge \gamma_0(r)$, respectively (see Lemma 2.1(i)), we see that γ must be chosen sufficiently large in order to have $\gamma 2^{-p} \ge \gamma_0'$. This does not present any theoretical difficulties, since γ can take on arbitrarily large values without having any result discussed in this work break down. On the other hand, the quality of the a priori and a posteriori estimates may suffer, as is the case with (3.1). (In this respect, see the discussion at the beginning of section 5 and Figures 5.5 and 5.6.) In practice, however, we anticipate that $r' = r + 1$ and/or $h' = h/2$ should be sufficient. This was indeed the case in all our numerical experiments.

For each $K \in \mathcal{T}_h$, we consider the "local" space $V'(K)$ obtained by restricting V' to K. By extending the elements of $V'(K)$ by zero to the rest of Ω, $V'(K)$ becomes a subspace of V'. Indeed, the latter is the direct sum of these local subspaces. On $V'(K) \times V'(K)$ we introduce the bilinear form $a'_K(\cdot,\cdot)$ as the restriction of $a'(\cdot,\cdot)$ to $V'(K) \times V'(K)$ (see (4.4) in [11]). As such, a'_K inherits the symmetry and coercivity of a' on $V'(K)$. In particular, for any $\gamma' \geq \gamma'_0$ there holds

$$a'_K(v,v) \geq c\|v\|^2_{1,K} \quad \forall v \in V'(K), \tag{4.2}$$

where $\|\cdot\|_{1,K}$ denotes the restriction of the $\|\cdot\|_{1,h'}$ norm to $V'(K)$. Adopting the terminology of [11], we consider $\mathcal{T}_h$ as the coarse mesh of $\mathcal{T}_{h'}$. Also, each $K \in \mathcal{T}_h$ is considered as a subdomain in $\mathcal{T}_{h'}$. In other words, $\mathcal{T}_h$ is both the coarse mesh and the subdomain partition of $\mathcal{T}_{h'}$.

Now let $u' := u^{\gamma'}_{h'} \in V'$ be the discontinuous Galerkin approximation of u in the space V'; i.e.,

$$a'(u',v) = (f,v) \quad \forall v \in V'. \tag{4.3}$$

At this point, we observe that, by virtue of (4.1), (2.5) and (4.3) imply the following orthogonality relation:

$$a'(u' - u^\gamma_h, v) = 0 \quad \forall v \in V^r_h. \tag{4.4}$$

Next, let the functions $\{\eta_K \in V'(K) \,|\, K \in \mathcal{T}_h\}$ be given as the solutions of the local problems

$$a'_K(\eta_K, v) = (f,v) - a'(u^\gamma_h, v) \quad \forall v \in V'(K). \tag{4.5}$$

The functions $\{\eta_K\}$ can be computed independently of each other and in parallel. Moreover, the function $\eta := \sum_{K\in\mathcal{T}_h} \eta_K$ approximates $\zeta := u' - u^\gamma_h$ in the following sense.

THEOREM 4.1. *There exist positive constants C_1 and C_2 such that*

$$C_1\,\|\eta\|_{1,h'} \leq \|\zeta\|_{1,h'} \leq C_2 \frac{h}{h'}\|\eta\|_{1,h'}. \tag{4.6}$$

Proof. Since $(f,v) = a'(u',v)$, $v \in V'$, from (4.5) we have

$$a'_K(\eta_K, v) = a'(\zeta, v) \quad \forall v \in V'(K). \tag{4.7}$$

Thus,

$$\sum_{K\in\mathcal{T}_h} a'_K(\eta_K,\eta_K) = a'(\zeta,\eta) \leq c\|\zeta\|_{1,h'}\|\eta\|_{1,h'}. \tag{4.8}$$

From (4.2) it follows that $\sum_{K\in\mathcal{T}_h} a'_K(\eta_K,\eta_K) \geq c\sum_{K\in\mathcal{T}_h} \|\eta_K\|^2_{1,K}$. On the other hand, it is easy to see that $\sum_{K\in\mathcal{T}_h} \|\eta_K\|^2_{1,K} \geq \|\eta\|^2_{1,h'}$. Thus, the first half of (4.6) follows.

To prove the second inequality, let $\zeta = \sum_{K\in\mathcal{T}_h} \zeta_K, \zeta_K \in V'(K)$. Let ζ_0 be the piecewise constant function on $\mathcal{T}_h$ defined by

$$\zeta_0\big|_K := \zeta_{K,0} = \frac{1}{|K|}\int_K \zeta_K\, dx.$$

Now since $\zeta_0 \in V_h^r$, it follows from (4.7) and (4.4) that

$$a'_K(\eta_K, \zeta_K - \zeta_{K,0}) = a'(\zeta, \zeta_K - \zeta_{K,0}) = a'(\zeta, \zeta_K).$$

Summing over $K \in \mathcal{T}_h$, we obtain

$$\sum_{K\in\mathcal{T}_h} a'_K(\eta_K, \zeta_K - \zeta_{K,0}) = a'(\zeta,\zeta) \geq c\|\zeta\|^2_{1,h'}. \tag{4.9}$$

Now, it follows from the Cauchy–Schwarz inequality that

$$\begin{aligned}\sum_{K\in\mathcal{T}_h} a'_K(\eta_K, \zeta_K - \zeta_{K,0}) &\leq \left(\sum_{K\in\mathcal{T}_h} a'_K(\eta_K,\eta_K)\right)^{1/2} \\ &\quad \times \left(\sum_{K\in\mathcal{T}_h} a'_K(\zeta_K - \zeta_{K,0}, \zeta_K - \zeta_{K,0})\right)^{1/2}.\end{aligned} \tag{4.10}$$

Also from (4.8) it follows that

$$\left(\sum_{K\in\mathcal{T}_h} a'_K(\eta_K,\eta_K)\right)^{1/2} \leq c\|\zeta\|^{1/2}_{1,h'}\|\eta\|^{1/2}_{1,h'}. \tag{4.11}$$

On the other hand, with the interface bilinear form $I'(\cdot,\cdot)$ defined in (4.5) of [11],

$$\begin{aligned}I'(u,v) = \sum_{e'\in\mathcal{E}^I} \Big\{ &\left\langle \{\partial_n u\}, v^-\right\rangle_{e'} + \left\langle \{\partial_n v\}, u^-\right\rangle_{e'} \\ &- \gamma'(h')^{-1}\left[\left\langle u^+, v^-\right\rangle_{e'} + \left\langle v^+, u^-\right\rangle_{e'}\right]\Big\} \qquad \forall u, v \in V',\end{aligned}$$

we have

$$\begin{aligned}\sum_{K\in\mathcal{T}_h} a'_K(\zeta_K - \zeta_{K,0}, \zeta_K - \zeta_{K,0}) &= a'(\zeta - \zeta_0, \zeta - \zeta_0) - I'(\zeta - \zeta_0, \zeta - \zeta_0) \\ &\leq 2a'(\zeta,\zeta) + 2a'(\zeta_0,\zeta_0) + \left|I'(\zeta - \zeta_0, \zeta - \zeta_0)\right|.\end{aligned} \tag{4.12}$$

In [11] it is proved that $a'(\zeta_0,\zeta_0)$ and $|I'(\zeta - \zeta_0, \zeta - \zeta_0)|$ are bounded by $c\frac{h}{h'}a'(\zeta,\zeta)$. Thus from (4.9)–(4.12) it follows that

$$\|\zeta\|^2_{1,h'} \leq c\left(\frac{h}{h'}\right)^{1/2}\|\zeta\|^{3/2}_{1,h'}\|\eta\|^{1/2}_{1,h'},$$

from which the second inequality of (4.6) follows. □

We shall use the equivalence just proved to obtain estimates for $e = u - u_h^\gamma$. Letting $e' = u - u'$, we have $e = e' + \zeta$. We now argue as follows: It is reasonable to expect that e' is much smaller than e, say in the energy norm; therefore e and ζ

are nearly equal. Since ζ is not computed, we shall approximate it, and hence e, by η, where the latter is obtained by solving local problems. To quantify matters, since $a'(e', v) = 0\ \forall v \in V'$ and $\zeta \in V'$, we obtain

$$a'(e, e) = a'(e', e') + a'(\zeta, \zeta). \tag{4.13}$$

It follows from this and (2.7) that $a'(e', e') = \epsilon\, a'(e, e)$ for some $0 < \epsilon < 1$. Based on a priori estimates, it is reasonable to expect that $\epsilon = O(\frac{(h')^{r'-1}}{h^{r-1}}) \ll 1$. Thus,

$$a'(e, e) = \frac{1}{1-\epsilon}\, a'(\zeta, \zeta).$$

In view of the equivalence between η and ζ provided by Theorem 4.1, we can use $a'(\eta, \eta)$ to obtain lower and upper bounds for $a'(e, e)$.

4.2. An overlapping approach. Let $\Omega = \cup_{e\in\mathcal{E}}\Omega_e$ be an overlapping decomposition of Ω, where each Ω_e is the following union of the triangles in $\mathcal{T}_h$: If $e \in \mathcal{E}^I$, then $e = \partial K^+ \cap \partial K^-$ and $\Omega_e = K^+ \cup K^-$; else if $e \in \mathcal{E}^B$, then $e = \partial K \cap \partial\Omega$ and $\Omega_e = K$.

On $\mathcal{T}_h$, we define a finite element space $V' := V_h^{r'}$ of discontinuous piecewise polynomial functions of degree less than or equal to $r' - 1$, where $r' \geq r + 1$; let us recall that $r \geq 2$ is fixed. We construct a subspace decomposition of this latter finite element space by defining the subspaces $\{V'_e\}_{e\in\mathcal{E}}$ associated with the subdomains $\{\Omega_e\}_{e\in\mathcal{E}}$ by

$$V'_e = \{v_h \in V', v_h = 0 \text{ in } \Omega \setminus \bar{\Omega}_e\}.$$

Thus the following decomposition, which is not direct, holds:

$$V' = V_h^r + \sum_{e\in\mathcal{E}} V'_e. \tag{4.14}$$

On $V' \times V'$, we define the bilinear form $a' := a_h^{\gamma'}$ as in the definition of a_h^γ in (2.3), and again it is crucial for the analysis that a_h^γ be the restriction of a' to V_h^r in the sense that

$$a_h^\gamma(v, w) = a'(v, w) \quad \forall v, w \in V_h^r. \tag{4.15}$$

For each edge $e \in \mathcal{E}$, we consider on $V'_e \times V'_e$ the symmetric and coercive bilinear form $a'_e(\cdot,\cdot)$ as the restriction of $a'(\cdot,\cdot)$ to $V'_e \times V'_e$ (see (4.4) in [11]),

$$a'_e(v, w) = a'(v, w) \quad \forall v, w \in V'_e. \tag{4.16}$$

Following [11], we are able to define the additive operator $T = T_0 + \sum_{e\in\mathcal{E}} T_e$, where T_0 is a projection operator from V' to V_h^r defined by

$$a_h^\gamma(T_0 u, v) = a'(T_0 u, v) = a'(u, v) \quad \forall v \in V_h^r \tag{4.17}$$

and T_e is a projection operator from V' to V'_e defined by

$$a'_e(T_e u, v) = a'(T_e u, v) = a'(u, v) \quad \forall v \in V'_e. \tag{4.18}$$

Lemmas 5.1–5.5 in [11] can easily be adapted to the present case with $H = h$, and $\delta \sim h$, and Theorem 5.7 in [11] then reads as follows.

THEOREM 4.2. *There exist positive constants c_1, c_2, which are independent of h and of the number of edges, such that there holds the estimate*

$$c_1 a'(v,v) \leq a'(Tv,Tv) \leq c_2 a'(v,v) \quad \forall v \in V'. \tag{4.19}$$

Now let $u' := u_h^{\gamma'} \in V'$ be the discontinuous Galerkin approximation of u in the space V'; i.e.,

$$a'(u',v) = (f,v) \quad \forall v \in V'. \tag{4.20}$$

Next, let the functions $\eta_e \in V'_e$ be given as the solutions of the local problems

$$a'_e(\eta_e, v) = (f,v) - a'(u_h^\gamma, v) \quad \forall v \in V'_e. \tag{4.21}$$

The functions $\{\eta_e\}_{e\in\mathcal{E}}$ can be computed independently of each other and in parallel, and the function $\eta := \sum_{e\in\mathcal{E}} \eta_e$ approximates $\zeta := u' - u_h^\gamma$ in the following sense.

THEOREM 4.3. *There exist positive constants C_1 and C_2 such that*

$$C_1 \|\zeta\|_{1,h} \leq \|\eta\|_{1,h} \leq C_2 \|\zeta\|_{1,h}. \tag{4.22}$$

Proof. Let us prove that $\eta = T\zeta$. Indeed, from (4.20) and (4.21) we get

$$a'_e(\eta_e, v) = a'(u' - u_h^\gamma, v) \quad \forall v \in V'_e, \tag{4.23}$$

which means from the definition (4.18) of T_e that $\eta_e = T_e\zeta$ $\forall e \in \mathcal{E}$.

Now, by virtue of (4.15), (2.5) and (4.20) imply the orthogonality relation

$$a'(u' - u_h^\gamma, v) = 0 \quad \forall v \in V_h^r, \tag{4.24}$$

which means from the definition of T_0 that $T_0\zeta = 0$.

Thus (4.22) follows from Theorem 4.2. □

We conclude, in a similar way as in section 4.1, that we can use $a'(\eta,\eta)$ to obtain lower and upper bounds for $a'(e,e)$.

Remark 4.2. As for the nonoverlapping approach, we could define the finite element space $V' := V_{h'}^{r'}$ with $h' = h/2^p$, where the mesh is obtained in dimension 2 by cutting p times every triangle into four equal triangles.

5. Numerical results in one dimension.

In this section, we present numerical results obtained from the one-dimensional (1-d) model problem

$$-\frac{d^2u}{dx^2} = f, \qquad 0 < x < 1, \quad u(0) = u(1) = 0, \tag{5.1}$$

with the exact solution $u(x) = e^{-\alpha(x-\frac{1}{2})^2}, \alpha = 100$.

For the sake of brevity, we shall consider in these numerical experiments only the weak formulation (2.3) due to Baker; for some comparisons with the Arnold formulation, we refer to the technical report [13].

5.1. Convergence with respect to γ and h.

For a number of years we have been interested in the behavior of the discontinuous Galerkin approximations as a function of the penalty parameter γ, and indeed we had a proof (unpublished) that

$$\lim_{\gamma\to\infty} u_h^\gamma = u_h^G, \tag{5.2}$$

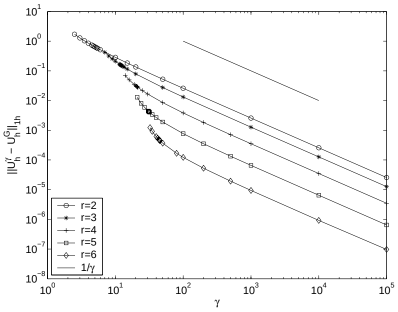


FIG. 5.1. $\|u_h^\gamma - u_h^G\|_{1,h}$ *versus* γ.

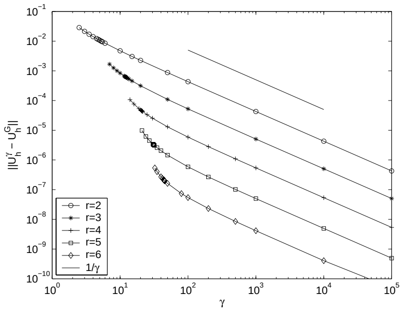


FIG. 5.2. $\|u_h^\gamma - u_h^G\|$ *versus* γ.

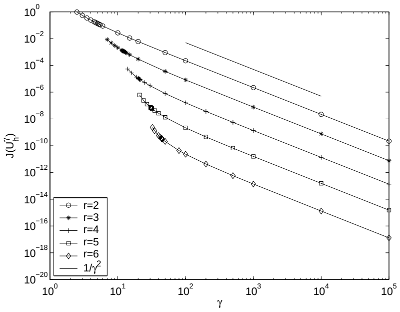


FIG. 5.3. $J(u_h^\gamma)$ *versus* γ.

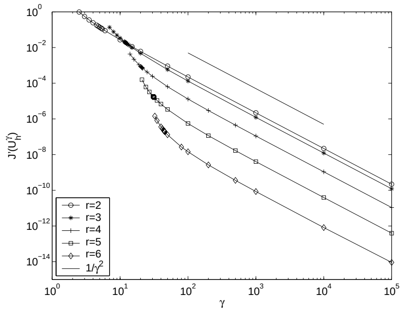


FIG. 5.4. $J'(u_h^\gamma)$ *versus* γ.

where u_h^G is the standard Galerkin approximation of u defined by

$$(\nabla u_h^G, \nabla \chi) = (f, \chi) \quad \forall \chi \in \overset{0}{V}{}_h^r. \tag{5.3}$$

A more recent proof can be found in [14]. Besides its intrinsic value, this result can be used to show that the discontinuous Galerkin method can yield more accurate results than the standard Galerkin version for a range of values of γ, as shown later in Figure 5.5.

In the first experiments, the domain $[0,1]$ is divided into a uniform mesh of 20 subintervals. The approximations u_h^γ, solution of (2.5), and u_h^G, solution of (5.3), are computed using piecewise polynomials of degree up to 5. Figures 5.1 and 5.2 show the difference between u_h^γ and u_h^G in the energy norm $\|\cdot\|_{1,h}$ and the L^2 norm, respectively, as a function of γ. These plots highlight the convergence (5.2) according to the rate $O(\frac{1}{\gamma})$. Similarly, Figure 5.3 shows the behavior of the jump in u_h^γ,

$$J(u_h^\gamma) \equiv \sum_{e\in\mathcal{E}^I} h_e^{-1} |[u_h^\gamma]|_e^2 + \sum_{e\in\mathcal{E}^B} h_e^{-1} |u_h^\gamma|_e^2, \tag{5.4}$$

as a function of γ. Note that $J(u_h^\gamma) = J(u_h^\gamma - u_h^G)$. We see that $J(u_h^\gamma)$ behaves as $\frac{1}{\gamma^2}$ when γ tends to infinity. In the same way, one can observe in Figure 5.4 that the jump of the derivative

$$J'(u_h^\gamma) \equiv \sum_{e\in\mathcal{E}^I} h_e \left|\{\partial_n (u_h^\gamma - u_h^G)\}\right|_e^2 + \sum_{e\in\mathcal{E}^B} h_e \left|\partial_n (u_h^\gamma - u_h^G)\right|_e^2 \tag{5.5}$$

behaves also as $\frac{1}{\gamma^2}$.

We now study the difference between the discontinuous Galerkin and exact solutions of the problem. For $h = \frac{1}{20}$ and $r = 2$, $\|u_h^\gamma - u\|$ and $|u_h^\gamma - u|_{1,h} \equiv (\sum_{K\in\mathcal{T}_h} ||\nabla(u_h^\gamma - u)||_K^2)^{1/2}$ are plotted in Figures 5.5 and 5.6, respectively. We see that these converge to values (represented by the dashed lines) which, in view of (5.2), must be $\|u_h^G - u\|$ and $\|\nabla(u_h^G - u)\|$, respectively. We also observe that there exists an optimal value γ_{opt} of the penalty parameter γ, for which $\|u_h^\gamma - u\|$ is minimized. From these and other numerical experiments not reported here (see the report [13] for other values of r), we claim that, in the case of the Baker formulation, this optimal value does not depend either on the mesh size or on u (or, from an equivalent point of view, on the function f), but depends on the degree of the polynomial approximations: It follows approximately the rule

$$\gamma_{opt} = (r-1)(r+3), \tag{5.6}$$

as can be seen in Figure 5.7, where the circles represent the numerical value of γ_{opt} for different values of $r-1$ and the continuous line represents the $(r-1)(r+3)$ function.

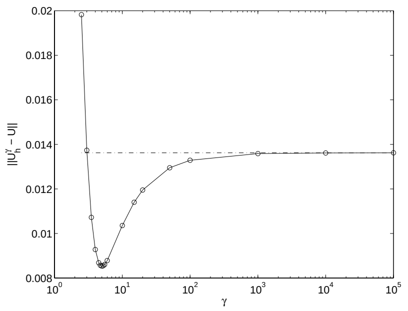


FIG. 5.5. $\|u_h^\gamma - u\|$ *versus* γ.

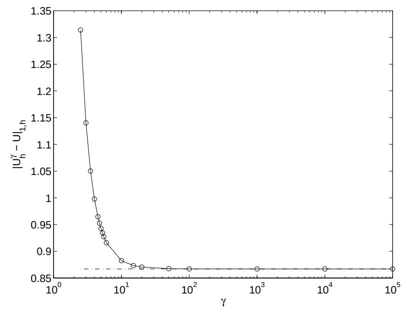


FIG. 5.6. $|u_h^\gamma - u|_{1,h}$ *versus* γ.

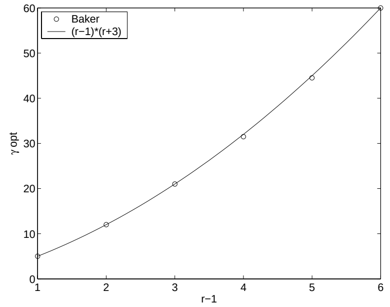


FIG. 5.7. γ_{opt} *versus* $r-1$.

For the parameter γ chosen approximately equal to γ_{opt}, we now investigate the convergence of u_h^γ to u on a sequence of uniformly refined meshes. The differences $\|u_h^\gamma - u\|$ and $\|u_h^\gamma - u\|_{1,h}$ are plotted in Figures 5.8 and 5.9, respectively, for piecewise polynomials of degree $r-1 = 1,2,3,4$. The observed rates of convergence of $O(h^r)$ and $O(h^{r-1})$, respectively, conform to the a priori estimates expressed in (2.12) and

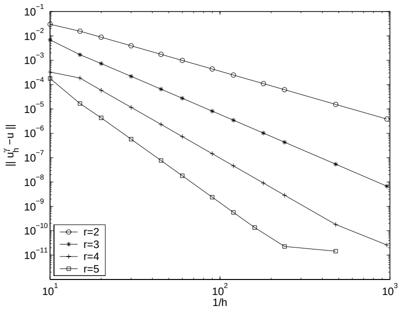


FIG. 5.8. $\|u_h^\gamma - u\|$ *versus* $1/h$.

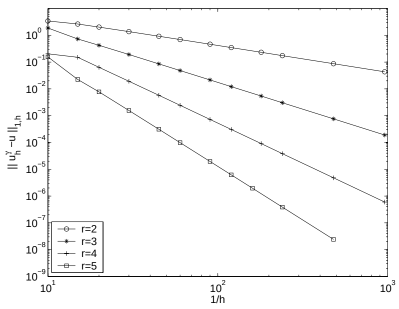


FIG. 5.9. $\|u_h^\gamma - u\|_{1,h}$ *versus* $1/h$.

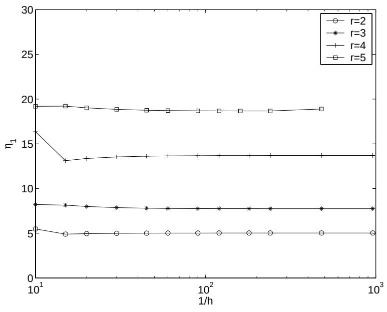


FIG. 5.10. η_1 *versus* $1/h$.

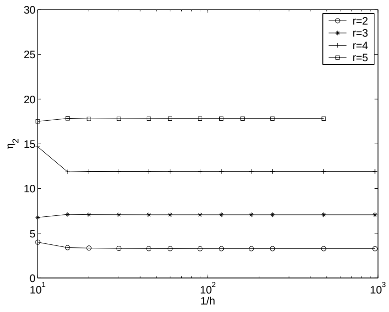


FIG. 5.11. η_2 *versus* $1/h$.

in (2.11): The distorted line for $r = 5$ is due to the limitations of computations in double precision.

5.2. Effectivity indices. In order to judge the quality of the various error estimators presented above, we compute for each an effectivity index, defined as the ratio of the estimator to the exact error. First, we study three estimators featured in Theorems 3.1 and 3.2. Specifically, in Figure 5.10, the effectivity index η_1 of the first estimator corresponding to Theorem 3.1 and formula (3.1) is plotted as a function of $1/h$ and for values of $r-1$ between 1 and 4:

$$\eta_1^2 = \frac{\displaystyle\sum_{K\in\mathcal{T}_h} h_K^2\|f+\Delta u_h^\gamma\|_K^2 + \sum_{e\in\mathcal{E}^I} h_e\big|\big[\partial_n u_h^\gamma\big]\big|_e^2 + \gamma^2\sum_{e\in\mathcal{E}^I} h_e^{-1}|[u_h^\gamma]|_e^2 + \gamma^2\sum_{e\in\mathcal{E}^B} h_e^{-1}|u_h^\gamma|_e^2}{\displaystyle\sum_{K\in\mathcal{T}_h}\|\nabla e\|_K^2}. \tag{5.7}$$

In Figure 5.11, we plot the effectivity index η_2 of the second estimator, which corresponds to Theorem 3.2(i) and to (3.11):

$$\eta_2^2 = \frac{\displaystyle\sum_{K\in\mathcal{T}_h} h_K^2\|f+\Delta u_h^\gamma\|_K^2}{\displaystyle\sum_{K\in\mathcal{T}_h}\|\nabla e\|_K^2}. \tag{5.8}$$

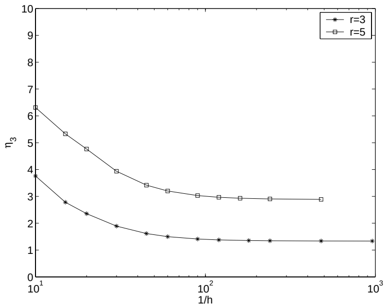


FIG. 5.12. *η_3 versus $1/h$ for even $r-1$.*

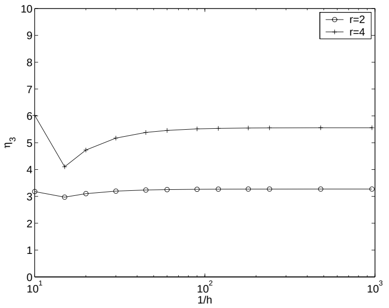


FIG. 5.13. *η_3 versus $1/h$ for odd $r-1$.*

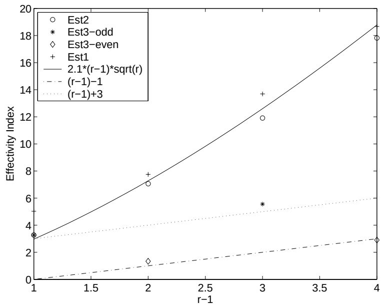


FIG. 5.14. *Effectivity indices versus $r-1$.*

Finally, Figures 5.12 and 5.13 represent for odd and even degrees of polynomials, respectively, the effectivity index η_3 of the estimator associated with Theorem 3.2(ii) and (3.12):

$$\eta_3^2 = \frac{\displaystyle\sum_{e\in\mathcal{E}^I} h_e \big|\big[\partial_n u_h^\gamma\big]\big|_e^2}{\displaystyle\sum_{K\in\mathcal{T}_h} \|\nabla e\|_K^2}. \tag{5.9}$$

It is seen that, as h decreases, these indices converge to values larger than 1. We also observe that η_1 and η_2 attain their respective asymptotic values rather quickly. On the other hand, while η_3 is somewhat slower in that respect, it is still nearly constant over a wide range of values of h.

Additionally, the asymptotic values depend strongly on r. Since it is desirable to have effectivity indices close to 1, we tried to find simple laws describing this dependence. As evidenced by Figure 5.14, the following functions seem to “fit” the asymptotic values reasonably well:

$$\eta_1 \sim \eta_2 \sim 2.1(r-1)\sqrt{r}, \tag{5.10}$$

$$\eta_3 \sim r-2 \quad \text{if } r-1 \text{ is even}, \tag{5.11}$$

$$\eta_3 \sim r+2 \quad \text{if } r-1 \text{ is odd}. \tag{5.12}$$

Indeed, dividing the above estimators by the corresponding asymptotic values should result in effectivity indices that are very close to 1.

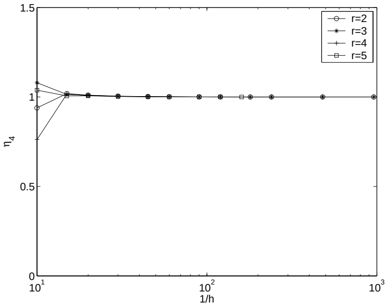


FIG. 5.15. *Effectivity index of the nonoverlapping approach–based estimator with $h' = h$ and $r' = r + 1$ versus h.*

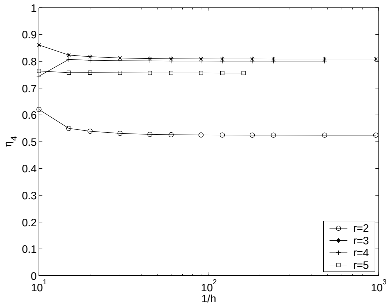


FIG. 5.16. *Effectivity index of the nonoverlapping approach–based estimator with $h' = h/2$ and $r' = r + 1$ versus h.*

The previous experiments are repeated with the two error estimators based on the nonoverlapping and overlapping domain decomposition approaches. More precisely, we define η_4 by

$$\eta_4 = \frac{\|\eta\|_{1,h'}}{\|e\|_{1,h'}}, \tag{5.13}$$

where $\eta = \sum_{K \in \mathcal{T}_h} \eta_K$ and η_K is the solution of the local problem (4.5). Among the various values of parameters that are possible, we chose the two combinations $h' = h$, $r' = r + 1$ and $h' = h/2$, $r' = r + 1$, the results being reported in Figures 5.15 and 5.16, respectively. In the former case, we observe that the index is exactly equal to 1 even for relatively large values of h and does not depend on r or on $r' > r$. In the latter case, the index is slightly less than 1 and depends on r and not on $r' \geq r + 1$.

In the same way,

$$\eta_5 = \frac{\|\eta\|_{1,h'}}{\|e\|_{1,h'}}, \tag{5.14}$$

where $\eta = \sum_{K \in \mathcal{T}_h} \eta_K$ and η_K is the solution of the local problem (4.21). The results for the case $h' = h$ and $r' = r + 1$ are reported in Figure 5.17. We observe that the effectivity index in this case is equal to 2, which is also the number of triangles in a subdomain Ω_e. In the case of $h' = h/2$ and $r' = r + 1$, this index is slightly higher than 1, as can be seen in Figure 5.18.

If any conclusions can be drawn after such a limited number of experiments, they would be that, while the estimators η_4 and η_5 based on the ideas of domain decomposition seem to be very robust, the estimators η_2 and η_3 are, in contrast, less expensive to implement, and offer the added advantage of being entirely local.

5.3. Adaptive mesh strategy. In order to gauge the efficiency of the a posteriori error estimates that we have derived, we present here two h-adaptive methods for approximating the solution of problem (5.1). We based our numerical experiments on the second estimate, which corresponds to Theorem 3.2(i) and to the effectivity index η_2 plotted in Figure 5.11.

Both strategies modify the mesh by refinement of some marked elements while keeping the degree of the polynomials constant. The goal is to generate a mesh in a finite number of steps such that a given tolerance is met by the approximate solution

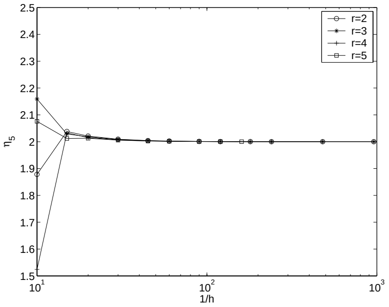


FIG. 5.17. *Effectivity index of the overlapping approach–based estimator with $h' = h$ and $r' = r + 1$ versus h.*

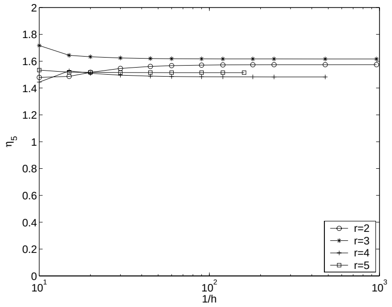


FIG. 5.18. *Effectivity index of the overlapping approach–based estimator with $h' = h/2$ and $r' = r + 1$ versus h.*

on this mesh. To do this, some optimality criteria have to be imposed. The first technique is based on the convergent adaptive algorithm proposed in [10] for solving Poisson's equation and used, for instance, in [17]. In order to minimize the total number of degrees of freedom, this strategy equidistributes the given tolerance (tol) on each element. Consequently, the local error of the optimal mesh $\mathcal{T}_h$ satisfies $\eta_K(u_h^\gamma)^2 \sim \frac{tol^2}{m_h}$, where m_h is the number of elements and u_h^γ the discontinuous Galerkin solution on $\mathcal{T}_h$. Since the number of iterations required to get this optimal mesh is quite large, we derived a second strategy, which turned out to require less cpu-time. We shall next describe these two strategies and apply them to problem (5.1), with the following exact solution:

$$u(x) = (1-x)\left(\tan^{-1}\left(\alpha\left(x-\frac{1}{2}\right)\right) + \tan^{-1}\left(\frac{\alpha}{2}\right)\right), \quad \alpha = 100.$$

Given an error tolerance tol and a coarse mesh $\mathcal{T}_0$, let u_0^γ denote the discontinuous Galerkin solution on $\mathcal{T}_0$. In this study, for simplicity reasons, we are not considering the effect of data oscillations as in [15]. Let $k = 0$. The first strategy involves the following steps:

(i) compute the local indicator $\eta_K^k(u_k^\gamma)$ such that

$$\eta_K^k(u_k^\gamma)^2 = \frac{{h_K^k}^2\|f + \Delta u_k^\gamma\|_K^2}{r(r-1)^2}\,; \tag{5.15}$$

(ii) compute the total error estimate

$$\eta^k(u_k^\gamma) = \left(\sum_{K\in\mathcal{T}_k} \eta_K^k(u_k^\gamma)^2\right)^{1/2}; \tag{5.16}$$

(iii) select a set $\hat{\mathcal{T}}_k$ of "marked" elements to be refined such that, for a given parameter θ (fixed in our experiments to 0.5),

$$\left(\sum_{K\in\hat{\mathcal{T}}_k} \eta_K^k(u_k^\gamma)^2\right)^{1/2} \geq \theta\eta^k(u_k^\gamma)\,; \tag{5.17}$$

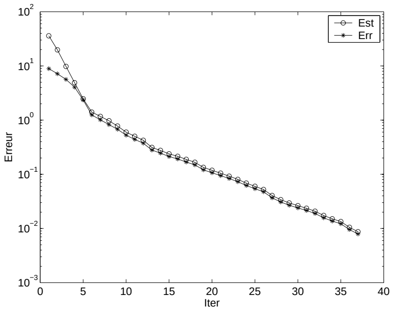


FIG. 5.19. *Strategy* 1: *estimate and exact error versus h-adaptive iterations (r* = 2*).*

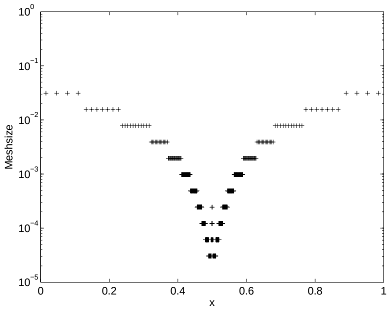


FIG. 5.20. *Strategy* 1: *mesh size versus x (r* = 2*).*

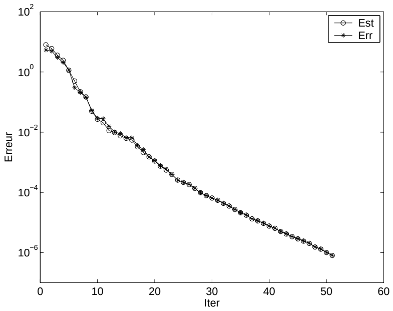


FIG. 5.21. *Strategy* 1: *estimate and exact error versus h-adaptive iterations (r* = 4*).*

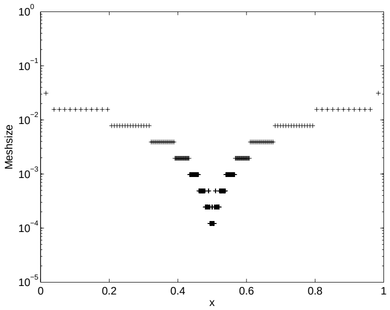


FIG. 5.22. *Strategy* 1: *mesh size versus x (r* = 4*).*

(iv) obtain a refined mesh $\mathcal{T}_{k+1}$ by dividing each element $K \in \hat{\mathcal{T}}_k$ (in two parts for a 1-d problem);

(v) compute the discontinuous Galerkin solution on $\mathcal{T}_{k+1}$;

(vi) $k \leftarrow k+1$ and go to step (i).

The algorithm is stopped when $\eta^k(u_k^\gamma) \leq tol$ in step (ii). In practice, for computing marked elements in step (iii), we follow the procedure proposed in [10]. Let us remark that, for changing to the other estimates, formula (5.15) just has to be adapted in step (i).

For $r = 2$ and $tol = 0.01$, this strategy required 37 iterations to reach the optimal mesh, which has 1411 elements and whose distribution of mesh size is plotted in Figure 5.20. For $r = 4$ and $tol = 10^{-6}$, 50 iterations were necessary, the final mesh has 471 elements, and the mesh size distribution is plotted in Figure 5.22. In both cases, the error estimate is an accurate approximation of the exact error (in energy norm), as can be seen in Figures 5.19 and 5.21.

Now let us observe that, to get $\sum_K \|\nabla e\|_K^2 \leq tol^2$, it is sufficient to distribute (not to equidistribute) the errors as follows:

$$\|\nabla e\|_K \leq \sqrt{\frac{|K|}{|\Omega|}}\, tol\,. \tag{5.18}$$

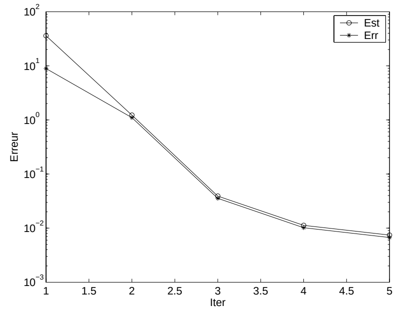


FIG. 5.23. *Strategy* 2: *estimate and exact error versus h-adaptive iterations (r* = 2*).*

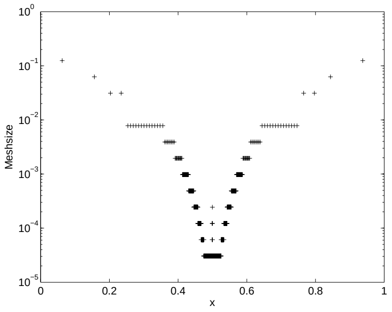


FIG. 5.24. *Strategy* 2: *mesh size versus x (r* = 2*).*

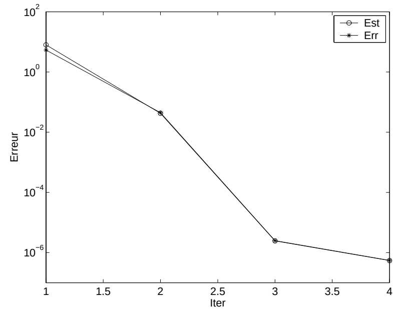


FIG. 5.25. *Strategy* 2: *estimate and exact error versus h-adaptive iterations (r* = 4*).*

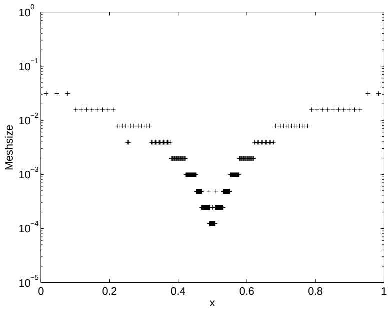


FIG. 5.26. *Strategy* 2: *mesh size versus x (r* = 4*).*

Therefore, we developed a strategy in which an element K, whose local error estimate η_K is larger than $\sqrt{|K|/|\Omega|}\,tol$, has to be refined as many times as necessary to reduce the local error by the amount $\sqrt{|K|}\,tol/\sqrt{|\Omega|}\,\eta_K$. Let us recall that from the a priori estimation (2.11) the rate of convergence in the energy norm is $O(h^{r-1})$. Thereafter, the number of times the element has to be divided can be estimated to be

$$nbr = \frac{\log\left(\dfrac{\sqrt{|\Omega|}\,\eta_K}{\sqrt{|K|}\,tol}\right)}{\log 2^{r-1}}. \tag{5.19}$$

In one dimension, this is equivalent to determining into how many segments, nbs, the element K has to be divided:

$$nbs = \left(\frac{\sqrt{|\Omega|}\,\eta_K}{\sqrt{|K|}\,tol}\right)^{\frac{1}{r-1}}. \tag{5.20}$$

The second strategy consists in the following steps:

(i) compute the local indicator $\eta_K^k(u_k^\gamma)$ given by (5.15),
(ii) compute the total error estimate $\eta^k(u_k^\gamma)$ according to (5.16),
(iii) compute for each element the nearest power of 2 of *nbs* defined in (5.20),
(iv) obtain a refined mesh $\mathcal{T}_{k+1}$ by dividing each element by this power of 2 for a 1-d problem,

(v) compute the discontinuous Galerkin solution on $\mathcal{T}_{k+1}$,

(vi) $k \leftarrow k+1$ and go to step (i).

The algorithm is stopped when $\eta^k(u_k^\gamma) \leq tol$ in step (ii).

For $r = 2$ and $tol = 0.01$, in only 5 steps this strategy reaches the mesh such that $\eta^k \leq tol$. The cpu-time is then significantly reduced. However, this time the number of elements is not optimal anymore and is equal to 2316 elements. For $r = 4$ and $tol = 10^{-6}$, we get the given tolerance in 4 iterations, and the final mesh has 573 elements. The distribution of mesh size plotted in Figures 5.24 and 5.26 is almost the same as in the first strategy, and, except for the first iteration, the a posteriori estimate gives a good approximation of the exact error, as shown in Figures 5.23 and 5.25.

REFERENCES

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

[2] D. N. Arnold, *An interior penalty finite element method with discontinuous elements*, SIAM J. Numer. Anal., 19 (1982), pp. 742–760.

[3] D. Arnold, F. Brezzi, B. Cockburn, and D. Marini, *Discontinuous Galerkin methods for elliptic problems*, in Proceedings of the International Symposium on the Discontinuous Galerkin Method, B. Cockburn, G. E. Karniadakis, C.-W. Shu, eds., Springer Lecture Notes in Comput. Sci. Engrg. 11, Springer-Verlag, Berlin, 2000, pp. 89–101.

[4] I. Babuška and W. C. Rheinboldt, *Error estimates for adaptive finite element computations*, SIAM J. Numer. Anal., 15 (1978), pp. 736–754.

[5] I. Babuška and W. C. Rheinboldt, *A posteriori error estimates for the finite element method*, Internat. J. Numer. Methods Engrg., 12 (1978), pp. 1597–1615.

[6] G. Baker, *Finite element methods for elliptic equations using nonconforming elements*, Math. Comp., 31 (1977), pp. 45–59.

[7] G. A. Baker, W. N. Jureidini, and O. A. Karakashian, *Piecewise solenoidal vector fields and the Stokes problem*, SIAM J. Numer. Anal., 27 (1990), pp. 1466–1485.

[8] R. Becker, P. Hansbo, and M. Larson, *Energy norm a posteriori error estimation for discontinuous Galerkin methods*, Comput. Methods Appl. Mech. Engrg., to appear.

[9] J. Douglas, Jr., and T. Dupont, *Interior Penalty Procedures for Elliptic and Parabolic Galerkin Methods*, Lecture Notes in Phys. 58, Springer, Berlin, 1976.

[10] W. Dörfler, *A convergent adaptive algorithm for Poisson's equation*, SIAM J. Numer. Anal., 33 (1996), pp. 1106–1124.

[11] X. Feng and O. A. Karakashian, *Two-level additive Schwarz methods for a discontinuous Galerkin approximation of second order elliptic problems*, SIAM J. Numer. Anal., 39 (2001), pp. 1343–1365.

[12] O. A. Karakashian and W. N. Jureidini, *A nonconforming finite element method for the stationary Navier–Stokes equations*, SIAM J. Numer. Anal., 35 (1998), pp. 93–120.

[13] O. Karakashian and F. Pascal, *A Priori and A Posteriori Estimates for Discontinuous Galerkin Method*, Technical report, in preparation.

[14] M. Larson and A. Niklasson, *Conservation Properties for the Continuous and Discontinuous Galerkin Methods*, Chalmers Finite Element Center Preprint 2000-08.

[15] P. Morin, R. H. Nochetto, and K. G. Siebert, *Data oscillation and convergence of adaptive FEM*, SIAM J. Numer. Anal., 38 (2000) pp. 466–488.

[16] B. Rivière and M. Wheeler, *A posteriori error estimates and mesh adaptation strategy for discontinuous Galerkin methods applied to diffusion problems*, Comput. Math. Appl., to appear.

[17] A. Schmidt and K. G. Siebert, *A posteriori estimators for the h-p version of the finite element method in* 1*D*, Appl. Numer. Math., 35 (2000), pp. 43–66.

[18] R. Verfürth, *A Review of A Posteriori Error Estimation and Adaptive Mesh Refinement Techniques*, Wiley-Teubner, New York, 1995.

[19] M. F. Wheeler, *An elliptic collocation-finite element method with interior penalties*, SIAM J. Numer. Anal., 15 (1978), pp. 152–161.

[20] B. I. Wohlmuth, *Hierarchical a posteriori error estimators for mortar finite element methods with Lagrange multipliers*, SIAM J. Numer. Anal., 36 (1999), pp. 1636–1658.